Bootcamp Data Science

Statystyka

Przemysław Spurek

O co chodzi z tym ciągiem zmiennych losowych?

Populacje i próbki

- W statystycznej analizie danych zazwyczaj wykorzystujemy dane z kilku wybranych **próbek**, aby wyciągnąć wnioski dotyczące **populacji**, z której pobrano te próbki.
- Właściwie zaprojektowana analiza powinien zapewnić, że dane dotyczące **próbek** są reprezentatywne dla **populacji**, z której pobrano próbki.

Populacje i próbki

Główna różnica między **populacją**, a **próbką** ma związek z przypisywaniem obserwacji do zbioru danych.

- **Populacja** – zawiera wszystkie elementy z zestawu danych.
- **Próbka** – składa się z jednej lub kilku obserwacji z populacji.

Można uzyskać więcej niż jedną próbkę z tej samej populacji.

Populacje i próbki

Przykład

Oceniając parametr populacji np. ciężar mężczyzn w Europie, zazwyczaj nie możemy zważyć wszystkich osób.

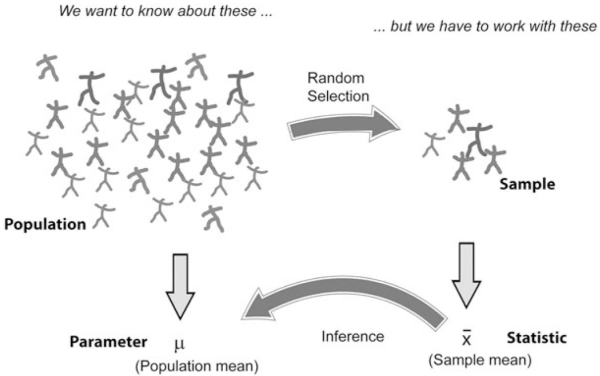
Musimy ograniczyć się do zbadania (przypadkowych reprezentantów) losowej próbki pobranej z tej grupy (populacji).

Populacje i próbki

Na podstawie statystyk próbki, czyli odpowiedniej wartości obliczonej na podstawie danych z próbki, wykorzystujemy wnioskowanie statystyczne, aby dowiedzieć się, co wiemy o odpowiednim parametrze w populacji.

- **Parametr** – charakterystyka populacji, np. średnie lub odchylenie standardowe.
- **Statystyka** – mierzalna charakterystyka próbki. Przykładem statystyki jest średnia z danych.

Populacje i próbki



Próbka

Definicja

Prostą próbą losową (lub krócej próbą losową) o liczności n nazywamy **ciąg niezależnych zmiennych losowych** $X_1, X_2, \ldots, X_n$ o takim samym rozkładzie.

Próbka

Definicja

Prostą próbą losową (lub krócej próbą losową) o liczności n nazywamy **ciąg niezależnych zmiennych losowych** $X_1, X_2, \ldots, X_n$ o takim samym rozkładzie.

Uwaga

Konkretny ciąg wartości $x_1, x_2, ..., x_n$ (prostej) próby losowej $X_1, X_2, \ldots, X_n$ nazywamy realizacją (prostej) próby losowej lub próbką.

Próbka

Definicja

Prostą próbą losową (lub krócej próbą losową) o liczności n nazywamy **ciąg niezależnych zmiennych losowych** $X_1, X_2, \ldots, X_n$ o takim samym rozkładzie.

Uwaga

Konkretny ciąg wartości $x_1, x_2, ..., x_n$ (prostej) próby losowej $X_1, X_2, \ldots, X_n$ nazywamy realizacją (prostej) próby losowej lub próbką.

Uwaga

Statystyką nazywamy każdą zmienną losową będącą ustaloną funkcją próby losowej $X_1, X_2, \ldots, X_n$.

Populacje i próbki

Przykład parametrów i statystyk

	Population parameter	Sample statistic
Mean	μ	$\bar{x}$
Standard deviation	σ	s

- **Parametry** – najczęściej oznaczane są za pomocą greckich liter.
- **Statystyki** – najczęściej oznaczane są za pomocą zwykłych liter.

Stopnie swobody (Degrees of Freedom)

Pojęcie stopni swobody (DOF), które w mechanice wydaje się być krystalicznie czyste, trudniej jest zrozumieć dla zastosowań statystycznych.

Przykład

W mechanice, jeśli cząstka poruszająca się na płaszczyźnie ma "2 DOF":

Stopnie swobody (Degrees of Freedom)

Pojęcie stopni swobody (DOF), które w mechanice wydaje się być krystalicznie czyste, trudniej jest zrozumieć dla zastosowań statystycznych.

