Analiza Wielowymiarowa

Statystyczne podstawy

Paweł Strawiński

Uniwersytet Warszawski

Zajęcia 2
12 października 2023

Rodzaje skal pomiarowych

- Skala nominalna
- Skala porządkowa (rangowa)
- Skala przedziałowa
- Skala ilorazowa
- Skala absolutna

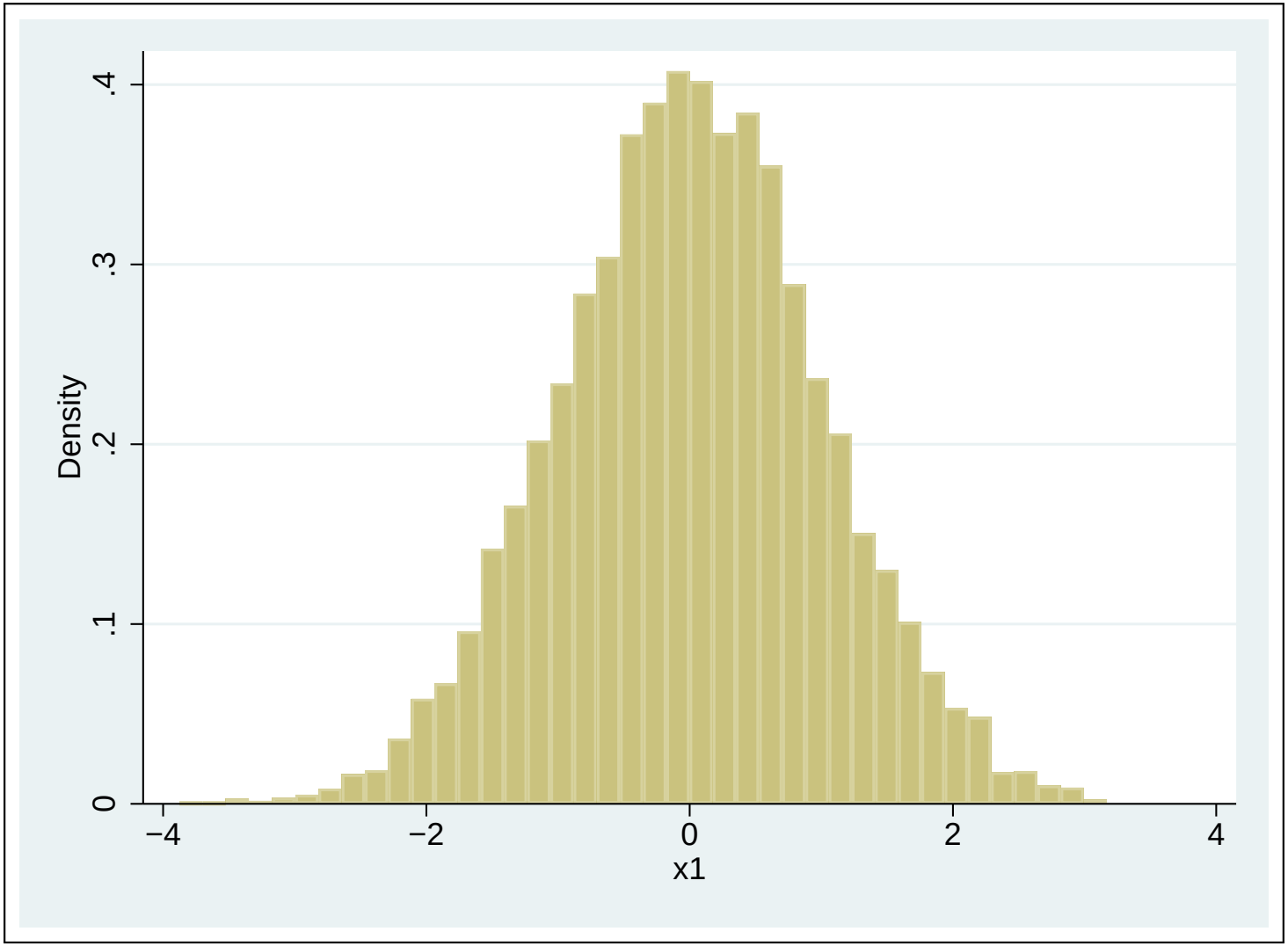
Skala Likerta

1. Nie zgadzam się
2. Raczej nie zgadzam się
3. Nie mam zdania
4. Raczej zgadzam się
5. Zgadzam się

Rozkład normalny

- Rozkład normalny jest jednym z najważniejszych rozkładów prawdopodobieństwa
- Jest często wykorzystywany w naukach przyrodniczych i społecznych do modelowania zmiennych losowych o wartościach rzeczywistych, których rozkłady nie są znane
- Przyczyną jest częstość występowania w naturze
- Z Centralnego Twierdzenia Granicznego wynika, że suma niezależnych czynników losowych ma rozkład zbliżony do rozkładu normalnego
- Cały rozkład opisany jest przez dwa parametry: średnią μ i odchylenie standardowe σ
- Rozkład oznaczany jest $X \sim N(\mu, \sigma^2)$

