COMP824 2023
Week 12
Describing graphs and statistical output

Overview

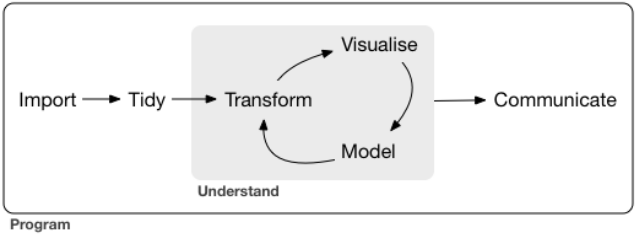
Data Types

Numeric Data

Categorical Data

OSEM Mnemonic

The Process of Analytics



Learning obectives

- Explain summary statistics and graphs of numerical data and categorical data
- Understand the OSEM model for describing data

Data Types

- ***Numeric/Quantitative***
  - **Continuous** (e.g. weight, time in seconds)
  - **Discrete** (e.g. number of cups of coffee, number of cylinders)
- ***Categorical/Qualitative***
  - **Ordinal** - categories have an order (e.g. small/medium/large, bad/okay/good)
  - **Nominal** - categories have no order(e.g. red/blue/yellow, Audi/Toyota/Honda)

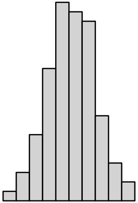
Numeric Data

With numerical data we can comment on:

- **Position** (usually a measure of centre, e.g. mean, median, mode)
- **Spread** (standard deviation, variance, IQR, range)
- **Shape** (unimodal/bimodal/multimodal, symmetric/left skewed/right skewed)
- **Special Features** (outliers, groups of points etc.)

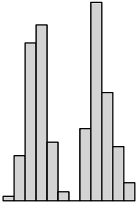
Shape of Numeric Data & Best Measure of Centre and Spread

Unimodal



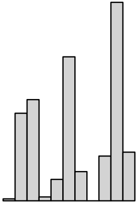
Mean
Standard Deviation

Bimodal



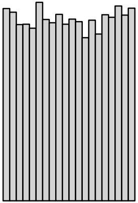
Median
IQR

Multimodal



Median
IQR

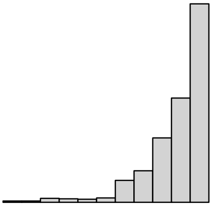
Uniform



Median
Range

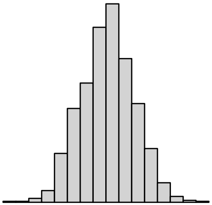
Shape of Numeric Data & Best Measure of Centre and Spread

Left Skewed



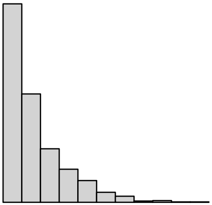
Median
IQR

Symmetric



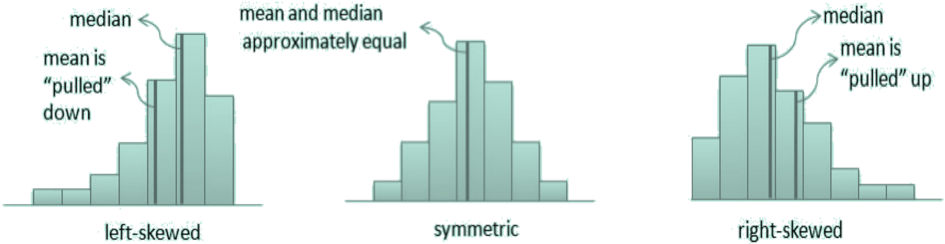
Mean
Standard Deviation

Right Skewed



Median
IQR

Which Measure of Centre and Spread?



- **Symmetric** use the mean as our measure of centre, since this uses all the data values.
- **Skewed, bimodal or multimodal** use the median, since the mean is pulled towards the tail (or extreme values).
- **Uniform** use the median, as the mode will cover all the possible values since the bar heights are approximately the same.

