



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

MRC

Laboratory of
Molecular Biology

Analysis of Variance (ANOVA)

Cancer Research UK – 12th of March 2021

D.-L. Couturier / R. Nicholls / M. Fernandes

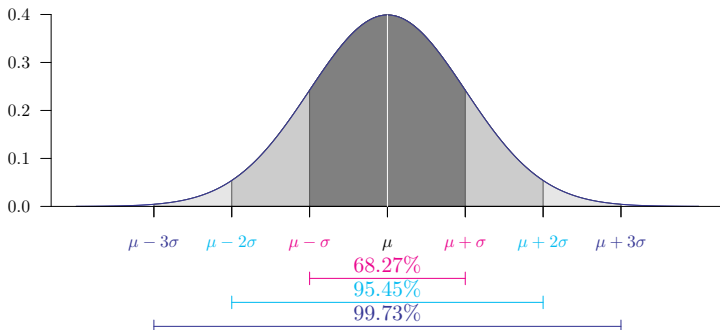
Quick review: Normal distribution

$$Y \sim N(\mu, \sigma^2), \quad f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$$E[Y] = \mu, \quad \text{Var}[Y] = \sigma^2,$$

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1), \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Probability density function of a normal distribution:



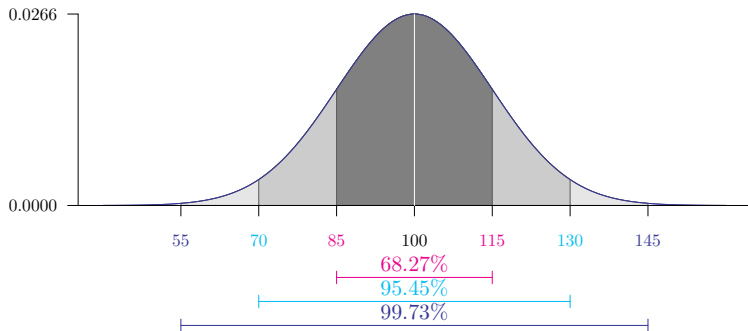
Quick review: Normal distribution

$$Y \sim N(\mu, \sigma^2), \quad f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

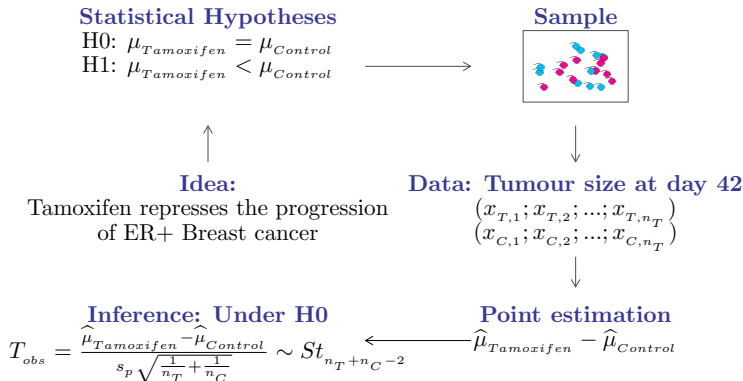
$$E[Y] = \mu, \quad \text{Var}[Y] = \sigma^2,$$

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1), \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Suitable modelling for a lot of phenomena: $\text{IQ} \sim N(100, 15^2)$.



Grand Picture of Statistics



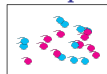
Grand Picture of Statistics

Statistical Hypotheses

$$H_0: \mu_{Tamoxifen} = \mu_{Control}$$

$$H_1: \mu_{Tamoxifen} < \mu_{Control}$$

Sample



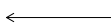
Data: Tumour size at day 42

$$\begin{pmatrix} x_{T,1}; x_{T,2}; \dots; x_{T,n_T} \\ x_{C,1}; x_{C,2}; \dots; x_{C,n_C} \end{pmatrix}$$



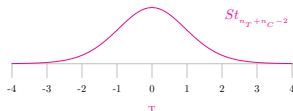
Point estimation

$$\hat{\mu}_{Tamoxifen} - \hat{\mu}_{Control}$$



Inference: Under H_0

$$T_{obs} = \frac{\hat{\mu}_{Tamoxifen} - \hat{\mu}_{Control}}{s_p \sqrt{\frac{1}{n_T} + \frac{1}{n_C}}} \sim St_{n_T + n_C - 2}$$



$$p\text{-value} = P(T < T_{obs})$$

One-sample Student's t-test

- Assumed model

$$Y_i = \mu + \epsilon_i,$$

where $i = 1, \dots, n$

and $\epsilon_i \sim N(0, \sigma^2)$.