Przykład

W mechanice, jeśli cząstka poruszająca się na płaszczyźnie ma "2 DOF": w każdym punkcie czasowym ruch opisany jest przez dwa parametry (współrzędne x i y określają położenie cząstki).

Stopnie swobody (Degrees of Freedom)

Pojęcie stopni swobody (DOF), które w mechanice wydaje się być krystalicznie czyste, trudniej jest zrozumieć dla zastosowań statystycznych.

Przykład

W mechanice, jeśli cząstka poruszająca się na płaszczyźnie ma “2 DOF”: w każdym punkcie czasowym ruch opisany jest przez dwa parametry (współrzędne x i y określają położenie cząstki).

Jeśli cząstka poruszająca się w przestrzeni, ma "3 DOF": współrzędne x, y i z.

Stopnie swobody (Degrees of Freedom)

W statystyce grupa n wartości $X = \{x_1, \ldots, x_n\}$ ma n stopni swobody.

Stopnie swobody (Degrees of Freedom)

W statystyce grupa n wartości $X = \{x_1, \ldots, x_n\}$ ma n stopni swobody. Jeśli policzymy wartość oczekiwaną, to możemy odjąć od każdego elementu wartości średnią próbki.

Stopnie swobody (Degrees of Freedom)

W statystyce grupa n wartości $X = \{x_1, \ldots, x_n\}$ ma n stopni swobody.
Jeśli policzymy wartość oczekiwaną, to możemy odjąć od każdego elementu wartości średnią próbki.
Tak otrzymane dane mają tylko $n - 1$ stopni swobody.

Stopnie swobody (Degrees of Freedom)

W statystyce grupa n wartości $X = \{x_1, \ldots, x_n\}$ ma n stopni swobody.
Jeśli policzymy wartość oczekiwaną, to możemy odjąć od każdego elementu wartości średnią próbki.
Tak otrzymane dane mają tylko $n - 1$ stopni swobody.

Przykład

W przypadku zbioru zawierającego dwa elementy $X = \{x_1, x_2\}$ $(n = 2)$ znamy średnią z danych oraz wartość x_1, to drugą wartość możemy uzyskać za pomocą wzoru

$$x_2 = 2 \cdot mean - x_1 .$$

Moda z próbki

Definicja

Moda (wartość modalna) jest to najczęściej występująca wartość zmiennej X. W przypadku, gdy kilka wartości jest osiąganych taką samą liczbę razy, wówczas każda z nich jest modą.

Przykład

Załóżmy, że rozważaną populacją jest zbiór samochodów znajdujących się w określonym czasie na pewnym parkingu, zaś cechą - nazwa producenta samochodu. Jej wartości mogą wyglądać, na przykład, tak:

Fiat, BMW, Ford, Ford, Fiat, Skoda, Fiat, Polonez, Toyota, Toyota, Toyota, Renault, Opel, Fiat, Opel, Opel, Toyota.

Nasza cecha ma dwie mody: Fiat i Toyota.

Średnia z próbki

https://github.com/przem85/statistic_4/blob/master/D03_Z01_descriptive_statistics.ipynb

Definicja

Jeżeli cecha X przyjmuje wartości $x_1, x_2, \ldots, x_n$, wówczas jej średnią arytmetyczną, lub krótko średnią, nazywamy:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x.$$

Średnia geometryczna

W niektórych sytuacjach średnia geometryczna może być użyteczna do opisania rozmieszczenia rozkładu. Można ją obliczyć za pomocą średniej arytmetycznej logarytmów wartości

$$mean_{geometric} = \left(\prod_{i=1}^{N} x_i\right)^{1/n} = exp\left(\frac{\sum_i ln(x_i)}{n}\right).$$

Zauważ, że wartości wejściowe dla średniej geometrycznej muszą być dodatnie.

