Analysis of Variance (ANOVA)

Cancer Research UK – 10th of March 2020
D.-L. Couturier / R. Nicholls / M. Fernandes
Quick review: Normal distribution

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

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

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

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: IQ \sim N(100, 15^2).
Quick review: Normal distribution

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

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

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

Central limit theorem (Lindeberg-Lévy CLT)

- Let \((X_1, \ldots, X_n)\) be \(n\) independent and identically distributed (iid) random variables drawn from distributions of expected values given by \(\mu\) and finite variances given by \(\sigma^2\),
- then

\[ \hat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right). \]

If \(X_i \sim N(\mu, \sigma^2)\), this result is true for all sample sizes.
Grand Picture of Statistics

**Statistical Hypotheses**

H0: $\mu_{\text{Tamoxifen}} = \mu_{\text{Control}}$

H1: $\mu_{\text{Tamoxifen}} < \mu_{\text{Control}}$

**Idea:**

Tamoxifen represses the progression of ER+ Breast cancer

**Inference: Under H0**

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

**Data: Tumour size at day 42**

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

**Sample**

**Point estimation**

$$\hat{\mu}_{\text{Tamoxifen}} - \hat{\mu}_{\text{Control}}$$
Grand Picture of Statistics

Statistical Hypotheses
H0: $\mu_{Tamoxifen} = \mu_{Control}$
H1: $\mu_{Tamoxifen} < \mu_{Control}$

Idea:
Tamoxifen represses the progression of ER+ Breast cancer

Sample

Data: Tumour size at day 42
$(x_{T,1}; x_{T,2}; \ldots; x_{T,n_T})$
$(x_{C,1}; x_{C,2}; \ldots; x_{C,n_T})$

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

Inference: Under H0
$T_{obs} = \frac{\hat{\mu}_{Tamoxifen} - \hat{\mu}_{Control}}{s_p \sqrt{\frac{1}{n_T} + \frac{1}{n_C}}} \sim St_{n_T+n_C-2}$

$p - value = P(T < T_{obs})$
Statistical hypothesis testing Process

Several-step process:

- Define H0 and H1 according to a theory
- Set $\alpha$, the probability of rejecting H0 when it is true (type I error),
- Define $n$, the sample size, allowing you to reject H0 when H1 is true with a probability $1 - \beta$ (Power),
- Determine the test statistic to be used,
- Collect the data,
- Perform the statistical test, define the $p$-value, and reject (or not) the null hypothesis.
Statistical hypothesis testing 4 possible outcomes

Conclude:
- if \( p\)-value > \( \alpha \) → do not reject H0.
- if \( p\)-value < \( \alpha \) → reject H0 in favour of H1.

<table>
<thead>
<tr>
<th>Unknown Truth</th>
<th>H0 true</th>
<th>H1 true</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0 true</td>
<td>( 1 - \alpha )</td>
<td>( \alpha )</td>
</tr>
<tr>
<td>H1 true</td>
<td>( \beta )</td>
<td>( 1 - \beta )</td>
</tr>
</tbody>
</table>

where
- \( \alpha \) is the type I error, the probability of rejecting H0 when it is correct,
- \( \beta \) is the type II error, the probability of not rejecting when H1 is true,
- \( 1 - \beta \) is the power, the probability of accepting H1 when it is true.
One-sample Student’s t-test

- **Assumed model**
  \[ Y_i = \mu + \epsilon_i, \]
  where \( i = 1, \ldots, n \)
  and \( \epsilon_i \sim N(0, \sigma^2) \).

- **Hypotheses**
  - **H0**: \( \mu = 0 \),
  - **H1**: \( \mu > 0 \).

- **Test statistic’s distribution under H0**
  \[ T = \frac{\bar{Y} - \mu_0}{s} \sim \text{Student}(n-1). \]

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 \( g = A, B, i = 1, \ldots, n_{g}, \)

\( \epsilon_{i(g)} \sim N(0, \sigma^{2}) \) and \( \sum n_{g}\delta_{g} = 0. \)

▶ Hypotheses

\( \textbf{H0: } \mu_{A} = \mu_{B}, \)

\( \textbf{H1: } \mu_{A} \neq \mu_{B}. \)

▶ Test statistic’s distribution under \( \textbf{H0} \)

\[ T = \frac{(\bar{Y}_{A} - \bar{Y}_{B}) - (\mu_{A} - \mu_{B})}{s_{p}\sqrt{\frac{1}{n_{A}} + \frac{1}{n_{B}}}} \sim \text{Student}(n_{A} + n_{B} - 2). \]

Two Sample t-test

data: dietA and dietB

\[ t = 0.0475, \text{ df } = 47, \text{ p-value } = 0.9623 \]