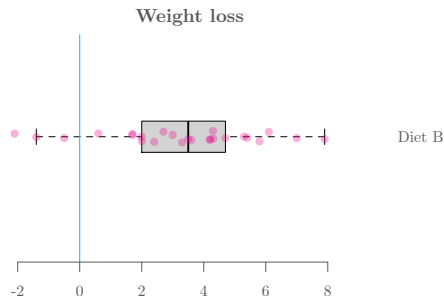
- Hypotheses

- $H_0: \mu = 0$,

- $H_1: \mu > 0$.

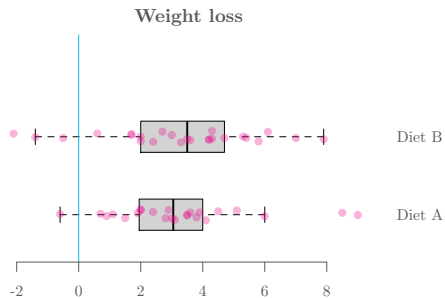
- Test statistic's distribution under H_0

$$T = \frac{\bar{Y} - \mu_0}{s} \sim Student(n - 1).$$



One Sample t-test

```
data: dietB
t = 6.6301, df = 24, p-value = 3.697e-07
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 2.424694      Inf
sample estimates:
mean of x
 3.268
```



Two-sample location tests:
t-tests and Mann-Whitney-Wilcoxon's test

Two independent sample Student's t-test

Assumed model

$$Y_{i(g)} = \mu_g + \epsilon_{i(g)},$$
$$= \mu + \delta_g + \epsilon_{i(g)},$$

where $g = A, B$, $i = 1, \dots, n_g$,

$\epsilon_{i(g)} \sim N(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

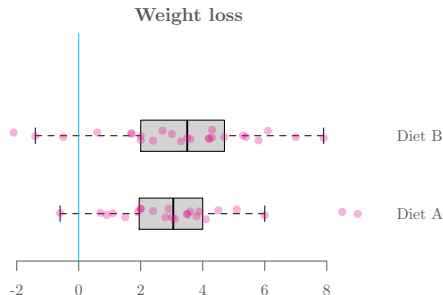
Hypotheses

▷ **H0:** $\mu_A = \mu_B$,

▷ **H1:** $\mu_A \neq \mu_B$.

Test statistic's distribution under H0

$$T = \frac{(\bar{Y}_A - \bar{Y}_B) - (\mu_A - \mu_B)}{s_p \sqrt{n_A^{-1} + n_B^{-1}}} \sim Student(n_A + n_B - 2).$$



Two Sample t-test

```
data: dietA and dietB
t = 0.0475, df = 47, p-value = 0.9623
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.323275  1.387275
sample estimates:
mean of x mean of y
  3.300    3.268
```


Two independent sample Welch's t-test

Assumed model

$$Y_{i(g)} = \mu_g + \epsilon_{i(g)},$$
$$= \mu + \delta_g + \epsilon_{i(g)},$$

where $g = A, B$, $i = 1, \dots, n_g$,

$\epsilon_{i(g)} \sim N(0, \sigma_g^2)$ and $\sum n_g \delta_g = 0$.

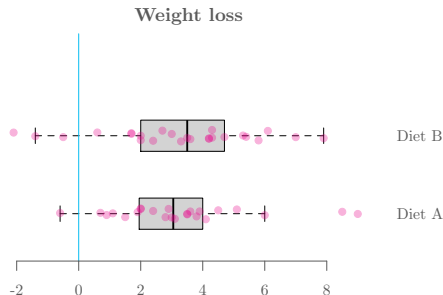
Hypotheses

▷ **H0:** $\mu_A = \mu_B$,

▷ **H1:** $\mu_A \neq \mu_B$.