Mediana z próbki

Dla danego ciągu liczb $x_1, \dots, x_n$, określamy ciąg $x_{(1)}, \dots, x_{(n)}$, który powstaje przez jego niemalejące uporządkowanie, czyli:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Mediana z próbki

Dla danego ciągu liczb $x_1, \ldots, x_n$, określamy ciąg $x_{(1)}, \ldots, x_{(n)}$, który powstaje przez jego niemalejące uporządkowanie, czyli:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Definicja

Medianą cechy X, przyjmującej wartości $x_1, \ldots, x_n$, nazywamy środkowy wyraz ciągu $x_{(1)}, \ldots, x_{(n)}$, gdy n jest liczbą nieparzystą, lub średnią arytmetyczną dwóch wyrazów środkowych, gdy n jest liczbą parzystą. Zatem:

$$me = \begin{cases} x_{(k+1)} & \text{dla} \quad n = 2k+1 \\ \frac{x_{(k)}+x_{(k+1)}}{2} & \text{dla} \quad n = 2k. \end{cases}$$

Rang

Zakres (rang) jest po prostu różnicą między najwyższą i najniższą wartością w danych.
W przypadku zakresu danych łatwo jest zauważyć dane odstające. Często takie punkty są spowodowane błędami w wyborze próbki lub w procedurze pomiaru.

Odchylenie standardowe oraz wariancja

Używa się dwóch estymatorów wariancji próbki:

- $$S^2 = var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
- $$S^{*2} = var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Odchylenie standardowe oraz wariancja

Używa się dwóch estymatorów wariancji próbki:

- $$S^2 = var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
- $$S^{*2} = var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Odchylenie standardowe jest pierwiastkiem kwadratowym wariancji:

$$S = \sqrt{var}.$$

Odchylenie standardowe oraz wariancja

Używa się dwóch estymatorów wariancji próbki:

- $$S^2 = var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
- $$S^{*2} = var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Odchylenie standardowe jest pierwiastkiem kwadratowym wariancji:

$$S = \sqrt{var}.$$

Czasami używa się

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Odchylenie standardowe oraz wariancja

W przeciwieństwie do innych języków, takich jak Matlab lub R, numpy domyślnie oblicza wariancję dla “n”. Aby uzyskać wariancję próbki należy ustawić ‘‘ddof = 1’’ :

```
import numpy as np
data = np.arange(7,14)
print(np.std(data, ddof=0))
print(np.std(data))
print(np.std(data, ddof=1))
```

Sample standard error

Dla próbki z rozkładu normalnego (SE lub SEM) jest:

$$SEM = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \cdot \frac{1}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n^2}},$$

$$SEM^* = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \cdot \frac{1}{\sqrt{n}}.$$

Próbka

Definicja

Prostą próbą losową (lub krócej próbą losową) o liczności n nazywamy **ciąg niezależnych zmiennych losowych** $X_1, X_2, \ldots, X_n$ o takim samym rozkładzie.

Uwaga

Konkretny ciąg wartości $x_1, x_2, ..., x_n$ (prostej) próby losowej $X_1, X_2, \ldots, X_n$ nazywamy realizacją (prostej) próby losowej lub próbką.

Uwaga

Statystyką nazywamy każdą zmienną losową będącą ustaloną funkcją próby losowej $X_1, X_2, \ldots, X_n$.

Rozkłady prawdopodobieństwa a testowanie hipotez

`https://github.com/przem85/statistic_4/blob/master/D03_Z02_hypothesis_testing_introduction.ipynb`

Aby zilustrować związek pomiędzy **rozkładami prawdopodobieństwa** a **testowaniem hipotez**, rozważmy następujący problem:

Rozkłady prawdopodobieństwa a testowanie hipotez

https://github.com/przem85/statistic_4/blob/master/D03_Z02_hypothesis_testing_introduction.ipynb

Aby zilustrować związek pomiędzy **rozkładami prawdopodobieństwa** a **testowaniem hipotez**, rozważmy następujący problem:

- Średnia masa noworodków w USA wynosi 3.5 kg, przy odchyleniu standardowym 0.76 kg.

Rozkłady prawdopodobieństwa a testowanie hipotez

https://github.com/przem85/statistic_4/blob/master/D03_Z02_hypothesis_testing_introduction.ipynb

Aby zilustrować związek pomiędzy **rozkładami prawdopodobieństwa** a **testowaniem hipotez**, rozważmy następujący problem:

- Średnia masa noworodków w USA wynosi 3.5 kg, przy odchyleniu standardowym 0.76 kg.
- Załóżmy, że chcemy **znaleźć** wszystkie dzieci znacznie różniące się od normy (aby móc monitorować ich rozwój).

