Data Science and Statistics in Research: unlocking the power of your data

Session 1.4:

Data and variables

OUTLINE

Types of data

What are data and variables?

- Data are the results of sampling values of some variables associated with a population or a process.
- A variable takes on one of a set of allowed values each time it is observed.
- Variables can be either qualitative or quantitative.
- Multiple measurements of a variable form a sample.

Examples of data

- **Heart rates of a patients –** heart rates taken at various times of day
    - heart rate is a variable
    - each measurement is an observation of that variable.
- **Car attributes –** collecting fuel consumption and 10 other aspects of car design and performance for 32 automobiles.
    - fuel consumption and the 10 other aspects are variables
    - each car tested is an observation of these variables.
- **Charateristics of iris flowers –** measurements of petals for 50 iris flowers
    - petal length and width are variables
    - each iris measured is an observation of these variables.

Types of variables

Qualitative Variables

- These take on distinct values or classes.
- **Categorical**: for example, whether someone travels to work by car/bus/train/foot/bicycle/motor cycle. These are called nominal data.
- **Ordered categorical**: for example whether someone is a non-smoker/light/moderate/heavy smoker or has low/medium/high blood pressure. These are called ordinal data as there is an order to the classes.
- **Binary**: a special case of variables which take on one of two possible values, for example true/false, male/female, survived/died.

QUANTITATIVE VARIABLES

- These take on numeric values and can be of two classes.
- **Discrete**: for example, number of patients in a study, number of cases of disease.
- **Continuous**: for example, temperature, blood pressure.
- Continuous quantitative variables can have their values grouped into classes and presented as discrete or ordered categorical variables.

Example: Motor Trend Car Road Tests

- In R, there is a dataset from the 1974 Motor Trend US magazine
  - comprises fuel consumption and 10 aspects of design and performance for 32 cars (1973–74 models).
- Dataset has 32 rows (observations) and 11 columns (variables).
- Let's look at a few variables to understand the types
  - `mpg` - Miles per Gallon
  - `cyl` – Number of cylinders
  - `hp` – Horsepower
  - `wt` – Weight (1000 lbs)
  - `am` – Automatic or manual
  - `gear` – Number of gears.

Example: Motor Trend Car Road Tests

```
head(mtcars)
```

```
                   mpg cyl  hp    wt am gear
Mazda RX4         21.0   6 110 2.620  1    4
Mazda RX4 Wag     21.0   6 110 2.875  1    4
Datsun 710        22.8   4  93 2.320  1    4
Hornet 4 Drive    21.4   6 110 3.215  0    3
Hornet Sportabout 18.7   8 175 3.440  0    3
Valiant           18.1   6 105 3.460  0    3
```

EXAMPLE: MOTOR TREND CAR ROAD TESTS

- `mpg` - Miles per Gallon.
- `cyl` – Number of cylinders.
- `hp` – Horsepower.
- `wt` – Weight (1000 lbs).
- `am` – Automatic or manual.
- `gear` – Number of gears.

EXAMPLE: MOTOR TREND CAR ROAD TESTS

- `mpg` - Miles per Gallon (Continuous).
- `cyl` – Number of cylinders (Categorical).
- `hp` – Horsepower (Discrete).
- `wt` – weight (1000 lbs) (Continuous).
- `am` – Automatic or manual (Binary).
- `gear` – Number of gears (Discrete).

Example: Motor Trend Car Road Tests

- We can also group continuous quantitative variables into classes and present as discrete/ordered categorical variables.
- Lets categorise the horsepower of the cars into three groups: (i) 1-100, (ii) 101-200 and (iii) 201+.

Horsepower	Frequency	Percentage
1 - 100	9	28.1%
101 - 200	16	50.0%
201+	7	21.9%
Total	32	100%

Presentation of data

PRESENTING DATA

- We may want to describe data using
  - tabulation
  - visualisation.
- The most appropriate type of presentation will depend on
  - the variable type (qualitative/quantitative)
  - number of variables being presented.

