Week 3 3.2 Discover Note

Lecture 7: Centre

---

1. Numerical Summaries

- Reduce all the date to **1 simple number** ("statistic")
  - This loses a lot of information...
  - ...but allows **easy communication and comparison**.
- Major Features:
  - Maximum
  - Minimum
  - Centre (mean, median)
  - Spread (standard deviation, range, IQR)

2. Mean and Median

> Mean（平均数）: the average of the data.

- Mean = Sum of data/Size of data
- Commands:

```
# Calculate the mean:
mean(data$Sold)
# To focus specifically on a variable/variables:
mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"])
# ↑ It means that only choose "houses" with "4" bedrooms in the data of all the property sold.
```

- Mean is the **balancing point** in the data. On a histogram:

```
# Create a histogram:
hist(data$Sold, main="Newtown Properties", xlab="Price (in 1000s)")
# Add a vertical (v) "green" (col=) line (adline) of mean:
abline(v=mean(data$Sold), col="green")
```

```
hist(data$Sold[data$Type=="House" & data$Bedrooms=="4"], main="Newtown 4 Bedrooms", xlab="Price (in 1000s)")
abline(v=mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="green")
```

> Median（中位数）: the **middle data point**, when the data is ordered from *smallest to largest*.

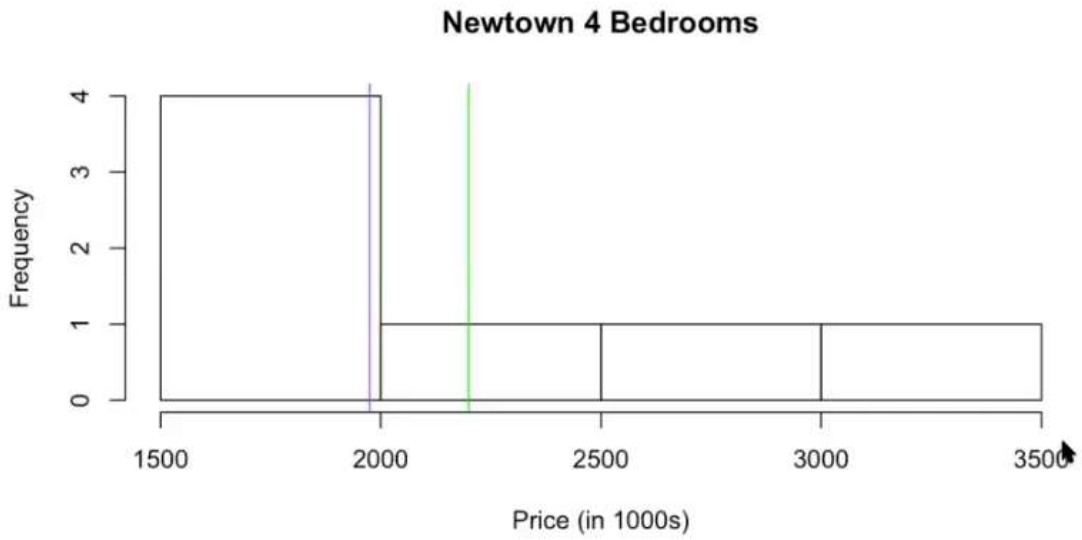
- Median =
  - the unique middle point (in an *odd* sized dataset)
  - the average of the 2 middle points (in an *even* sized dataset)
- Commands:

```
# Order the data (ranked data):
sort(data$Sold)
# Measure the length:
length(data$Sold)
# Calculate the median:
median(data$Sold)
# To focus specifically on a variable/variables:
median(data$Sold[data$Type=="House" & data$Bedrooms=="4"])
```

- Median is the **half way point** in the data. On a histogram:

```
# Create a histogram:
hist(data$Sold)
# Add a vertical (v) "purple" (col=) line (adline) of median:
abline(v=median(data$Sold), col="purple")
```