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

\(-1.323275 \quad 1.387275 \)

sample estimates:

mean of x mean of y

3.300 \quad 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 \( g = A, B, i = 1, \ldots, n_g, \)
  \( \epsilon_{i(g)} \sim N(0, \sigma_g^2) \) and \( \sum n_g \delta_g = 0. \)

- **Hypotheses**
  - \( H_0: \mu_A = \mu_B, \)
  - \( H_1: \mu_A \neq \mu_B. \)

- **Test statistic’s distribution under \( H_0 \)**
  \[ T = \frac{(Y_A - Y_B) - (\mu_A - \mu_B)}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}} \sim \text{Student}(df). \]

*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
  \[ Y_{i(g)} = \theta_g + \epsilon_{i(g)}, \]
  \[ = \theta + \delta_g + \epsilon_{i(g)}, \]
  where \( g = A, B, i = 1, \ldots, n_g, \epsilon_{i(g)} \sim iid(0, \sigma^2) \) 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
Model assumptions Normality – Heteroscedasticity

Simulate 2500 samples with

- \( X_i \sim Uniform(1.5, 2.5), \ i = 1, \ldots, n_X \),
- \( Y_i \sim Uniform(0, 4), \ i = 1, \ldots, n_Y \),

so that \( E[X_i] = E[Y_i] = 2 \) (i.e., same mean, same median).

Assume

- \( X_i \sim iid(\mu_X, \sigma^2), \ i = 1, \ldots, n_X \),
- \( Y_i \sim iid(\mu_X + \delta, \sigma^2), \ i = 1, \ldots, n_Y \).

Test \( H_0: \delta = \delta_0 \) against \( H_1: \delta \neq \delta_0 \), at the 5% level, by means of

- Mann-Whitney-Wilcoxon test (MWW),
- T-test,
- Welch-test.

<table>
<thead>
<tr>
<th>( \hat{\alpha} )</th>
<th>Tests</th>
<th>( n_X = 200, n_Y = 70 )</th>
<th>( n_X = 20, n_Y = 7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MWW</td>
<td>Student’s t-test</td>
<td>Welch’s test</td>
</tr>
<tr>
<td>Sample size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.145</td>
<td>0.202</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.148</td>
<td>0.240</td>
<td>0.062</td>
</tr>
</tbody>
</table>
**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 \( g = 1, \ldots, G, i = 1, \ldots, n_g, \)

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

► Hypotheses

▷ H0: \( \mu_1 = \mu_2 = \ldots = \mu_G, \)

▷ 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

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

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

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

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>diet.type</td>
<td>2</td>
<td>60.5</td>
<td>30.264</td>
<td>5.383</td>
</tr>
<tr>
<td>Residuals</td>
<td>73</td>
<td>410.4</td>
<td>5.622</td>
<td>0.001 **</td>
</tr>
</tbody>
</table>

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

- Hypotheses
  - \textbf{H0:} \( \mu_1 = \mu_2 = \ldots = \mu_G, \)
  - \textbf{H1:} \( \mu_k \neq \mu_l \) for at least one pair \( (k, l). \)

- Test statistic’s distribution under \textbf{H0}
  \[ F^* = \frac{s_{\hat{Y}}^2}{1 + \frac{2(G-2)}{3\Delta}} \sim Fisher(G-1, \Delta), \]
  where
  - \( s_{\hat{Y}}^2 = \frac{1}{G-1} \sum_{g=1}^G w_g \left( Y_g - \bar{Y}^* \right)^2, \)
  - \( \Delta = \left[ \frac{3}{G^2-1} \sum_{g=1}^G \frac{1}{n_g} \left( 1 - \frac{w_g}{\sum w_g} \right) \right]^{-1}, \)
  - \( w_g = \frac{n_g}{s_g^2}, \ \bar{Y}^* = \sum_{g=1}^G \frac{w_g Y_g}{\sum w_g}. \)

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

\begin{align*}
\text{data: weight.diff and diet.type} \\
F = 5.2693, \ \text{num df} = 2.00, \ \text{denom df} = 48.48, \ p\text{-value} = 0.008497
\end{align*}
More than two sample case: Kruskal-Wallis test

- Assumed model

\[ Y_{i(g)} = \theta_g + \epsilon_{i(g)}, \]
\[ = \theta + \delta_g + \epsilon_{i(g)}, \]

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

- Hypotheses

\( \triangleright \) \textbf{H0:} \( \theta_1 = \theta_2 = \ldots = \theta_G, \)
\( \triangleright \) \textbf{H1:} \( \theta_k \neq \theta_l \) for at least one pair \((k, l)\).

