Krystyna Sulka lista 1

107101 Zootechnika

1. ***Podać definicję oraz po dwa przykłady zmiennej losowej ciągłej oraz zmiennej losowej dyskretnej.***
2. **zmienna losowa dyskretna-** zmienną losową nazywamy dowolną funkcję X, określoną na przestrzeni zdarzeń elementarnych E, o własnościach ze zbioru liczb rzeczywistych i mierzalną względem ciała zdarzeń Z.

Zmienna losowa X dana jest zbiorem X: E→R

Zmienne losowe oznaczamy dużymi literami np.: S, T, X, Y, Z, ich własności zaś odpowiednimi małymi literami: s, t, x, y, z, często ze wskaźnikami.

Jeżeli zbiór wartości, jakie przyjmuje funkcja X, jest zbiorem policzalnym, wtedy zmienną losową nazywamy zmienną losową dyskretną lub skokową. Zmienna losowa ma charakter skokowy, jeżeli zbiór możliwych jej wartości jest skończony lub nieskończony, lecz przeliczalny. Przyjmują one najczęściej wartości liczb naturalnych.

Przykłady.

Ilość dziewczynek urodzonych w 2019 roku U(x)=[0,1,2,3,4,..]

ilość osób, które wyzdrowiały z korono-wirusa w Polsce. C(x)= [0,1,2,3,4..]

1. **Zmienna losowa ciągła-** zmienną losową nazywamy dowolną funkcję X, określoną na przestrzeni zdarzeń elementarnych E, o własnościach ze zbioru liczb rzeczywistych i mierzalną względem ciała zdarzeń Z.

Zmienna losowa X dana jest zbiorem X: E→R

Zmienne losowe oznaczamy dużymi literami np.: S, T, X, Y, Z, ich własności zaś odpowiednimi małymi literami: s, t, x, y, z, często ze wskaźnikami.

Jeżeli zbiór wartości, jakie przyjmuje funkcja X, przybiera dowolne wartości z pewnego przedziału liczbowego, skończonego lub nieskończonego to jest to zmienna losowa ciągła.

Przykłady: Dokładny czas w jakim kot łapie mysz[s] , lub też dokładna waga[kg] (np.57,8kg) mieszkańców danej ulicy.

1. **Podaj definicję dystrybuanty zmiennej losowej.**

**Dystrybuanta zmiennej losowej- to prawdopodobieństwo uzyskania wartości nie większej lub równej od ustalonej wartości.**

X jest funkcją określoną na całym zbiorze R= (−∞, +∞), i jest dana wzorem **F(x)=P(X<x),**

Własności dystrybuanty F zmiennej X są następujące:

* 0≤F(x)≤1,, dla każdego x∈R,
* F(x), jest funkcją niemalejącą,
* F(x), jest funkcją co najmniej lewostronnie ciągłą, czyli F(x0−0)=F(x0), dla każdego x∈R,
* limx→−∞F(x)=0, oraz limx→+∞F(x)=1,

1. Dla podanego wektora liczb, wyznaczyć średnią, odchylenie standardowe, wariancję, wartość minimalną, wartość maksymalną, medianę oraz rozstęp (różnica między wartością maksymalną oraz minimalną). Dla wektora X wyznaczyć wykres boxplot oraz go opisać.

Polecenia :

- mean (x) – średnia

- sd (x) – odchylenie standardowe

- var (x) – wariancja

- max (x) – wartość minimalna

- max (x) – wartość maksymalna

- median (x) – mediana

- max(x) – min(x) - rozstęp

**Liczby** 🡪 X = (4.41, 3.92, 5.38, 2.77, 4.60, 5.14, 3.93, 3.15, 4.67, 2.98, 0.001, 10.45)

liczby = c(4.41, 3.92, 5.38, 2.77, 4.60, 5.14, 3.93, 3.15, 4.67, 2.98, 0.001, 10.45)

print(paste("Średnia: ",mean(liczby)))

**Średnia: 4.28341666666667**

print(paste("odchylstand", sd(liczby)))

**Odchylenie standardowe: 2.41395000544499**

print(paste("wariancja: ", var(liczby)))

