Statystyczna Analiza Danych – laboratorium

Testowanie hipotez statystycznych

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 4
23/24 marca 2023

Idea zajęć – co i po co będziemy robić?

- Zajmiemy się dziś wnioskowaniem statystycznym
- Warto rozumieć, jak działa wnioskowanie statystyczne i czego możemy oczekiwać od testu statystycznego, żeby chociażby mieć więcej dystansu do podawanych w mediach informacji
- Istnieją co najmniej 3 metodologie wnioskowania statystycznego, w tym labie zajmujemy się najprostszą z nich (z reguły spotykaną)

Hipotezy – **co** jest przedmiotem testu

- Hipoteza statystyczna – przypuszczenie dotyczące populacji (weryfikowane na podstawie próby)
  1. Hipoteza zerowa H_0 – stwierdzenie (falsyfikowalne), które jest przedmiotem testu statystycznego
  2. Hipoteza alternatywna H_1 – konkurencyjna hipoteza (przyjmowana, jeśli odrzucone H_0)
- Dla każdego testu statystycznego musimy sformułować obie

Statystyki testowe i obszary krytyczne – **czy prawdopodobne** to, co mamy

- Statystyka testowa – funkcja danych, na jej podstawie sprawdzamy, czy H_0 jest sprzeczne z zaobserwowanymi danymi
- Statystyka testowa ma swój (teoretyczny) rozkład F przy założeniu prawdziwości H_0
- Sprawdzamy, jak bardzo prawdopodobne jest uzyskanie takiej wartości statystyki testowej, jaką uzyskaliśmy z naszych danych na tle rozkładu teoretycznego
- Obszar krytyczny – służy do podjęcia decyzji, że odrzucamy H_0. Zależy od postaci H_1!
- Wartość krytyczna $k_\alpha : F(k_\alpha) = 1 - \alpha$

Schemat procesu wnioskowania (tradycyjny)

1. Postaw H_0 i H_1
2. Ustal poziom istotności α – jak bardzo godzisz się na odrzucenie prawdziwego H_0 (0,01; 0,05; 0,1)
3. Oblicz wartość statystyki testowej k
4. Znajdź wartość krytyczną k_α dla rozkładu statystyki testowej i w oparciu o H_1 skonstruuj obszar krytyczny
   - k wpada do obszaru krytycznego – H_0 odrzucamy
   - k nie wpada do obszaru krytycznego – nie ma podstaw do odrzucenia H_0

Schemat procesu wnioskowania (p-wartość)

1. Postaw H_0 i H_1
2. Ustal poziom istotności α – jak bardzo godzisz się na odrzucenie prawdziwego H_0 (0,01; 0,05; 0,1)
3. Oblicz wartość statystyki testowej k
4. Znajdź **p-wartość** p na podstawie k i postaci H_1
   - $p \leqslant \alpha$ – H_0 odrzucamy
   - $p > \alpha$ – nie ma podstaw do odrzucenia H_0

Uwaga na p-value!

- *Problemem nie jest samo p-value tylko to, jak badacze z niego korzystają...*
- ASA Statement on p-values (2016): `https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108`
- Moving to a World Beyond $p < 0.05$ (2019) `https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913`
- W obliczu obecnych rozmiarów zbiorów danych próg 0.05 jest z reguły ZNACZNIE za duży, p-value musi być oceniane w kontekście
- (AFAIK!) Alternatywa/wskazówki, co robić są jeszcze polem dyskusji
- Ale nawet gdyby już je ustalono, to niezbędny czas na dostosowania – trzeba zmienić podręczniki, "nauczyć uczących"...

Hipotezy jedno- i dwustronne

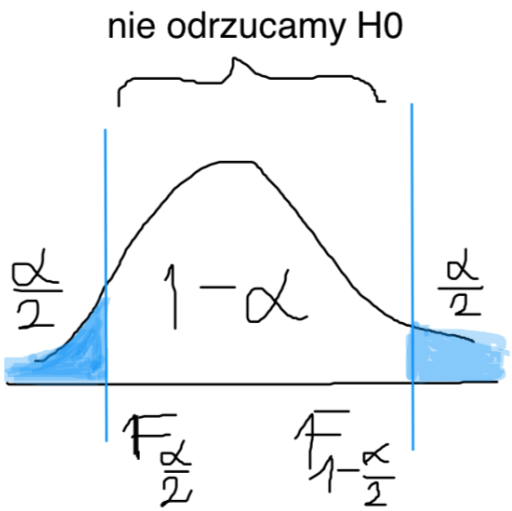
- Jednostronna hipoteza alternatywna zakłada kierunek zachowania się badanego zjawiska (np. $\theta < 5$)
- Dwustronna nie precyzuje kierunku (np. $\theta \neq 5$ – ale nie wiemy, czy mniejsze czy większe)
- Hipotezę dwustronną możemy zapisać jako sumę hipotez jednostronnych: $\theta < 5 \vee \theta > 5$

Obszary krytyczne w zależności od H_1

obszar krytyczny zaznaczony na niebiesko

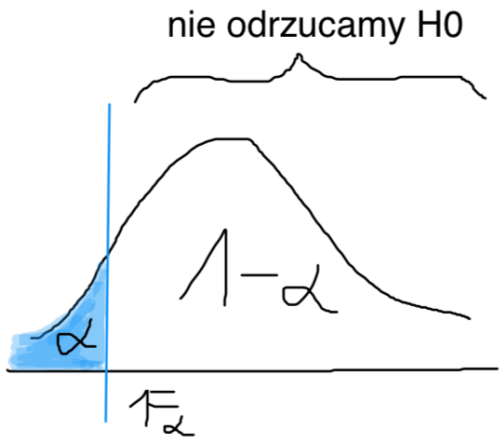
α – poziom istotności, F – dystrybuanta rozkładu statystyki testowej, θ – badany parametr

