**Statystyka opisowa** to dział statystyki zajmujący się metodami opisu danych statystycznych uzyskanych podczas badania statystycznego. Celem stosowania metod statystyki opisowej jest podsumowanie zbioru danych i wyciągnięcie pewnych podstawowych wniosków i uogólnień na temat zbioru**.**

**Statystyka matematyczna** - Zajmuje się metodami wnioskowania statystycznego, które polegają na tym, że na podstawie wyników uzyskanych z próby formułujemy wnioski o całej zbiorowości. Wnioskowanie statystyczne obejmuje estymacje i weryfikację hipotez statystycznych.

**Statystyka opisowa + rachunek prawdopodobieństwa = statystyka matematyczna**

**Doświadczenie** – obserwacja lub eksperyment.

**Losowa** – pomimo spełnienia wszystkich warunków nie możemy przewidzieć wyniku z całą pewnością.

**Wariancja** – jest miernikiem zróżnicowania wartości zmiennej.

**Zdarzenie elementarne** – wynik doświadczenia losowego.

**Własności zdarzeń elementarnych:**

- dane zdarzenie elementarne może zaistnieć lub nie.

- w dane doświadczeniu losowym jedno zdarzenie elemtenarne na pewno zaistnieje.

- wystąpienie jednego zdarzenia elementarnego wyklucza inne.

**Przestrzeń zdarzeń elementarnych** – zbiór zdarzeń elementarnych. Może być rodzaju:

- skończona(ma skończoną ilość zdarzeń elem.) , -nieskończona ale przeliczalna, -nieskończona i nie przeliczalna.

**Prawdopodobieństwo** – miara zaistnienia danego zdarzenia losowego.

**Prawdopodobieństwo (klasyczna definicja wg Plata)** – jeżeli zdarzenie E rozkłada się na n wykluczających się parami i jednakowo możliwych zdarzeń spośród których m sprzyja zaistnienia zdarzenia A to prawdopodobieństwo tego zdarzenia wynosi P(A) = m/n.

**Definicja Geometryczna -** Jeżeli Q oraz q są dwoma zbiorami w przestrzeni R wymiarowej i jeżeli q zawiera się w Q to

prawdopodobieństwo tego, że jakiś punkt który należy do Q będzie należał również do q równa się

stosunkowi miary q do Q.

**Jednostka statystyczna (x)** – element zbiorowości statystycznej. Jednostką stat może być człowiek, rzecz lub zjawisko. Zbiór jednostek stat nazywamy zbiorowość statystyczną (populacja). Każda jednostka statystyczna ma właściwości tzw. Cechy statystycznej.

**Cecha statystyczna** **(y)** – właściwość populacji, która jest przedmiotem badania statystycznego. Zgodnie z definicją[1] cecha statystyczna jest to funkcja przypisująca elementom populacji elementy zbioru wartości cechy statystycznej. Dzielimy na:

1. ilościowe (mierzalne):

* ciągłe – np. wzrost, masa, wiek (w rozumieniu ilości dni między datą urodzin a datą badania)
* porządkowe (quasi-ilościowe) – np. klasyfikacja wzrostu: (niski,średni,wysoki)
* skokowe (dyskretne) – np. ilość posiadanych dzieci, ilość gospodarstw domowych, wiek (w rozumieniu ilości skończonych lat)

1. jakościowe (niemierzalne) – np. kolor oczu, płeć, grupa krwi

**Cechy jednostki statystycznej:**

-stałe (ich wartości są takie same dla każdej jednostki danej zbiorowości).

-zmienne (przyjmują różne wartości dla różnych jednostek).

**Stałe i zmienne** mogą być skokowe (dyskretne) albo cechy ciągłe.

**Zmienna skokowa** – przyjmuje tylko niektóre wartości ze zbioru liczb rzeczywistych.

**Zmienna ciągła** – może przyjąć dowolna wartość z danego przedziału liczbowego np. czas życia, mierzenie długości.

**Zmienna** – liczbowy obraz cechy.

**Pomiar –** proces empiryczny w ramach którego przyporządkowuje się liczby poszczególnym kategoriom cechy w taki sposób aby odzwierciedlały one relacje między obiektami (jednostkami) o danych kategoriach cechy.