Categorical Data - Example

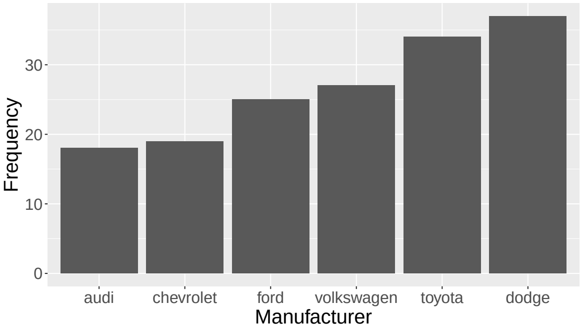
```r
mpg_count <- mpg %>% count(manufacturer) %>%
              arrange(-n) %>% slice_head(n = 6)

mpg_count %>% kableExtra::kable()
```

manufacturer	n
dodge	37
toyota	34
volkswagen	27
ford	25
chevrolet	19
audi	18

Categorical Data - Example

```
mpg_count %>% ggplot() +
  geom_col(mapping = aes(x = reorder( manufacturer, n),
                         y = n)) +
  labs(x = "Manufacturer",
       y = "Frequency")
```



Categorical Data - Description

Categorical data has no centre, spread or shape.

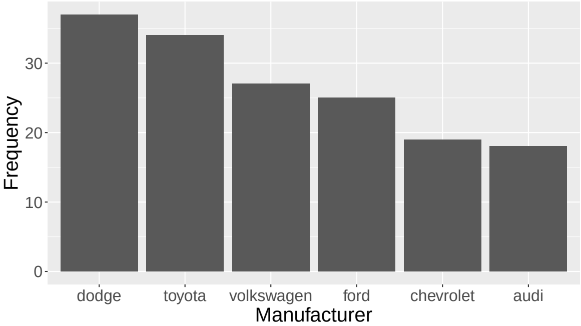
With categories we talk about:

- The Mode or most common category
- Count (or percent) for each category

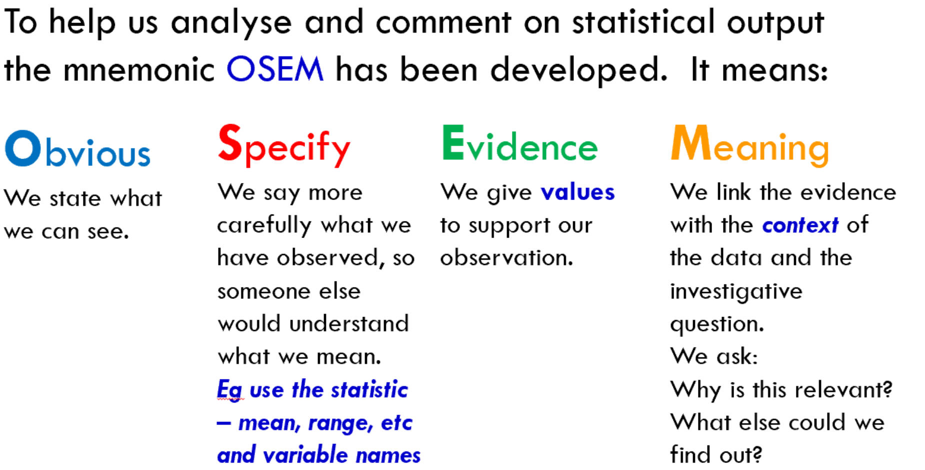
Discussion: Why does categorical data have no shape?

Categorical Data - Example

```
mpg_count %>% ggplot() +
  geom_col(mapping = aes(x = reorder(manufacturer, -n),
                         y = n)) +
  labs(x = "Manufacturer",
       y = "Frequency")
```



OSEM Mnemonic



To help us analyse and comment on statistical output the mnemonic OSEM has been developed. It means:

Obvious

We state what we can see.

Specify

We say more carefully what we have observed, so someone else would understand what we mean.

Eg use the statistic – mean, range, etc and variable names

Evidence

We give **values** to support our observation.

Meaning

We link the evidence with the ***context*** of the data and the investigative question.

