THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Analysis of Engineering and Scientific Data

Semester 1 – 2019

Sabrina Streipert s.streipert@uq.edu.au

Descriptive Statistics

- Visualisation of the data.
- Analysis and presentation of characteristics of the data.

Data types

Possible data types:

1. *Continuous quantitative* data ⟶ values in continuous range (height, width, length, temperature, humidity, volume, area, and price)

Data types

Possible data types:

1. *Continuous quantitative* data $\longrightarrow$ values in continuous range (height, width, length, temperature, humidity, volume, area, and price)
2. *Discrete qualitative* data (*factor / categorical variable*) $\longrightarrow$ values in discrete range (number of family members, gender (male or female), count of objects).

Data types

Possible data types:

1. *Continuous quantitative* data ⟶ values in continuous range (height, width, length, temperature, humidity, volume, area, and price)
2. *Discrete qualitative* data (*factor* / *categorical variable*) ⟶ values in discrete range (number of family members, gender (male or female), count of objects).

   Discrete Sub-types:

Data types

Possible data types:

1. *Continuous quantitative* data $\longrightarrow$ values in continuous range (height, width, length, temperature, humidity, volume, area, and price)
2. *Discrete qualitative* data (*factor* / *categorical variable*) $\longrightarrow$ values in discrete range (number of family members, gender (male or female), count of objects).

   Discrete Sub-types:

   - *Nominal* factors = variables without order, such as males and females.

Data types

Possible data types:

1. *Continuous quantitative* data ⟶ values in continuous range (height, width, length, temperature, humidity, volume, area, and price)
2. *Discrete qualitative* data (*factor* / *categorical variable*) ⟶ values in discrete range (number of family members, gender (male or female), count of objects).

   Discrete Sub-types:

   - *Nominal* factors = variables without order, such as males and females.
   - *Ordinal* factors = variable with a certain order, such as *age group*.

Data configurations

- Each configuration will consist of continuous and discrete (ordinal and nominal) variables.

Data configurations

- Each configuration will consist of continuous and discrete (ordinal and nominal) variables.
- **Major configuration types:**

Data configurations

- Each configuration will consist of continuous and discrete (ordinal and nominal) variables.
- **Major configuration types:**
  - A single sample configuration consists of m scalars: $\mathcal{D} = \{x_1, x_2, \ldots, x_m\}$.
    Nr of fisherman per day; $m = 365$ and $x_i = 0, 1, \ldots$.

Data configurations

- Each configuration will consist of continuous and discrete (ordinal and nominal) variables.
- **Major configuration types:**
  - A single sample configuration consists of m scalars: $\mathcal{D} = \{x_1, x_2, \ldots, x_m\}$.
    Nr of fisherman per day; $m = 365$ and $x_i = 0, 1, \ldots$.
  - Two (or more) sets of samples:
    $$\mathcal{D} = \left\{ \left\{x_1^1, \ldots, x_{m_1}^1\right\}, \left\{x_1^2, \ldots, x_{m_2}^2\right\}, \ldots, \left\{x_1^k, \ldots, x_{m_k}^k\right\} \right\}.$$
    Nr of fisherman per day in k different regions.

Data configurations

- Each configuration will consist of continuous and discrete (ordinal and nominal) variables.
- **Major configuration types:**
  - A single sample configuration consists of m scalars: $\mathcal{D} = \{x_1, x_2, \ldots, x_m\}$.
    Nr of fisherman per day; $m = 365$ and $x_i = 0, 1, \ldots$.
  - Two (or more) sets of samples:
    $$\mathcal{D} = \left\{ \left\{x_1^1, \ldots, x_{m_1}^1\right\}, \left\{x_1^2, \ldots, x_{m_2}^2\right\}, \ldots, \left\{x_1^k, \ldots, x_{m_k}^k\right\} \right\}.$$
    Nr of fisherman per day in k different regions.
  - Data tuples: $\mathcal{D} = \{(x_{1,1}, x_{1,2}), (x_{2,1}, x_{2,2}), \ldots, (x_{m,1}, x_{m,2})\}$.
    $x_{i,1} =$ Nr of fisherman at ith day, $x_{i,2}$ is the number of fishing nets used at day i.

Data configurations

- Each configuration will consist of continuous and discrete (ordinal and nominal) variables.
- **Major configuration types:**
  - A single sample configuration consists of m scalars: $\mathcal{D} = \{x_1, x_2, \ldots, x_m\}$.
    Nr of fisherman per day; $m = 365$ and $x_i = 0, 1, \ldots$.
  - Two (or more) sets of samples:
    $$\mathcal{D} = \left\{ \left\{x_1^1, \ldots, x_{m_1}^1\right\}, \left\{x_1^2, \ldots, x_{m_2}^2\right\}, \ldots, \left\{x_1^k, \ldots, x_{m_k}^k\right\} \right\}.$$
    Nr of fisherman per day in k different regions.
  - Data tuples: $\mathcal{D} = \{(x_{1,1}, x_{1,2}), (x_{2,1}, x_{2,2}), \ldots, (x_{m,1}, x_{m,2})\}$.
    $x_{i,1} =$ Nr of fisherman at ith day, $x_{i,2}$ is the number of fishing nets used at day i.
  - Generalization of tuples to vectors:
    $$\mathcal{D} = \{(x_{1,1}, \ldots, x_{1,n}), \ldots, (x_{m,1}, \ldots, x_{m,n})\}$$
    $x_{i,1} =$ Nr of fisherman at ith day, $x_{i,2}$ is the number of fishing nets used at day i, $x_{i,3} =$ Sea-Surface temperature at day $i, \ldots$

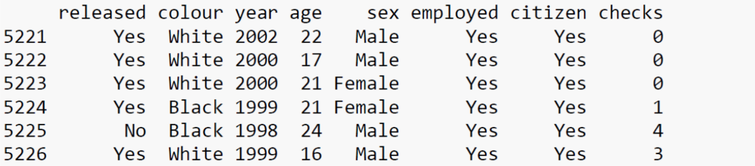
1. Data tables

The table **rows** represent observed measurements for *independent* variables (**columns**).

Observ.	variable 1	variable 2	$\cdots$	variable i	$\cdots$	variable n
1	·	·	·	·	·	·
2	·	·	·	·	·	·
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	·	·	·	·	·	·

```
library(carData)
D <- Arrests
tail(D)

#or alterantive:
library(data.table)
print(data.table(D))
```



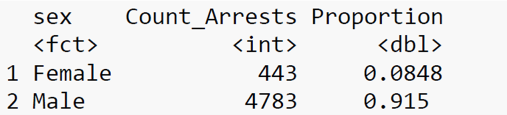
```
     released colour year age    sex employed citizen checks
5221      Yes  White 2002  22   Male      Yes     Yes      0
5222      Yes  White 2000  17   Male      Yes     Yes      0
5223      Yes  White 2000  21 Female      Yes     Yes      0
5224      Yes  Black 1999  21 Female      Yes     Yes      1
5225       No  Black 1998  24   Male      Yes     Yes      4
5226      Yes  White 1999  16   Male      Yes     Yes      3
```

Figure: Data on police arrests in Toronto for possession of marijuana.

Data summarization

A *statistic* is a numerical quantity, such as the proportion, that is computed from a sample $x_1, \ldots, x_m$.

```r
library(dplyr)
D1 <- D %>% group_by(sex) %>% summarize(Count_Arrests = n(), Proportion = Count_Arrests/nrow(D))
D1
```



```
  sex    Count_Arrests Proportion
  <fct>          <int>      <dbl>
1 Female           443     0.0848
2 Male            4783     0.915
```