**wariancja: 5.8271546287878**

print(paste("wartość", "minimalna:", min(liczby)))

**wartość minimalna: 0.001**

print(paste("wartość", "maxymalna:", max(liczby)))

**wartość maksymalna: 10.45**

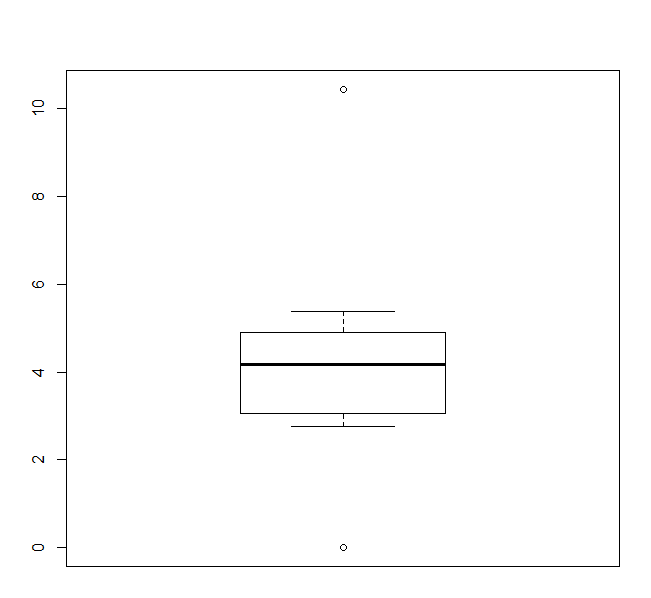
print(paste("mediana", median(liczby)))

**mediana 4.17**

print(paste("rozstęp:", max(liczby)-min(liczby)))

**rozstęp: 10.449**

**Boxplot**



Przedstawiony powyżej wykres pudełkowy mówi nam, że:

* Górna obserwacja odstająca (wartość maksymalna) wynosi 10,45 i spełnia ona warunek ≥ Q3 + IQR \* 1,5, czyli, że wynik jest większy niż rezultat dodania do wartości trzeciego kwartyla wartości uzyskanej poprzez pomnożenie rozstępu kwartylowego i 1,5.
* Dzięki temu, że mam podaną wartość kwartyla Q1 i Q3 mogę obliczyć rozstęp ćwiartkowy, czyli 1,68
* Wartość zamykająca górę pudełka to trzeci kwartyl 4.7875, czyli wynik 75% rozkładu wszystkich wyników
* Mediana, czyli kwartyl drugiego stopnia wynosi 4,17
* Wartość zamykająca dół pudełka, czyli pierwszy kwartyl ma wartość 3.1075, czyli wartości 25% rozkładu wszystkich wyników.
* Dolna obserwacja odstająca (wartość minimalna), wynosi 0.001, i spełnia warunek ≤ Q1 – IQR \* 1,5, informując o tym, że wynik, który otrzymałam jest niższy niż rezultat odejmowania iloczynu rozstępu ćwiartkowego i 1,5 od wartości pierwszego kwartyla.

Na pierwszy rzut oka widać, że istnieją dwie skrajne wartości, które znacznie odbiegają od pozostałych wartości. Wartości te mogą zaburzyć analizę wszystkich elementów, dlatego chcąc uzyskać najbardziej prawdopodobne wyniki należałoby ich nie wliczać do analizy. Mając tak mały zbiór danych należałoby się zastanawiać skąd wzięły się tak odstające wartości czy są one istotne do przebiegu dalszej analizy. Zakładając, że nasza obserwacja dotyczyłaby zwartości tłuszczu w mleku wartości te miały by znaczenie dla analizy, ponieważ pokazywały by negatywny wynik (choroby osobnika itp.) ale badając np. średnie ciśnienie krwi studentów tak odbiegające wartości mogły wyniknąć z błędu przy pomiarze lub zapisywaniu, dlatego w takim wypadku nie powinny być brane pod uwagę, aby nie burzyć analizy całego badania.

4.Które obserwacje z wektora X możesz uznać za wartości odstające? Odstające wartości

wystandaryzować.