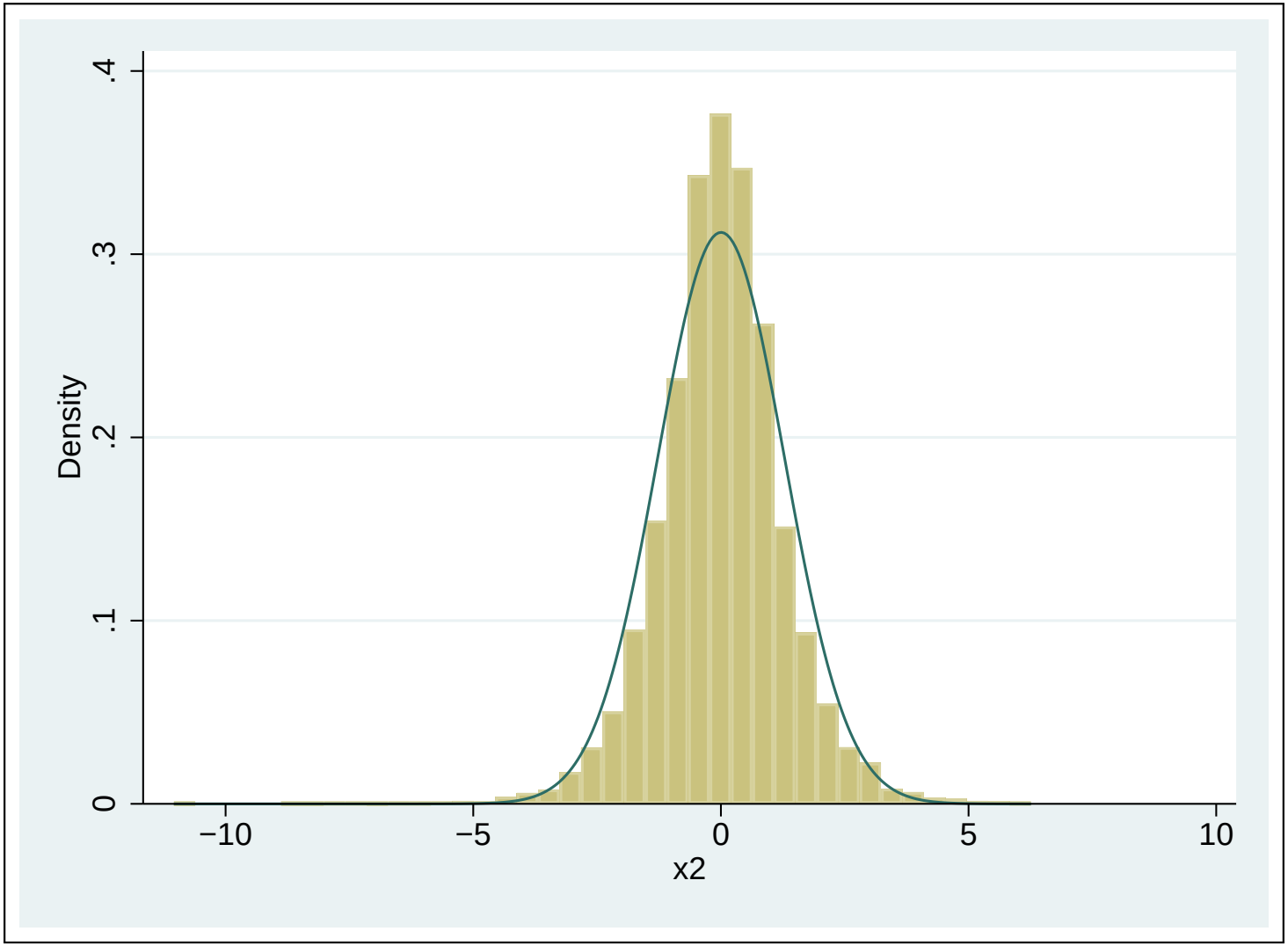
Rozkład normalny



Rozkład t-Studenta

- Rozkład opracowano dla oszacowania przedziału ufności dla średniej, gdy prawdziwa wartość średniej i wariancji nie jest znana
- Gdy liczba stopni swobody jest niewielka rozkład ma grubsze ogony i niższą kurtozę niż rozkład normalny
- Gdy liczba stopni swobody rośnie, gęstość rozkładu t-Studenta staje się podobna do gęstości rozkładu normalnego
- Rozkład jest stosowany w estymacji przedziałowej, testach parametrycznych, oraz w testach istotności parametrów statystycznych – dla prób o małej liczbie obserwacji, mniejszej niż 30

