Metody statystyczne – przypomnienie. Statystyki opisowe. Testowanie hipotez.

1. Zainstaluj pakiet **sp**. Wczytaj dane **gleby** zawierające informacje na temat kwasowości gleb pewnego obszaru w Kanadzie. Używając odpowiedniej funkcji wyświetl nazwy kolumn w danych. Przedstaw statystyki dla zmiennej **pH**. Oblicz wartość minimalną, maksymalną, wariancję, odchylenie standardowe oraz kwartyle: I, II i III dla zmiennej **pH**.
2. Histogram jest graficzną reprezentacją rozkładu danych. Wartości danych są łączone w przedziały (oś pozioma) a na osi pionowej jest ukazana liczba punktów (obserwacji) w każdym przedziale. Histogram (jak i tablica częstości) są uzależnione od liczebności klas i wartości początkowej pierwszej klasy. Przedstaw histogram zmiennej pH.
3. Podobną funkcję do histogramu spełnia estymator jądrowy gęstości. Przypomina on wygładzony wykres histogramu i również służy graficznej reprezentacji rozkładu danych. Przedstaw estymator jądrowy gęstości dla zmiennej pH.
4. Wykres Q-Q jest wykresem punktowym dla danych uporządkowanych w szereg rozdzielczy i pozwala on na porównywanie rozkładów dwóch parametrów. Wyświetl wykres kwantyl-kwantyl dla pH. Jak można go interpretować?
5. Dystrybuanta wyświetla prawdopodobieństwo, że wartość zmiennej przewidywanej jest mniejsza lub równa określonej wartości. Przedstaw dystrybuantę dla pH za pomocą funkcji **stat_ecdf().**

Weryfikacja hipotez statystycznych

Statystyczna hipoteza jest założeniem o parametrze populacji. To założenie może się okazać prawdziwe lub nie. Testowanie hipotezy odnosi się do formalnych procedur używanych przez statystyków by przyjąć albo odrzucić hipotezę statystyczną. Najlepszym sposobem aby ustalić czy hipoteza statystyczna jest prawdziwa, byłoby przebadanie całej populacji. Jako, że jednak jest to często niepraktyczne, badacze zazwyczaj badają losową próbkę z populacji. Jeżeli próbka nie potwierdza hipotezy, wtedy hipoteza zostaje odrzucona.

Typy hipotez statystycznych:

- **Hipoteza zerowa** – oznaczana jako H_0, jest zwykle hipotezą, gdzie wyniki z próbek obserwacyjnych wynikają z czystego przypadku.
- **Hipoteza alternatywna** – oznaczano jako H_1, jest hipotezą, gdzie wyniki z próbek obserwacyjnych są zdefiniowane przez jakąś nielosową przyczynę.

Hipotezy statystyczne – błąd pierwszego rodzaju

Z błędem pierwszego rodzaju mamy do czynienia wtedy, kiedy odrzucamy hipotezę zerową, a była ona prawdziwa.

Poziom istotności

Prawdopodobieństwo popełnienia błędu pierwszego rodzaju w sytuacji, gdy hipoteza zerowa była prawdziwa określamy symbolem α (alfa) i nazywamy poziomem istotności. Poziom istotności określa maksymalne ryzyko popełnienia błędu pierwszego rodzaju, jakie jesteśmy skłonni zaakceptować.

Hipotezy statystyczne – błąd drugiego rodzaju
Błąd drugiego rodzaju występuje wtedy, kiedy nie odrzucimy hipotezy fałszywej.
Prawdopodobieństwo popełnienia błędy drugiego rodzaju określamy symbolem **β** (beta).

Moc testu

Moc testu (prawdopodobieństwo, że prawidłowo odrzucimy hipotezę zerową) to **1-β**. Inaczej mówiąc jest to prawdopodobieństwo niepopełnienia błędu II rodzaju.

Przykład- Testowanie o normalności rozkładu

Shapiro-Wilk Test – sprawdza czy losowa próbka pochodzi z populacji o rozkładzie normalnym

H_0- populacja ma rozkład normalny

H_1- populacja nie ma rozkładu normalnego

Kolmogorov-Smirnov Test – sprawdza czy rozkład w populacji dla pewnej zmiennej losowej różni się od założonego rozkładu teoretycznego.

H_0- rozkład w badanej populacji zgadza się z rozkładem teoretycznym

H_1- rozkład w badanej populacji różni się od rozkładu teoretycznego

6. Wykonaj test normalności rozkładu zmiennej pH za pomocą testu Shapiro-Wilka używając funkcji **shapiro.test()**. Odrzuć hipotezę zerową gdy p-value < 5%.
7. Wykonaj test Kolmogorova-Smirnova aby sprawdzić czy zmienna pH ma rozkład normalny. Użyj funkcji **ks.test()** z argumentem ‘pnorm’.