- z obserwacji z wektora X, jako wartości odstające można uznać wartość maksymalną, czyli 10,45 oraz wartość minimalną, czyli 0,001. Aby wykonać standaryzacje muszę odjąć od każdej wartości zmiennej (maksymalnej i minimalnej) wartości średniej i podzielić przez odchylenie standardowe.

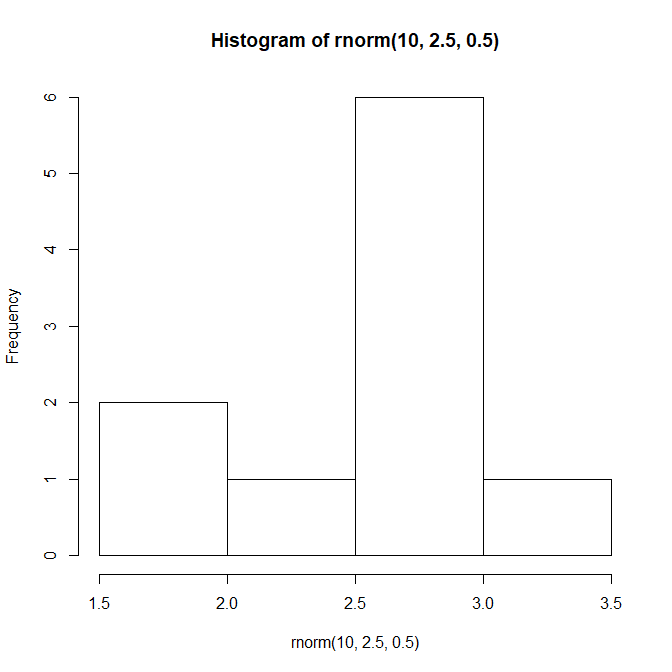


=

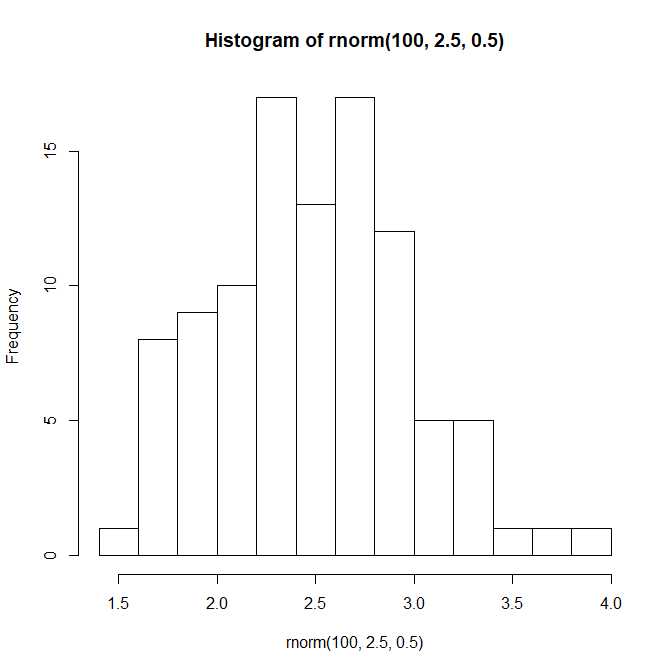
=

5. Przy pomocy funkcji rnorm() wygeneruj wektor dla próby 10, 100 oraz 1000 elementowej ze średnią równą 2.5 oraz odchyleniem standardowym równym 0.5. Dla wygenerowanych wektorów stworzyć wykresy histogramy i je omówić.