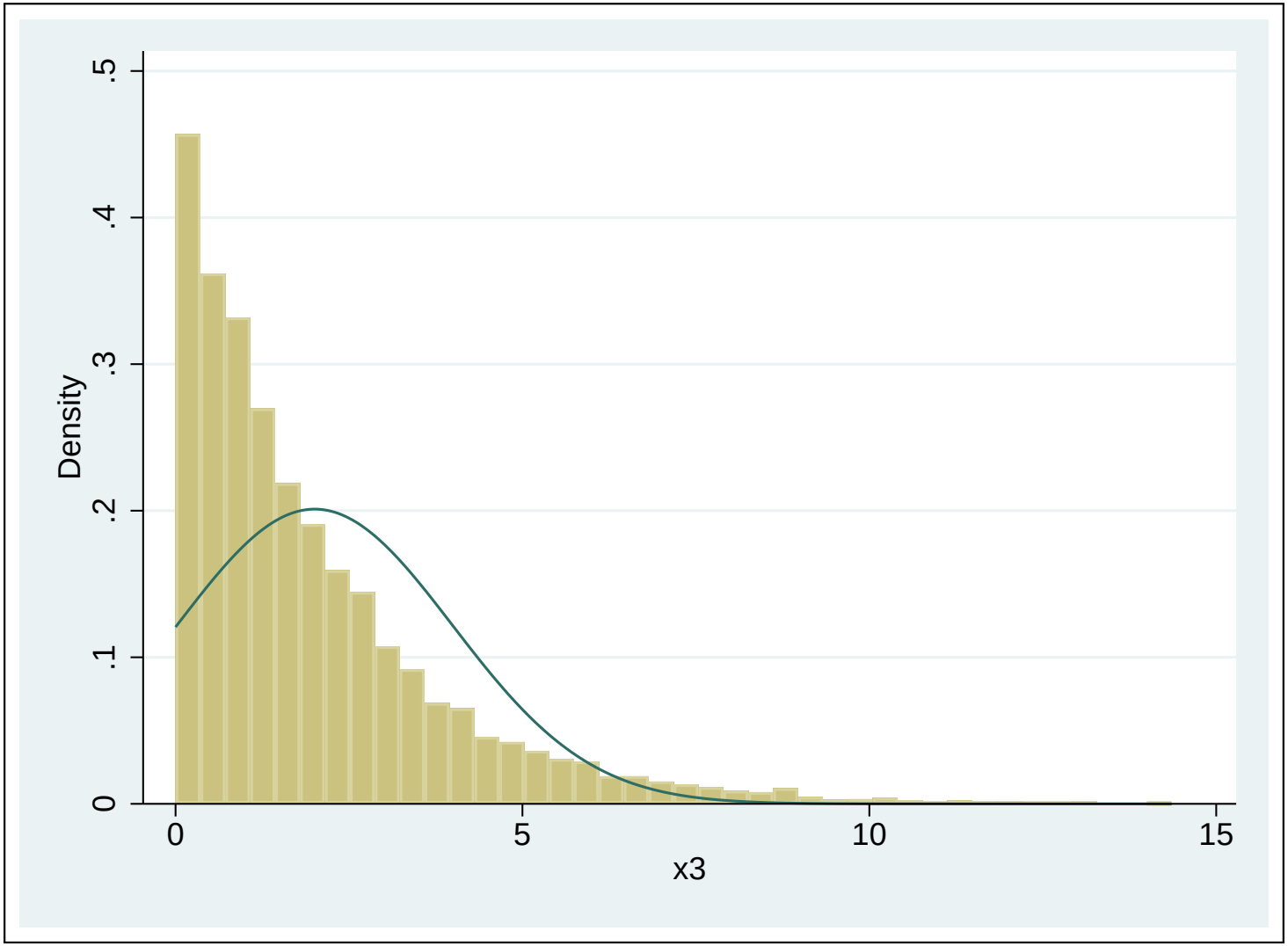
Rozkład t-Studenta



Rozkład chi-kwadrat

- Rozkład chi-kwadrat z k stopniami swobody to rozkład sumy kwadratów k niezależnych zmiennych losowych o standardowym rozkładzie normalnym
- Rozkład jest wykorzystywany w testach niezależności rozkładów i testach zgodności z zadanym rozkładem
- Średnia rozkładu wynosi k, wariancja $2k$.

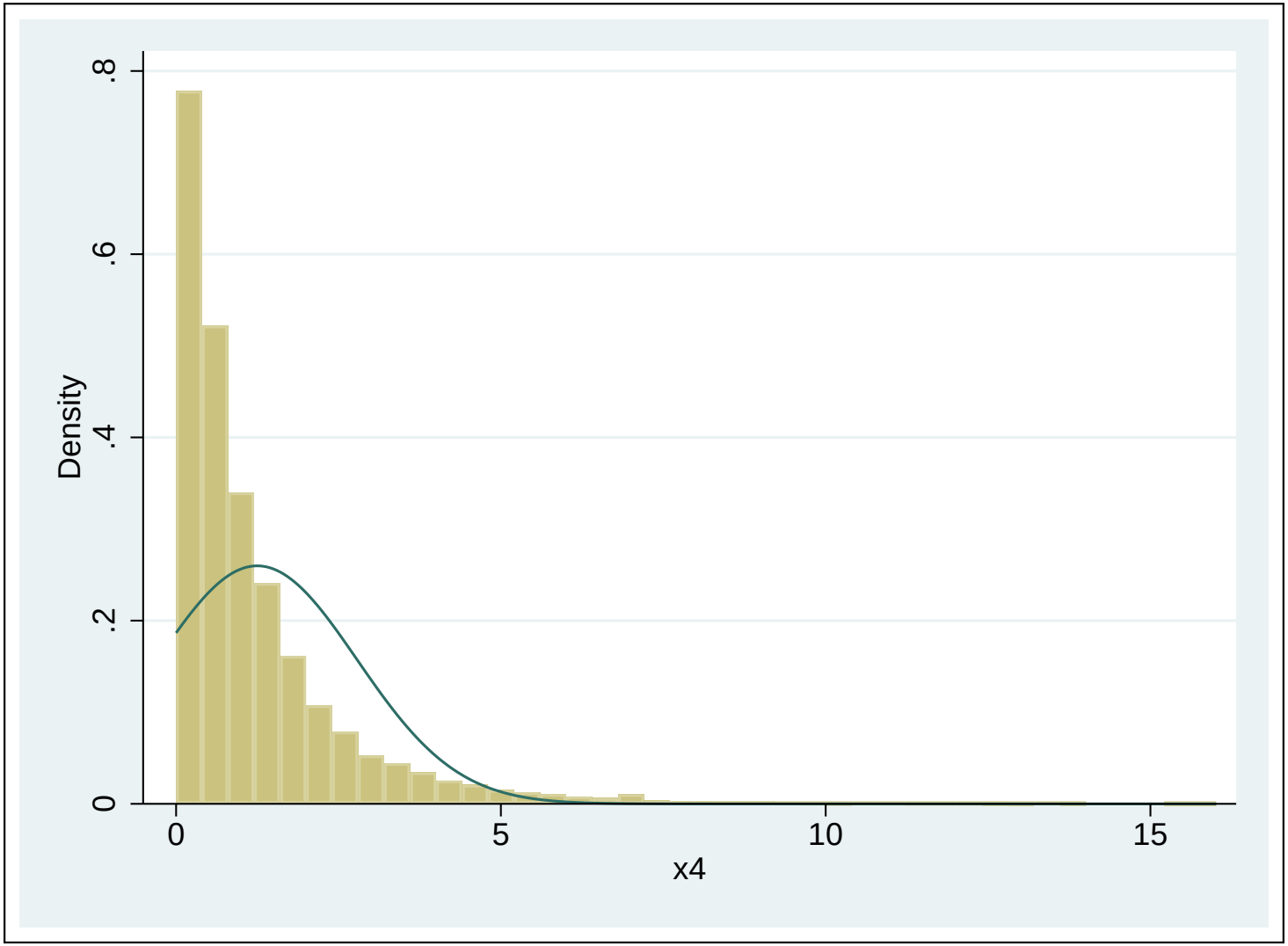
Rozkład chi-kwadrat



Rozkład F

- Rozkład F Fishera-Snedecora jest zdefiniowany jako iloraz dwóch niezależnych zmiennych losowych o rozkładzie chi-kwadrat
- Rozkład F również można zdefiniować jako rozkład kwadratu zmiennej losowej o rozkładzie t-Studenta.
- Jest wykorzystywany w testach statystycznych