- Test statistic’s distribution under \( \textbf{H0} \)

\[ H = \frac{12}{N(N+1)} \sum_{g=1}^{G} \frac{R_g}{n_g} - 3(N - 1) \]
\[ 1 - \frac{\sum_{v=1}^{V} t_v^3 - t_v}{N^3 - N} \sim \chi(G - 1), \]

where

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

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
\[ W = 0.99175, \text{ p-value} = 0.9088 \]

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
  \[ H_0^1: \mu_1 = \mu_2, \quad H_0^2: \mu_1 = \mu_3, \quad \ldots \quad H_0^K: \mu_{G-1} = \mu_G, \]
  leading to a total of \( K = G(G-1)/2 \) pairwise comparisons.

- Family-wise type I error for \( K \) tests, \( \alpha_K \)
  For each test, the probability of rejecting \( H_0 \) when \( H_0 \) is true equals \( \alpha \).
  For \( K \) independent tests, the probability of rejecting \( H_0 \) at least 1 time when \( H_0 \) is true, \( \alpha_K \), is given by
  \[
  \alpha_K = 1 - (1 - \alpha)^K.
  \]
  \[
  \begin{align*}
  \alpha_1 &= 0.05, \\
  \alpha_2 &= 0.0975, \\
  \alpha_{10} &= 0.4013.
  \end{align*}
  \]

- 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).

---

![Diagram showing 95% family-wise confidence level for differences in mean levels of diet type]
Two or more sample location tests: two-way ANOVA
More than one factor: Fisher’s two-way ANOVA

- Assumed model

\[ Y_{i(g)} = \mu gk + \epsilon_{i(gk)}, \]
\[ = \mu + \delta_g + \delta_k + \delta_{gk} + \epsilon_{i(gk)}, \]

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

- Hypotheses

- \( H_{01}: \delta_g = 0 \; \forall \; g, \)
- \( H_{11}: H_{01} \) is false.
- \( H_{02}: \delta_k = 0 \; \forall \; k, \)
- \( H_{12}: H_{02} \) is false.
- \( H_{03}: \delta_{gk} = 0 \; \forall \; g, k, \)
- \( H_{13}: H_{03} \) is false.
More than one factor: Fisher's two-way ANOVA

Assumed model

\[ Y_{i(g)} = \mu_{gk} + \epsilon_{i(gk)}, \]
\[ = \mu + \delta_g + \delta_k + \delta_{gk} + \epsilon_{i(gk)}, \]
\[ g = 1, \ldots, G, \ k = 1, \ldots, K, \]
\[ i = 1, \ldots, n_g, \]
\[ \epsilon_{i(gk)} \sim N(0, \sigma^2) \]
\[ \sum n_g \delta_g = \sum n_k \delta_k = \sum n_{gk} \delta_{gk} = 0. \]

Hypotheses

\[ H_{01}: \delta_g = 0 \ \forall \ g, \]
\[ H_{11}: H_{01} \text{ is false.} \]
\[ H_{02}: \delta_k = 0 \ \forall \ k, \]
\[ H_{12}: H_{02} \text{ is false.} \]
\[ H_{03}: \delta_{gk} = 0 \ \forall \ g, k, \]
\[ H_{13}: H_{03} \text{ is false.} \]
More than one factor: Fisher’s two-way ANOVA

Assumed model

\[ Y_{i(g)} = \mu_{gk} + \epsilon_{i(gk)}, \]
\[ = \mu + \delta_g + \delta_k + \delta_{gk} + \epsilon_{i(gk)}, \]

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

Hypotheses

- \( \textbf{H0}_1: \delta_g = 0 \ \forall \ g, \)
- \( \textbf{H1}_1: \textbf{H0}_1 \text{ is false.} \)
- \( \textbf{H0}_2: \delta_k = 0 \ \forall \ k, \)
- \( \textbf{H1}_2: \textbf{H0}_2 \text{ is false.} \)
- \( \textbf{H0}_3: \delta_{gk} = 0 \ \forall \ g, k, \)
- \( \textbf{H1}_3: \textbf{H0}_3 \text{ is false.} \)

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>diet.type</td>
<td>2</td>
<td>60.5</td>
<td>30.264</td>
<td>5.629</td>
<td>0.00541**</td>
</tr>
<tr>
<td>gender</td>
<td>1</td>
<td>0.2</td>
<td>0.169</td>
<td>0.031</td>
<td>0.85991</td>
</tr>
<tr>
<td>diet.type:gender</td>
<td>2</td>
<td>33.9</td>
<td>16.952</td>
<td>3.153</td>
<td>0.04884*</td>
</tr>
<tr>
<td>Residuals</td>
<td>70</td>
<td>376.3</td>
<td>5.376</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Summary