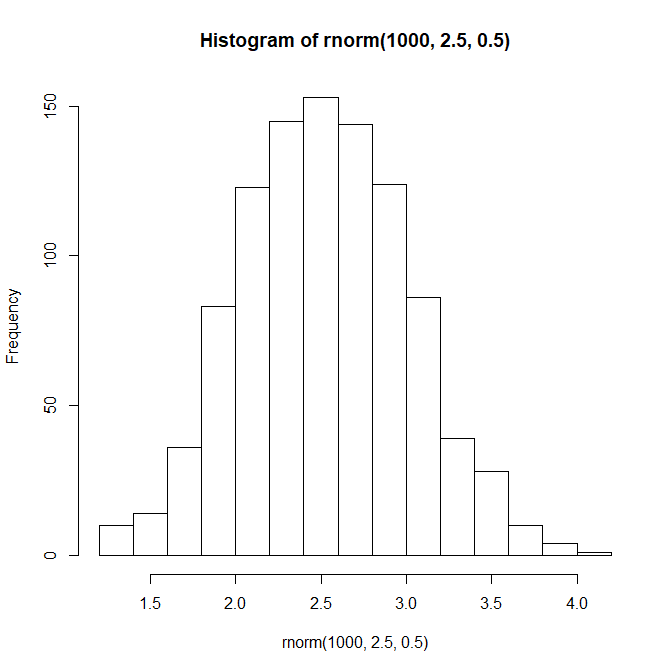
hist(rnorm(10,2.5,0.5))



hist(rnorm(100,2.5,0.5))



hist(rnorm(1000,2.5,0.5))



wykresy przedstawiają liczebność badanej cechy w populacji -oś X, oraz rozpiętość przedziałów klasowych- oś Y.

Histogram jest potrzebny, aby określić charakter rozkładu zmiennej, w tym wypadku rozkładu normalnego. Oba histogramy przedstawiają rozkład normalny, najlepiej widać to na histogramie trzecim, na drugim jest to trochę mniej widoczne a na pierwszym prawie wcale.

wnioski:

Im więcej zmiennych tym bardziej rozkład zbliża się do rozkładu normalnego. Widać to dokładnie na podstawie pierwszego wykresu, gdzie próba miała mało elementów (10) przez co nie było widać rozkładu normalnego. Na drugim wykresie natomiast liczba elementów zwiększyła się do 100 dzięki czemu wykres bardzo zbliżył się do rozkład normalnego, wykres trzeci z liczbą elementów 1000 bardzo dobrze ukazuje rozkład normalny.

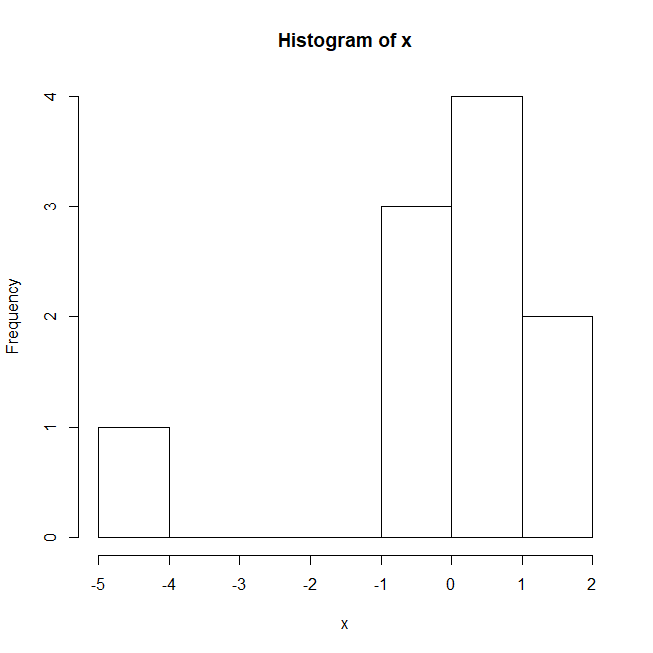
6. Za co odpowiada funkcja rt(), za jej pomocą funkcji rt() wygeneruj wektor dla próby 10 oraz 100 elementowej z 9 stopniami swobody. Dla wygenerowanych wektorów stworzyć histogramy i je omówić. rt() generuje n losowych wartości dla rozkładu t-studenta , przyjmuje ona dwa parametry : liczbę obserwacji (n) i liczbę stopni swobody (dt)

x = rt(10,9)

y = rt(100,9)

print(x)

hist(x)



[1] 0.73139968 1.31577350 0.41972439 -0.02552798 -0.09497599 0.41474654

[7] 1.34563798 0.77785506 -0.83503260 -4.37266531

x = rt(10,9)

