統計諮詢：Mid-term Project 1

國立成功大學統計學系暨數據科學研究所

陳溫茹（R26091040)、廖傑恩（RE6094028)、戴庭筠（R26091032）

2021-05-30

1 問題敘述

已知相較於其他肉，家禽肉做成的熱狗比較健康，例如：相較全牛肉與豬牛混合肉，家禽肉的熱狗卡路里（calorie）比較低。不過，有人好奇家禽肉熱狗的鈉含量（sodium content）是否高於其他肉做成的熱狗。

本研究的問題為：相較於全牛肉與混合肉者的熱狗，家禽肉的熱狗鈉含量是否比較高。

2 資料集敘述

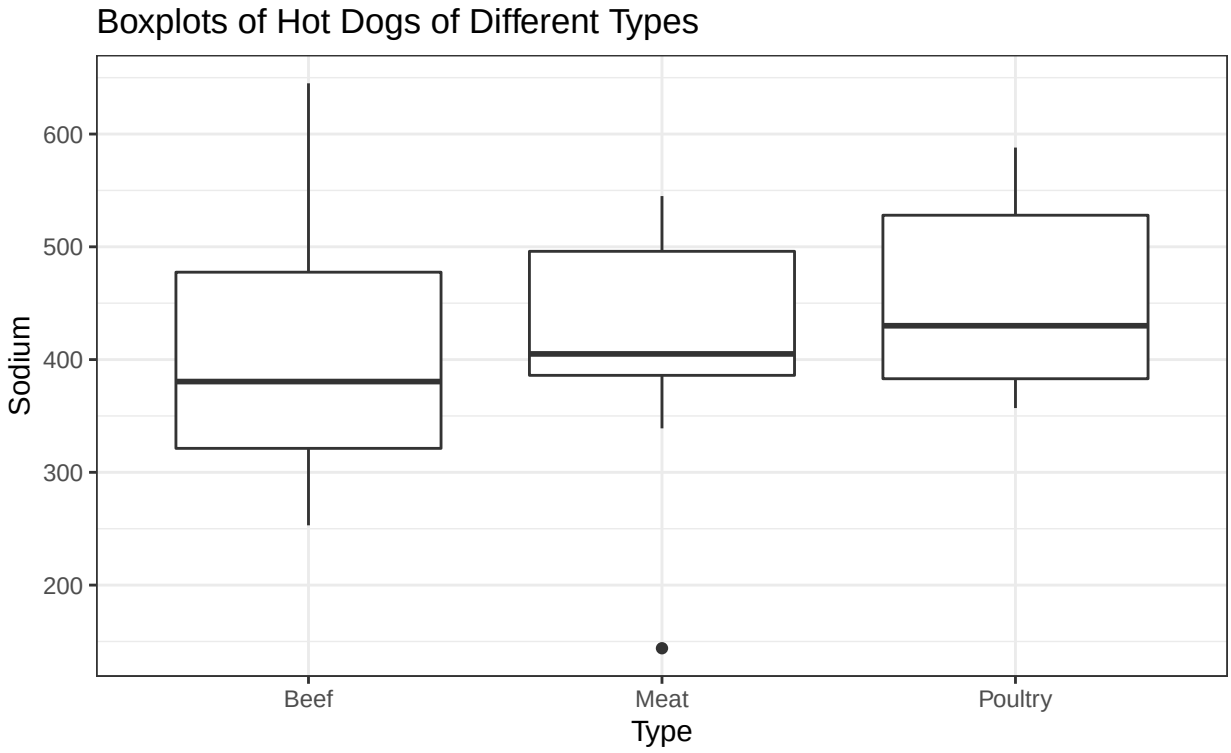
關於熱狗的資料集 `hotdog` 來自於 1986 年的 Consumer Reports，Moore 與 McCabe 在 1989 年時也有使用這份資料進行研究。資料有 54 筆，每一筆為一個熱狗的資料，包含 3 個變項：

- `Type`: 熱狗型態，包含：牛肉（beef)、混合肉（meat）與家禽肉（poultry）三種
- `Calories`: 熱狗卡路里含量
- `Sodium`: 熱狗鈉含量（單位：毫克）

