Politechnika Wrocławska

Lab 1

Statystyka w zastosowaniach

04.04.2025

Mateusz Nasewicz

Spis treści

1 Analiza wpływu liczebności próby na oszacowania średniej i odchylenia standardowego

Celem tego eksperymentu było zbadanie, jak liczebność próby wpływa na trafność oszacowań parametrów populacji — średniej i odchylenia standardowego.

Opis populacji

Wygenerowano populację składającą się ze 100 000 elementów z rozkładu normalnego o średniej $\mu = 50$ oraz odchyleniu standardowym $\sigma = 10$.

Opis prób

Z populacji wylosowano próbki o liczebności $n = 10$, $n = 50$ oraz $n = 1000$. Dla każdej liczebności wylosowano wiele próbek, a następnie obliczono średnie oraz odchylenia standardowe tych próbek.

Wyniki

Na poniższym rysunku przedstawiono histogramy rozkładów średnich (lewy wykres) oraz odchyleń standardowych (prawy wykres) próbek.

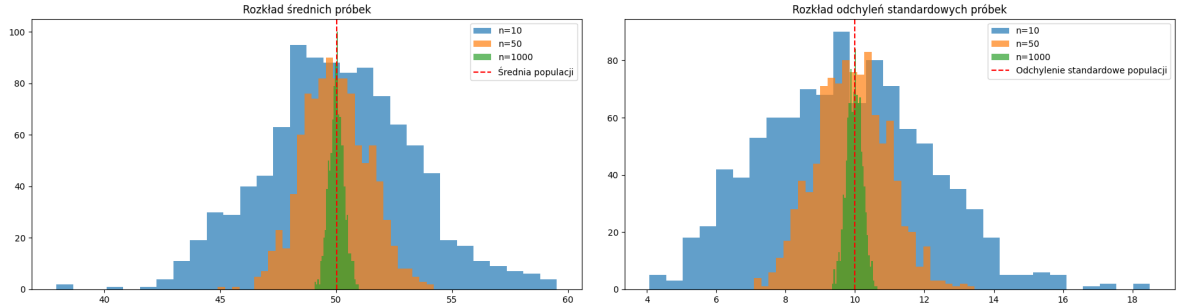

Fig... 1: Rozkład średnich (lewo) oraz odchyleń standardowych (prawo) próbek o różnych liczebnościach ($n = 10$, $n = 50$, $n = 1000$). Czerwona przerywana linia przedstawia prawdziwą średnią i odchylenie standardowe populacji.

Wnioski

Na podstawie wykresów można zaobserwować, że:

- Dla małych prób (np. $n = 10$), rozrzut wartości średnich i odchyleń standardowych jest większy — oszacowania są mniej precyzyjne.
- Wraz ze wzrostem liczby obserwacji ($n = 50$ i szczególnie $n = 1000$), rozkłady średnich i odchyleń standardowych stają się węższe i bardziej skoncentrowane wokół wartości populacyjnych.
- Średnie próbek przy dużych n zbliżają się do średniej populacji, co ilustruje działanie prawa wielkich liczb.
- Podobnie, odchylenia standardowe próbek coraz lepiej odwzorowują rzeczywiste odchylenie populacji.

Otrzymane wyniki potwierdzają, że większe próby dostarczają lepszych (bardziej precyzyjnych i mniej zmiennych) oszacowań parametrów populacyjnych.

2 Wpływ liczebności próby na wiarygodność rezultatów

Celem tej części analizy było zbadanie, jak liczebność próby wpływa na szerokość przedziału ufności dla średniej populacji.

Opis populacji

Wygenerowano populację liczącą 100 000 elementów z rozkładu normalnego $\mathcal{N}(100, 15)$.

Opis prób

Z populacji losowano próbki o różnych liczebnościach: $n = 10$, $n = 50$ oraz $n = 500$. Dla każdej liczebności wyznaczono średnią oraz 95% przedział ufności dla średniej na podstawie wielu powtórzeń (np. 1000 prób).