**Skale pomiarowe:**

-skala nominalna (o równości lub różności danych obiektów).

-skala porządkowa (dodatkowo porządek informacji (hierarchia).

-skala interwałowa (dodatkowa informacja nt. wielkości róznicy).

-skala ilorazowa (wielkość stosunku).

-skala absolutna (dodatkowo uzyskane informacje są absolutnie niezmienne (zmiennych nie można przekształcić na inne liczby).

**Moc skali** – zakres informacji jakie możemy uzyskać z liczb pozyskanych na inne w danej skali

Pomiar skali silniejszej można zredukować, zmienić na skale słabszą ale nigdy na odwrót !!!

**Zmienna losowa** – wielkość którą może w wyniku doświadczenia losowego przyjąd jakąś wartośd przy czym jaka to będzie wartośd nie jesteśmy pewni przed wykonaniem doświadczenia losowego)

**Zmienna skokowa –** zawiera tylko niektóre wartości ze zbioru liczb rzeczywistych

**Formalna definicja zmiennej losowej -** zmienna losową X nazywamy funkcję X=X(e) określoną na zbiorze zdarzeo elementarnych E przyjmującą wartości ze zbioru liczb rzeczywistych taką, że dla każdej liczby rzeczywistej X zbiór A zdarzeń elementarnych dla których X(e) < x spełnia warunek, że A należy do Z jest zdarzeniem losowym.

**Rozkład można opisywać za pomocą mierników** ( charakterystyk rozkładu ) **.**

**Wartość oczekiwana** ( dla zmiennej losowej ) ( inaczej nazywana nadzieją matematyczną ).

**Mediana –** jest to miara położenia rozkładu, jest wartością ze środka rozkładu.

**Modalna –** wartość cechy dla której obserwujemy największe prawdopodobieństwo jej zaistnienia.

**Współczynnik zmienności** – zaletą tego współczynnika jest względnośd. Pozwala porównywad poziomy zmienności. Może być wyznaczony na podstawie średniej arytmetycznej i odch standardowego.

**Rozkłady zmiennej losowej skokowej:**

- rozkład sformalizowany – zapisany za pomocą wzoru

- rozkład zero-jedynkowy ( zmienna Binesha )

- rozkład Bernoulliego

**Rozkład Poissona** ( Aby użyć ten rozkład musimy znać lambdę )

Warunki Rozkładu Poissona są takie same jak dla rozkładu Bernoulliego. Ponadto są jeszcze dwa dodatkowe warunki.

Jeżeli da się zastosować Rozkład Poissona to da się zastosować Bernoulliego. Odwrotna operacja nie jest zawsze możliwa.

**Dystrybuanta** – w [rachunku prawdopodobieństwa](http://pl.wikipedia.org/wiki/Teoria_prawdopodobie%C5%84stwa), [statystyce](http://pl.wikipedia.org/wiki/Statystyka) i dziedzinach pokrewnych, [funkcja](http://pl.wikipedia.org/wiki/Funkcja) [rzeczywista](http://pl.wikipedia.org/wiki/Liczby_rzeczywiste) jednoznacznie wyznaczająca [rozkład prawdopodobieństwa](http://pl.wikipedia.org/wiki/Rozk%C5%82ad_prawdopodobie%C5%84stwa)

Funkcje gęstości dla zmiennej ciągłej powinna być funkcją ciągłą ale może być skończona liczba

punktów nieciągłych.

