Testowanie hipotez statystycznych- wstęp

Hipoteza zerowa i alternatywna

Hipotezą statystyczną jest dowolne przypuszczenie co do rozkładu populacji generalnej (jego postaci funkcyjnej lub wartości parametrów). Prawdziwość tego przypuszczenia jest oceniana na podstawie wyników próby losowej.

Testem statystycznym nazywamy regułę postępowania, która każdej możliwej próbie przyporządkowuje decyzję przyjęcia lub odrzucenia hipotezy. Oznacza to, że test statystyczny jest regułą rozstrzygającą, jakie wyniki próby pozwalają uznać sprawdzaną hipotezę za prawdziwą, a jakie – za fałszywą.

Testy statystyczne dzielimy zasadniczo na:

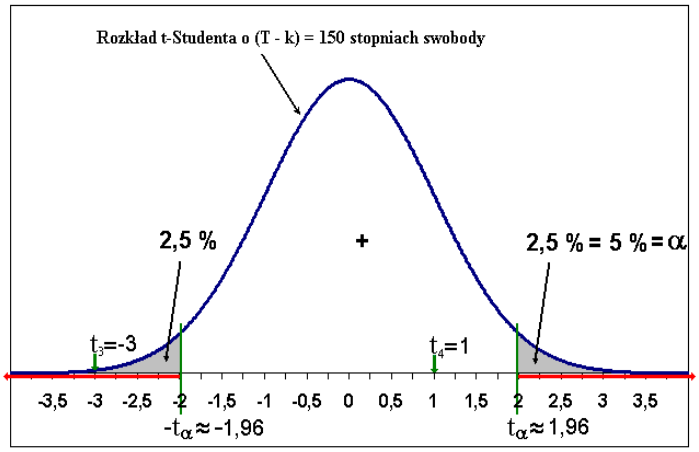
- parametryczne, czyli dotyczące wartości parametrów statystycznych populacji, takich jak np. średnia,
- nieparametryczne, czyli dotyczące postaci rozkładu zmiennej lub losowości próby.

Hipotezę, która podlega weryfikacji nazywamy hipotezą zerową, a jej przeciwieństwo - hipotezą alternatywną.

Każdy test statystyczny rozpoczynamy od sformułowania hipotezy zerowej H_0, czyli hipotezy podlegającej sprawdzeniu, oraz hipotezy konkurencyjnej H_1, którą jesteśmy w stanie przyjąć, gdy odrzucimy hipotezę zerową. Ponadto musimy określić poziom istotności α , czyli maksymalne ryzyko błędu, jakie badacz jest skłonny zaakceptować.

Następnie wybieramy odpowiednią statystykę testową (wybór statystyki uzależniony od informacja jaką posiadamy o próbie oraz od postaci hipotezy zerowej i alternatywnej) oraz obliczamy wartość tej funkcji dla badanej próby. Jeśli prawdopodobieństwo osiągnięcia otrzymanej bądź jeszcze bardziej ekstremalnej wartości statystyki jest niskie to wątpimy, że nasze dane są zgodne z hipotezą zerową i jesteśmy skłonni przyjąć hipotezę alternatywną:

- Jeżeli $p <= \alpha \Rightarrow$ odrzucamy H_0 przyjmując H_1,
- Jeżeli $p > \alpha \Rightarrow$ nie ma podstaw, aby odrzucić H_0.
- Inaczej: jeżeli wartość statystyki wpada do obszaru krytycznego to odrzucamy H_0 przyjmując H_1, w przeciwnym przypadku nie ma podstaw, aby odrzucić H_0.



Uwaga: Testy statystyczne w zależności od wyniku pozwalają nam hipotezę zerową odrzucić i wtedy przyjąć hipotezę konkurencyjną lub nie dają podstaw do odrzucenia H_0, co nie jest równoznaczne z jej przyjęciem. Używając testów, którymi dysponuje MATLAB, należy sprawdzić, jaka hipoteza przyjmowana jest w tym teście.

Testowanie hipotez na temat średniej

Przykład 1. Każda linia komunikacji miejskiej ma określony czas przejazdu (od pętli do pętli). Przeprowadźmy test hipotezy dla wybranej linii. Podany czas na rozkładzie wynosi 28 minut.

$$H_0: \mu=28$$

$$H_1: \mu\neq 28$$

Wybieramy losową próbę 100 przejazdów tej linii i obliczamy średni czas dla tej próby $\bar{x}=31{,}5$ minuty. i odchylenie standardowe próby s = 5 minut. Przy założeniu normalności rozkładu, średnia arytmetyczna próby pobranej z populacji o rozkładzie $N(\mu,\sigma)$ ma rozkład $N(\mu,\frac{\sigma}{\sqrt{n}})$. Jeśli prawdziwa jest hipoteza zerowa, to statystyka o postaci $Z=\frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ ma rozkład $N(0,1)$.

