# Statystyka opisowa

Prof. PK Dr Marek Malinowski

**Wykład kończy się oceną**. **Na test należy przyjść z laptopem** (nie tablet, nie telefon), ponieważ test zostanie przeprowadzony na platformie MOODLE.
Pozytywna ocena z wykładu potwierdza uzyskanie przez studenta efektu uczenia się:

1. "EP-1: Student zna i rozumie w zaawansowanym stopniu wybrane zagadnienia z zakresu statystyki opisowej przydatne do formułowania i rozwiązywania zadań praktycznych związanych z informatyką oraz ma wiedzę w zakresie interpretowania wyników tych analiz."

   Wykład - skala ocen:
   - 51%-60% ocena 3.0,
   - 61%-70% ocena 3.5,
   - 71%-80% ocena 4.0,
   - 81%-90% ocena 4.5,
   - 91%-100% ocena 5.0
2. Drugi efekt uczenia się EP-2 (zgodny z sylabusem) zostanie oceniony w ramach Ćwiczeń.

# Podstawowe pojęcia

**Statystyka** (łac. *status* - *państwo*, *statisticus* - *polityczny*) to nauka zajmująca się zbieraniem, opisem i analizą danych w celu identyfikacji i ilościowego wyrażenia prawidłowości zjawisk masowych.

**Zjawiska masowe** powtarzają się często, badane na dużej liczbie jednostek (o podobnych, ale niekoniecznie tych samych właściwościach) wykazują pewne charakterystyczne dla siebie tendencje, np. urodzenia, wypadki samochodowe, formy spędzania wolnego czasu.

Badaniami statystycznymi objęte są określone zbiorowości statystyczne-**populacje generalne**, np. osoby, rzeczy, zjawiska.

Elementy populacji nazywamy **jednostkami statystycznymi**. Bywa, że w populacji można je wyodrębnić na kilka sposobów, np. w badaniu demograficznym jednostką statystyczną może być osoba, rodzina, gospodarstwo domowe.

Jednostki statystyczne mogą mieć różne właściwości, które podlegają obserwacji statystycznej. Nazywamy je **cechami statystycznymi**, np. w przypadku populacji mieszkańców Polski można mówić o cechach takich jak płeć, wiek, wzrost, kolor oczu itp.

Niektóre cechy mają charakter ilościowy (wyrażony liczbami), np. wiek, wzrost, i nazywamy je **cechami mierzalnymi**, inne mają charakter jakościowy (wyrażony słowami), np. płeć, kolor oczu, i te nazywamy **cechami niemierzalnymi**.

Wartości cechy uzyskane z obserwacji lub pomiaru nazywamy wynikami, obserwacjami lub **danymi**.

**Przejdź na moment $--> $ Prezentacja danych.**

Jednostki statystyczne różnią się na ogół wartościami badanej cechy. Badanie zróżnicowania wartości cechy w populacji jest istotnym elementem każdego **badania statystycznego**.

Rozróżnia się **dwa zasadnicze rodzaje badań statystycznych**, które pozwalają na pozyskanie danych, a mianowicie:

1. **pełne** (obejmują wszystkie jednostki statystyczne populacji, np. spisy, ewidencje),
2. **częściowe** (obejmują wybrane celowo lub losowo jednostki statystyczne z populacji, tzw. **próby**)
   - **ankietowe** (dane zbierane są za pomocą ankiet przeprowadzonych wśród ściśle określonych osób, zazwyczaj mają charakter panelowy, powtarzane w pewnych odstępach czasu, dotyczą tej samej zbiorowości i tego samego problemu, firmy CBOS, Pentor+OBOP=TNS Polska),
   - **monograficzne** (szczegółowy opis i analiza wybranej jednostki lub grupy jednostek, np. badanie warunków życia grup),
   - **reprezentacyjne** (oparte na próbie pobranej z populacji w sposób losowy, reprezentatywny w tym sensie, że każda jednostka ma różne od zera prawdopodobieństwo znalezienia się w tej **próbie losowej**).