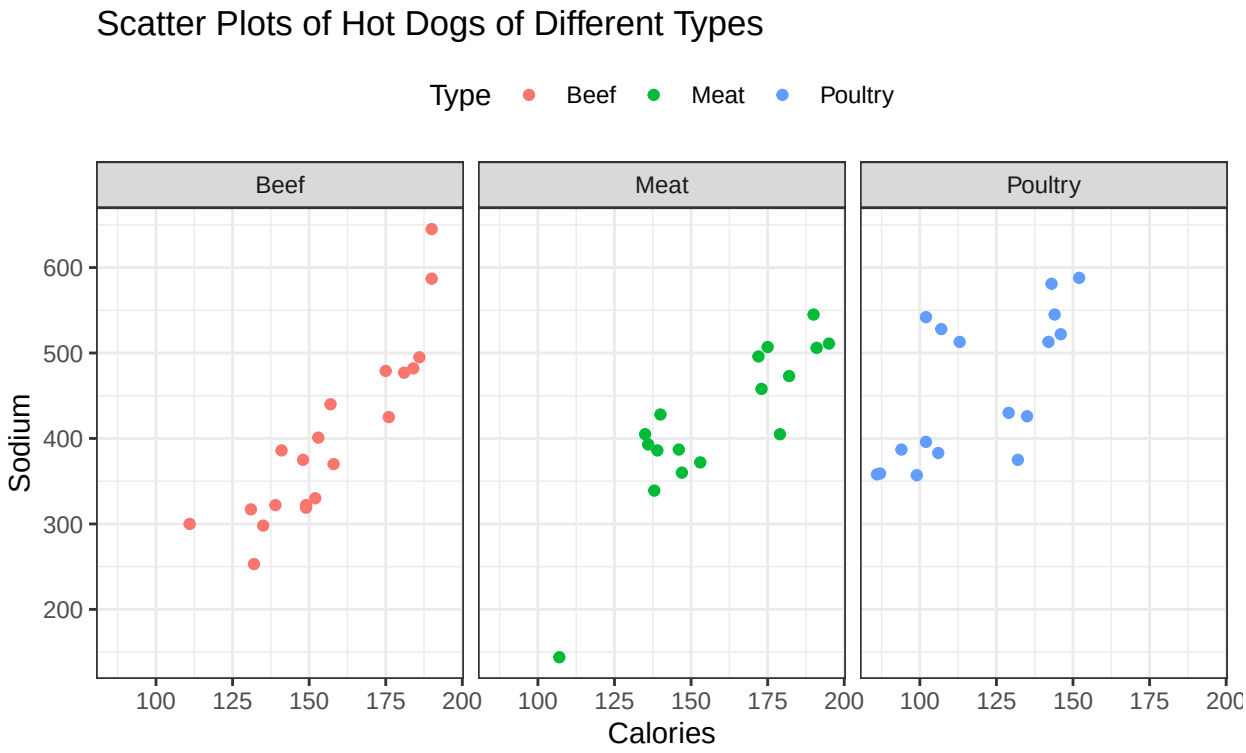
三種型態的熱狗數量分布如下：

表 1: 熱狗各型態數量表

Type	數量
Beef	20
Meat	17
Poultry	17



由上盒鬚圖（box plot）可以發現，三種型態的熱狗鈉含量的中位數都位在 400 左右，牛肉熱狗鈉含量的變異較大，混合肉熱狗鈉含量的變異較小。



上圖是熱狗卡路里與鈉含量的散佈圖（scatter plot），可以觀察到不論型態為何，熱狗卡路里與鈉含量皆似乎呈現正相關。

3 統計分析

3.1 獨立樣本單因子變異數分析 (one-way anylysis of variance, one-way ANOVA)

3.1.1 研究假說

我們欲以 one-way ANOVA 檢驗不同型態的熱狗鈉含量是否有所差異，其中獨變項（independent variable）為熱狗型態，依變項為熱狗鈉含量。令 μ_a 為型態 a 的熱夠的鈉含量母體平均數。檢定的虛無假設（null hypothesis）與對立假設（alternative hypothesis）如下：

- 虛無假設：家禽肉、全牛肉與混合肉的三種型態的熱狗鈉含量均相同

$$H_0 : \mu_{poultry} = \mu_{beef} = \mu_{meat}$$

- 對立假設：家禽肉、全牛肉與混合肉的熱狗三者，其中至少有兩種的鈉含量不同

$$H_1 : \mu_{poultry} \neq \mu_{beef} \ or \ \mu_{poultry} \neq \mu_{meat} \ or \ \mu_{beef} \neq \mu_{meat}$$

3.1.2 檢驗 ANOVA 的前提假設是否被滿足

One-way ANOVA 有若干個前提假設（assumption），在進行檢驗之前，我們先檢查資料是否符合這些假設。

1. 獨變項須為類別變數（categorical variable），依變項必須是連續變數（continuous variable）

此分析獨變項為熱狗型態，為含有 3 個類別的類別變數；依變項為熱狗鈉含量，為連續變項。符合。

2. 各組樣本依變項獨立

此分析中，各組依變項為牛肉熱狗鈉含量、混合肉熱狗鈉含量、家禽肉熱狗鈉含量，此三者互不影響彼此，符合前提假設。

3. 變異數同質（homogeneity of variance）：各組依變項的變異數必須相等。

- 針對依變項進行常態檢定

在常態分布下，以 Bartlett 檢定變異數同質性會有較高的統計檢定力（power）（Lim & Loh，1996），因此在進行變異數同質性檢定前，我們先以 Shapiro-Wilk 常態檢定法對依變項（Y）進行常態檢定，檢定的假說如下，顯著水準設定為 0.05。