Test statistic's distribution under H0

$$T = \frac{(\bar{Y}_A - \bar{Y}_B) - (\mu_A - \mu_B)}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}} \sim Student(df).$$



Welch Two Sample t-test

```
data: dietA and dietB
t = 0.047594, df = 46.865, p-value = 0.9622
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.320692  1.384692
sample estimates:
mean of x mean of y
  3.300    3.268
```

Two independent sample Mann-Whitney-Wilcoxon test

► Assumed model

$$\begin{aligned}Y_{i(g)} &= \theta_g + \epsilon_{i(g)}, \\ &= \theta + \delta_g + \epsilon_{i(g)},\end{aligned}$$

where $g = A, B$, $i = 1, \dots, n_g$,

$\epsilon_{i(g)} \sim iid(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

► Hypotheses

▷ **H0:** $\theta_A = \theta_B$,

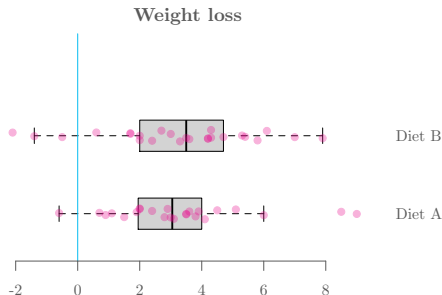
▷ **H1:** $\theta_A \neq \theta_B$.

► Test statistic's distribution under H0

$$z = \frac{\sum_{i=1}^{n_B} R_{i(g)} - [n_B(n_A + n_B + 1)/2]}{\sqrt{n_A n_B (n_A + n_B + 1)/12}},$$

where

► $R_{i(g)}$ denotes the global rank of the i th observation of group g .

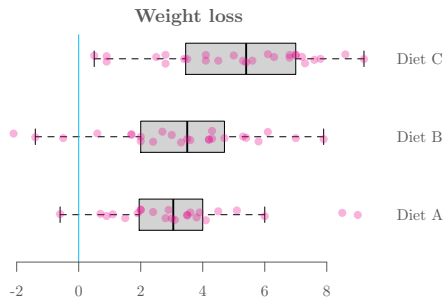


Wilcoxon rank sum test with continuity correction

data: dietA and dietB

W = 277, p-value = 0.6526

alternative hypothesis: true location shift is not equal to 0



Two or more sample location tests:
one-way ANOVA & multiple comparisons

More than two sample case: Fisher's one-way ANOVA

Assumed model

$$Y_{i(g)} = \mu_g + \epsilon_{i(g)},$$

$$= \mu + \delta_g + \epsilon_{i(g)},$$

where $g = 1, \dots, G$, $i = 1, \dots, n_g$,
 $\epsilon_{i(g)} \sim N(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

Hypotheses

- ▷ **H0:** $\mu_1 = \mu_2 = \dots = \mu_G$,
- ▷ **H1:** $\mu_k \neq \mu_l$ for at least one pair (k, l) .

Test statistic's distribution under H0

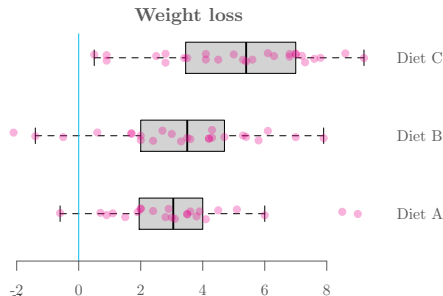
$$F = \frac{N s_Y^2}{s_p^2} \sim \text{Fisher}(G - 1, N - G),$$

where

$$s_Y^2 = \frac{1}{G-1} \sum_{g=1}^G \frac{n_g}{N} (\bar{Y}_g - \bar{\bar{Y}})^2,$$

$$s_p^2 = \frac{1}{N-G} \sum_{g=1}^G (n_g - 1) s_g^2,$$

$$N = \sum n_g, \quad \bar{\bar{Y}} = \frac{1}{N} \sum_{g=1}^G n_g \bar{Y}_g.$$



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diet.type	2	60.5	30.264	5.383	0.0066 **
Residuals	73	410.4	5.622		

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

More than two sample case: Welch's one-way ANOVA

Assumed model

$$Y_{i(g)} = \mu_g + \epsilon_{i(g)},$$

$$= \mu + \delta_g + \epsilon_{i(g)},$$

where $g = 1, \dots, G$, $i = 1, \dots, n_g$,
 $\epsilon_{i(g)} \sim N(0, \sigma_g^2)$ and $\sum n_g \delta_g = 0$.