Rozkład F



Średnia i wariancja

- Średnia arytmetyczna to inaczej pierwszy moment zwykły, nazywana jest również wartością przeciętną lub wartością oczekiwaną
- Średnia arytmetyczna jest miarą położenia rozkładu i jednocześnie miarą tendencji centralnej
- Wariancja to inaczej drugi moment centralny
- Wariancja jest miarą rozproszenia wartości cechy wokół jej wartości średniej

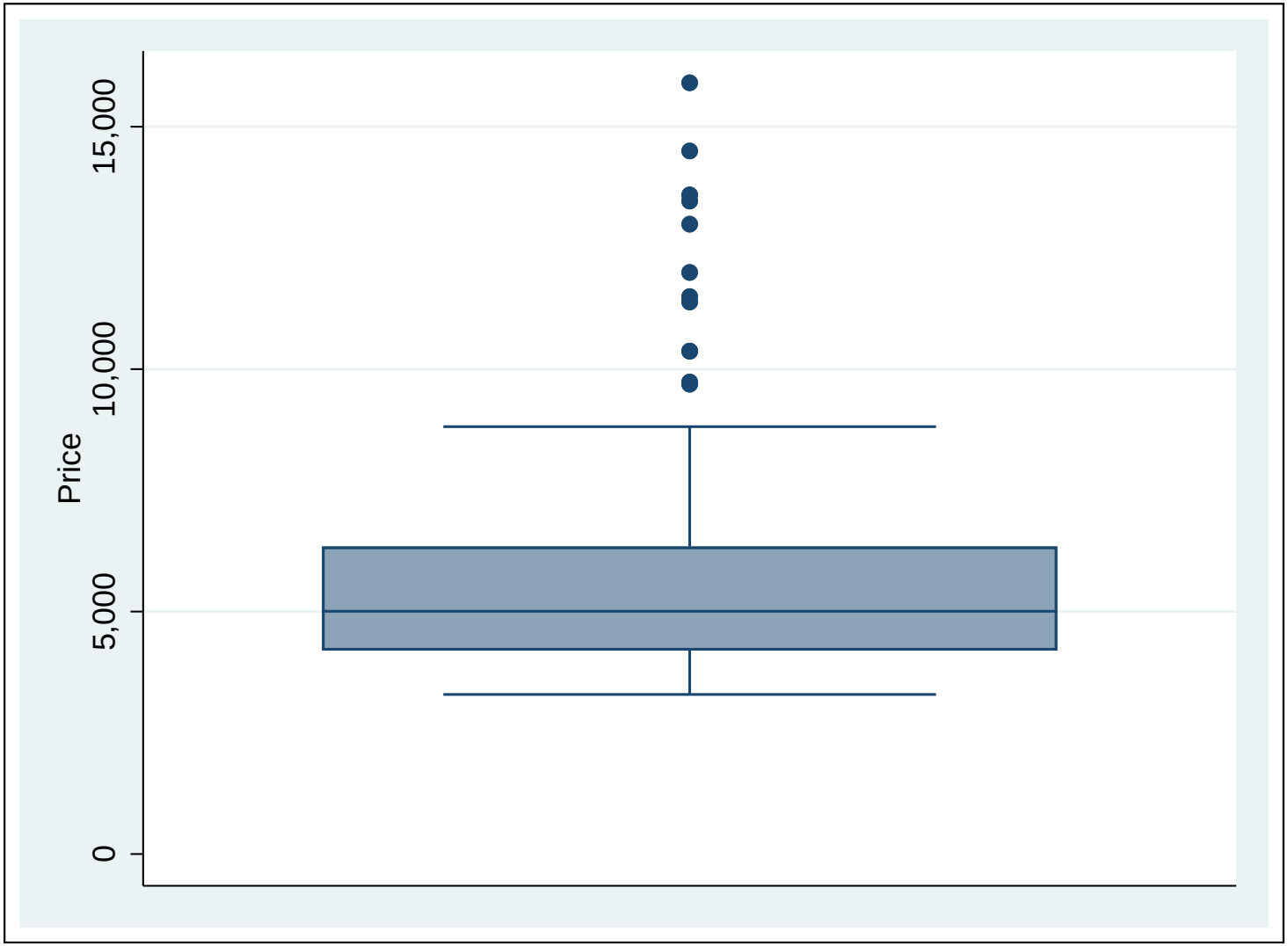
Kwantyle

- Kwantyle to punkty przecięcia dzielące dziedzinę rozkładu prawdopodobieństwa na ciągłe przedziały o równych prawdopodobieństwach lub dzielące obserwacje w próbie na podzbiory o równej liczebności
- Liczba kwantyli jest o jeden mniejsza liczba utworzonych grup
- Często wykorzystywane kwantyle mają specjalne nazwy, takie jak kwartyle (cztery grupy), decyle (dziesięć grup) i percentyle (100 grup). Tworzone grupy są określane jako połówki, ćwiartki itp., chociaż często terminy dotyczące kwantyla są używane do tworzonych grup, a nie do punktów cięcia

Rozstęp międzykwantylowy

- Rozstęp międzykwantylowy (ćwiartkowy) to różnica między trzecim a pierwszym kwartylem
- Pokazuje zróżnicowanie wartości cechy
- Graficznym przedstawieniem pierwszego kwartla, mediany i trzeciego kwartyla oraz 95% przedziału ufności jest wykres pudełkowy (ang. *boxplot*)