Wizualizacja wyników

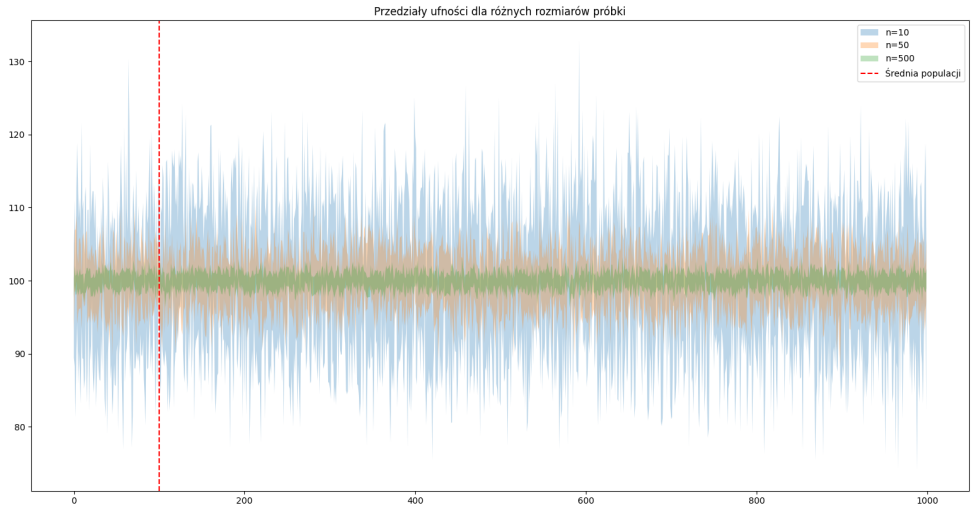

Fig... 2: Przedziały ufności dla średnich z próbek o różnych liczebnościach ($n = 10$, $n = 50$, $n = 500$). Czerwona przerywana linia przedstawia rzeczywistą średnią populacji ($\mu = 100$).

Wnioski

Na podstawie wykresu można zauważyć wyraźny wpływ liczebności próby na szerokość przedziałów ufności:

- Dla najmniejszych prób ($n = 10$), przedziały ufności są najszersze — występuje największa niepewność w oszacowaniu średniej populacyjnej.
- Dla średnich prób ($n = 50$), szerokość przedziałów znacznie się zmniejsza, a większość z nich nadal zawiera prawdziwą średnią.
- Przy dużej liczbie obserwacji ($n = 500$), przedziały ufności są bardzo wąskie i skoncentrowane wokół średniej populacji.

- Zmniejszająca się szerokość przedziałów wraz ze wzrostem n potwierdza teoretyczne oczekiwania wynikające z centralnego twierdzenia granicznego i własności estymatora średniej.

Podsumowując, im większa próba, tym bardziej precyzyjne oszacowania średniej oraz węższe przedziały ufności, co przekłada się na większą wiarygodność rezultatów statystycznych.

3 Wpływ reprezentatywności próby na wyniki analizy

Celem tej części eksperymentu było zbadanie, w jaki sposób brak reprezentatywności próby wpływa na oszacowanie parametrów populacyjnych, takich jak średnia i odchylenie standardowe.

Opis populacji

Wygenerowano populację o bimodalnym rozkładzie — składającą się z dwóch podpopulacji pochodzących z dwóch różnych rozkładów normalnych. Dzięki temu uzyskano wyraźnie dwuszczytowy rozkład.

Parametry całej populacji:

- Średnia: 50.00
- Odchylenie standardowe: 11.19

Opis prób

Pobrano dwie różne próbki:

- **Próbka losowa** — zawiera dane z całej populacji, dobrane losowo.
- **Próbka nielosowa** — zawiera dane tylko z jednej z podpopulacji (jednego ze szczytów rozkładu), przez co nie odzwierciedla pełnej struktury populacji.

Porównanie wyników

Rodzaj danych	**Średnia**	**Odchylenie standardowe**
Populacja (referencyjna)	50.00	11.19
Próbka losowa	50.40	10.80
Próbka nielosowa	39.62	4.83

Tabelka 1: Porównanie statystyk populacji z próbą losową i nielosową

Wnioski

Analiza pokazuje, że:

- Próbka losowa dobrze odwzorowuje właściwości populacji – zarówno średnia, jak i odchylenie standardowe są bardzo zbliżone do rzeczywistych wartości.
- Próbka nielosowa, pochodząca tylko z jednej części populacji, znacznie zaniża zarówno średnią, jak i odchylenie standardowe.
- Brak reprezentatywności prowadzi do poważnych błędów w estymacji parametrów populacyjnych i może skutkować błędnymi wnioskami w analizie statystycznej.

Wniosek: reprezentatywność próby jest kluczowa dla wiarygodnych i rzetelnych analiz — losowy dobór obserwacji minimalizuje ryzyko systematycznych zniekształceń wyników.

4 Wpływ metody pobierania próbek na oszacowania statystyczne

Celem tej części było porównanie różnych metod pobierania próbek i ich wpływu na estymację parametrów populacyjnych, takich jak średnia oraz odchylenie standardowe.

Opis populacji

Wygenerowano populację liczącą 100 000 elementów, przy czym zastosowano rozkład normalny jako bazowy model danych.

Metody pobierania próbek

W analizie zastosowano trzy różne techniki:

- **Losowanie proste** – dane losowano z całej populacji przy pomocy funkcji `numpy.random.choice()`.
- **Losowanie warstwowe** – dane podzielono na grupy (warstwy), z których losowano proporcjonalnie.
- **Losowanie systematyczne** – dane pobierano w stałych odstępach (np. co 10. element).

Wyniki porównania

Metoda	Średnia	Odchylenie standardowe
Prosta	50.01	10.23
Warstwowa	49.97	9.75
Systematyczna	50.85	10.12

Tabelka 2: Porównanie oszacowań średnich i odchyleń standardowych dla różnych metod pobierania próbek

Wnioski

Z przedstawionych wyników można wyciągnąć następujące obserwacje:

- Wszystkie trzy metody dają zbliżone wyniki, jednak metoda systematyczna nieco zawyżyła średnią — może to być wynikiem struktury danych w populacji (np. występowania regularności).
- Losowanie warstwowe dało najbardziej precyzyjne (najmniejsze) odchylenie standardowe, co potwierdza jego skuteczność w kontrolowaniu zmienności, gdy populacja jest heterogeniczna.
- Metoda losowania prostego jest ogólnie skuteczna i uniwersalna, jednak może nie być optymalna, gdy dane są silnie zróżnicowane.