Zatem: $z=\frac{\bar{x}-\mu}{\sigma}\sqrt{n}=\frac{31{,}5-28}{5}\sqrt{100}=7>z_{\alpha/2}=1{,}96$

Dlatego odrzucamy H_0 na poziomie α=0,05.

Do badania hipotez dotyczących średniej służą funkcje: ttest, ttest2, tcdf, tinv. Więcej informacji na temat założeń ttestu będzie można znaleźć w kolejnych listach.

```
[h,p] = ttest(x,m)
tcdf(t,n-1)
```

Ćwiczenie 1: Sprawdź powyższe rachunki korzystając z ttestu.

Wskazówka: Wylosuj próbę z rozkładu normalnego o zadanych parametrach i przeprowadź analizę.

Ćwiczenie 2: Chcemy sprawdzić, czy czas oczekiwania na dostarczenie przesyłki przez pewna firmę kurierską to przeciętnie 3 dni (m = 3). W tym celu z populacji klientów tej firmy wylosowano próbę liczącą 22 osoby i zapisano informacje o ilości dni, jakie minęły od dnia nadania przesyłki do jej dostarczenia, były to następujące wielkości: (1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7). Ilość dni oczekiwania na przesyłkę w badanej populacji spełnia założenie normalności

Ćwiczenie 3: Agencja nieruchomości podała, że ceny gruntu w centrum miasta wzrosły o 49% w ciągu 3 lat. Inwestor chcąc przetestować te dane, znajduje próbę 18 nieruchomości w centrum, dla których zna cenę obecną i sprzed 3 lat. Dla każdej nieruchomości oblicza procentowy wzrost wartości a następnie znajduje średnią i odchylenie standardowe z próby. Statystyki próby wynoszą m=38% i s = 14%. Przeprowadź test na poziomie istotności α = 0.01.

Wskazówka: Statystyka próby jest mała a odchylenie standardowe populacji nieznane, więc należy skorzystać z rozkładu t o n – 1 = 17 stopni swobody (funkcja tcdf). Sprawdź czy takie same wyniki uzyskasz przy zastosowaniu ttest.

Testowanie hipotez na temat wariancji

Jednym z założeń testów parametrycznych (np. t-testu dla dwóch prób) jest homogeniczność wariancji. Homogeniczność możemy tutaj rozumieć jako równość, jednolitość. Dokładniej, porównywane ze sobą grupy, za pomocą testów parametrycznych powinny mieć podobne wariancje. Oznacza to, różnorodność uzyskanych wyników w poszczególnych grupach powinna być podobna. Do badania wariancji w grupie bądź w kilku grupach służą następujące testy:

1. Do testowania hipotez na temat wariancji używamy statystyki chi-kwadrat o n – 1 stopniach swobody (`vartest(x,v)`):

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

   Gdzie s jest wariancją n-elementowej próby wylosowanej z populacji, a σ_0^2 jest wartością wariancji podaną w H_0.

   Test ten pozwala ocenić czy wariancja w próby jest równą założonej wariancji v.

2. Do testowania równości wariancji w dwóch populacjach stosuje się test F (`vartest2(x,y)`):

$$F_{(n_1-1,n_2-1)} = \frac{s_1^2}{s_2^2}$$

gdzie s_1^2 i s_2^2 są wariancjami próby odpowiednio o n_1 i n_2 elementach wylosowanymi z populacji.

3. Przydatna może być również statystyka porównująca wariancję w większej ilości grup np. test Barletta (`vartestn(x,group)`) czy test Levene'a (`Levenetest(x)`)

Ćwiczenie 4: Chcemy sprawdzić, czy odchylenie standardowe w rozkładzie czasu montowania elementu w pralce rzeczywiście wynosi v=1,5 minuty. Wygeneruj 25-elementową próbę rozkładu normalnego o dowolnej średniej i zadanym odchyleniu. Sprawdź stosując odpowiedni test (z właściwie dobranymi parametrami) czy wariancja wygenerowanej próby jest mniejsza niż 1,6min?

Sprawdź, czy wynik testowania zmienia się gdy przyjmiemy poziom istotności równy 0,1?

Ćwiczenie 5: Przypuszcza się, że młodsze osoby łatwiej decydują się na zakup nowych nieznanych produktów. Badanie przeprowadzone wśród przypadkowych 20 nabywców nowego produktu i 22 nabywców znanego już wyrobu pewnej firmy dostarczyło następujących informacji o wieku klientów:

- Nabywcy nowego produktu: średnia 27,7; odchylenie standardowe 5,5.
- Nabywcy znanego produktu: średnia 32,1; odchylenie standardowe 6,3.

Wygeneruj dwa zestawy danych o zadanych parametrach. Czy różnica pomiędzy odchyleniami standardowymi jest statystycznie znacząca? Jak sformułujesz wnioski wynikające z przeprowadzonej analizy?