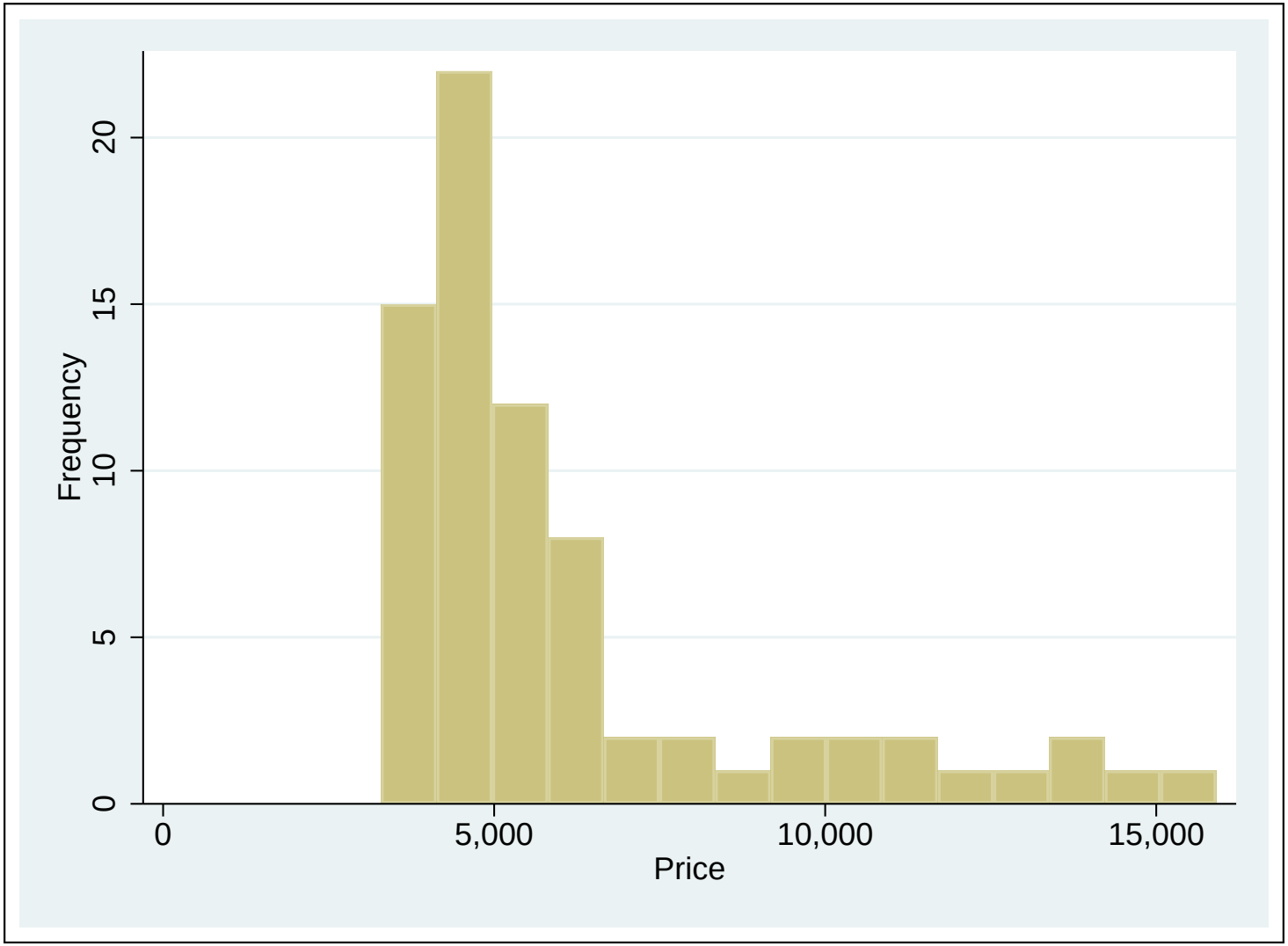
Wykres pudełkowy



Histogram

- Histogram jest to graficzny sposób przedstawiania rozkładu empirycznego wartości cechy
- Składa się z prostokątów umieszczonych na osi współrzędnych
- Szerokość prostokątów jest stała i wyznacza podział na przedziały klasowe wartości cechy
- Wysokość prostokątów jest określona przez liczebności lub częstości elementów należących do określonego przedziału klasowego