$$H_0 : Y \sim ND \;\; v.s. \;\; H_1 : Y \; does \; not \sim ND$$

檢定結果：檢定統計量為 0.9796，其 p 值為 0.4836，不小於顯著水準，因此我們不拒絕 H_0，也就是說我們沒有足夠的證據證明母體分配不服從常態分佈。

- 針對依變項進行變異數同質檢定

我們以 Bartlett 檢定檢驗變異數同質是否成立。令 σ_x^2 為型態 x 的熱狗鈉含量的母體變異數，研究假說如下：

$$\begin{cases} H_0 : \sigma_{beef}^2 = \sigma_{meat}^2 = \sigma_{poultry}^2 \\ H_1 : \sigma_{beef}^2 \neq \sigma_{meat}^2 \; or \; \sigma_{beef}^2 \neq \sigma_{poultry}^2 \; or \; \sigma_{meat}^2 \neq \sigma_{poultry}^2 \end{cases}$$

我們同樣令顯著水準為 0.05。檢定結果的檢定統計量 $Barlett's \; k^2$ 為 0.6006，其 p 值為 0.7406，不小於顯著水準，因此我們不拒絕 H_0，意味著我們無法證明至少有一組母體變異數與其他組不同，也就是說我們沒有足夠的證據證明變異數同質性不存在，通過此前提假設。

4. 殘差（residuals）服從常態分配。

待配適完模型後診斷。

以上步驟顯示，在我們的資料中，ANOVA 的前提假設均滿足（殘差常態假設待檢驗），因此我們可以進行 ANOVA。

3.1.3 檢定統計量與檢定結果

One-way ANOVA 的檢定統計量為 F 值，其服從自由度為 $k-1$ 與 $N-k$ 的 F 分配，數學式如下：

$$F_{TS} = \frac{explained \; variation}{unexplained \; variation} = \frac{\sum_{j=1}^{k} n_j(\bar{Y_j} - \bar{Y})/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y_j})^2/(N-k)} \sim F(k-1, N-k)$$

其中，k 為獨變項組別數；n_j 為第 j 組的觀察值（observations）個數；$N = \sum_{j=1}^{k} n_j$，也就是總觀察值數；$\bar{Y_j}$ 為第 j 組依變項的樣本平均數，$\bar{Y}$ 為依變項樣本平均數；Y_{ji} 為第 j 組的第 i 個依變項觀察值。

我們令顯著水準為 0.05，檢定統計量 F_{TS} 為 1.778，P 值為 0.179，不小於顯著水準，因此我們不拒絕 H_0，意味著我們無法證明「家禽肉、全牛肉與混合肉的三種型態的熱狗鈉含量均相同」這個宣稱是錯的，所以對於研究問題「相較於全牛肉與混合肉者的熱狗，家禽肉的熱狗鈉含量是否比較高」，以目前分析結果的回答是「否」。不過，在前面的資料描述中，我們已發現卡路里與鈉含量呈現些微的相關，而且我們已知不同型態的熱狗卡路里含量不同，顯示卡路里可能與鈉含量有所關聯，進而混淆了我們先前在 one-way ANOVA 中看到的熱狗型態對於鈉含量變異的效果，也就是說，當我們嘗試以

熱狗型態來對鈉含量進行預測時，其預測力可能包含卡路里的貢獻。我們想嘗試排除熱狗卡路里對於鈉含量變異的效果，單純來看熱狗型態對於鈉含量變異的效果，因此需再要進行共變異數分析。

3.1.4 檢驗殘差是否為常態分配

完成 One-way ANOVA 分析後，我們須檢驗殘差是否為常態分配，來確定是否符合 ANOVA 的前提假設。我們以 Shapiro-Wilk 常態檢定法對殘差進行常態檢定，檢定的假說如下:

$$H_0: \ Residuals \sim ND \quad v.s. \quad H_1 : not \ H_0$$

我們令信心水準為 0.95，透過資料算出統計量 W 為 0.9653，p 值為 0.1187，大於 0.05，我們無顯著證據顯示殘差不服從常態分配，因此我們通過殘差項的假設。

3.2 共變異數分析（analysis of covariance, ANCOVA）

3.2.1 潛在共變項

共變異數分析旨在排除與主要獨變項存在共線性（collinearity）的「共變項」對依變項的效果，以淨化主要獨變項的效果。我們想嘗試排除熱狗卡路里對於鈉含量變異的效果，單純來看熱狗型態對於鈉含量變異的效果。

3.2.2 ANCOVA 模型假設

ANCOVA 為線性迴歸與 ANOVA 模型之結合，其模型假設為:

$$Y_{ji} = \mu + a_j + \beta_j(X_{ji} - \bar{X}_{..}) + \epsilon_{ji} = \mu_j + \beta_j(X_{ji} - \bar{X}_{..}) + \epsilon_{ji}$$

$$\forall j = 1, 2, ..., k; \ i = 1, 2, ..., n_j, \ \epsilon_{ji} \sim N(0, \sigma_j^2)$$

