Wykład 10: Elementy statystyki

dr Mariusz Grządziel

20 grudnia 2010

Podstawowe pojęcia
Biolodzy: -badają pojedyńcze rośliny lub zwierzęta; -chcemy rozszerzyć wnioski na wszystkich przedstawicieli gatunku lub odmiany: -można to osiągnąć wykonując analizę statystyczną.

Definicja 1. *Populacją nazywamy zbiorowość (skończoną lub nieskończoną), w stosunku do której mają być formułowane wnioski. Próbą jest natomiast skończona część populacji.*

Stosowane podejścia

- metody analizy danych: statystyki opisowe, metody graficzne
- wnioskowanie statystyczne- oparte na metodach i pojęciach teorii prawdopodobieństwa

Podstawowe pojęcia
Zazwyczaj jesteśmy zainteresowani kilkoma konkretnymi cechami populacji.
Cechy: -jakościowe (kolor oczu, kolor włosów itd.) -ilościowe (masa zwierząt, liczba jaj składanych przez określony gatunek ptaków itd.)
Jeśli w badanich ograniczamy się do jednej cechy- populacja jednocechowa. Np. badając badając zwierzęta określonego gatunku możemy być zainteresowani tylko ich temperaturą itd.

Przykład
Dane *normtemp*— zebrane w celu weryfikacji hipotezy mówiącej, że średnia wartość temperatury zdrowego człowieka jest równa $98{,}6$ stopni w skali Fahrenheita ($37{,}0$ stopni w skali Celsjusza). Dane nt. temperatury i tętna (temperatura- stopnie Fahrenheita) Można je pobrać z odpowiedniego repozytorium a następnie zapisać do zbioru o nazwie np. **t** (tzw. „data frame"). Zbiór **t** składa się z trzech zmiennych: **temperature**, **gender** i **hr**. Aby uczynić naszą prezentację bardziej czytelną, zmieniamy nazwy zmiennych na odpowiednio: **temp**, **plec** i **tetno**. Odpowiednie polecenia systemu R są zapisane w pliku $t.R$. Wydruk tego pliku zamieszczamy poniżej (na następnym slajdzie):

Pobieranie zbioru danych z repozytorium systemu R

```
library(utils)
install.packages(c("xlsReadWrite","UsingR"),
```

```
repo="http://cran.r-project.org")
library(xlsReadWrite)
library(UsingR)
t<-normtemp
names(t)<-c("temp","plec","tetno")
write.xls(t,'c:/t.xls') # zbior t zapisany
                        # do pliku t.xls
```

System R 2.8 można pobrać pod adresem http://r.meteo.uni.wroc.pl/bin/windows/base/

Pobieranie zbioru danych z repozytorium systemu R

```
> names(t)
[1] "temp"  "plec"  "tetno"
> t[1:10,]
   temp plec tetno
1  96.3    1    70
2  96.7    1    71
3  96.9    1    74
4  97.0    1    80
5  97.1    1    73
6  97.1    1    75
7  97.1    1    82
8  97.2    1    64
9  97.3    1    69
10 97.4    1    70
> sort(t$temp)
  [1]  96.3  96.4  96.7  96.7  96.8  96.9  97.0  97.1
  [9]  97.1  97.1  97.2  97.2  97.2  97.3  97.4  97.4
                 ...
[129] 100.0 100.8
```

Szereg rozdzielczy

Dla zbioru danych liczbowych $\{y_1, y_2 \ldots, y_N\}$ niech: $MIN1$ oznacza liczbę mniejszą od najmniejszej z liczb $\{y_1, y_2 \ldots, y_N\}$ $MAX1$ oznacza liczbę większą lub równą od największej z liczb $\{y_1, y_2 \ldots, y_N\}$ $MIN1$ i $MAX1$ mogą być odpowiednimi zaokrągleniami wartości, odpowiednio, minimalnej i maksymalnej naszego zbioru danych $MIN1 < MAX1$.
Podzielmy odcinek $(MIN1, MAX1]$ na k przedziałów (zwanych klasami) o równej długości:

$$(x_0, x_1], (x_1, x_2], \ldots, (x_{k-1}, x_k],$$

gdzie $x_0 = MIN1$, $x_k = MAX1$. Funkcję przyporządkowującą poszczególnym przedziałom liczbę elementów naszego zbioru danych do nich należących będziemy nazywać szeregiem rozdzielczym.
Liczbę klas k w szeregu rozdzielczym wyznaczamy korzystając z wzorów:

$$k \approx log_2 n \quad \text{lub} \quad k \approx \frac{3\sqrt{n}}{4}.$$

Szereg rozdzielczy— dane NT

Przyjmujemy: $MIN1 = 96$, $MAX1 = 101$, $k = 10 \approx \frac{3\sqrt{130}}{4}$.