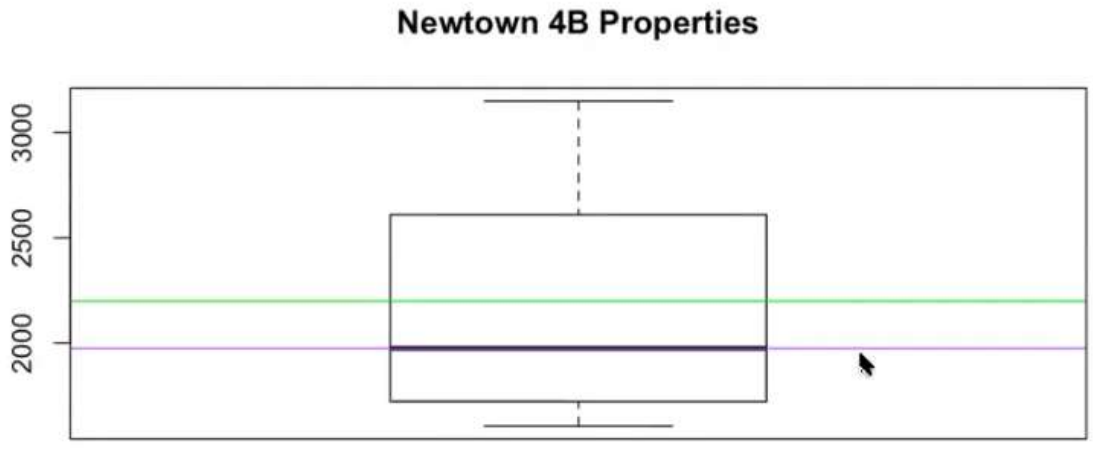
```
hist(data$Sold[data$Type=="House" & data$Bedrooms=="4"], main="Newtown 4 Bedrooms", xlab="Price (in 1000s)")
abline(v=mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="green")
abline(v=median(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="purple")
```



- Both on a boxplot:

```
boxplot(data$Sold, main="Newtown Properties")
abline(v=mean(data$Sold), col="green")
abline(v=median(data$Sold), col="purple")
```

```
boxplot(data$Sold[data$Type=="House" & data$Bedrooms=="4"], main = "Newtown 4B Properties")
abline(h=mean(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="green")
abline(h=median(data$Sold[data$Type=="House" & data$Bedrooms=="4"]),col="purple")
```



3. Robustness and Comparisons

> Robustness（顽健性）: The median is said to be **robust** and is a good summary for skewed data as it is *not affected by* **outliers**.

> **Comparison: The difference between the mean and the median can be an indication of the shape of the data.**

Data Types	Mean Compared to Median
Symmetric	Same
Left Skewed	Smaller
Right Skewed	Larger

Lecture 8: Spread

---

1. Standard Deviation

- To measure the spread, we can calculate the `gaps`.
- Commands:

```
# Measure all the gaps:
gaps = data$Sold - mean(Data$Sold)
# To check the maximum in the gaps:
max(gaps)
```

```
gaps = data$Sold - mean(data$Sold)
gaps
```

```
##  [1]   567.857143  -157.142857  -127.142857  -627.142857  -757.142857
##  [6]   692.857143  -732.142857  -667.142857  -782.142857   542.857143
## [11]   -32.142857   167.857143  -408.142857  -452.142857  -547.142857
## [16]   197.857143   182.857143  -167.142857 -1037.142857   532.857143
## [21]  -687.142857  -452.142857  -487.142857   442.857143   192.857143
## [26]  -652.142857    -7.142857   145.857143  1402.857143   192.857143
## [31]   792.857143   372.857143   398.857143   293.857143   -98.142857
## [36]  -307.142857  -472.142857  -762.142857    52.857143   -37.142857
## [41]  -715.142857  1742.857143  1002.857143  -637.142857   254.857143
## [46]   827.857143   592.857143   382.857143   342.857143   302.857143
## [51]   192.857143  -546.142857  -667.142857   -92.142857   892.857143
## [56]  -595.142857
```

```
max(gaps)
```

```
## [1] 1742.857
```

> **RMS(Root Mean Square)（均方根）: the average of a set of numbers, regardless of the signs.**

- Steps (in *S-M-R* order): *square* the numbers, then *mean* the result, then *root* the overall result:
- Commands:

```
# Apply RMS to the gaps:
sqrt(mean(gaps^2))
```

> SD (Standard Deviation)（标准差）: measures the spread of the data.