Tabulation

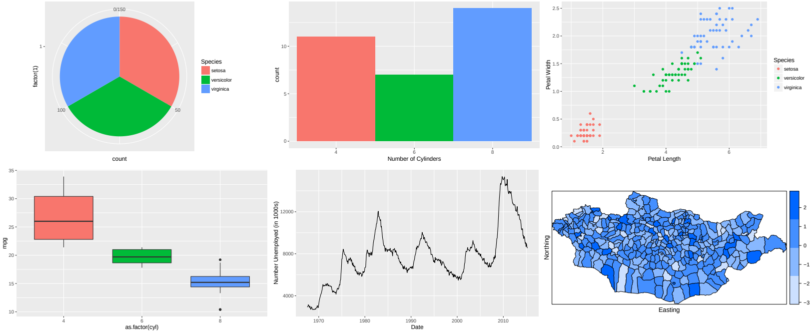
- Tables can be used to describe almost any type of data (provided there is not too much!).

Table: Years of smoking and lung capacity (on a scale 0-100) for emphysema patients.

Patient	Years Smoked	Lung Capacity
1	25	55
2	36	60
3	22	50
4	15	30
5	48	75
6	39	70

VISUALISATION

- You can visualise your data using pie charts, bar charts, histograms, scatter plots, box plots, line plots and maps.



Tables

USING TABLES TO PRESENT DATA

- The way that you present data in tables are very important.
- Readers are often drawn towards tables and figures, because it is an efficient way of obtaining information, as compared to reading a written account of the same content.
- Tables and figures add value to an analysis, if they can portray the relevant information and are concise.
- Tables and figures can provide readers with a large amount of information in a short time span.

Using Tables to present data

- Ensure that tables are self-explanatory by using clear, informative captions and titles.
- Be careful consistent in the way that you display information
  - remove repetition
  - set amount of decimal places
  - be careful of scientific notation.
- Make sure your table only contains information that adds value to your analysis.
- Always review a table as if you are a non-expert!

EXAMPLE: USING TABLES TO PRESENT DATA

- Consider an analysis that tests whether a new pesticide affects the growth of wheat plants.
- Half of the wheat plants are given the new pesticide (Treatment) with the other half not given any (Control).
- Some plants regardless of their treatment are given 12 hours of light per day and the rest 16 hours of light per day.
- The height of the wheat plants are measured after 5 and 10 days of treatment.
- For the initial data analysis, the means and variance of the wheat plants height are produced.
- The results are presented in a table.

Example: Using Tables to Present Data

Table: Height after treatment

Group	light	5 days	10 days
control	12	70.3 (2)	90 (5)
Control	16	75.7 (8)	100 (3)
treatment	12	60.4 (1.5)	78 (7.9)
Treatment	16	52.2 (2.01)	81 (6.7)

- Is this the clearest way of portraying this information?

Example: Using Tables to Present Data

Comments

- Labels are not consistent – capitalised in some places but not others.
- There are too many borders in the table
  - Many journals will not accept vertical borders.
- The way they are ordered suggests we should compare the affect of light on the height not the treatment.
- The number of decimal places are not consistent.
- We cannot see what type of descriptive statistics are being used.
- The amount of light is repeated.
- The caption does not give enough information to clearly understand the table without knowing the study information.

Example: Using Tables to Present Data

Comments

- Labels are not consistent – capitalised in some places but not others.
- There are too many lines in the table.
- The way they are ordered suggests we should compare the affect of light on the height not the treatment.
- The number of decimal places are not consistent.
- We cannot see what type of descriptive statistics are being used.
- The amount of light is repeated.
- The caption does not give enough information to clearly understand the table without knowing the study information.

Example: Using Tables to Present Data

Table: Means and variances of the height (in centimetres) of wheat plants after 5 and 10 days; for control and treatment groups.

	5 Days		10 Days	
Group	**Mean**	**Variance**	**Mean**	**Variance**
12 hours of light				
Control	70.3	2.2	90.2	5.0
Treatment	60.4	1.5	78.0	7.9
16 hours of light				
Control	75.7	7.6	99.9	2.9
Treatment	52.2	2.0	81.1	6.7

Summarising Data

Descriptive Statistics

- A statistic is calculated from the values of variable(s) in a sample.
- Various statistics are routinely used to describe samples.
- The following data refer to the total cost of drugs (in Burundi francs) received by 84 adults aged 20-29 visiting five different health centres in the Myinga province of Burundi in 1991-2.