Zbieranie, opis i analiza danych z wykorzystaniem badania reprezentacyjnego jest przedmiotem statystyki. Jednak **w badaniu częściowym występują błędy** wynikające stąd, że struktura próby pod względem badanych cech odchyla się (w sposób trudny do określenia) od odpowiedniej struktury w populacji generalnej. Wnioskowanie z próby niekoniecznie musi być prawdziwe dla całej populacji. Powstaje pytanie: w jakim stopniu wnioskowanie prawdziwe dla próby wolno uogólnić na własności dla całej populacji? W tym uogólnianiu istotną rolę odgrywa **rachunek prawdopodobieństwa** i oparte na nim modele. Zastosowanie ich umożliwia **określenie popełnionego błędu** i rozwiązuje podstawowy problem, że niemal każda próba daje różne wyniki.

**Pobieranie próby losowej ze skończonej populacji (podstawowe schematy losowania)**.

1. **Losowanie proste zależne** (bez zwracania): losujemy po kolei $m$ jednostek z populacji, przy czy za każdym razem wylosowana jednostka nie bierze udziału w kolejnym dalszym losowaniu oraz prawdopodobieństwo w każdym losowaniu dowolnego, dostępnego elementu jest takie samo dla każdego z nich.
2. **Losowanie systematyczne**: załóżmy, że liczebność populacji $n$ jest podzielna przez $m$ i $q = \frac{n}{m} > 1$. Do próby wybieramy co $q$-ty element począwszy od pewnego miejsca.
3. **Losowanie warstwowe**: populację dzieli się na subpopulacje (podpopulacje), zwane warstwami, a następnie z poszczególnych subpopulacji niezależnie pobiera się w sposób losowy elementy do próby. Stosuje się w niejednorodnych populacjach.

4. **Losowanie proste niezależne** (ze zwracaniem): nadajemy każdej jednostce statystycznej jednakowe prawdopodobieństwo dostania się do próby. Jednostka raz wylosowana do badania bierze udział w dalszym losowaniu. Na tym schemacie losowania oparta jest cała teoria klasycznego wnioskowania statystycznego. Modelem takiego schematu losowania może być losowanie kul z urny i odkładanie ich z powrotem do urny. Próba uzyskana tą metodą nosi nazwę **próby losowej prostej**.

W praktyce pobieranie próby prostej z populacji skończonej odbywa się **za pomocą tablic liczb losowych** na podstawie pełnej listy ponumerowanych jednostek populacji generalnej.

**Etapy badania statystycznego.**

1. **Przygotowanie badania**: określenie celu i metody badania, określenie populacji i jednostki statystycznej, określenie cechy lub zbioru cech.
2. **Obserwacja**: polega na rejestracji wartości badanej cechy lub kilku cech.
3. **Opracowanie i prezentacja graficzna materiału**: np. grupowanie i zliczanie danych, szeregi statystyczne, histogram, diagram, wykres przebiegu (w czasie), wykres korelacyjny.
4. **Opis lub wnioskowanie statystyczne**: opis dotyczy sytuacji, gdy dane pobrane są w badaniu pełnym albo częściowym, ale nie losowym, a wnioskowanie sytuacji, gdy mamy do czynienia z badaniem reprezentacyjnym.

**Opracowanie materiału - Skale pomiarowe.**
W wyniku obserwacji statystycznych otrzymujemy pewne dane, niekoniecznie liczbowe. Aby analizować te dane należy przyporządkować im pewne liczby. W tym celu często grupujemy (klasyfikujemy) wyniki obserwacji, a otrzymane dane stają się danymi statystycznymi. Wyróżniamy cztery typy grupowania (klasyfikacji) danych, czyli cztery **skale pomiarowe**: nominalne, porządkowe, przedziałowe, ilorazowe.

**Skala nominalna** występuje, gdy obiekty podzielone są na grupy. Elementy z tej samej grupy reprezentowane są przez wspólną wartość. Np. w aktach personalnych możemy użyć 1 dla reprezentacji mężczyzn i 2 dla reprezentacji kobiet. Dzielimy zbiorowość na dwie grupy i możemy policzyć liczbę elementów każdej z nich. Najmniej parametrów statystycznych potrafimy obliczyć właśnie dla skali nominalnej. Możemy wyznaczyć modę, ale nie wyznaczymy średniej.

**Skala porządkowa** występuje, gdy obiekty zostały pogrupowane tak jak dla danych nominalnych, ale dodatkowo wprowadzona jest relacja porządku między grupami, tzn. ustalona jest kolejność kategorii. Np. ranking szkół wyższych w Polsce, ranking kandydatów na studia w trakcie rekrutacji (podawane są tylko nazwiska, nie wiadomo o ile punktów jeden kandydat wyprzedza drugiego). Zmienne są na skali porządkowej, gdy przyjmują wartości, dla których dane jest uporządkowanie (kolejność), jednak nie da się w sensowny sposób określić różnicy ani ilorazu między dwiema wartościami.