其中 j 為處理數，本研究中表示家禽肉、牛肉與混和肉，i 為處理內的樣本數，Y_{ji} 為依變項，X_{ji} 為共變項，$\bar{X}_{..}$ 為共變項的總體平均，亦即 $\bar{X}_{..} = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ji}$ （$N = \sum_{j=1}^{k} n_j$ ），μ 為整體期望值，α_i 為處理效應，β_i 為該處理的依變項與共變項做簡單線性迴歸後的斜率，ϵ_{ji} 為隨機效應，服從常態分配。

3.2.3 檢驗 ANCOVA 的前提假設是否被滿足

若要進行 ANCOVA 檢定必須滿足下列的假設:

1. 共變項 X 必須是連續變項，並與依變項 Y 具線性關係:

本次檢定的共變項為卡路里，為連續變項，且分別對家禽肉、牛肉、混和肉的鈉含量（Y_{ji}）與卡路里（X_{ji} ）做簡單線性迴歸，調整後的 R^2 分別為 0.73、0.78 與 0.73，表示這三筆資料皆可弭合簡單線性模型。

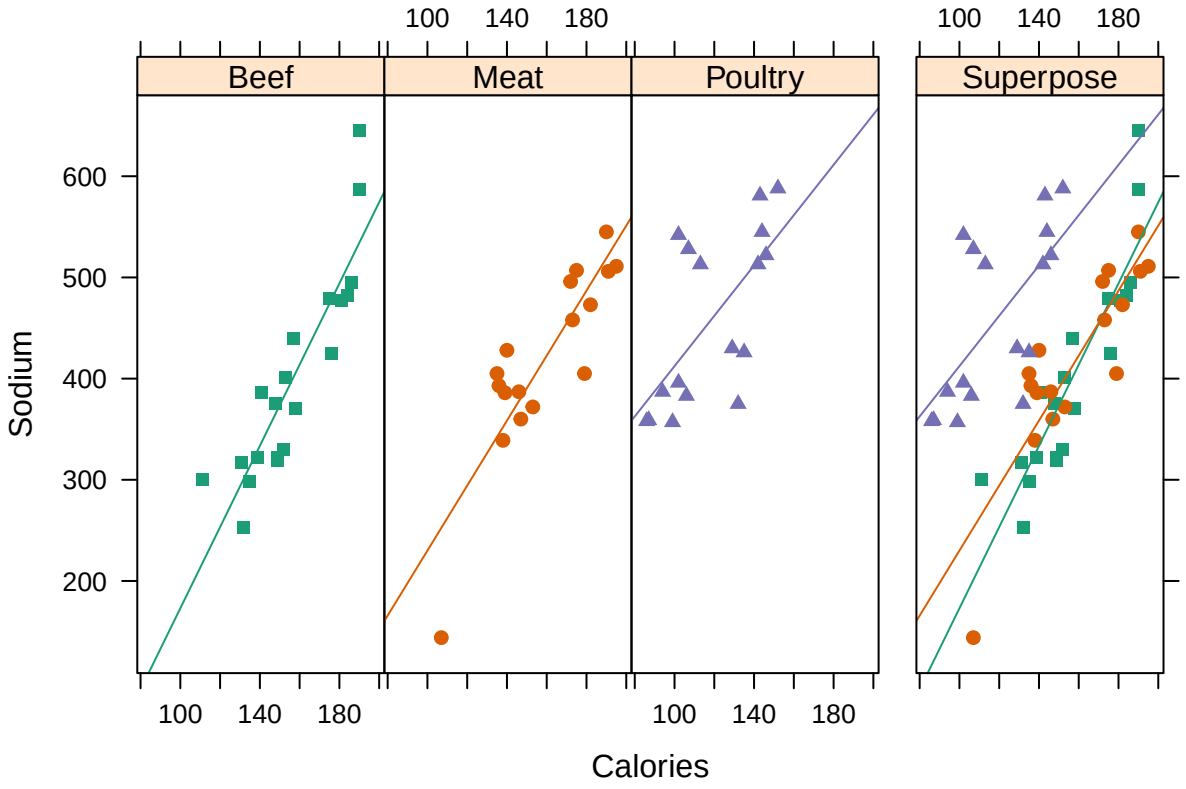
2. 殘差（ϵ_{ji}）具獨立性，且皆服從變異數相同，期望值 μ 為 0 的常態分配，亦即 $\epsilon_{ji} \overset{iid}{\sim} N(0, \sigma_i^2)$ ： 待配適完模型後診斷。