Podsumowując, wybór metody pobierania próbek może istotnie wpływać na dokładność estymacji – warto dostosować strategię do charakterystyki populacji, zwłaszcza gdy istnieją naturalne podziały na grupy.

5 Statystyki opisowe i interpretacja wyników

W tej części analizy wykorzystano rzeczywisty zbiór danych z biblioteki `seaborn`, zawierający informacje o rachunkach i napiwkach w restauracji. Analizie poddano trzy zmienne: `total_bill` (wartość rachunku), `tip` (napiwek) oraz `size` (liczba osób przy stoliku).

Podstawowe statystyki opisowe

	total_bill	tip	size
Liczba obserwacji	244	244	244
Średnia	19.79	3.00	2.57
Odchylenie standardowe	8.90	1.38	0.95
Min	3.07	1.00	1.00
25% kwartyl (Q1)	13.35	2.00	2.00
Mediana (Q2)	17.80	2.90	2.00
75% kwartyl (Q3)	24.13	3.56	3.00
Maksimum	50.81	10.00	6.00

Tabelka 3: Podstawowe statystyki opisowe dla zmiennych w zbiorze `tips`.

Wizualizacja danych

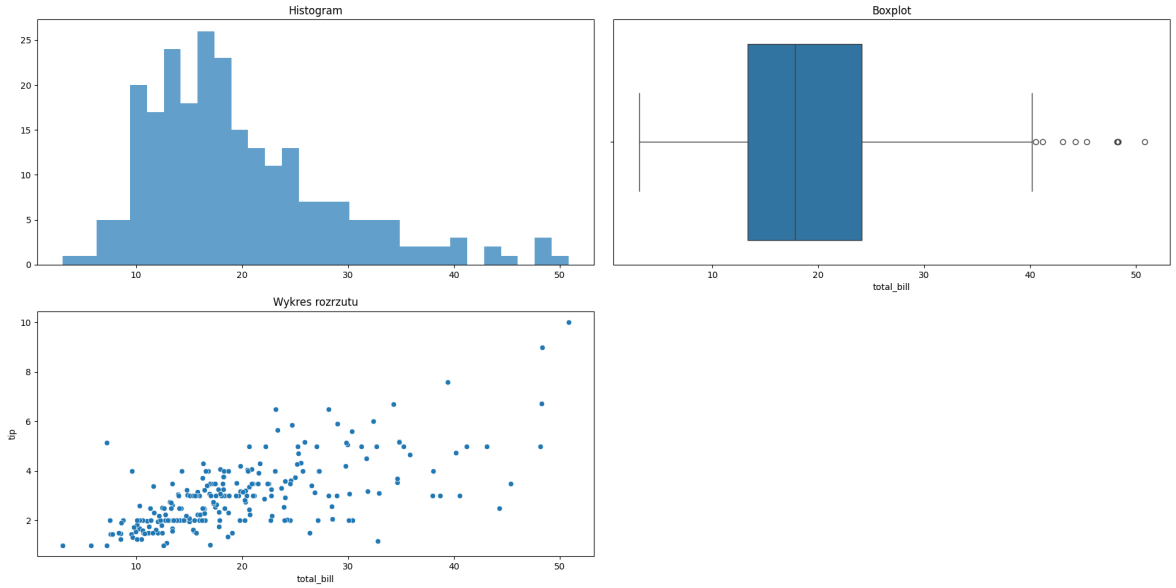

Fig... 3: Histogram, boxplot oraz wykres rozrzutu dla zmiennych `total_bill` i `tip`.

Interpretacja wyników

- **Rozkład wartości rachunku (`total_bill`)** jest prawoskośny – większość wartości koncentruje się wokół niższych kwot, a kilka wysokich rachunków tworzy „ogon” po prawej stronie.
- **Wartości odstające** są widoczne zarówno na wykresie pudełkowym, jak i w danych – szczególnie rachunki przekraczające 40 USD oraz napiwki powyżej 6 USD.
- **Wykres rozrzutu** wskazuje na pozytywną korelację pomiędzy wysokością rachunku a napiwkiem – większe rachunki zazwyczaj skutkują wyższymi napiwkami.
- **Zmienne `tip` i `size`** również wykazują niewielką skośność. Zmienna `size` ma wartości całkowite, najczęściej od 1 do 4, co odpowiada typowej liczbie osób przy stoliku.

Podsumowując, dane są reprezentatywne dla typowej restauracji, jednak niektóre wartości odstające (duże rachunki lub wysokie napiwki) mogą mieć wpływ na wyniki analizy statystycznej. Użycie miar odpornych na odstępstwa (np. mediany, kwartyle) jest więc uzasadnione.