W przypadku skali porządkowej nie wyznacza się średniej, ale można wyznaczyć modę.

Zarówno skala nominalna jak i porządkowa stosowana jest do pomiaru cech niemierzalnych (jakościowych), będących często przedmiotem badań socjologicznych, marketingowych czy psychologicznych.

**Skala przedziałowa** ma wszystkie cechy skali porządkowej, ale dodatkowo grupy danych odwzorowane są na jedną skalę z równymi jednostkami przedziałowymi. Znamy nie tylko kolejność, ale również różnice wartości między kolejnymi grupami. Cecha jest na skali przedziałowej, gdy różnice między dwiema jej wartościami dają się obliczyć i mają interpretację w świecie rzeczywistym, jednak nie ma sensu dzielenie dwóch wartości zmiennej przez siebie. Innymi słowy określona jest jednostka miary, jednak punkt zero jest wybrany umownie. Np. lista rankingowa kandydatów na studia, w której umieszczono liczbę uzyskanych punktów, wiemy o ile punktów jedna osoba wyprzedza drugą. Np. wyniki sondaży przedwyborczych w % poparcia. Np. data urodzenia. Np. temperatura w stopniach Celsjusza.

Ta skala stosowana jest do pomiaru cech ilościowych.

**Skala ilorazowa** ma wszystkie cechy skali przedziałowej, ale dodatkowo ma wyraźnie określone zero, stanowiące początek skali. Tutaj znamy nie tylko odległość między dwoma dowolnymi wartościami, ale także stosunek między dwoma dowolnymi punktami (bez względu na zastosowanie jednostki miary). Cecha jest na skali ilorazowej, gdy stosunki między dwiema jej wartościami mają interpretację w świecie rzeczywistym. Dla skali ilorazowej możemy wyznaczyć wszystkie parametry statystyczne, np. średnią. Skala ilorazowa, w odróżnieniu od uboższych skal, nie nakłada ograniczeń w stosowaniu operacji matematycznych i metod statystycznych.

Np. napięcie elektryczne, inflacja, bezrobocie, długość przedmiotów (np. w cm albo calach), temperatura w kelwinach. Wszystkie skale mają zero.

Skala Kelvina jest bezwzględną, termodynamiczną skalą temperatury, używającą jej punktu bazowego jako absolutnego zera. Ta skala nie ma jednostek ujemnych. Bezwzględne zero jest punktem zerowym, poniżej którego temperatura nie istnieje, a energia molekularna jest minimalna. Absolutne zero w skali Kelvina przekłada się na -273,15°w skali Celsjusza.

**Prezentacja danych**.
Zebrano informacje o wartości pewnej cechy dla wszystkich jednostek badanej populacji. Jeśli elementów jest $n$, a wartość cechy dla $i$-tej jednostki oznaczymy przez $x_i$, to otrzymamy zbiór wartości $x_1, x_2, \ldots, x_n$ zwany **szeregiem statystycznym**. Zwykle nie jest to szereg przetworzony w żaden sposób i jest to **szereg szczegółowy nieuporządkowany**. Np. wiek 10 mieszkańców Krakowa:
12,45,32,32,14,56,25,78,65,88.

Dane będą łatwiejsze do interpretacji, gdy je uporządkujemy (np. rosnąco).
Uzyskujemy **szereg szczegółowy uporządkowany**: 12,14,25,32,32,45,56,65,78,88.

Wartości wybranej cechy można gromadzić nie tylko w obrębie pewnej populacji, ale w kolejnych jednostkach czasu i wówczas otrzymany ciąg wartości nazywamy **szeregiem czasowym**.

| Dzień | 1-07 | 2-07 | 3-07 | 4-07 | 5-07 | 6-07 | 7-07 | 8-07 | 9-07 | 10-07 | 11-07 | 12-07 | 13-07 | 14-07 | 15-07 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Temperatura o godz. 8:00 | 15 | 16 | 20 | 22 | 25 | 25 | 26 | 20 | 18 | 17 | 16 | 16 | 20 | 25 | 25 |

Jeśli cecha jest badana w pewnym terytorialnym (przestrzennym) rozmieszczeniu, to szereg nazywany jest **przestrzennym**.