3. 共變項與獨變項無交互作用，亦即 $\beta_1 = \beta_2 = ... = \beta_k = \beta$（$k = 3$ here）。

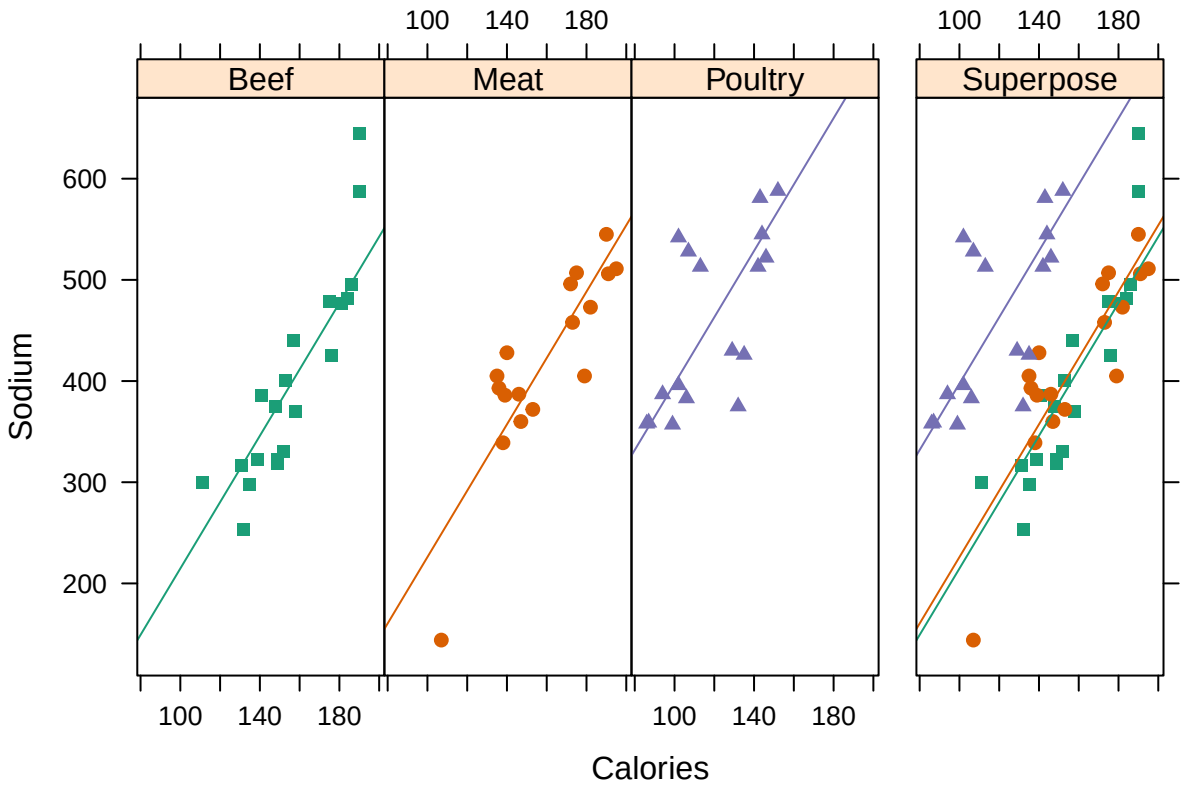
對交互作用變數做 ANOVA 的 F 檢定，其假設為:

- H_0: 卡路里（共變項）與肉的種類（獨變項）無交互作用
- H_1: 卡路里（共變項）與肉的種類（獨變項）具交互作用

下圖為卡路里與鈉含量之散佈圖與簡單線性迴歸模型，該模型考慮了交互作用項，所得之 β_{Beef}、β_{Beef} 與 $\beta_{Meat} = \beta_{Poultry}$ 三者並不全相同。



我們令顯著水準為 0.05。我們將對卡路里與肉種對鈉含量做線性迴歸後的係數做 ANOVA 的 F 檢定，其交互作用項的 F 值為 2，服從 $F(2, 48)$ 之分配，算得 p 值為 0.185，大於 0.05，無法拒絕 H_0，即無足夠證據顯示兩者具有交互作用，即 $\beta_{Beef} = \beta_{Meat} = \beta_{Poultry} = \beta$，如下圖，為不考慮交互作用的情況，卡路里與鈉含量之關係，可看出三條線皆平行，只有截距項之差異。



在不考慮卡路里與肉種類的交互作用的情況下，我們可得以下估計式：

$$\hat{Y_{ji}} = \mu_j + 3.2797(X_{ji} - \bar{X}_{..})$$

3.2.4 進行 ANCOVA 檢定

我們欲以 one-way ANCOVA 檢驗考慮卡路里含量後之不同型態的熱狗鈉含量是否有所差異，其中獨變項（independent variable）為熱狗型態，依變項為熱狗鈉含量，共變項（covariate variable）為卡路里。