Hypotheses

- ▷ **H0**: $\mu_1 = \mu_2 = \dots = \mu_G$,
- ▷ **H1**: $\mu_k \neq \mu_l$ for at least one pair (k, l) .

Test statistic's distribution under H0

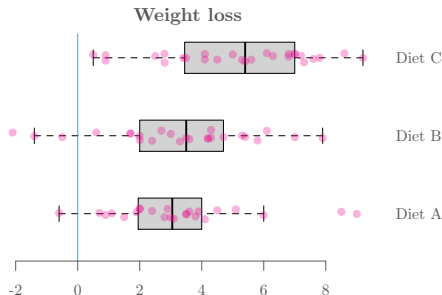
$$F^* = \frac{s_Y^{*2}}{1 + \frac{2(G-2)}{3\Delta}} \sim \text{Fisher}(G-1, \Delta),$$

where

$$\text{▷ } s_Y^{*2} = \frac{1}{G-1} \sum_{g=1}^G w_g (\bar{Y}_g - \bar{\bar{Y}}^*)^2,$$

$$\text{▷ } \Delta = \left[\frac{3}{G^2-1} \sum_{g=1}^G \frac{1}{n_g} \left(1 - \frac{w_g}{\sum w_g} \right) \right]^{-1},$$

$$\text{▷ } w_g = \frac{n_g}{s_g^2}, \quad \bar{\bar{Y}}^* = \frac{\sum_{g=1}^G w_g \bar{Y}_g}{\sum w_g}.$$



One-way analysis of means (not assuming equal variances)

data: weight.diff and diet.type

F = 5.2693, num df = 2.00, denom df = 48.48, p-value = 0.008497

More than two sample case: Kruskal-Wallis test

► Assumed model

$$\begin{aligned}Y_{i(g)} &= \theta_g + \epsilon_{i(g)}, \\ &= \theta + \delta_g + \epsilon_{i(g)},\end{aligned}$$

where $g = 1, \dots, G$, $i = 1, \dots, n_g$,

$\epsilon_{i(g)} \sim iid(0, \sigma^2)$ and $\sum n_g \delta_g = 0$.

► Hypotheses

► **H0:** $\theta_1 = \theta_2 = \dots = \theta_G$,

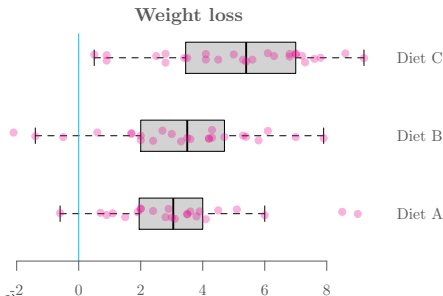
► **H1:** $\theta_k \neq \theta_l$ for at least one pair (k, l) .

► Test statistic's distribution under **H0**

$$H = \frac{\frac{12}{N(N+1)} \sum_{g=1}^G \frac{\bar{R}_g}{n_g} - 3(N-1)}{1 - \frac{\sum_{v=1}^V t_v^3 - t_v}{N^3 - N}} \sim \chi(G-1),$$

where

- $\bar{R}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} R_{i(g)}$ and $R_{i(g)}$ denotes the global rank of the i th observation of group g ,
- V is the number of different values/levels in \mathbf{y} and t_v denotes the number of times a given value/level occurred in \mathbf{y} .



