Rozkłady empiryczne

Statystyka opisowa

- Niech $x = (x_1, x_2, \dots , x_n)'$ będzie próbką, tzn. $x_1, x_2, \dots, x_n$ będą obserwacjami zmiennej losowej X
- Zadaniem **statystyki opisowej** jest prezentacja rozkładu zmiennej X w próbce x (**rozkładu empirycznego**), za pomocą tabeli lub wykresu
- Często wystarczy podać tylko kilka liczb (**parametry rozkładu**) charakteryzujących ten rozkład

Metody opisu rozkładu empirycznego:

- Tabelaryczny
- Graficzny
- Statystyki opisowe

Statystyki opisowe

- Klasyczne – uśredniające wartości próbki, np. momenty zwykłe (rzędu r:
$$m_r = \frac{1}{n}\sum_{k=1}^{n} x_k^r$$
- Pozycyjne – bazujące na posortowanych (rosnąco) wartości próbki, np. dolny kwartyl:
$$Q_1 = \frac{1}{2}\left(x_i + x_j\right)$$
gdzie
$$i = \left\lceil\frac{n+1}{4}\right\rceil, j = \left\lceil\frac{n}{4}\right\rceil$$
lub górny kwartyl:
$$Q_3 = \frac{1}{2}\left(x_i + x_j\right)$$
gdzie
$$i = \left\lceil\frac{3}{4}(n+1)\right\rceil, j = \left\lceil\frac{3}{4}n\right\rceil$$

Statystyki opisowe

Charakterystyki tendencji centralnej rozkładu empirycznego:

- Średnia

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

- Mediana

$$Me = x_i, \qquad i = \frac{n+1}{2}, \qquad n \ -\text{nieparzyste}$$

$$Me = \frac{1}{2}\left(x_i + x_j\right), \qquad i = \frac{n}{2}, \qquad j = \frac{n}{2} + 1, \qquad n \ -\text{parzyste}$$

Statystyki opisowe

Charakterystyki rozrzutu rozkładu empirycznego:

- Odchylenie standardowe

$$s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

- Współczynnik zmienności

$$v = \frac{s}{\bar{x}}100\%$$

Funkcje w R związane ze statystyką opisową

- **table** – szereg rozdzielczy (liczebności)
- **prop.table** – szereg rozdzielczy (proporcje, częstości)
- **cut** – dla cechy ilościowej ciągłej podział na przedziały klasowe
- **barplot** – wykres słupkowy (cecha jakościowa lub ilościowa dyskretna)
- **pie** – wykres kołowy (cecha jakościowa lub ilościowa dyskretna)
- **hist** – histogram (cecha ilościowa ciągła)
- **mean** – średnia z próby
- **median** – mediana z próby
- **sd** – odchylenie standardowe z próby

Model statystyczny

Model statystyczny

- Jeżeli próba $x = (x_1, x_2, \dots, x_n)'$ jest **reprezentatywna**, to stanowi ona podstawę do wnioskowania o **populacji**, z której pochodzi
- Wnioskowanie takie wymaga zbudowania **modelu** "zachowania się" zmiennej (cechy) X w populacji
- Budowa modelu polega na przyjęciu założenia o rozkładzie (teoretycznym) zmiennej X w populacji oraz traktowaniu obserwacji jako wartości tej zmiennej
- Zatem, budując model statystyczny traktujemy wektor
$$x = (x_1, x_2, \dots, x_n)'$$
jako realizacja wektora losowego
$$X = (X_1, X_2, \dots, X_n)'$$
z nieznanego (lub częściowo nieznanego) rozkładu

Model statystyczny

- **Statystyka**, to każda (mierzalna) funkcja próby. Zatem w modelu statystyka jest wielkością losową
- Modelujemy wyniki doświadczenia w którym dokonujemy n niezależnych obserwacji badanej zmiennej X na losowo wybranych z populacji jednostkach eksperymentalnych

Model statystyczny – przykład 1

- W celu określenia czasu bezawaryjnej pracy urządzeń po wykonaniu kapitalnego remontu, wybrano 50 urządzeń i obserwowano czas ich bezawaryjnej pracy. Wyniki (w godz.) są następujące:
  629, 325, 215, ... ,612, 841, 492.
- Budując model statystyczny tego eksperymentu zakładamy, że (w populacji) czas bezawaryjnej pracy urządzenia (cecha X) ma rozkład wykładniczy z nieznanym parametrem λ
- Model ten ma jeden parametr: λ

Model statystyczny – przykład 2

- Przeprowadzono 50 niezależnych eksperymentów polegających na hamowaniu badanego typu samochodu wyposażonego w nowy typ układu hamulcowego (na suchym asfalcie, przy prędkości 40 km/h, itd.).
- Notowano długość drogi hamowania z dokładnością do jednego centymetra. Otrzymane wyniki to:
  18.66, 17.81, 18.96, ... ,17.62, 18.61, 17.99
- Budując model statystyczny tego eksperymentu zakładamy, że (w populacji) długość drogi hamowania (cecha X) ma rozkład normalny z nieznanymi parametrami μ i σ^2

Model statystyczny – przykład 2

- Model ten często zapisujemy w następującej postaci:

$$X_i = \mu + \varepsilon_i, \qquad i = 1,2,\dots,n$$

  gdzie:
- X_i – i-ta obserwacja badanej cechy X
- μ – wartość oczekiwana (średnia) badanej cechy X
- ε_i – błędy (reszty) – niezależne zmienne losowe o jednakowym rozkładzie $N(0, \sigma^2)$
- Model ten ma dwa parametry: μ i σ^2