**Standaryzacja rozkładu jest to doprowadzenie do porównywalności.**

**Badanie statystyczne** – ciąg czynności zmierzających do okreslenia prawidłowości badanej zbiorowości. Dzielimy na:

* + ciągłe - cały czas badanie jest przeprowadzane
  + okresowe – badanie przeprowadzane jest przez określony czas
  + doraźne

**Badanie pełne** – badana jest cała zbiorowość ( np. spis powszechny )

**Badanie częściowe** – badana jest część zbiorowości ( np. badanie reprezentacyjne ) pozwala

wyciągnąć wnioski n apodstawie anaizy części zbiorowości

**Badanie szacunkowe** – badamy coś, czego nie da się zbadać, wnioskujemy na podstawie jakiegoś zjawiska, które jest obserwowalne i na tej podstawie mówimy o drugim zjawisku którego nie możemy zbadać, warunkiem jest ŁĄCZNOŚĆ tych zjawisk ( np. badanie szarej strefy ) Innymi słowami : na podstawie innych informacji które mamy badamy inne zjawisko które łączy się z informacjami które posiadamy.

**Etapy badania statystycznego**

**1. Przygotowanie badania**

◦ określenie celu badania

◦ określamy co/kogo będziemy badać (czyli określamy jednostkę statystyczną)

◦ wybieramy cechy statystyczne (zakres badania który będziemy badać)

◦ określamy rodzaj badania (pełne, okresowe, doraźne)

**2. Zbieramy materiał statystyczny którego finalnym rezultatem będzie zbiór danych**

**statystycznych**

• materiał pierwotny – został pozyskany na potrzeby naszego, konkretnego badania

• materiał wtórny – materiały statystyki publicznej ( materiały, których sami nie

gromadzimy, tylko np. znaleźliśmy gdzieś, nie są to własne materiały )

**3. Opracowanie materiału statystycznego**

• polega na grupowaniu i kontroli materiału statystycznego

• musimy sprawdzić, czy nie mamy błedów

**Rodzaje błędów:**

* błąd statystyczne ( np. odmowa w badaniu, źle skonstruowane pytanie i w wyniku tego błędne odpowiedzi na to pytanie).
* błąd losowy - dane muszą być wysokiej jakości poprzez eliminacje błędów.

**4. Prezentacja materiału statystycznego**

a) za pomocą tabel statystycznych ( tabela musi posiadać numer i tytuł ! )

Każde pole w tabeli musi być wypełnione ( słownie, liczbą, znaki)

b) za pomocą wykresów statystycznych – prezentują one w uproszczony sposób dane, mają ukazać strukturę, pokazywać tendencję itp.

c) szeregi statystyczne – ciąg wielkości statystycznych wedle przyjętego kryterium

➔ szeregi szczegółowe ( wyliczające ) :

➢ prosty.

➢ ważony - porządkujemy i wypisujemy wartości cechy i przyporządkowujemy ilości wystąpień.

➔ szeregi geograficzne – uporządkowane obserwacje ze względu na jednostki terytorialne.

➔ szeregi czasowe ( chronologiczne/dynamiczne) uporządkowują wartości ze względu na jednostkę czasu.

➔ szeregi strukturalne – pokazują strukturę zbiorowości ze względu na cechę, szeregi te używamy dla cechy jakościowej lub ilościowej:

➢ dla jakościowej mamy

✔ szeregi rozdzielcze punktowe – rozdziela zbiorowość na wartości.

✔ przedziałowy.

**Znaki spotykane w tabelach**

„-” oznacza, że zjawisko nie wystąpiło

„x” oznacza, że wypenienie danego pola nie jest celowe lub niemożliwe

„●” oznacza, że jest brak danych lub brak wiarygodnych danych

„(0,0)”oznacza, że zjawisko wystąpiło, ale na marginalnym poziomie

w tym „>” oznacza, że nie podajemy wszystkich elementów sumy

„#” oznacza, że podanie danych jest objęte tajemnicą ( np. państwową )

**5. Analiza statystyczna**

a) badanie struktur i rozkładów zjawisk lub procesów

b) badanie zależności między zjawiskami i procesami

c) badanie dynamiki zjawisk

**Charakterystyki rozkładu cechy** – mierniki opisujące ten rozkład

**Podziały charakterystyki :**

– miary położenia rozkładu ( miary tendencji centralnej )