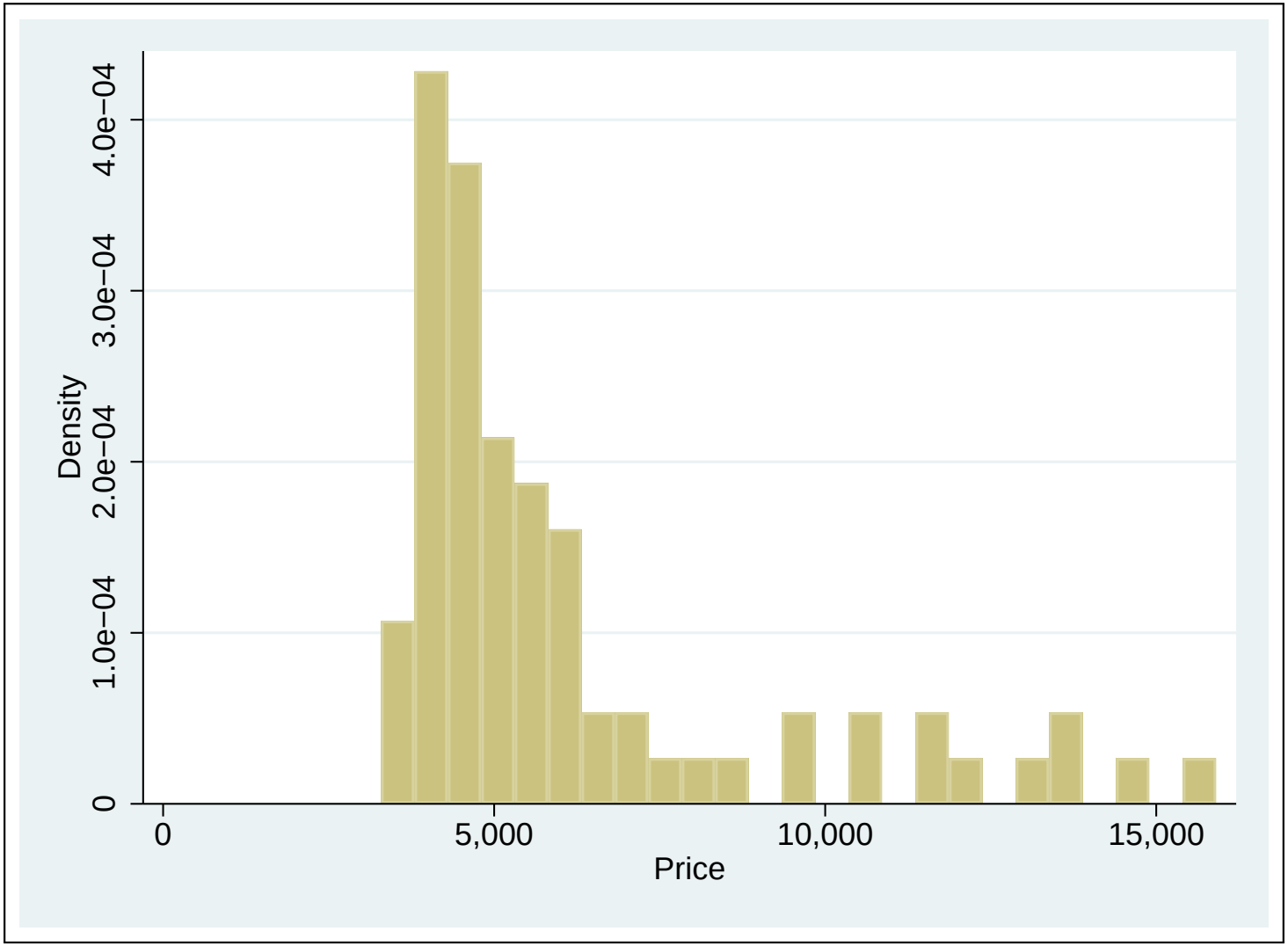
Histogram



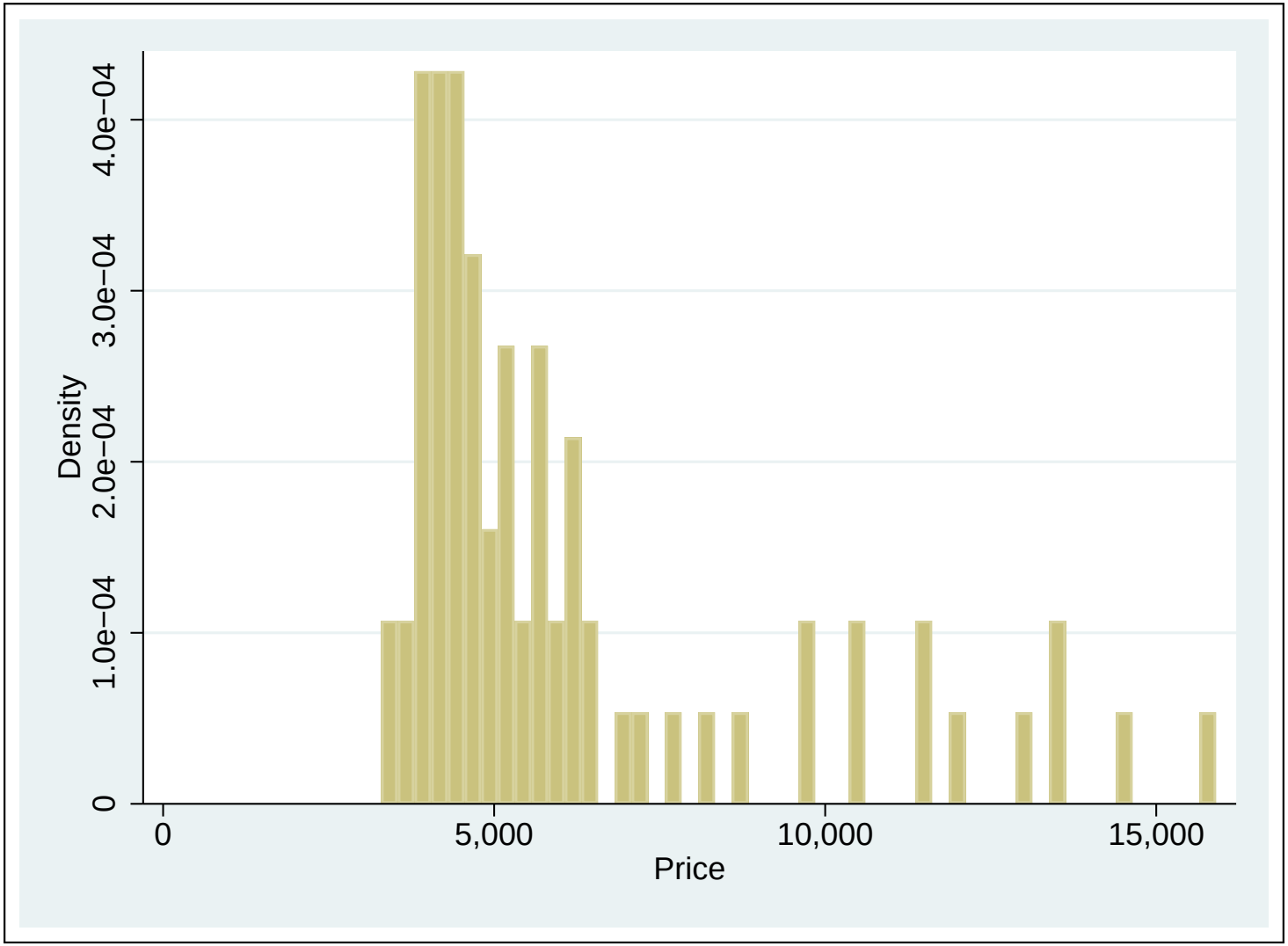
Estymator jądrowy

- Estymator jądrowy gęstości jest to nieparametryczny estymator przeznaczony do wyznaczania gęstości rozkładu zmiennej losowej na podstawie próby
- Nie wymagana wiedzy *a priori* o typie występującego rozkładu
- Najprostszym nieparametrycznym estymatorem jądrowym gęstości jest histogram
- Estymator jądrowy w pewnym stopniu przypomina odpowiednio wygładzony wykres histogramu o małej szerokości słupków