We ask:

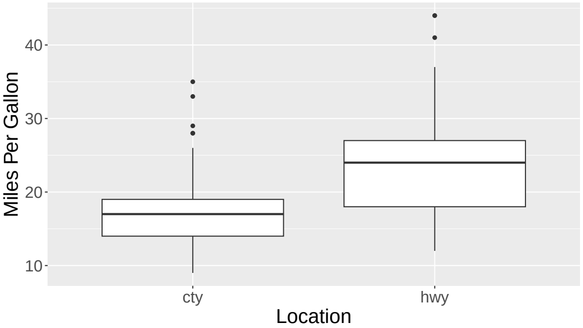
Why is this relevant?

What else could we find out?

OSEM provides a useful framework to get you started. It is not a definitive rule.

Example: Cars - Highway Miles Per Gallon

```
mpg %>% pivot_longer(hwy:cty) %>%
  ggplot() + geom_boxplot(mapping = aes(x = name, y = value)) +
  labs(x = "Location", y = "Miles Per Gallon")
```



Example: Cars - Highway Miles Per Gallon

```r
mpg %>% select(cty, hwy) %>% summary()
```

```
      cty             hwy       
 Min.   : 9.00   Min.   :12.00  
 1st Qu.:14.00   1st Qu.:18.00  
 Median :17.00   Median :24.00  
 Mean   :16.86   Mean   :23.44  
 3rd Qu.:19.00   3rd Qu.:27.00  
 Max.   :35.00   Max.   :44.00  
```

Cars - Miles Per Gallon - OSEM (position)

- **Obvious -** cty is lower than hwy
- **Specific -** the median for cty is less than the lower quartile for hwy
- **Evidence -** the median for miles per gallon in the city is 17 which is slightly less than the lower quartile for miles per gallon on a highway, which is 18.
- **Meaning -** For the vehicles in the mpg dataset, the fuel efficiency when driving in the city is lower than the fuel efficency when driving on a highway.

Final statement

The median for miles per gallon in the city is 17 which is slightly less than the lower quartile for miles per gallon on a highway, which is 18. For the vehicles in the mpg dataset, the fuel efficiency when driving in the city is lower than the fuel efficency when driving on a highway.

Cars - Miles Per Gallon - OSEM (spread)

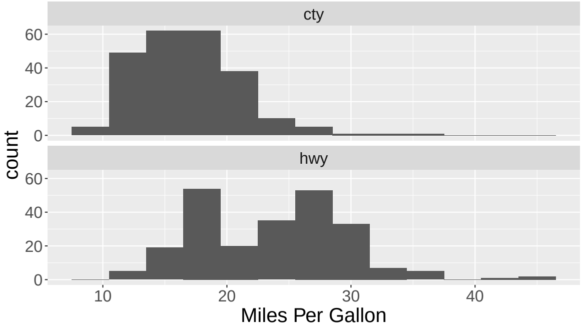
- **Obvious -** cty smaller box, hwy bigger box
- **Specific** - cty has a smaller IQR than hwy
- **Evidence** - The miles per gallon in the city is has an interquartile range of 5 which is almost half the interquartile range of miles per gallon on the highway, which is 9.
- **Meaning -** For the vehicles in the mpg dataset, there is more variation in the miles per gallon for vehicles driving on a highway, compared with driving in the city.

Final Statement

The miles per gallon in the city has an interquartile range of 5 which is almost half the interquartile range of miles per gallon on the highway, which is 9. There is more variation in the miles per gallon for vehicles driving on a highway, compared with driving in the city.

Cars - Miles Per Gallon - OSEM (shape)

```
mpg %>% pivot_longer(hwy:cty) %>% ggplot() +
  geom_histogram(mapping = aes(x = value), binwidth = 3) +
  labs( x = "Miles Per Gallon") +
  facet_wrap(vars(name), nrow = 2)
```



Cars - Miles Per Gallon - OSEM (shape)

- **Obvious** - cty one peak, hwy two peaks
- **Specific** - cty is unimodal and right skewed, hwy is bimodal
- **Evidence** - The miles per gallon in the city is unimodal and slightly right skewed, with a peak around the median of 17. The miles per gallon on a highway is bimodal with peaks around 18 and 27.
- **Meaning** - This suggests that some vehicles are less efficient, having a low highway miles per gallon, and some vehicles are more efficient, having a higher highway miles per gallon. This difference is not as apparaent when driving in the city.