– miary zróżnicowania/zmienności/dyspresji

– miary asymetrii rozkładu

– miary koncentracji i rozkładu ( spłaszczenia i koncentracji )

**Miary położenia dzielą się na**

– przeciętne

– kwantyle

**Przeciętne dzielą się na**

– średnie klasyczne

– przeciętne pozycyjne

**Średnie klasyczne**

– średnia arytmetyczna

– średnia harmoniczna

– średnia geometryczna

– średnia potęgowa

**Przecietne pozycyjne**

– modalna

– mediana

**Kwantyle**

– kwartyle

– kwintye

– decyle

– centyle

**Średnia Harmoniczna** ( stosujemy ją do zjawisk, które są wyrażone w jednostkach na jednostkę np. km/h, sztuki/godzinę )

**Średnia geometryczna (** wykorzystujemy przy wyznaczaniu przeciętnej dynamiki zjawisk)

**Kwantyle** – rząd kwantyla przyjmuje wartości [0,1]

**Miary zmienności** ( miary zróżnicowania, nierówności, dyspersji )

Różnica między zróżnicowaniem a nierównością :

– np. dochody można rozdzielać równo bądź nierówno, możemy zmieniać te nierówności

– np. wzrostu nie możemy zmienić, nie ma rozdziału wzrostu, co zostało rozdzielone to jest niezmienne, więc jest zróżnicowanie.

**Nierówność** – poprzez jakieś działanie można zmieniać nierówności.

**zróżnicowanie** – nie możemy zmienić tego co zostało podzielone.

**Miary zmienności:**

– miary bezwzględne (absolutne) (coś (np. badania) wyrażone są w jednostce)

– miary względne (odnoszą wynik pomiaru do jakiejś innej wartości, względem czegoś w stosunku do czegoś. Nie ma jednostki gdyż jednostki się redukują).

Aby porównać miary bezwzględne, potrzebujemy takie same jednostki, tą samą średnią. Wtedy możemy porównywać wyniki.

**R -rozstęp** ( różnica między wartościa maksymalną a minimalną dla cechy )

**Współczynnik Giniego –** miara względna zmienności ( taki sam jest współczynnik koncentracji Lorenza, chodzi o to samo ) W tym współczynniku porównujemy np. wszystkie osoby ze sobą. Jeżeli korzystamy z miar względnych, musi być coś mierzone na skali ILORAZOWEJ.

**Cecha diagnostyczna** – cecha która ma coś diagnozować

**Własności :**

– w miare wysoki poziom zróżnicowania

– powinna charakteryzować się asymetrią prawostronną gdy jest stymulantą. Nie wszystkie cechy są stymulantą.

**Miary koncentracji** Współczynnik skupienia (koncentracji) (kurtoza) K- jest miarą skupienia poszczególnych obserwacji wokół średniej. Im wyższa wartość współczynnika tym bardziej wysmukła krzywa liczebności, większa koncentracja wartości cech wokół średniej.

**Dwa rozumienia koncentracji :**

– zgrupowanie obserwacji wokół średniej.

– nierówne rozdysponowanie ogólnej sumy cechy statystycznej.

Moment centralny rzędu 2 jest wariancją. Moment centralny rzędu 4 jest miarą koncentracji.

**Kowariancja – miernik zależności/niezależności między zmiennymi**

Zależność i korelacja to nie do końca to samo ( choć w niektórych sytuacjach się pokrywają)

**Badanie zależności od strony statystycznej :**

1. Zastanówmy się czy jest logiczny sens naszych badań

**zależność funkcyjna –** występuje gdy zmiany wartości jednej zmiennej są dokładnie

zdeterminowane przez zmianę wartości drugiej zmiennej.

np.

P=a^2

y=ax+b < dokładna zależność funkcyjna ( dokładna czyli nie ma żadnych odchyleń)

Prawie zawsze nie mamy dokładnej zależności

**Zależność stochastyczna** – występuje gdy zmiany wartości jednej cechy określają zmiany rozkładu prawdopodobieństwa drugiej cechy.

