R：Statistical Programming Methods
R：程式、機率與統計

Exploratory Data Analysis (2)

Descriptive Statistics

敘述性統計

Which chicken is fatter?

- A farmer named wanted to explore the effects of two new types of chicken feed. To do this, he gathered two groups of chickens as samples and carefully observed any differences that emerged between them.

- Mr Rogers also happens to have a degree in statistics, so he writes out his experimental information using statistics and hypotheses:
  - H0: There is no difference in mean weights between the chicken groups
  - H1: There is a differences between the mean weights of the two chicken groups

```r
chicken <- read.csv("chicken.csv", header=TRUE)

chickenA <- subset(chicken, Diet=="A")
chickenB <- subset(chicken, Diet=="B")


 mean(chickenA$weight)

 ## [1] 142.95

 mean(chickenB$weight)

 ## [1] 135.2627
```

Results

- The average weight of chicken with Feed A (Diet 3 in your data) and Feed B (Diet 4 in your data) :

Feed A	Feed B
Mean Weight = 142.95g	Mean weight = 135.27g

- Looks like Feed A can make more profit than Feed B?

Why maybe?

- Now RANDOMLY split one of the experimental groups into two sub-groups.
- If these chickens were fed the same, the weight of every chicken is expected to be the no difference.
- This difference is a naturally occurring random effects. The means of two measurements from two groups on a continuous distribution (weight is continuous), will always be different.
- In summary, the chickens that eat feed A have a larger weight than the chickens that eat feed B, but it may due to **naturally occurring variation** (or unsystematic random effects). 只是剛好，還是Feed A真的比較好?
- So how do we solve this problem? How can we be sure?

Inferential Statistics 推論統計

- There is no 100% certain that Feed A is better than Feed B.
- We can only give a probability that this is NOT due to chance.
  - 有多少的機率不是"剛好"

- "Statistically Significant" 統計上顯著
  - 90%, 95%, 99% confident that the **difference is due to the experimental condition**, and not naturally occurring variation. The higher the confidence, the lower the error rate, the harder it is to 'prove' statistically.

```r
#sample function
sample(1:15, 10, replace=FALSE)
##  [1] 13  2  7  3 15  9  8  5 14 12

sample(1:15, 10, replace=TRUE)
##  [1] 12  2 11 12 10  9  9  7  1  2
```

#randomly split A into two groups

```
#randomly split A into two groups
sampleA1 <- sample(nrow(chickenA), (nrow(chickenA)/2), replace=FALSE)
sampleA1
```

```
##  [1]  14  12  50  48  42   5 106  59  71  65  34 104 100  13  38
10  26  76  70

## [20]  87  15  81  39  49  74  36  73 107  93  79  96  31  77  47
4  43 116  44

## [39]  52 108 114  61  16   2  83  33  11  30  84  56  19  68  89
23  37 103 109

## [58]  90  75   8
```

```
chickenA1_sample <- chickenA[sampleA1, ]

chickenA2_sample <- chickenA[-sampleA1, ];
```

Non-A1 Sample

Are two sample mean the same?

```
mean(chickenA1_sample$weight)
## [1] 139.5667
mean(chickenA2_sample$weight)
## [1] 146.3333
```

Practice

- Download the data “Customer.csv”
- This data reflects the customer details in a retail store
  - Customer ID
  - Gender (Male/Female)
  - Income
  - Score (i.e., spending score)
  - Profession (e.g., Lawyer/Engineer, Healthcare)
  - Work (i.e., years of experience)
  - Family (i.e., number of family member)
- Calculate the average and standard deviation of spending score for both Male and Female
- Calculate the average working years experience for both Engineer and Lawyer.