另設一個調整後的依變項

$$Y_{ji.adj} = Y_{ji} - \hat{\beta}(X_{ji} - \bar{X}_{..})$$

，使用調整後的 $Y_{ji.adj}$ 與肉種做 ANOVA 檢定，令 $\mu_{a.adj}$ 為型態 a 的熱狗的調整後的鈉含量母體平均數。檢定的虛無假設與對立假設如下：

- 虛無假設：家禽肉、全牛肉與混合肉的三種型態的熱狗調整後鈉含量均相同

$$H_0 : \mu_{poultry.adj} = \mu_{beef.adj} = \mu_{meat.adj}$$

- 對立假設：家禽肉、全牛肉與混合肉的熱狗三者調整後的鈉含量不全相同

$$H_1 : \mu_{poultry.adj} \neq \mu_{beef.adj} \ or \ \mu_{poultry.adj} \neq \mu_{meat.adj} \ or \ \mu_{beef.adj} \neq \mu_{meat.adj}$$

3.2.4.1 檢驗 ANOVA 前提假設是否成立

這裡我們同樣要檢查 One-way ANOVA 的 4 個前提假設是否有被滿足。

1. 獨變項須為類別變數，依變項必須是連續變數

此部分的分析中，獨變項為熱狗型態，為含有 3 個類別的類別變數；依變項為調整後的熱狗鈉含量，仍為連續變項。符合。

2. 各組樣本依變項獨立

此分析中，各組依變項為牛肉熱狗鈉含量、混合肉熱狗鈉含量、家禽肉熱狗鈉含量，此三者互不影響彼此，符合前提假設。

3. 變異數同質：各組依變項的變異數必須相等。

- 針對調整後的依變項進行常態檢定

我們以 Shapiro-Wilk 常態檢定法對調整過後的依變項（排除共變項的依變項，$Y_{.adj}$）進行常態檢定，檢定的假說如下，顯著水準設定為 0.05。

$$H_0 : Y_{.adj} \sim ND \;\; v.s. \;\; H_1 : Y_{.adj} \; does \; not \sim ND$$

針對調整過後的依變項（$Y_{.adj}$）的檢定結果：檢定統計量為 0.9327，其 p 值為 0.0047，小於顯著水準，因此我們拒絕 H_0，也就是說我們有足夠的證據證明母體分配不服從常態分佈。統計檢定力較高的 Bartlett 檢定必須在常態分布下才能使用，因此我們在針對 $Y_{.adj}$ 進行變異數同質性檢定時改採 Levene 檢定法。

- 針對調整後的依變項進行變異數同質性檢定

我們以 Levene 變異數同質性檢定檢驗變異數同質是否成立。令 σ_x^2 為型態 x 的熱狗鈉含量的母體變異數，研究假說如下：

$$\begin{cases} H_0 : \sigma^2_{beef.adj} = \sigma^2_{meat.adj} = \sigma^2_{poultry.adj} \\ H_1 : \sigma^2_{beef.adj} \neq \sigma^2_{meat.adj} \; or \; \sigma^2_{beef.adj} \neq \sigma^2_{poultry.adj} \; or \; \sigma^2_{meat.adj} \neq \sigma^2_{poultry.adj} \end{cases}$$

我們同樣令顯著水準為 0.05。檢定結果的檢定統計量 F 為 0.4836，其 p 值為 0.6193，不小於顯著水準，因此我們不拒絕 H_0，意味著我們沒有足夠的證據證明變異數同質性不存在，也就是說我們無法證明至少有一組母體變異數與其他組不同，通過此前提假設。