Rozkłady prawdopodobieństwa a testowanie hipotez

https://github.com/przem85/statistic_4/blob/master/D03_Z02_hypothesis_testing_introduction.ipynb

Aby zilustrować związek pomiędzy **rozkładami prawdopodobieństwa** a **testowaniem hipotez**, rozważmy następujący problem:

- Średnia masa noworodków w USA wynosi 3.5 kg, przy odchyleniu standardowym 0.76 kg.
- Załóżmy, że chcemy **znaleźć** wszystkie dzieci znacznie różniące się od normy (aby móc monitorować ich rozwój).
- Co zrobić z dzieckiem, które urodziło się z wagą 2.6 kg?

Rozkłady prawdopodobieństwa a testowanie hipotez

https://github.com/przem85/statistic_4/blob/master/D03_Z02_hypothesis_testing_introduction.ipynb

Aby zilustrować związek pomiędzy **rozkładami prawdopodobieństwa** a **testowaniem hipotez**, rozważmy następujący problem:

- Średnia masa noworodków w USA wynosi 3.5 kg, przy odchyleniu standardowym 0.76 kg.
- Załóżmy, że chcemy **znaleźć** wszystkie dzieci znacznie różniące się od normy (aby móc monitorować ich rozwój).
- Co zrobić z dzieckiem, które urodziło się z wagą 2.6 kg?
- Możemy wypowiedzieć ten problem w formie testu hipotez:
  - Nasza hipoteza polega na tym, że dziecko pochodzi z populacji "zdrowych" niemowląt.
  - Czy możemy odrzucić hipotezę, czy też ciężar dziecka sugeruje, że nie ma podstaw do odrzucenia takiej hipotezy?

Aby odpowiedzieć na to pytanie, możemy postępować w następujący sposób:

Rozkłady prawdopodobieństwa a testowanie hipotez

Aby odpowiedzieć na to pytanie, możemy postępować w następujący sposób:

- Załóżmy, że urodzenia są modelowane rozkładem normalnym o parametrach $\mu = 3.5$, $\sigma = 0.76$.

Rozkłady prawdopodobieństwa a testowanie hipotez

Aby odpowiedzieć na to pytanie, możemy postępować w następujący sposób:

- Załóżmy, że urodzenia są modelowane rozkładem normalnym o parametrach $\mu = 3.5$, $\sigma = 0.76$.
- Znajdźmy dystrybuantę (CDF) tej zmiennej losowej oraz wyznacz $CDF(2.6)$.

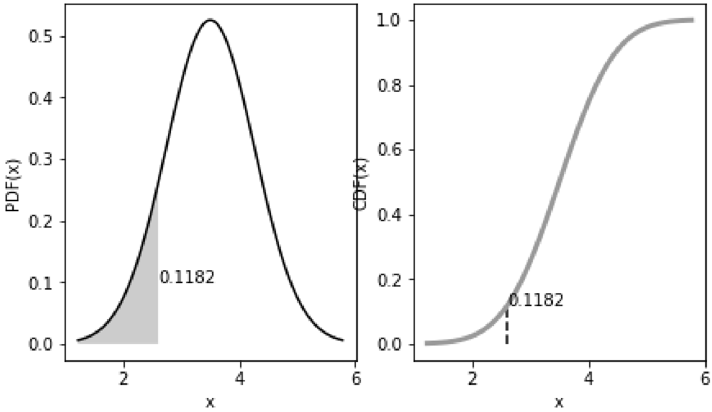
Rozkłady prawdopodobieństwa a testowanie hipotez

Aby odpowiedzieć na to pytanie, możemy postępować w następujący sposób:

- Załóżmy, że urodzenia są modelowane rozkładem normalnym o parametrach $\mu = 3.5$, $\sigma = 0.76$.
- Znajdźmy dystrybuantę (CDF) tej zmiennej losowej oraz wyznacz $CDF(2.6)$.
- Innymi słowy, prawdopodobieństwo, że zdrowe dziecko jest co najmniej o 0.9 kg lżejsze od przeciętnego dziecka:

$$P(X < 2.6) = CDF(2.6) = 0.118.$$

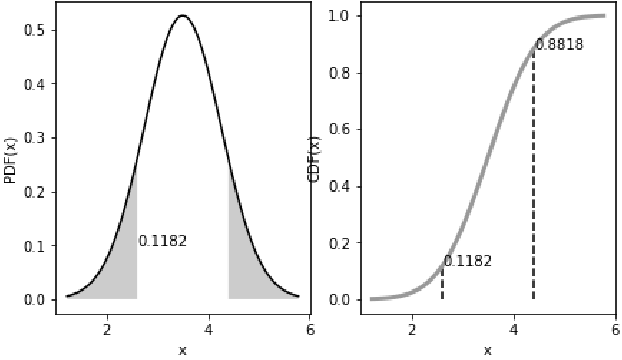
Rozkłady prawdopodobieństwa a testowanie hipotez