| Województwo | Warszawskie | Bielskie | Bydgoskie | Gdańskie | Katowickie | Kieleckie | Radomskie |
|---|---|---|---|---|---|---|---|
| Liczba ciągników | 14 693 | 17 632 | 38,494 | 22 937 | 28 689 | 53 550 | 40 163 |

Gdy wrócimy do szeregu szczegółowego, to zwrócimy uwagę, że są tam obserwacje powtarzające się. Szereg, który zawiera informacje o liczbie jednostek, dla których cecha przyjmuje określoną wartość nazywa się **szeregiem rozdzielczym punktowym**. Np.

| liczba dzieci | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| liczba małżeństw | 6 | 30 | 24 | 4 | 1 |

Szereg, który zawiera informacje o liczbie jednostek, dla których cecha przyjmuje wartość z określonego przedziału liczbowego nazywa się **szeregiem rozdzielczym przedziałowym (klasowym)**.

| wzrost 22-latki | [150,160) | [160,170) | [170,180) | [180,190) | [190,200) |
|---|---|---|---|---|---|
| liczba osób | 16 | 30 | 54 | 24 | 2 |

Gdy dysponujemy szeregiem szczegółowym uporządkowanym, to nie ma uniwersalnego sposobu na dobry szereg rozdzielczy klasowy. Trzeba mieć na uwadze, że kolejne klasy nie mogą na siebie nachodzić i powinny obejmować wszystkie obserwacje.

Gdy **budujemy szereg o jednakowej rozpiętości klas**, to postępujemy następująco:

1. obliczamy rozstęp $R = x_{\max} - x_{\min}$,
2. ustalamy liczbę klas, np. $k \approx \sqrt{n}$ (często wykorzystywane), $k \approx \frac{3}{4}\sqrt{n}$, $k \leqslant 5 \log n$, $k$ odczytane z tablic statystycznych,
3. ustalamy rozpiętość (długość) $l$ klasy jako $l = \frac{R}{k}$ przy warunku, aby $k \cdot l \leqslant R$,
4. dobranie początku szeregu klasowego, np. początek pierwszej klasy to $x_1$ albo $x_1$ jest środkiem pierwszej klasy.

**Przykład.**
Zarejestrowano wiek 20 pracowników zgłaszających się na badania okresowe. Po uporządkowaniu danych otrzymano szereg szczegółowy uporządkowany:
22,25,27,30,31,31,32,33,33,34,35,36,36,36,37,38,38,39,41,47.
Wtedy

1. $R = 47 - 22 = 25$,
2. $k \approx \sqrt{20} \approx 5$ (zaokrąglamy z nadmiarem),
3. $l = \frac{R}{k} = \frac{25}{5} = 5$
4. początek pierwszej klasy jest równy $x_1 = 22$

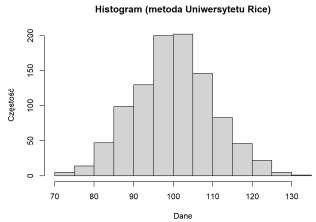
i szereg rozdzielczy klasowy wygląda następująco

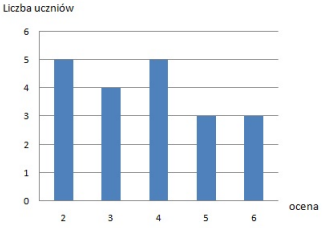
| Lp. | klasy | liczebność $n_i$ |
|---|---|---|
| 1 | [22,27) | 2 |
| 2 | [27,32) | 4 |
| 3 | [32,37) | 8 |
| 4 | [37,42) | 5 |
| 5 | [42,47] | 1 |

**Graficzna prezentacja danych**.

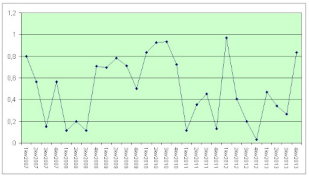
**Dane ilościowe**:

1. histogram, diagram opisują one rozkład cechy mierzalnej w prostokątnym układzie współrzędnych.

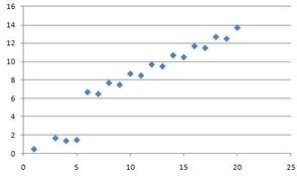



2. wykres przebiegu (dynamiczny)



3. wykres korelacyjny (punktowy) prezentuje zależność pomiędzy dwiema cechami




**Dane jakościowe**.

1. wykres słupkowy, kołowy, piramida