$H_1 : \theta \neq \theta_0$



$H_1 : \theta > \theta_0$

$H_1 : \theta < \theta_0$



p-wartość (p-value)

- Policzony poziom istotności
- Prawdopodobieństwo, że statystyka testowa osiągnie wielkość większą lub równą wartości uzyskanej z próby
- Dla hipotez jednostronnych $H_1 : \theta > \theta_0 : p = 1 - F(k)$
- Dla hipotez jednostronnych $H_1 : \theta < \theta_0 : p = F(k)$
- Dla hipotez dwustronnych: $p = 2min(F(k), [1 - F(k)])$

(k – wartość statystyki testowej)

Jednopróbkowy test t – znana wariancja

- Najprostszym sposobem porównania średnich jest wykorzystanie testu opartego na statystyce o rozkładzie t-Studenta
- Niech zbiór X ma n obserwacji
- Wówczas przy prawdziwej H_0, że średnia jest równa μ

$$t = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \rightarrow N(0,1)$$

- $\bar{X}$ – średnia arytmetyczna ze zmiennej X (oszacowanie próbkowe), σ – odchylenie standardowe dla zmiennej X.
- H_1 – jedna z trzech możliwości

Jednopróbkowy test t – nieznana wariancja

- Niech zbiór X ma n obserwacji
- Wówczas przy prawdziwej H_0, że średnia jest równa μ

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}} \sqrt{n-1} \rightarrow t(n-1)$$

- $\bar{X}$ – średnia arytmetyczna ze zmiennej X (oszacowanie próbkowe), $\hat{\sigma}$ – nieobciążony estymator odchylenia std dla zmiennej X.
- H_1 – jedna z trzech możliwości

Haczyk #1

- Test t jest testem parametrycznym, wymaga spełnienia pewnych założeń dotyczących rozkładów zmiennych
- Próba X musi pochodzić z rozkładu normalnego. Jeśli to założenie nie jest spełnione, to rozkład statystyki testowej może się różnić od zakładanego
- Założenie o postaci rozkładu **szczególnie** ważne dla prób o niskich liczebnościach
- Co z próbami o wysokiej liczebności? Jakie inne problemy (związane z rozkładem) mogą Państwo dojrzeć?

Haczyk #2

- Przyjmuje się, że minimalna liczebność pojedynczej próby to 30 obserwacji
- Zmienne powinny mieć rozkład (quasi)ciągły. Czemu?
- ... *od kiedy już mówimy o próbie o wysokiej liczebności?*"

Jak sobie poradzić?

- Z wymogiem minimalnej liczebności trudno walczyć.
- Niewłaściwy rozkład zmiennych:
  - Próba o niskiej liczebności – sprawdź obecność outlierów (średnia jest nieodporna na obserwacje odstające). Jak to nie pomaga, zmień test na nieparametryczną alternatywę (np. test U-Manna-Whitneya)
  - Próba o wysokiej liczebności, rozkład jednomodalny – prawdopodobnie możesz rozważyć interpretację testu (CTG!). Jeśli masz wątpliwości – rozważ nieparametryczną alternatywę

Zadanie 3

- Zweryfikuj hipotezę, że średnie zadłużenie gminy w województwie mazowieckim jest równe 25% przy hipotezie alternatywnej, że jest mniejsze.
- Samodzielnie oblicz wartość statystyki testowej oraz p-value
- Porównaj uzyskane wyniki z wynikami funkcji `t.test`
- Utwórz przedział ufności dla badanej średniej. Porównaj z wynikiem uzyskanym w teście.

Test istotności dla wariancji

- Najprostszym sposobem porównania wariancji jest wykorzystanie testu opartego na statystyce o rozkładzie χ^2
- Niech zbiór X ma n obserwacji
- Wówczas przy prawdziwej H_0, że wariancja jest równa σ_0^2

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

- S_n^2 – nieobciążony estymator wariancji dla zmiennej X.
- H_1 – jedna z trzech możliwości

Zadanie 4

- Wybierz dane dotyczące województw łódzkiego i pomorskiego
- Przy założeniu, że rozkład zadłużenia jest normalny, przetestuj hipotezę, że wariancja zadłużenia w każdej z tych gmin jest równa $\sigma_0^2 = 226$, przy hipotezie alternatywnej $H_1 : \sigma^2 \neq 15$.
- Oblicz samodzielnie wartość statystyki testowej i p-value.
- Porównaj wyniki z wynikami funkcji `EnvStats::varTest`

Test t dla prób niezależnych

- Niech zbiór X ma n obserwacji, a zbiór Y m obserwacji
- Wówczas przy prawdziwej H_0 o równości średnich

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma^2_X}}{n} + \frac{\hat{\sigma^2_Y}}{m}}} \rightarrow t(n + m - 2)$$

- $\bar{X}$ – średnia arytmetyczna ze zmiennej X (oszacowanie próbkowe), $\hat{\sigma^2_X}$ – nieobciążony estymator wariancji dla zmiennej X. Analogicznie dla Y
- H_1 – średnie nie są sobie równe (test dwustronny)

Zadanie 5

- Wybierz dane dotyczące województw łódzkiego i pomorskiego
- Przy założeniu, że rozkład zadłużenia jest normalny, przetestuj hipotezę, że średnie zadłużenie w tych województwach jest sobie równe, przeciwko hipotezie alternatywnej, że jest różne
- Oblicz samodzielnie wartość statystyki testowej i p-value.
- Porównaj wyniki z wynikami funkcji `t.test`. Zwróć uwagę na argument `var.equal`