Rozkłady prawdopodobieństwa a testowanie hipotez

Aby odpowiedzieć na to pytanie, możemy postępować w następujący sposób:

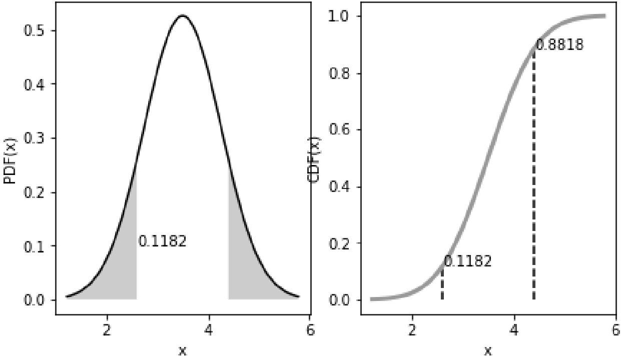
- My zakładamy że zjawisko to jest modelowane rozkładem normalny więc prawdopodobieństwo, że zdrowe dziecko jest co najmniej o 0.9 kg cięższe od przeciętnego dziecka, wynosi również 11.8%.



Rozkłady prawdopodobieństwa a testowanie hipotez

Interpretacja wyników:

- Jeśli dziecko jest zdrowe, prawdopodobieństwo, że jego masa odbiega o co najmniej 0.9 kg od średniej wynosi $2 \cdot 11,8\% = 23,6\% = 0.236$. To nie jest znaczące, więc nie mamy wystarczających dowodów na odrzucenie naszej hipotezy, a nasze dziecko uważa się za zdrowe.



Rozkłady prawdopodobieństwa a testowanie hipotez II

https://github.com/przem85/statistic_4/blob/master/D03_Z03_hypothesis_testing_introduction.ipynb

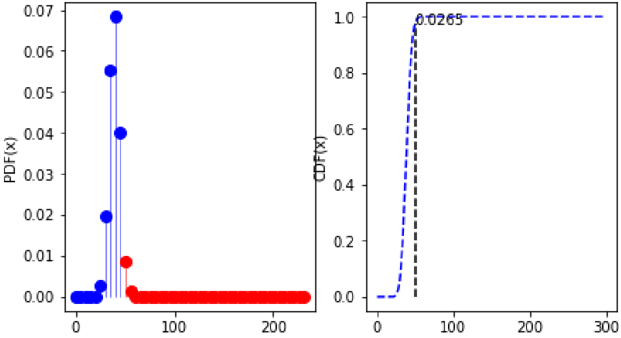
Aby zilustrować związek pomiędzy **rozkładami prawdopodobieństwa** a **testowaniem hipotez**, rozważmy następujący problem:

- Załóżmy, że mamy grę planszową, która zależy od rzutu kostką. Oczywiście jak dostaniemy wynik 6 to poruszamy się najszybciej. W danej grze 6 wypadła 51 razy w ciągu 235 rzutów.
- Jeśli kostka jest uczciwa, oczekiwalibyśmy, że 6 wypadnie $235/6 = 39.17$ razy.
- Czy kostka aby na pewno jest uczciwa?

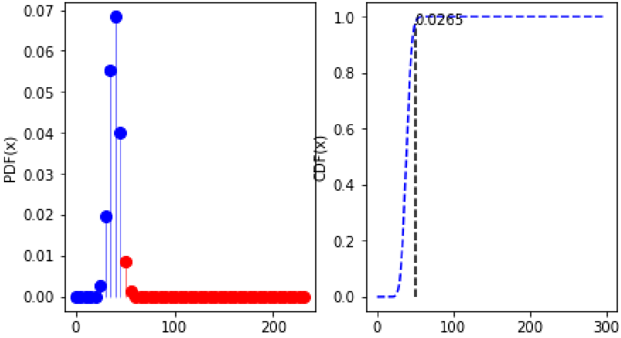
Aby znaleźć odpowiedź na to pytanie:

- skontrujmy rozkład dwumianowy z parametrami $n = 235$ i $p = \frac{1}{6}$. Podobnie jak wcześniej zakładamy, że kostka jest uczciwa.
- obliczymy prawdopodobieństwo wypadnięcia dokładnie 51 razy 6, 52 razy i itd. Następnie dodajmy te wyniki. W ten sposób obliczymy prawdopodobieństwo wypadnięcia dokładnie 51 razy 6 lub wyniku większego $P(X \geq 51)$

Rozkłady prawdopodobieństwa a testowanie hipotez II



Rozkłady prawdopodobieństwa a testowanie hipotez II



W tym przykładzie wynik wynosi 0.0265, co wskazuje, że obserwowanie 51 szóstek jest mało prawdopodobne (poniżej 5%). Kostka najprawdopodobniej nie jest uczciwa.