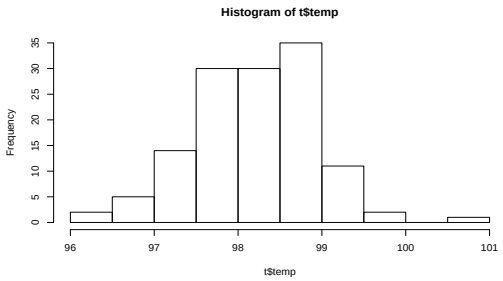
Zakładając, że dane dotyczące temperatury zdrowych ludzi znajdują się w zmiennej `t$temp` konstruujemy szereg rozdzielczy w środowisku R:

```
> table(cut(t$temp, breaks = c(96,96.5,97,97.5,98,
+ 98.5,99,99.5,100,100.5,101)))

(96,96.5]   (96.5,97]   (97,97.5]   (97.5,98]
        2           5          14          30
  (98,98.5]   (98.5,99]   (99,99.5]  (99.5,100]
       30          35          11           2
(100,100.5] (100.5,101]
        0           1
```

Histogram- dane NT

Wykres słupkowy odpowiadający szeregowi rozdzielczemu-histogram liczebności



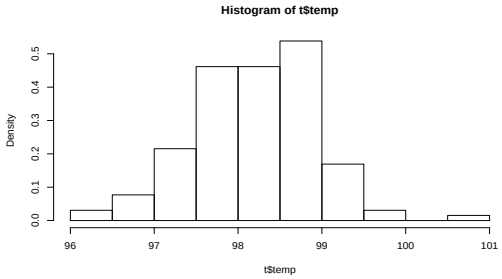
Rysunek 1: Histogram dla danych NT odpowiadający szeregowi rozdzielczemu z poprzedniego slajdu

Histogram probabilistyczny- dane NT

Histogram probablistyczny-tak wyskalowany, aby "pole pod nim było równe 1": wyskości słupków: $\frac{2}{130\times 0{,}5}, \frac{5}{130\times 0{,}5}, \ldots$; histogram probabilistyczny- przez niektórych definiowany jako funkcja przedziałami ciągła (stała), której wykres "pokrywa się" z zdefiniowanym wyżej wykresem słupkowym

Histogram probabilistyczny-funkcja przedziałami ciągła

$$h(x) = \begin{cases} \frac{2}{130\times 0{,}5}, & \text{dla } x \in (96; 96{,}5] \\ \frac{5}{130\times 0{,}5}, & \text{dla } x \in (96{,}5, 97] \\ \vdots & \\ \frac{1}{130\times 0{,}5}, & \text{dla } x \in (100{,}5; 101] \end{cases}$$



Rysunek 2: Histogram probabilistyczny dla danych NT

Frakcja obserwacji należących do $(96; 97]$;

$$\int_{96}^{97} h(x)dx = \frac{1}{2}\Big(\frac{2}{130 \times 0{,}5} + \frac{5}{130 \times 0{,}5}\Big) = \frac{7}{130} \approx 0{,}054$$

Konstrukcja histogramu probabilistycznego— przypadek ogólny
W ogólnym przypadku wysokość k-tego słupka histogramu probabilistycznego (wartość funkcji dla argumentów należących do k-tej klasy) jest równa $\frac{n_k}{nd}$, gdzie

- n_k jest liczebnością $k - tej$ klasy,
- d jest długością klasy,
- n jest liczbą obserwacji.

Funkcja h — „histogram probabilistyczny” — zdefiniowana wzorem:

$$h(x) = \begin{cases} \frac{n_k}{nd}, & \text{x należy do } k\text{-tej klasy,} \\ 0, & \text{x nie należy do żadnej z klas.} \end{cases}$$

Dziedziną funkcji h jest zbiór liczb rzeczywistych $\mathbb{R}$.

Histogram probabilistyczny i gęstość cechy
Chcemy obliczyć frakcję osób mających temperaturę w przedziale $(96; 97]$ w populacji wszystkich dorosłych zdrowych ludzi (a także w innych przedziałach $[a, b]$).
Zamiast histogramu probabilistycznego do obliczeń można użyć funkcję ciągłą f, która go sensownie „wygładza”; szukana frakcja:

$$\int_{96}^{97} f(x)dx$$

Wielobok częstotliwości
Prosta recepta na wygładzenie histogramu probabilistycznego: wielobok częstotliwości.
W ogólnym przypadku: odcinek $(MIN1, MAX1]$ na k przedziałów (zwanych klasami) o równej długości:

$$(x_0, x_1], (x_1, x_2], \ldots, (x_{k-1}, x_k],$$

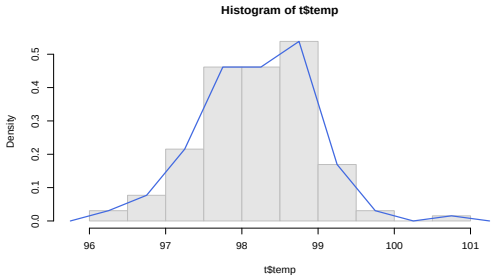
gdzie $x_0 = MIN1, x_k = MAX1$
Niech: $H = x_1 - x_0$,

$$\bar{x}_i = x_i - H/2, \quad i = 0, \ldots, k; \qquad \bar{x}_{k+1} = x_0 + H/2$$

Wielobok czętotliwości powstaje z połączenia:

$$(\bar{x}_0, 0), (\bar{x}_1, h(\bar{x}_1)), \ldots, (\bar{x}_k, h(\bar{x}_k)), (\bar{x}_{k+1}, 0)$$

Histogram+wielobok częstotliwości



Rysunek 3: Histogram+wielobok częstotliwości

Uwagi o wykresach dla danych jakościowych
Dla danych jakościowych, np. dotyczących składu wyznaniowego Warszawy w roku 1864, można sprządzić odpowiednie wykresy słupkowe lub kołowe— por. J. Koronacki, J. Mielniczuk, Statystyka dla kierunków technicznych i przyrodniczych, WNT 2001, paragraf 1.2.1, str. 14–17.