**Zależność korelacyjna** – dla każdego z rozkłądów liczymy średnią w rozkłądach warunkowych (występuje gdy zmiany wartości jednej cechy determinują średnią drugiej cechy).

**Zależność może być dodatnia bądź ujemna.**

**Dodatnia –** wzrost jednej cechy powoduje wzrost drugiej cechy

**ujemna –** wzrost jednej cechy powoduje spadek wartości drugiej cechy

**Jednokierunkowe i dwukierunkowe**

**jednokierunkowe** – zależność idzie tylko w jedną strone, jedna zmienna jest zależna a druga się zmienia

**dwukierunkowy** – raz jedna , a raz druga zmienna jest zależna ( są zależne wymiennie )

**Regresja** – metoda analizy zależności, dzielimy na:

* **pierwszego rzędu** – uzyskujemy poprzez wyznaczanie wartości przeciętnych w rozkładach warunkowych
* **drugiego rzędu** – uzyskujemy poprzez wykorzystanie funkcji teoretycznych opisujących zależność (za pomocą modeli regresyjnych).

**Estymacja oraz Weryfikacja hipotez.**

**Estymacja –** szacowanie, ocenianie, przybliżanie – estymujemy bo nie mamy pełnych danych. Wyniki statystyki matematycznej nie są 100% prawdziwe, ale są wiarygodne. Miarą wiarygodności jest prawdopodobieństwo.

**Co możemy Estymować?**

– parametry rozkładu

– postać funkcyjną

**Estymacja punktowa i przedziałowa**

Podstawą analiz w statystyce matematycznej jest próba statystyczna (nie może być to dowolna

grupa). Musi być to tzw. grupa reprezentatywna - grupa jest reprezentatywna gdy jej struktura na badane cechy jest zlbiżona do populacji na te cechy.

Warunki aby uzyskać grupę reprezentatywną :

* próba powinna być losowa.
* musi być odpowiednio liczna.
* losowa – znane jest prawdopodobieństwo dostania się do próby każdej jednostki statystycznej ( każdego zespołu jednostek statystycznych).
* odpowiednia liczebność.

**Różne losowania:**

– zależne, niezależne

– jednostopniowe, wielostopniowe

– indywidualne, zespołowe

– nieograniczone, ograniczone

**Próba prosta** – każda jednostka ma takie samo prawdopodobieństwo na dostanie się do próby.

Uzyskujemy ją np. w losowaniu indywidualnym, nieograniczonym, niezależnym.

Podstawowym narzędziem statystyki matematycznej jest statystyka jako funkcja próby statystycznej. ( Coś chcemy uzyskać na podstawie czegoś ) Operujemy na zmiennych losowych, mamy realizację zmiennej losowej. Sama statystyka ma rozkład, który zależy od postaci funkcyjnej, liczebności próby, rozkładu zmiennych losowych ( zmiennej losowej x )

**Mamy dwa rodzaje rozkładu statystyki**

– rozkład dokładny

– rozkład graniczny (statystyka może nie mieć rozkładu normalnego, ale ma rozkład graniczny).

**Rozkład dokładny** – określony dla dowolnej liczebności próby, na jego podstawie mogę określić dokłady rozkład statystyki.

**Rozkład graniczny** – rozkład dla liczebności grupy zmierzającej do nieskończoności, dla bardzo dużych grup ( n > 30, korzystamy wtedy z rozkładu granicznego).

Dlaczego rozkład jest tak ważny ? Jeżeli nie znamy rozkładu statystyki, to jest ona bezużyteczna,

nie mamy punktu odniesienia. Mamy jakiś wynik ale nie wiadomo co on symbolizuje.

**Stopień swobody** – liczba niezależnych obserwacji w próbie (niezależność obliczeń)

**Estymacja przedziałowa** – przedział który pokrywa nieznany obszar. Statystyki które wykorzystujemy do estymacji nazywamy estymatorami. Estymator to taka statystyka której rozkład zależy od szacowanego parametru.