Statystyki testowe mogą mieć najróżniejsze rozkłady.

Rozkład chi kwadrat

Rozkład chi-kwadrat wiąże się z rozkładem normalnym w prosty sposób: jeśli zmienna losowa X ma rozkład normalny $X \sim N(0,1)$, to X^2 ma rozkład chi-kwadrat z jednym stopniem swobody $X \sim \chi_1^2$.

Rozkład chi kwadrat

Rozkład chi-kwadrat wiąże się z rozkładem normalnym w prosty sposób: jeśli zmienna losowa X ma rozkład normalny $X \sim N(0,1)$, to X^2 ma rozkład chi-kwadrat z jednym stopniem swobody $X \sim \chi_1^2$.

Suma kwadratów n niezależnych zmiennych losowych o standardowym rozkładzie normalnych ma rozkład chi-kwadrat z n stopniami swobody:

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

Rozkład chi kwadrat

Rozkład chi kwadrat – to rozkład zmiennej losowej, która jest sumą n kwadratów niezależnych zmiennych losowych o standardowym rozkładzie normalnym. Liczbę naturalną n nazywa się liczbą stopni swobody rozkładu zmiennej losowej.
Jeżeli ciąg niezależnych zmiennych losowych $X_i \sim N(0,1)$ oraz:

$$Y = \sum_{i=1}^{n} (X_i)^2,$$

to:

$$Y \sim \chi_n^2,$$

czyli słownie: Zmienna losowa Y ma rozkład chi kwadrat o n stopniach swobody.

Rozkład chi kwadrat

Rozkład chi kwadrat ma gęstość

$$f_X(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{gdy } x \geq 0 \\ 0 & \text{gdy } x < 0 \end{cases},$$

gdzie Γ oznacza funkcję Gamma.
https://pl.wikipedia.org/wiki/Funkcja_%CE%93

Rozkład chi kwadrat

https://github.com/przem85/statistic_4/blob/master/D03_Z04_chi_2.ipynb

Zadanie

Proszę napisać skrypt w Pythonie, w którym:

- zdefiniujesz zmienną losową o rozkładzie chi kwadrat,
- narysujesz dla niej gęstość i dystrybuantę,
- wylosujesz próbkę i narysujesz histogram (na jednym rysunku),
- narysujesz kilka gęstości rozkładu chi kwadrat z różnymi parametrami,
- wylosujesz kilka próbek dla zmiennej losowej o rozkładzie chi kwadrat. (Czemu się od siebie różnią?),
- policz skośność i kurtozę dla zdefiniowanej zmiennej.

Rozkład chi kwadrat

Jeżeli $(X_1, \ldots, X_n)$ jest próbka prostą z rozkładu $N(\mu, \sigma^2)$, to zmienna losowa

$$\frac{(n-1)S^{*2}}{\sigma^2},$$

gdzie

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

ma rozkład χ^2_{n-1} (chi kwadrat z $n-1$ stopniami swobody).

Rozkład chi kwadrat

`https://github.com/przem85/statistic_4/blob/master/D03_Z05_chi_2.ipynb`

Zadanie (dodatkowe)

Producent pigułek na ból głowy zobowiązał się dostarczyć pigułki z odchyleniem standardowym $\sigma = 0.05$. Z następnej partii pigułek wybrano próbkę $n = 13$ pigułek o wagach
$3.04, 2.94, 3.01, 3.00, 2.94, 2.91, 3.02, 3.04, 3.09, 2.95, 2.99, 3.10, 3.02$ g.

Pytanie: Czy odchylenie standardowe jest większe niż dozwolone?

Rozkład chi kwadrat

- Stawiamy hipotezę, że odchylenie jest mniejsze od ustalonego σ.

Rozkład chi kwadrat

- Stawiamy hipotezę, że odchylenie jest mniejsze od ustalonego σ.
- Zauważmy, że:

$$\frac{(n-1)S^{*2}}{\sigma^2} = \frac{(n-1)\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}.$$

ma rozkład chi kwadrat.

