LITERATURA:
[1] J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwa Naukowo-Techniczne, Warszawa, 2006.
[2] A.M. Mood, F.A. Graybill, D.C. Boes, *Introduction to the theory of statistics*, McGraw-Hill Publishing Company, 1983.
[3] P. Biecek, *Przewodnik po pakiecie R*, Oficyna Wydawnicza GIS, Wrocław, 2008.
[4] P. Dalgaard, *Introductory Statistics with R*, Springer, 2008

Wykład 1: Analiza danych a wnioskowanie statystyczne

Statystyka obejmuje dwa nurty:

1. analizę danych,
2. wnioskowanie statystyczne.

Celem **analizy danych** jest prezentacja konkretnego zbioru danych, w sposób ukazujący jego własności; w szczególności syntetyczny opis podstawowych jego cech. Otrzymujemy wówczas wnioski, które dotyczą **wyłącznie analizowanego zbioru danych**. Na przykład mamy zebrane informacje na temat słuchaczy studiów doktoranckich w PW, którzy rozpoczęli owe studia w 2010 roku. Dokładniej, mamy listę tych słuchaczy wraz z następującymi danymi:

- płeć słuchacza,
- czas (w miesiącach) kiedy słuchacz był doktorantem,
- wysokość pobranego stypendium,
- czy studia zakończyły się uzyskaniem dyplomu doktora,
- czy student został pracownikiem PW, innej jednostki naukowej lub czy podjął pracę gdzie indziej.

Na podstawie tych danych możemy stwierdzić np.

- jaki procent słuchaczy, rozpoczynających studia doktorskie w PW w 2010 r., zakończył studia uzyskaniem dyplomu doktora;
- jaki procent owych słuchaczy stanowiły kobiety;
- jakie łączne wydatki poniesiono na stypendia studentów, którzy nie uzyskali dyplomu.

Otrzymamy wyniki **dokładne** i **pewne**, ale dotyczyć będą one **jedynie** słuchaczy studiów doktoranckich w PW, którzy rozpoczęli te studia w 2010 roku.

Teraz wyobraźmy sobie, że chcemy wiedzieć:

- jaki procent studentów doktorantów, studiujących w Polsce, to kobiety;
- ile wynosi średnie miesięczne stypendium doktorantów studiujących w Polsce.

Aby uzyskać dokładną i pewną odpowiedź na powyższe pytania, potrzebowalibyśmy zebrać dane dotyczące wszystkich doktorantów studiujących obecnie w Polsce. To trudne zadanie - czasochłonne i kosztowne, a nawet niekoniecznie wykonalne, bo niektóre jednostki mogą odmówić nam współpracy lub zwlekać z dostarczeniem stosownych danych. Pozostaje wtedy pójść na kompromis - zebrać dane dotyczące tylko wybranych studentów i na ich podstawie wyciągać wnioski o wszystkich studentach. Mamy wtedy do czynienia z **wnioskowaniem statystycznym**. Musimy w nim zwrócić uwagę na dwa aspekty.

1. Bardzo ważny jest odpowiedni wybór studentów do naszego badania - ogólniej - **odpowiedni wybór obserwacji do próby**, tak by dobrze reprezentowały one całą populację.
2. Jeśli tylko jako próby nie weźmiemy całej populacji (a tak we wnioskowaniu statystycznym postępujemy), to **uzyskane wyniki nie będą ani dokładne, ani pewne - pozostaną obarczone błędem**.

Analiza danych

Jak już wspomnieliśmy, celem analizy danych jest opis podstawowych cech konkretnego zbioru danych. Często, aby taki opis uzyskać, musimy najpierw dane, zawarte w zbiorze, uporządkować i uprościć. Porządkowanie danych rozpoczynamy od ustalenia jakiego są one typu. Możemy mieć:

- **dane ilościowe**, czyli dane w postaci liczb; np. wysokości miesięcznego stypentium studentów w PLN, czas (w miesiącach) od rozpoczęcia studiów doktoranckich do ich zakończenia;
- **dane jakościowe** opisujące cechę jakościową, jak np. płeć, kolor oczu, zawód itp.

Do opisu danych ilościowych możemy użyć miar liczbowych.

MIARY LICZBOWE DLA DANYCH ILOŚCIOWYCH

1). Miary położenia:

- miary tendencji centralnej:
  1. średnia (mean): $\bar{x} := \frac{\sum_{i=1}^{n} x_i}{n}$
     > mean(wektor) lub > mean(wektor,na.rm=TRUE) gdy są obserwacje brakujące
  2. mediana (median) - wartość środkowa
     > median(wektor)
  3. moda (dominanta) (moda) - wartość najczęściej pojawiająca się w próbie;
- miary pozycji:
  1. dolny kwartyl (lower quartile): Q_1
     > quantile(wektor,0.25)
  2. górny kwartyl (upper quartile): Q_3
     > quantile(wektor,0.75)
  3. decyle, percentyle i kwantyle (deciles, percentiles and quantiles): q_p
     > quantile(wektor,c(0.1,0.99,0.85))
     Powyższa funkcja wyznacza pierwszy decyl, 99-ty percentyl i kwantyl rzędu 0.85.

2). Miary rozproszenia:

1. rozstęp (range): $Max - Min$
   > max(wektor)-min(wektor)
2. rozstęp międzykwartylowy (interquartile range): $IQR := Q_3 - Q_1$
   > IQR(wektor)
3. wariancja (variance): $S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
   > var(wektor)
4. odchylenie standardowe (standard devriance): $S := \sqrt{S^2}$
   > sd(wektor).