Kruskal-Wallis rank sum test

data: weight.loss by diet.type

Kruskal-Wallis chi-squared = 9.4159, df = 2, p-value = 0.009023

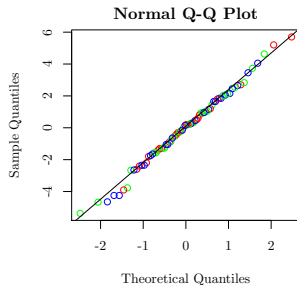
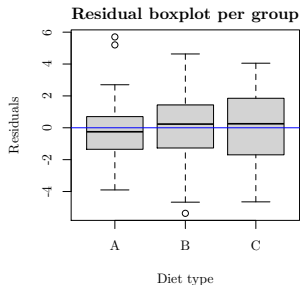
Model check: Residual analysis

$$Y_{i(g)} = \theta_g + \epsilon_{i(g)}$$

$$\hat{\epsilon}_{i(g)} = Y_{i(g)} - \hat{\theta}_g,$$

where

- ▶ $\hat{\epsilon}_{i(g)} \sim N(0, \hat{\sigma}^2)$ for Fisher's ANOVA
- ▶ $\hat{\epsilon}_{i(g)} \sim N(0, \hat{\sigma}_g^2)$ for Welch's ANOVA
- ▶ $\hat{\epsilon}_{i(g)} \sim iid(0, \hat{\sigma}^2)$ for Kruskal-Wallis' ANOVA



Shapiro-Wilk normality test

```
data: diet$resid.mean  
W = 0.99175, p-value = 0.9088
```

Bartlett test of homogeneity of variances

```
data: diet$resid.mean by as.numeric(diet$diet.type)  
Bartlett's K-squared = 0.21811, df = 2, p-value = 0.8967
```

Finding different pairs: Multiple comparisons

► All-pairwise comparison problem:

Interested in finding which pair(s) are different by testing

► H_{01} : $\mu_1 = \mu_2$, ► H_{02} : $\mu_1 = \mu_3$, ... ► H_{0K} : $\mu_{G-1} = \mu_G$,
leading to a total of $K = G(G-1)/2$ pairwise comparisons.

► Family-wise type I error for K tests, α_K

For each test, the probability of rejecting H_0 when H_0 is true equals α .
For K independent tests, the probability of rejecting H_0 at least 1 time
when H_0 is true, α_K , is given by

$$\alpha_K = 1 - (1 - \alpha)^K.$$

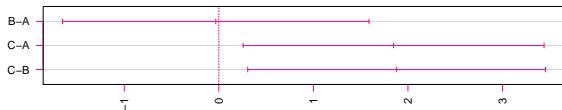
$$\begin{aligned}\triangleright \alpha_1 &= 0.05, \\ \triangleright \alpha_2 &= 0.0975, \\ \triangleright \alpha_{10} &= 0.4013.\end{aligned}$$

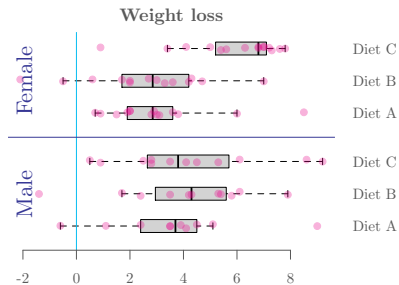
► Multiplicity correction

Principle: change the level of each test so that $\alpha_K = 0.05$, for example:

- Bonferroni's correction (indep. tests): $\alpha = \alpha_K / K$,
- Dunn-Sidak's correction (indep. tests): $\alpha = 1 - (1 - \alpha_K)^{1/K}$,
- Tukey's correction (dependent tests).

95% family-wise confidence level





Two or more sample location tests:
two-way ANOVA

More than one factor: Fisher's two-way ANOVA

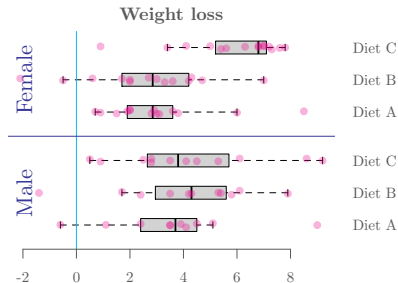
► Assumed model

$$\begin{aligned}Y_{i(g)} &= \mu_{gk} + \epsilon_{i(gk)}, \\ &= \mu + \delta_g + \delta_k + \delta_{gk} + \epsilon_{i(gk)},\end{aligned}$$

- $g = 1, \dots, G, k = 1, \dots, K,$
- $i = 1, \dots, n_g,$
- $\epsilon_{i(gk)} \sim N(0, \sigma^2)$
- $\sum n_g \delta_g = \sum n_k \delta_k = \sum n_{gk} \delta_{gk} = 0.$

► Hypotheses

- $H0_1: \delta_g = 0 \forall g,$
- $H1_1: H0_1$ is false.
- $H0_2: \delta_k = 0 \forall k,$
- $H1_2: H0_2$ is false.
- $H0_3: \delta_{gk} = 0 \forall g, k,$
- $H1_3: H0_3$ is false.