Rozkład chi kwadrat

- Rozkład chi kwadrat opisuje rozkład sumy kwadratów zmiennych losowych z rozkładu normalnego więc musimy znormalizować nasze dane, zanim obliczymy odpowiednią wartość CDF:

$$SF_{\chi^2_{(n-1)}}\left(\sum\left(\frac{x-\bar{x}}{\sigma}\right)^2\right) = 1 - CDF_{\chi^2_{(n-1)}}\left(\sum\left(\frac{x-\bar{x}}{\sigma}\right)^2\right) = 0.1929$$

Rozkład chi kwadrat

- Rozkład chi kwadrat opisuje rozkład sumy kwadratów zmiennych losowych z rozkładu normalnego więc musimy znormalizować nasze dane, zanim obliczymy odpowiednią wartość CDF:

$$SF_{\chi^2_{(n-1)}}\left(\sum\left(\frac{x-\bar{x}}{\sigma}\right)^2\right) = 1 - CDF_{\chi^2_{(n-1)}}\left(\sum\left(\frac{x-\bar{x}}{\sigma}\right)^2\right) = 0.1929$$

- Jeśli partia pigułek pochodzi z rozkładu z odchyleniem standardowym mniejszym od $\sigma = 0.05$ to prawdopodobieństwo otrzymania większej niż obserwowana wartość chi kwadrat wynosi około 19%.

Rozkład chi kwadrat

- Rozkład chi kwadrat opisuje rozkład sumy kwadratów zmiennych losowych z rozkładu normalnego więc musimy znormalizować nasze dane, zanim obliczymy odpowiednią wartość CDF:

$$SF_{\chi^2_{(n-1)}}\left(\sum\left(\frac{x-\bar{x}}{\sigma}\right)^2\right) = 1 - CDF_{\chi^2_{(n-1)}}\left(\sum\left(\frac{x-\bar{x}}{\sigma}\right)^2\right) = 0.1929$$

- Jeśli partia pigułek pochodzi z rozkładu z odchyleniem standardowym mniejszym od $\sigma = 0.05$ to prawdopodobieństwo otrzymania większej niż obserwowana wartość chi kwadrat wynosi około 19%.
- Innymi słowy, partia pasuje do oczekiwanego odchylenia standardowego.

Rozkład chi kwadrat

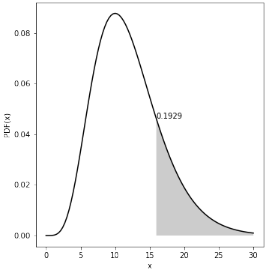
- Rozkład chi kwadrat opisuje rozkład sumy kwadratów zmiennych losowych z rozkładu normalnego więc musimy znormalizować nasze dane, zanim obliczymy odpowiednią wartość CDF:

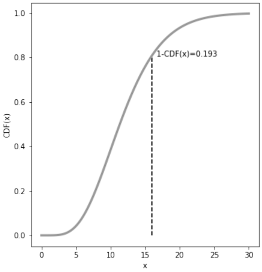
$$SF_{\chi^2_{(n-1)}}\left(\sum\left(\frac{x-\bar{x}}{\sigma}\right)^2\right) = 1 - CDF_{\chi^2_{(n-1)}}\left(\sum\left(\frac{x-\bar{x}}{\sigma}\right)^2\right) = 0.1929$$

- Jeśli partia pigułek pochodzi z rozkładu z odchyleniem standardowym mniejszym od $\sigma = 0.05$ to prawdopodobieństwo otrzymania większej niż obserwowana wartość chi kwadrat wynosi około 19%.
- Innymi słowy, partia pasuje do oczekiwanego odchylenia standardowego.
- Liczba DOF (stopni swobody) wynosi $n-1$, ponieważ interesuje nas tylko gęstość rozkładu, a średnia wartość jest odejmowana od wszystkich punktów danych.

Rozkład chi kwadrat

```
import numpy as np
from scipy import stats
data = np.r_[3.04, 2.94, 3.01, 3.00, 2.94, 2.91, 3.02,
        3.04, 3.09, 2.95, 2.99, 3.10, 3.02]
sigma = 0.05
chi2Dist = stats.chi2(len(data)-1)
statistic = sum( ((data-np.mean(data))/sigma)**2 )
chi2Dist.sf(statistic)
```





Rozkład t-Studenta

- W 1908 roku W. S. Gosset, który pracował dla browaru Guinness w Dublinie, interesował się problemami małych próbek (na przykład właściwości chemicznych jęczmienia, w których rozmiary próbek mogły być tak niskie, jak 3).