3). Miary kształtu:

1. skośność (współczynnik asymetrii) (skewness): $A := \frac{n}{(n-1)(n-2)S^3} \sum_{i=1}^{n}(x_i - \bar{x})^3$
   Jeśli obserwacje są symetrycznie rozłożone względem średniej (która w tej sytuacji równa się medianie), to $A = 0$.
2. kurtoza (współczynnik spłaszczenia) (kurtosis):
   $K := \frac{n(n+1)}{(n-1)(n-2)(n-3)S^4} \sum_{i=1}^{n}(x_i - \bar{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$
   Wskazuje czy dane zawierają więcej i bardziej skrajne obserwacje odstające ($K > 0$) czy ich mniej i mniej skrajne ($K < 0$) niż byśmy oczekiwali od danych z rozkładu normalnego.

```
> install.packages("e1071")
> library(e1071)
> skewness(wektor)
> kurtosis(wektor)
```

GRAFICZNA PREZENTACJA DANYCH ILOŚCIOWYCH

1. wykres skrzynkowy (wykres typu *skrzynka z wąsami*) (boxplot)
   `> boxplot(wektor,range=1.5,horizontal=FALSE)`
2. histogram liczności i histogram częstości (histograms)
   `> hist(wektor,freq=TRUE)` i `> hist(wektor,freq=FALSE)`
3. jądrowy estymator gęstości (kernel density estimator) - wygładzona wersja histogramu częstości
   `> plot(density(wektor))`

GRAFICZNA PREZENTACJA DANYCH JAKOŚCIOWYCH

1. wykres słupkowy (barchart, barplot)
   `> barplot(licznosci,col=c("green",...,"red"))`
2. wykres kołowy (piechart)
   `> pie(licznosci,col=c("blue",...,"yellow"))`

Wnioskowanie statystyczne

We wnioskowaniu statystycznym z populacji pobieramy próbę i na jej podstawie wyciągamy wnioski dotyczące całej populacji. Bardzo ważny jest wybór owej próby, tak by zawierała jak najwięcej informacji o badanej populacji. Jedną z metod jest wybór tzw. **prostej próby losowej**. Aby zdefiniować to pojęcie przyjrzyjmy się bliżej postawionemu problemowi.

Niech X oznacza badaną cechę populacji. Na przykład

- populacją może być zbiór wszystkich 10-cio letnich dzieci mieszkających w Polsce, a X - wzrostem dziecka;
- populacją mogą być wszystkie żarówki energooszczędne produkowane przez pewnien zakład, a X - czasem świecenia żarówki;

- populacją mogą być wszystkie szklane abażury produkowane przez pewnien zakład, a X - informacją czy abażur posiada wady czy nie.

X jest zmienną losową, bo jego wartość zależy od zdarzenia losowego: w przykładzie ze wzrostem 10-cio letnich dzieci X zależy od wybranego dziecka; w przykładzie z żarówkami X zależy od wybranej żarówki. Naszym celem jest opisanie rozkładu X. Aby go osiągnąć, pobieramy próbę, którą oznaczamy

$$X_1, X_2, \ldots, X_n.$$

Przed zebraniem danych elementy próby to zmienne losowe. Zakładamy o nich, że mają ten sam rozkład, co badana cecha populacji X. Jeśli dodatkowo przyjmiemy, że są one niezależne, to będziemy mieć prostą próbę losową, często zwaną po prostu próbą losową.

Definicja. Jeśli $X_1, X_2, \ldots, X_n$ są niezależne i mają ten sam rozkład co cecha populacji X, to $X_1, X_2, \ldots, X_n$ nazywamy *(prostą) próbą losową* z X.

Oczywiście założenie, że pracujemy z prostą próbą losową, musi mieć swoje odzwierciedlenie podczas procesu zbierania danych - do próby powinniśmy wybierać niezależne od siebie obserwacje i każda z nich powinna dobrze reprezentować badaną populację.

W wyniku zebrania danych otrzymujemy **realizację próby losowej**, czyli n ustalonych wartości, które oznaczamy $x_1, x_2, \ldots, x_n$.

Na podstawie próby losowej $X_1, X_2, \ldots, X_n$ chcemy opisać rozkład X. Możliwe są dwa podejścia.

1. **Podejście parametryczne** - zakładamy, że X ma rozkład o dystrybuancie o znanej postaci a nie znamy jedynie parametrów tej dystrybuanty. Na przykład zakładamy, że
   - X ma rozkład wykładniczy z parametrem λ, $Exp(\lambda)$, gdzie $\lambda > 0$, i nie znamy jedynie wartości parametru λ;
   - X ma rozkład normalny o średniej $\mu \in \mathbb{R}$ i wariancji $\sigma^2 > 0$, $\mathcal{N}(\mu, \sigma^2)$ i nie znamy jedynie parametrów μ i σ^2.
2. **Podejście nieparametryczne** - nic nie zakładamy o postaci rozkładu X.

Najpierw skupimy się na podejściu parametrycznym. Będziemy zatem zakładać, że X ma rozkład o dystrybuancie F z nieznanym parametrem θ, co symbolicznie zapisujemy $X \sim F_\theta$, gdzie θ może być zarówno jedno- jak i wielowymiarowym parametrem. Ponadto Θ oznaczać będzie zbiór wszystkich możliwych wartości parametru θ: $\theta \in \Theta$.

Reasumując, nasze założenia to:

$$\boxed{\begin{array}{l} X \sim F_\theta, \text{ gdzie } \theta \in \Theta \\ \qquad\qquad \text{i} \\ X_1, X_2, \ldots, X_n \text{ jest próbą losową z } X. \end{array}}$$

Przy tych założenich przyjrzymy się dokładniej dwóm aspektom wnioskowania statystycznego

- estymacji punktowej,
- weryfikacji hipotez.