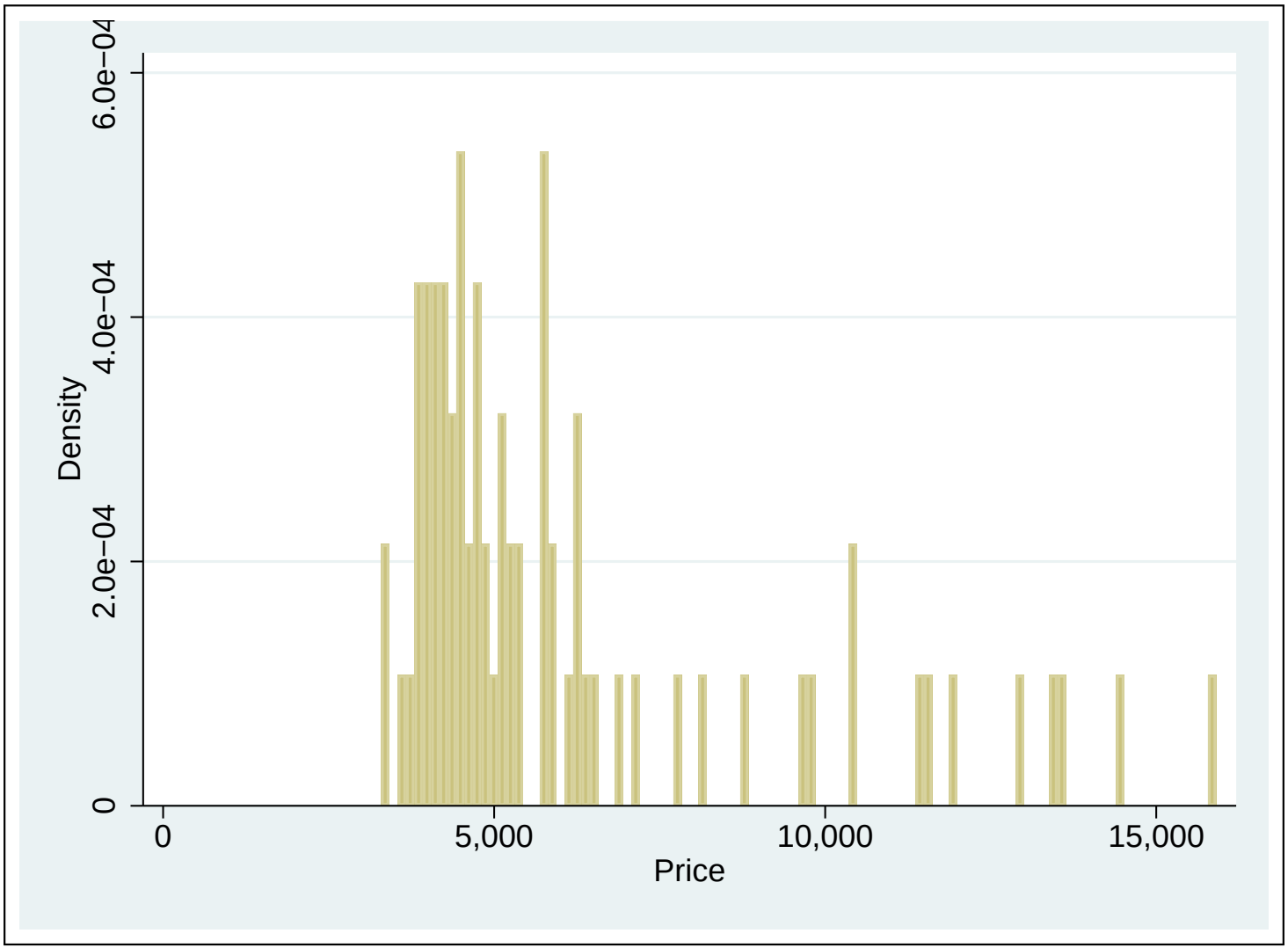
Estymator jądrowy



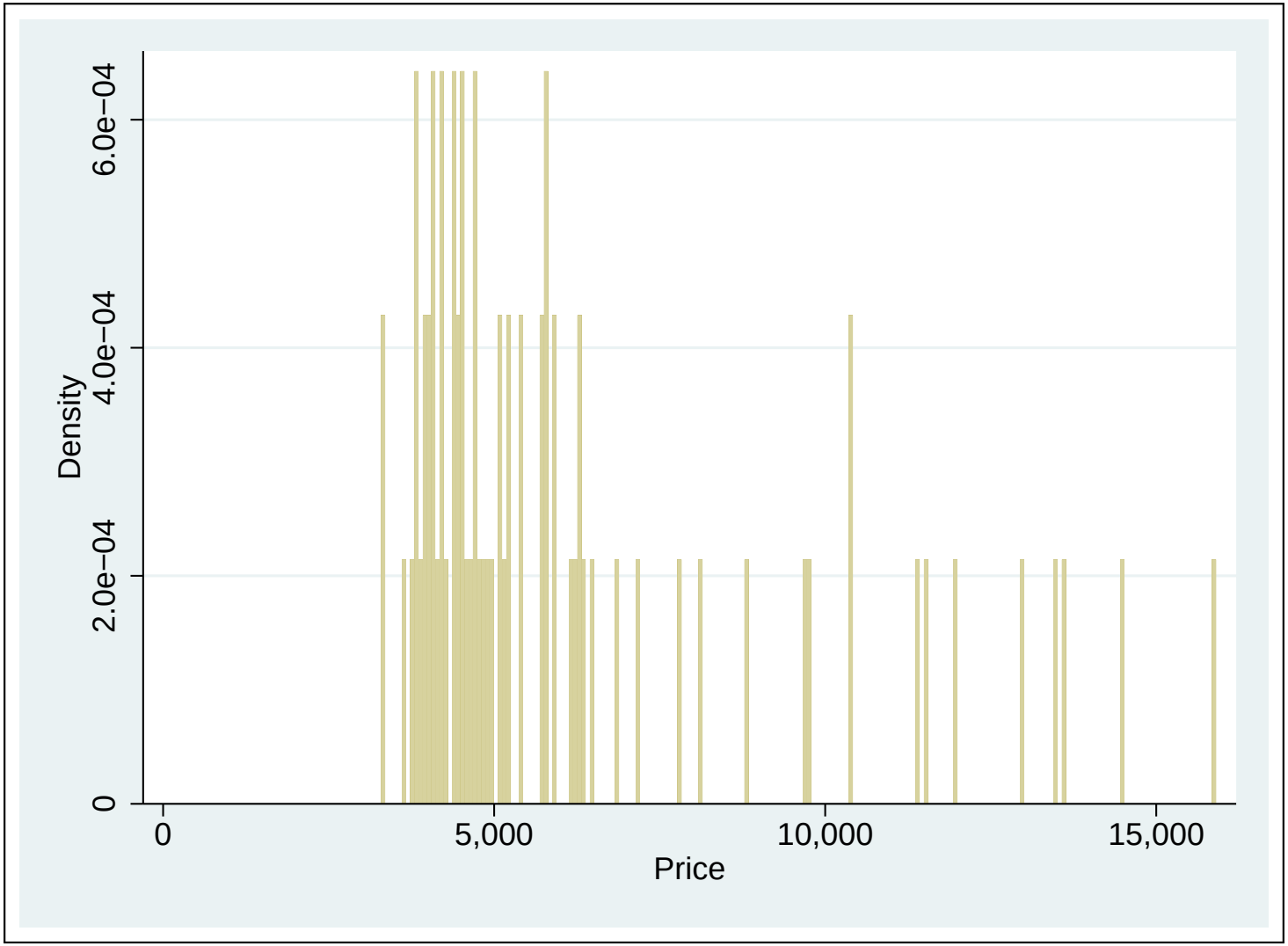
Estymator jądrowy



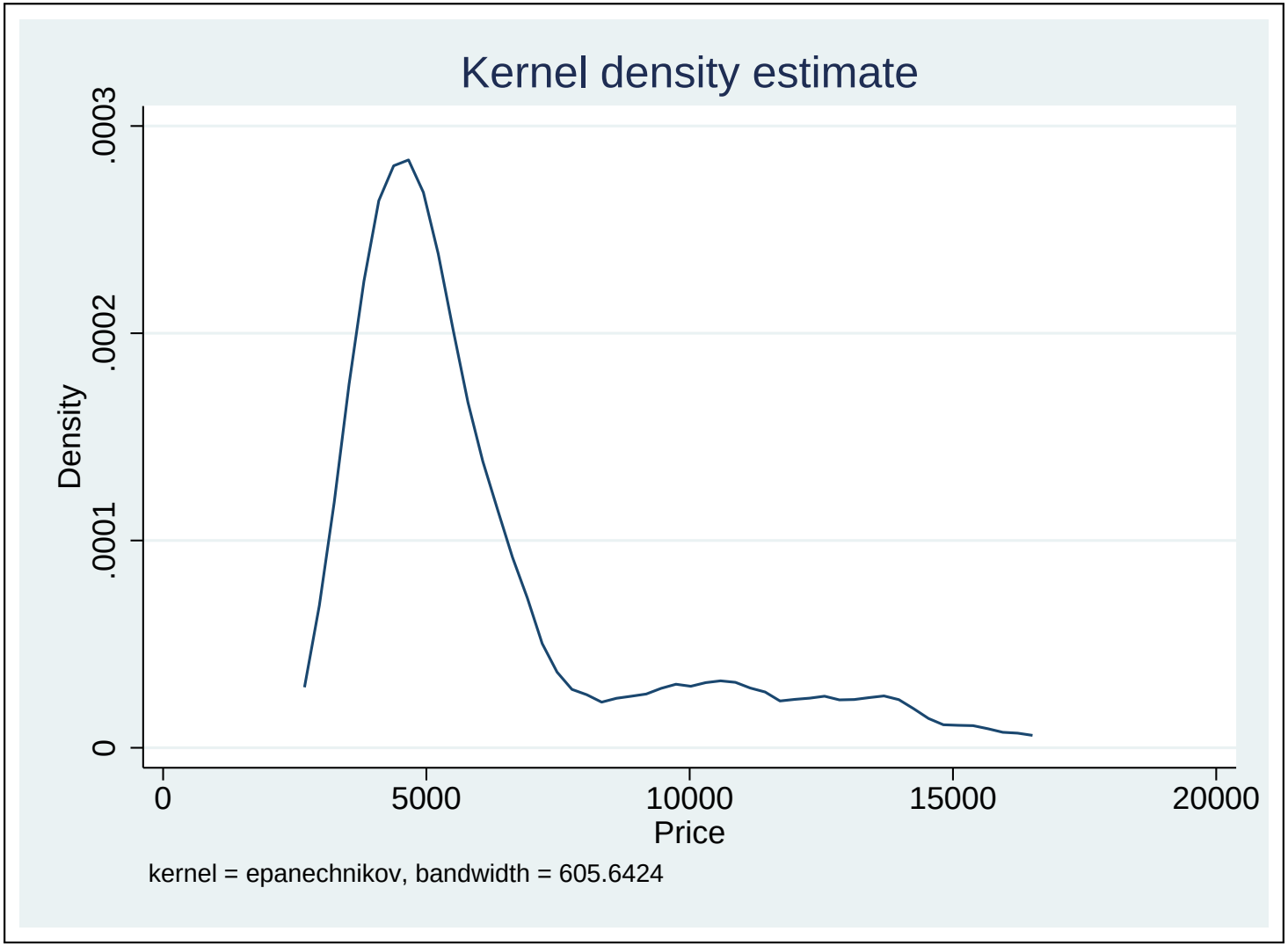
Estymator jądrowy



Estymator jądrowy



Estymator jądrowy



Jednowymiarowa tabela częstości

- Jednowymiarowa tabela rozkładu częstości to rozkład wartości cechy w próbie (populacji)
- Każdy wpis w tabeli zawiera liczbę wystąpień wartości lub częstotliwość w określonej grupie lub przedziale
- W ten sposób tabela podsumowuje rozkład wartości w próbce.

samochód	liczba	częstość
krajowy	52	70,3%
zagraniczny	22	29,7%
razem	74	100%