Rozkład t-Studenta

- W 1908 roku W. S. Gosset, który pracował dla browaru Guinness w Dublinie, interesował się problemami małych próbek (na przykład właściwości chemicznych jęczmienia, w których rozmiary próbek mogły być tak niskie, jak 3).
- Ponieważ w tych pomiarach nie było znane prawdziwe odchylenie standardowe więc przybliżył je za pomocą standardowego błędu SE.

Rozkład t-Studenta

- W 1908 roku W. S. Gosset, który pracował dla browaru Guinness w Dublinie, interesował się problemami małych próbek (na przykład właściwości chemicznych jęczmienia, w których rozmiary próbek mogły być tak niskie, jak 3).
- Ponieważ w tych pomiarach nie było znane prawdziwe odchylenie standardowe więc przybliżył je za pomocą standardowego błędu SE.
- Stosunek między średnią próbki, a błędem standardowym miał rozkład, który był nieznany do czasu, gdy Gosset pod pseudonimem "Student" rozwiązał ten problem.

Rozkład t-Studenta

- W 1908 roku W. S. Gosset, który pracował dla browaru Guinness w Dublinie, interesował się problemami małych próbek (na przykład właściwości chemicznych jęczmienia, w których rozmiary próbek mogły być tak niskie, jak 3).
- Ponieważ w tych pomiarach nie było znane prawdziwe odchylenie standardowe więc przybliżył je za pomocą standardowego błędu SE.
- Stosunek między średnią próbki, a błędem standardowym miał rozkład, który był nieznany do czasu, gdy Gosset pod pseudonimem "Student" rozwiązał ten problem.
- Szukanym rozkładem prawdopodobieństwa był t-Distribution. Ze względu na pseudonim Gosseta rozkład ten nosi nazwę t-Studenta.

Rozkład t-Studenta (t-Distribution)

Ponieważ w większości przypadków nie jest znana średnia populacji i jej wariancja, zazwyczaj analizuje się dane dotyczące próbek z rozkładem t-Studenta.

Rozkład t-Studenta (t-Distribution)

Ponieważ w większości przypadków nie jest znana średnia populacji i jej wariancja, zazwyczaj analizuje się dane dotyczące próbek z rozkładem t-Studenta.

Rozkład t-Studenta z n stopniami swobody jest rozkładem zmiennej losowej T postaci:

$$T = \frac{U}{\sqrt{Z}}\sqrt{n}$$

gdzie:

- U jest zmienną losową mającą standardowy rozkład normalny N(0,1),
- Z jest zmienną losową o rozkładzie chi kwadrat o n stopniach swobody,
- U i Z są niezależne.

Rozkład t-Studenta

Rozkład t-Studenta z n stopniami swobody ma gęstość:

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1 + \frac{x^2}{n}\right)^{-(\frac{n+1}{2})},$$

gdzie Γ oznacza funkcję Gamma.
https://pl.wikipedia.org/wiki/Funkcja_%CE%93

Rozkład t-Studenta

https://github.com/przem85/statistic_4/blob/master/D03_Z06_t_distribution.ipynb

Zadanie

Proszę napisać skrypt w Pythonie, w którym:

- zdefiniujesz zmienną losową o rozkładzie t-Studenta,
- narysujesz dla niej gęstość i dystrybuantę,
- wylosujesz próbkę i narysujesz histogram (na jednym rysunku),
- narysujesz kilka gęstości rozkładu t-Studenta z różnymi parametrami,
- wylosujesz kilka próbek dla zmiennej losowej o rozkładzie t-Studenta. (Czemu się od siebie różnią?),
- policzysz skośność i kurtozę dla zdefiniowanej zmiennej.

Rozkład t-Studenta

Jeżeli $(X_1, \ldots, X_n)$ jest próbka prostą z rozkładu $N(\mu, \sigma^2)$, to zmienna losowa:

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S^*},$$

gdzie

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

ma rozkład t-Studenta z $n-1$ stopniami swobody.

Rozkład t-Studenta

https://github.com/przem85/statistic_4/blob/master/D03_Z07_t_distribution.ipynb

Zadanie (dla chętnych)

Zmienna losowa X ma rozkład t-Studenta o $n = 15$ stopniach swobody. Oblicz prawdopodobieństwa:

- $P(|X| >= 1,753)$,
- $P(|X| < 2,13)$,
- $P(X >= 2,95)$.