Note: Additional analysis could be done to find out if there is anything special about the vehicles with a lower fuel efficiency (low `hwy`) - perhaps exploring the `class`.

Cars - Miles Per Gallon - OSEM (specials)

- **Obvious** - cty and hwy both have outliers
- **Specific** - cty 4 outliers, hwy 2 outliers
- **Evidence** - The miles per gallon in the city has 4 outliers with values above about 26 and miles per gallon on a highway has two outliers with values above 36.
- **Meaning** - This suggests that there a small number of vehicles that have a very high fuel effiency.

Cars - Classes

```r
mpg %>%
  count(class) %>%
  arrange(n) %>% print(n=7)
```

```
# A tibble: 7 x 2
  class          n
  <chr>      <int>
1 2seater        5
2 minivan       11
3 pickup        33
4 subcompact    35
5 midsize       41
6 compact       47
7 suv           62
```

Cars - Classes

```
mpg %>% count()
```

```
# A tibble: 1 x 1
      n
  <int>
1   234
```

- **Obvious** - most SUV, least 2 seater
- **Specific** - 234 vehices, 62 SUV, 5 2-seaters
- **Evidence** - The dataset contains information on 234 vehicles of a variety of types. The most frequent types were SUVs (62 vehicles) and the least common were 2-seaters (5 vehicles) and minivans (11 vehicles). The remaining types appeared between 33 and 47 times each in the dataset.
- **Meaning** - Various answers - would depend on what else was being discussed.

Bringing it all together (1)

The dataset contains information on 234 vehicles of a variety of classes. The most frequent vehicle classes were SUVs (62 vehicles) and the least common were 2-seaters (5 vehicles) and minivans (11 vehicles). The remaining 4 classes appeared between 33 and 47 times each in the dataset. The miles per gallon in the city and on a highway were reported for all vehicles.

As shown in the table of summary statistics, the median for miles per gallon in the city (17) is slightly less than the lower quartile for miles per gallon on a highway (18). Interquartile range for miles per gallon in the city is 5 which is almost half the interquartile range of miles per gallon on the highway, which is 9. For the vehicles in the mpg dataset, the fuel efficiency when driving in the city is lower and less variable than the fuel efficency when driving on a highway.

Bringing it all together (2)

As shown in the histogram, the miles per gallon in the city is unimodal and slightly right skewed, with a peak around the median of 17. The miles per gallon on a highway is bimodal with peaks around 18 and 27. This suggests that some vehicles are less efficient, having a low highway miles per gallon, and some vehicles are more efficient, having a higher highway miles per gallon. This difference is not as apparent when driving in the city. The boxplot shows that the miles per gallon in the city has 4 outliers with values above about 26 and miles per gallon on a highway has two outliers with values above 36. This suggests that there a small number of vehicles that have a very high fuel efficiency.

Points to note

- As you get more experience writing about data you will be able to combine multiple observations into a single paragraph.
- You don’t necessarily need to mention every aspect (position, shape, spread, specials) every time. Think about the “story” that you are telling and choose the appropriate statistics and graphs.
- The summary on the previous slide mentions “the histogram”, “the boxplox”, “the table” etc. In a report, you should use “Figure 1”, “Figure 2” etc to refer to figures and “Table 1” etc to refer to tables.
- OSEM provides a useful framework to get you started. It is not a definitive rule.

OSEM Resources

- Video: https://www.youtube.com/watch?v=L-ur3pRYKFk
- Resource: https://new.censusatschool.org.nz/wp-content/uploads/2014/12/Auckland-Stats-Teachers-Day-Brocklehurst-OSEM.pptx

Summary

Data Types

Numeric Data

Categorical Data

OSEM Mnemonic

Learning obectives

- Explain summary statistics and graphs of numerical data and categorical data
- Understand the OSEM model for describing data

References