2.0	4.0	6.2	8.0	8.0	8.0	12.0	15.0	15.0	18.0
18.2	20.0	20.0	20.0	21.0	22.0	24.2	27.0	27.0	27.0
28.0	29.7	29.7	29.7	29.7	29.7	30.0	30.0	31.0	41.4
42.3	45.4	45.4	45.4	45.4	45.4	45.4	45.4	45.4	45.4
49.4	50.8	51.4	53.4	56.0	57.0	57.4	59.0	59.4	60.0
61.5	65.4	65.4	65.4	65.4	67.0	90.8	92.0	94.0	94.0
94.0	94.0	94.0	94.0	94.7	105.0	125.0	126.0	130.0	130.0
130.0	151.2	160.0	177.0	187.0	187.0	194.4	194.4	194.4	212.4
213.0	233.0	267.0	320.4						

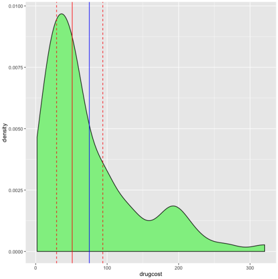
Example: Drug Costs

- There are many statistics that could be calculated from these data.
- The values the more common ones discussed earlier are listed in the following table.

Table: Sample statistics for the drug cost data.

Statistic	Value
Sample Size	84
Mean	75.1
Median	51.1
Variance	4494.9
Standard Deviation	67.0
Minimum	2
Maximum	320.4
Range	318.4
Lower Quartile	28
Upper Quartile	99.4
Interquartile Range	71.4

Figure: Density plot of the drug costs data.



Median and Quartiles

- The **median** is a measure of the central value of the distribution of data. It halves the distribution; 50% of the values are below and 50% of the values above.
- The median by itself is of limited use, so we also find the, **minimum**, **lower quartile** (Q_u), **upper quartile** (Q_l) and **maximum** which with the median (the middle quartile) split the data into four intervals.

Statistic		Quantile	List Position	R code
Minimum	min	0%	1	`min(x)`
Lower quartile	Q_l	25%	$\frac{1}{4}(N+1)$	`quantile(x,probs=0.25)`
Median	median	50%	$\frac{1}{2}(N+1)$	`median(x)`
Upper quartile	Q_u	75%	$\frac{3}{4}(N+1)$	`quantile(x,probs=0.75)`
Maximum	max	100%	N	`max(x)`

MEDIAN AND QUARTILES

- Where the list position is not a whole number, the values above and below should be averaged together to give the relevant value.
- An idea of the spread is given by calculating the inter-quartile range

$$\text{IQR} = Q_u - Q_l = \texttt{IQR(x)}$$

  or just by calculating the range

$$\begin{aligned} \text{Range} &= \max - \min \\ &= \texttt{max(x)} - \texttt{min(x)} \end{aligned}$$

- Advantage – extreme values do not affect the quartiles.
- Disadvantage – can be difficult to find if you have big data!

MEAN

- The **mean** is the most commonly used measure of the central value of a distribution.
- It is the sum of the observations divided by N the number of observations

$$\text{mean} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \texttt{sum(x)/length(x)}$$

- When the data is distributed symmetrically the mean will generally close to the mean.
- The median is far more robust to extreme values in your data.

VARIANCE

- When using the mean, the we categorise the spread of the data using the **standard deviation**, which is based on the difference of the observations from the mean.
- The **variance** is calculated by dividing the sum of squares of these deviations by $N - 1$

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \texttt{sum((x - mean(x))}^2\texttt{)/(length(x) - 1)}$$

- The standard deviation is equal to the square root of the variance.
- Advantage – uses all the information available and is therefore extensively used).
- Disadvantage – extreme values can affect the mean and standard deviation.

Any Questions?