- *Difference* between RMS and SD: RMS is based on **population**, while SD is based on **samples**. So, SD may need to multiply `sqrt(n-1/n)` to make the result on population (n).
- Commands:

```r
# Calculate the standard deviation:
sd(data$Sold)
# Adjusting (to make the sd equal to RMS)
sd(data$Sold)*sqrt(55/56)

# Another convenient way by installing multicon package:
install.packages("multicon")
library(multcon)
popsd(data$Sold) # popsd means sd on population
```

2. Standard Units (Z Score)

- Standard Units = (data point - mean) / SD
- For many data sets, we find that roughly:

percentage of Data	Distance from Mean
68%	Within 1 SD
95%	Within 2 SDs
99.7%	Within 3 SDs

> **Standard Unit** of A Data Point = how many standard deviations are below the mean.
>
> ### IQR (Interquartile Range)（四分位距）: Range of the middle 50% of the data.
>
> - IQR = Q3 (3rd Quartile) - Q1 (1st Quartile)
> - The median is the 50% or 2nd quartile (Q2)
> - Commands:

```r
# List all the quartiles of the data:
quantile(data$Sold)
# Calculate the IQR:
quantile(data$Sold)[4] - quantile(data$Sold)[2]
```

```r
quantile(data$Sold)
```

```
##      0%     25%     50%     75%    100%
##  370.00  860.75 1387.50 1782.50 3150.00
```

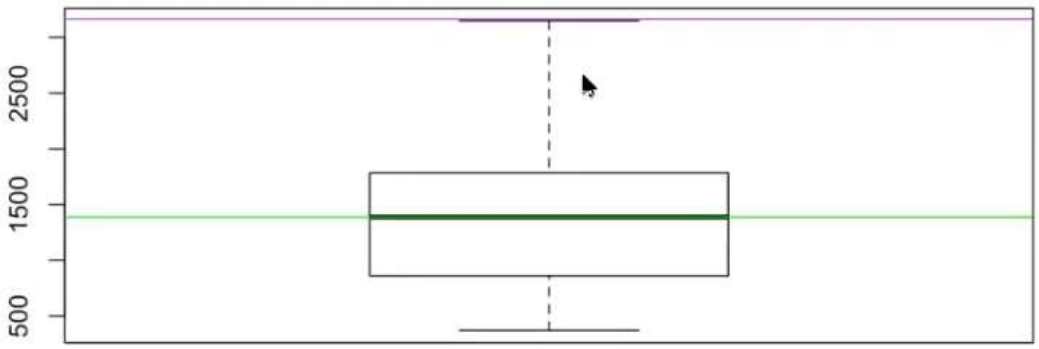
```r
quantile(data$Sold)[4] - quantile(data$Sold)[2]
```

```
##    75%
## 921.75
```

> IQR on a boxplot: The **length of the box** in the boxplot, representing the span of **50%**.

- The **lower** and **upper thresholds** are a distance of 1.5 from the quartiles:
  - LT: Q1 - 1.5IQR
  - UT: Q3 + 1.5IQR
- Data outside these thresholds is considered an **outlier** (Extreme reading).

```
boxplot(data$Sold)
iqr=quantile(data$Sold)[4] - quantile(data$Sold)[2]
abline(h=median(data$Sold),col="green")
abline(h=quantile(data$Sold)[2]- 1.5*iqr,col="purple")
abline(h=quantile(data$Sold)[4]+ 1.5*iqr,col="purple")
```



3. Reporting

- IQR and median are both **robust**, so they are suitable for **skewed data**.
- We report in pairs: (mean, SD) or (median, IQR)

4. Coefficient of Variation

> CV (Coefficient of Variation)（变异系数）: __Combines the SD and mean++ into 1 summary.

- CV = SD/Mean
- Commands:

```
m = mean(data$Sold)
sd = sd(data$Sold)
sd/m
```

Lecture 9: Data Wrangling

> Data Wrangling: Whatever is needed to get the data for analysis, also known as data munging or data janitor work.

- Steps:
  - Sourcing data
  - Scarping data
  - Cleaning and tidying data
  - Reshaping data
  - Splitting data
  - Combining data
  - Summarising data