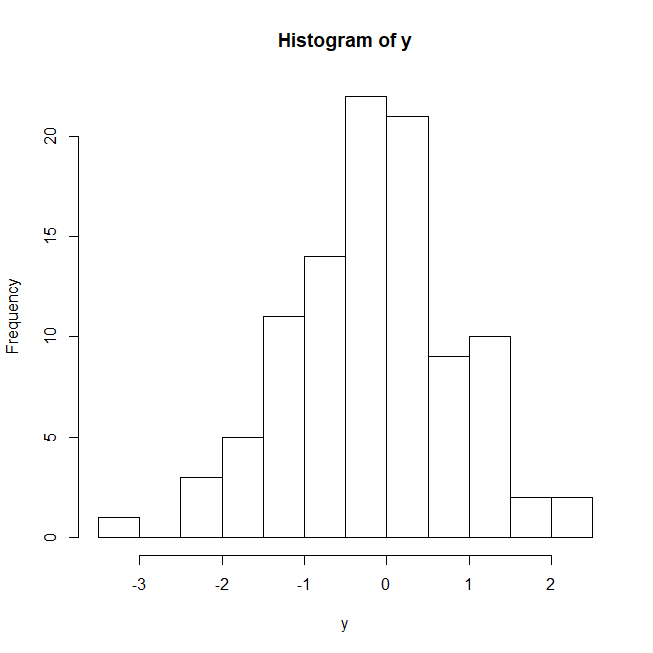
y = rt(100,9)

print(x)

hist(x)

print(y)

hist(y)



[1] -0.770018334 2.297027652 -0.177305510 -1.686194578 1.394008452

[6] -0.298956946 0.006704828 -1.702163832 -0.803153162 1.402098396

[11] -0.114665565 2.308345903 1.547811484 -1.268984313 -1.443904205

[16] 1.016725555 -0.172627741 -0.570780357 -0.791682484 1.328031113

[21] 0.540253867 -0.318223508 -0.019678574 0.747998233 0.375958636

[26] -0.886557015 -0.884953057 -1.811233113 -0.344818224 -0.398627661

[31] -0.733831649 -0.675434281 -0.642941305 -0.196376165 0.112617145

[36] -0.311301724 0.206793531 -2.479564796 0.487854681 1.761849767

[41] -0.172938173 -1.023062996 -2.198299681 -0.220740042 -0.224171874

[46] 0.037718015 0.228604342 0.356593802 0.346167079 1.411474746

[51] -2.031061875 0.346267761 -0.172636230 -1.430972259 0.918505523

[56] 1.398906085 1.400065230 0.871638635 -1.266666537 -1.071310709

[61] 0.169965590 -1.027318787 0.063630908 1.004071514 0.741065764

[66] -0.521013363 0.069756378 -1.395405037 -0.220614460 0.727875536

[71] -0.864272351 -1.825304662 -0.476716193 -1.101094558 -0.202810481

[76] -1.719144792 -1.269595032 0.026575168 0.279273521 0.284877260

[81] -0.153772019 -0.926789037 0.124613653 1.370955153 1.096168543

[86] 0.094642245 0.754620788 -0.149182654 0.340507722 0.732591245

[91] 0.241308134 -0.920901764 -1.489689012 0.460720002 -3.038105557

[96] -0.158989254 -0.835993940 -0.431166933 0.947471923 -0.217582947

Wnioski: Im większa liczba stopni swobody tym wykres bardziej przypomina rozkład normalny. Rozkład T Dla dużych prób (n > 30) praktycznie pokrywa się z rozkładem normalnym, dlatego drugi wykres wygląda jak rozkład normalny. Dzięki dużej ilości liczby stopni swobody wyniki które zostały otrzymane mogą mieć wpływ na niewystąpienie błędu statystycznego.

Gdy n > 30 rozkład t studenta niewiele różni się od rozkładu normalnego więc w zadaniach obliczeniowych przyjmuje się wartości z tablicy rozkładu normalnego. Dodając większą liczbę obserwacji kształt wykresu będzie bardzo zbliżony ro rozkładu normalnego.

Ponadto podobnie jak w zadaniu poprzednim osie X i Y opisują liczebność badanej cechy w populacji -oś X, oraz rozpiętość przedziałów klasowych- oś Y.