4. 殘差（residuals）服從常態分配。

待配適完模型後診斷。

以上步驟顯示，在我們的資料中，ANOVA 的前提假設均滿足（殘差常態假設待檢驗），因此我們可以進行 ANOVA。

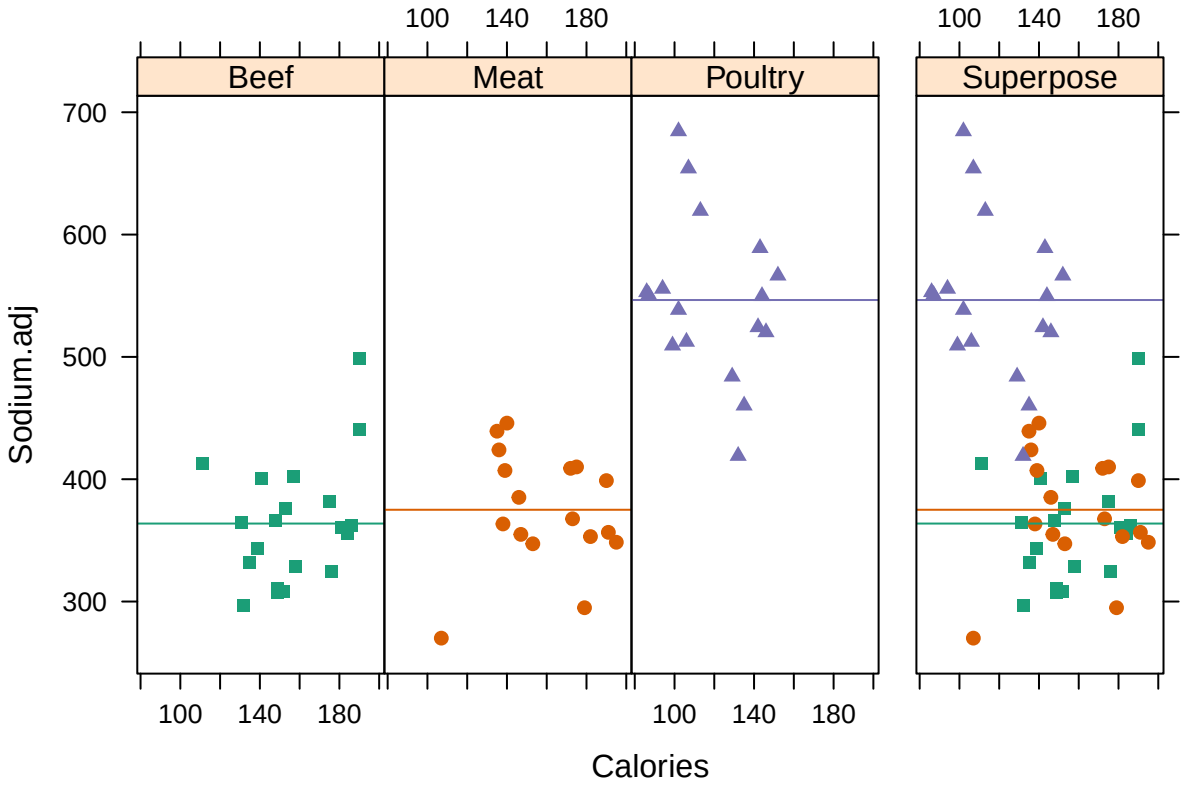
其統計量 F 為:

$$F_{ancova} = \frac{adjusted\ explained\ variation}{adjusted\ unexplained\ variation} = \frac{\sum_{j=1}^{k} n_j(\bar{Y}_{j.adj} - \bar{Y}_{.adj})/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji.adj} - \bar{Y}_{j.adj})^2/(N-k)} \sim F(k-1, N-k)$$

其中，k 為獨變項組別數；n_j 為第 j 組的觀察值（observations）個數；$N = \sum_{j=1}^{k} n_j$，也就是總觀察值數；$\bar{Y}_{j.adj}$ 為第 j 組依變項的調整後的樣本平均數，$\bar{Y}_{.adj}$ 為調整後的依變項樣本平均數；$Y_{ji.adj}$ 為第 j 組的第 i 個依變項觀察值。

我們令顯著水準為 0.05，檢定統計量 F_{ancova} 為 61.28，P 值為 $2.74e-14$，小於顯著水準，因此我們拒絕 H_0，意味著我們家禽肉、全牛肉與混合肉的三種型態的熱狗鈉含量在排除共變項卡路里的影響後，鈉含量的平均數明顯不同。

下圖為考慮共變項卡路里後的鈉含量平均數:



3.2.5 檢驗殘差是否為常態分配

完成 ANCOVA 分析後，我們須檢驗殘差是否為常態分配，來確定是否符合 ANCOVA 的前提假設。我們以 Shapiro-Wilk 常態檢定法對殘差進行常態檢定，檢定的假說如下:

$$H_0: residuals \sim ND \quad v.s. \quad H_1: not\ H_0$$

我們令信心水準為 95%，透過資料算出統計量 W 為 0.9819，p 值為 0.5865，大於顯著水準 $1 - 95\% = 0.05$，我們無顯著證據顯示 ϵ_{ji} 不服從常態分配，通過殘差項的假設。

3.2.6 多重比較

經由上述分析可得知移除共變量影響後，三種肉類的熱狗平均鈉含量中，至少有一組平均與其餘不同。我們希望更進一步了解家禽肉的平均鈉含量是否高於全牛肉和混和肉，我們可利用事後檢驗逕行多重比較。我們選用 Dunnett 檢定（Dunnett, 1955），並以家禽肉作為控制組，看混和肉和全牛肉的平均那含量是否低於家禽肉，將整體的信心水準訂為 95%，檢定的假說如下：

$$H_0 \mu_{meat.adj} - \mu_{poultry.adj} \geq 0 \quad v.s \quad H_1 \mu_{meat.adj} - \mu_{poultry.adj} < 0$$

$$H_0 \mu_{poultry.adj} - \mu_{meat.adj} \geq 0 \quad v.s \quad H_1 \mu_{poultry.adj} - \mu_{meat.adj} < 0$$