Może być wiele estymatorów jednego parametru. Mówimy o **trzech własnościach estymatorów**:

– **powinien być nieobciążony** – wtedy gdy jego wartość oczekiwana równa się estymowanemu parametrowi. Nieobciążoność jest właściwością kluczową. Jeżeli jest obciażony, z góry narażamy się na pomyłkę. Obciążony estymator jest narażony na błąd systematyczny, czyli regularnie nasze wyniki są w jakimś stopniu złe.

– **zgodność estymatora.**

E**fektywność estymatora –** ma najmniejszą ( jest najefektywniejszy) wariancję spośród wszystkich nieobciążonych estymatorów danego parametru ( ma najniższy rozrzut )

**METODY ESTYMACJI:**

**-Metoda najmniejszych kwadratów** ( omawiana na poprzednim wykładzie )

**-Metoda momentów**

**-Metoda największej wiarygodności ( MNW)**

**Estymacja Baysowska** – nie jest to klasyczne podejście, jest to pewna alternatywa.

**Wyznaczanie minimalnej liczebności próby** – należy o tym zagadnieniu znaleźć informacje np. w internecie.

**Przedziałem ufności nazywamy przedział spełniający warunki**

– jego krańce są funkcjami próby losowej i nie zależą od szacowanego parametru

– prawdopodobieństwo pokrycia przez ten przedział wartości nieznanego parametru nazywamy poziomem ( współczynnikiem) ufności.

**Długość przedziału zależy od:**

– im wyższy poziom ufności tym dłuższy jest przedział

– im większa liczebność próby tym krótszy jest przedziałem

**Weryfikacja hipotez statystycznych:**

Hipoteza to jakiś sąd na temat rzeczywistości. Hipoteza statystyczna to sąd który można sprawdzić na podstawie próby statystycznej.

**W stosunku do czego stawiamy hipotezy ?**

– hipotezy na temat parametrów przedziału – hipoteza parametryczna

– wszystkie hipotezy nie dotyczące przedziału – hipoteza nieparametryczna ( np. badanie

zależności)

**Inny podział hipotez**

– prosta – hipoteza określa dokładnie rozkład

– złożona – hipoteza nie określa rozkładu

**Kolejny podział**

– zerowa – hipoteza weryfikacyjna

– hipoteza alternatywna – jesteśmy skłonni przyjąć coś, gdy odrzucimy hipotezę zerową

Narzędziem weryfikacji hipotez jest test statystyczny ( pewne procedury) które na podstawie każdej

losowej próby przyporządkowują decyzje przyjęcia lub odrzucenia weryfikowanej hipotezy

**Elementy procedury**

1. Określenie hipotezy zerowej i alternatywnej

2. przyjęcie tzw. statystyki testowej

3. Obliczanie wartości statystyki testowej na podstawie próby statystycznej

4. określenie tzw przedziałów krytycznych na podstawie rozkładu statystyki testowej i

przyjętego prawdopodobieństwa ( musi być znany rozkład)

5. Wniosek w stosunku do hipotezy

**Błąd pierwszego rodzaju** – odrzucenie prawdziwej hipotezy ( prawdopodobieństwo to oznaczaomy jako alfa > prawdopodobieństwo popełnienia błędu)

**Błąd drugiego rodzaju** – przyjęcie fałszywej hipotezy ( prawdopodobieństwo to oznaczamy jako beta)

**Moc testy statystycznego** – prawdopodobieństwo tego, że test statystyczny odrzuci fałszywą hipotezę przy założonym prawdopodobieństwie popełnienia błędu pierwszego rodzaju Hipotezę możemy weryfikować poprzez wiele testów. Wybieramy test najmocniejszy.

**Testy istotności** – mówią o istotności różnicy

**Cecha szczególna testów istotności** – NIE przyjmujemy hipotezy zerowej.