More than one factor: Fisher's two-way ANOVA

Assumed model

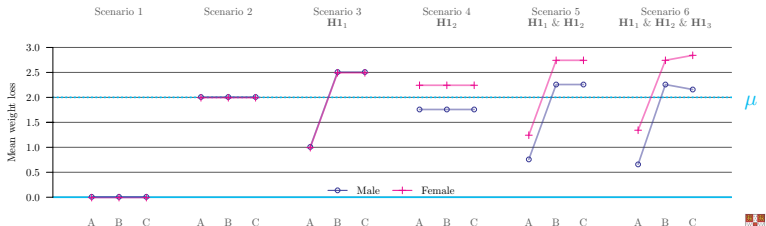
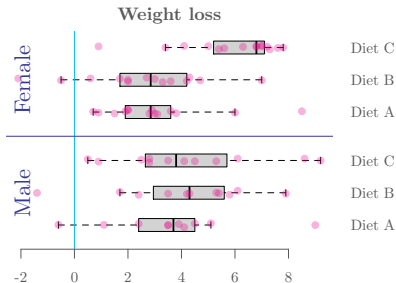
$$Y_{i(g)} = \mu_{gk} + \epsilon_{i(gk)},$$

$$= \mu + \delta_g + \delta_k + \delta_{gk} + \epsilon_{i(gk)},$$

- ▶ $g = 1, \dots, G, k = 1, \dots, K,$
- ▶ $i = 1, \dots, n_g,$
- ▶ $\epsilon_{i(gk)} \sim N(0, \sigma^2)$
- ▶ $\sum n_g \delta_g = \sum n_k \delta_k = \sum n_{gk} \delta_{gk} = 0.$

Hypotheses

- ▶ $H0_1: \delta_g = 0 \forall g,$
- ▶ $H1_1: H0_1$ is false.
- ▶ $H0_2: \delta_k = 0 \forall k,$
- ▶ $H1_2: H0_2$ is false.
- ▶ $H0_3: \delta_{gk} = 0 \forall g, k,$
- ▶ $H1_3: H0_3$ is false.



More than one factor: Fisher's two-way ANOVA

Assumed model

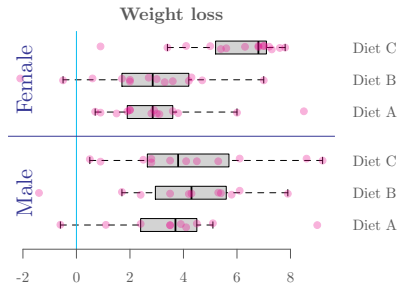
$$Y_{i(g)} = \mu_{gk} + \epsilon_{i(gk)},$$

$$= \mu + \delta_g + \delta_k + \delta_{gk} + \epsilon_{i(gk)},$$

- ▶ $g = 1, \dots, G, k = 1, \dots, K,$
- ▶ $i = 1, \dots, n_g,$
- ▶ $\epsilon_{i(gk)} \sim N(0, \sigma^2)$
- ▶ $\sum n_g \delta_g = \sum n_k \delta_k = \sum n_{gk} \delta_{gk} = 0.$

Hypotheses

- ▶ $H0_1: \delta_g = 0 \forall g,$
- ▶ $H1_1: H0_1$ is false.
- ▶ $H0_2: \delta_k = 0 \forall k,$
- ▶ $H1_2: H0_2$ is false.
- ▶ $H0_3: \delta_{gk} = 0 \forall g, k,$
- ▶ $H1_3: H0_3$ is false.



	Df	Sum Sq	Mean Sq	F	value	Pr(>F)
diet.type	2	60.5	30.264	5.629	0.00541	**
gender	1	0.2	0.169	0.031	0.85991	
diet.type:gender	2	33.9	16.952	3.153	0.04884	*
Residuals	70	376.3	5.376			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Summary