根據資料算出兩種肉類的平均鈉含量差異的 95% 信賴區間，混和肉和家禽肉的平均鈉含量差的信賴區間為 $(-\infty, -134.6923]$，全牛肉和家禽肉的平均鈉含量差的信賴區間為 $(-\infty, -147.3908]$，兩個區間皆遠小於 0，表示混和肉和全牛肉的平均鈉含量低於家禽肉，最後統整他們的關係：

$$\mu_{meat.adj} < \mu_{poultry.adj}$$

$$\mu_{beef.adj} < \mu_{poultry.adj}$$

4 建議（Comments）

1. 從資料集和說明無法得知樣本的來源，三種肉類的熱狗鈉含量和卡路里會受到製造國家和品牌的影響，若諮詢者可以再進行一次實驗，可以將品牌或是製造國家等可控因子納入考慮，再利用 Two-way ANCOVA 進行分析。

2. 由於此資料集各組肉類只有約 20 個樣本，且三組樣本數皆不相同。以實驗設計角度來說，若可以增加其實驗成本，建議增加其樣本數，且讓三組樣本數皆相同，可以獲得更精確的結果。

5 附錄

5.1 不同的事後比較檢定法

在事後多重比較的分析中，除了 Dunnett 檢定，我們也使用 Bonferroni 法校正檢定的顯著水準 α，進行兩兩比較的 t 檢定。然而，必須在檢定變項服從常態分佈的情況下，才可以使用此檢定法（Bland & Altman, 1995），而在之前以 Shapiro-Wilk 常態檢定法對 Y_{adj} 常態檢定時，我們已知 Y_{adj} 不服從常態分配，因此需要做變數轉換。我們先

選用 log 轉換法對 Y_{adj} 進行轉換，並對轉換後的變項 $Y_{adj.log}$ 再進行一次 Shapiro-Wilk 常態檢定，顯著水準同樣設定為 0.05。檢定假說如下:

$$H_0 : Y_{adj.log} \sim ND \ v.s. \ H_1 : Y_{adj.log} \ does \ not \ \sim ND$$

從資料算出統計量 $W = 0.9655$，p 值為 0.122，大於顯著水準，因此我們不拒絕 H_0，表示我們無顯著的證據顯示 $Y_{adj.log}$ 不服從常態分配，通過此假設。接著我們開始進行兩兩 log 平均鈉含量檢定，我們令 $\mu_{a.adj.log}$ 為型態 a 的 log 平均鈉含量。以下為 3 組兩兩比較的虛無假設與對立假設:

$$\begin{cases} H_0 : \mu_{meat.adj.log} - \mu_{beef.adj.log} \leq 0 \quad v.s \quad H_1 : \mu_{meat.adj.log} - \mu_{beef.adj.log} > 0 \\ H_0 : \mu_{poultry.adj.log} - \mu_{beef.adj.log} \leq 0 \quad v.s \quad H_1 : \mu_{poultry.adj.log} - \mu_{beef.adj.log} > 0 \\ H_0 : \mu_{poultry.adj.log} - \mu_{meat.adj.log} \leq 0 \quad v.s \quad H_1 : \mu_{poultry.adj.log} - \mu_{meat.adj.log} > 0 \end{cases}$$

檢定結果如上表。從資料得到第 1 組為 $\mu_{meat.log} - \mu_{beef.log}$，第 2 組為 $\mu_{poultry.log} - \mu_{beef.log}$，第 3 組為 $\mu_{poultry.log} - \mu_{meat.log}$。第 1 組得到統計量 $t = 0.73$，其調整後的 p 值為 0.703，大於調整後 $\alpha = 0.05$，因此我們不拒絕 H_0，表示沒有顯著證據顯示混和肉的 log 平均鈉含量大於全牛肉的 log 平均鈉含量。第 2 組的檢定統計量 $t = 9.61$，調整後的 p 值為 $7.38e-13$，遠小於顯著水準，因此我們拒絕 H_0，表示家禽肉的 log 平均鈉含量顯著高於全牛肉的 log 平均鈉含量。第 3 組的檢定統計量 $t = 8.54$，p 值為 $3.15e-11$，遠小於顯著水準，因此我們拒絕 H_0，表示家禽肉的 log 平均鈉含量顯著高於混和肉的 log 平均鈉含量。結果和由 Dunnett 檢定得到的相似。最後整理三種肉類熱狗 log 平均鈉含量的關係:

$$\mu_{beef.adj.log} \leq \mu_{meat.adj.log} < \mu_{poultry.adj.log}$$

6 參考資料

1. Heiberger, R. M. & Burt Holland, B. H. (2015). Statistical Analysis and Data Display An Intermediate Course with Examples in R. Springer.

2. Rutherford, A. (2001). Introducing ANOVA and ANCOVA: a GLM approach. Sage.

3. Lim, T. S., & Loh, W. Y. (1996). A comparison of tests of equality of variances. Computational Statistics & Data Analysis, 22(3), 287-301.

4. Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50(272), 1096-1121.

5. Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. Bmj, 310(6973), 170.