Completely randomised design (CRD)

Model	$y_{ij} = \mu_i + e_{ij}$ i=1,2,…t, j= 1, 2,…r
Assumptions	• Data are independent (study design) • The residuals are normally distributed (check histogram, normal probability plot). • The residuals have equal variances (boxplots, test of equal variances)
Estimated parameters	$\mathrm{SST} = \sum_{i=1}^{t}\sum_{j=1}^{r_i}(\bar{y}_{i.} - \bar{y}_{..})^2$ $SSE_f = \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{i.})^2$
Standard error	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ $\hat{S}.E.(\bar{y}_{i.}) = \sqrt{s^2/r_i}$ $s = \sqrt{MSE}$
95% CI	$\bar{y}_{i.} \pm t_{N-t}^{\alpha/2} \cdot \hat{SE}(\bar{y}_{i.}) = \bar{y}_{i.} \pm t_{N-t}^{\alpha/2} \cdot \sqrt{MSE/r_i}$

Contrasts

$C = \sum_{i=1}^{t} k_i\mu_i,\ \sum_i^t k_i = 0$

- k_i indicates the coefficient of the contrast.

$c = \sum_{i=1}^{t} k_i\bar{y}_{i.}$ Var (c) = $\sigma^2 \sum_i \frac{k_i^2}{r_i}$

$t_{crit} = \frac{c-C}{\sqrt{MSE \times \sum_i \frac{k_i^2}{r_i}}}$, v= N-t (df of SSE)

Sum of squares: $SSC = \frac{c^2}{\sum_i \frac{k_i^2}{r_i}}$

$$\sum SSC = Treatment\ SS$$

Orthogonal contrast

Contrasts that convey independent information

If $C_1 = k_1\mu_1 + k_2\mu_2 + \cdots + k_t\mu_t$

$C_2 = l_1\mu_1 + l_2\mu_2 + \cdots + l_t\mu_t$

C & D are orthogonal if

$$\sum_{i=1}^{t} \frac{k_i l_i}{r_i} = 0$$

You can drop the i for equal groups.

Complete set of orthogonal contrasts

- t- 1 mutually orthogonal contrast
- Each pair of contrasts is orthogonal

Multiple comparisons

1. Bonferroni

- suitable for ad hoc comparisons
- small number of tests

$\alpha_c = overall\ significance\ level$

$\alpha_e = individual = \frac{\alpha_c}{k}$, $k = t(t-1)$= no. of comparisons

95% CI:

$$\bar{y}_{i.} - \bar{y}_{j.} \pm t_v^{1-\alpha/2}\sqrt{MSE \cdot \left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$$

v=N-t

$$s_{pooled}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

2. Tukey

- suitable for balanced study
- less conservative

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

we reject H_0 if

$$\frac{\omega}{\sqrt{MSE/r}} = \frac{\max_{i=1,\dots,k}(\bar{Y}_{i.}) - \min_{i=1,\dots,k}(\bar{Y}_{i.})}{\sqrt{MSE/r}} > q_{\alpha,k,v}$$

k=t treatment, v = df of SSE

To find individual differences between means

- Find threshold value $q_{\alpha,k,v} \times \sqrt{MSE/r}$
- If $|\bar{y}_{i.} - \bar{y}_{j.}|$ exceeds threshold value, then reject H_0

3. Scheffe

- suitable for post hoc comparisons
- many tests
- can only conduct two-tailed tests as we are using F-values.

$S = \sqrt{(t-1)F_{\alpha,t-1,N-t}}$

$s_c = \sqrt{MSE \times \sum_i \frac{k_i^2}{r_i}}$

CI: c±S × s_c

If data is non-normal

- Transformation: log or sqrt
- Non-parametric test (Kruskal-Wallis)
  Test the different between treatment medians
  Assumption: all groups have similar shape

If variances are not equal

- Use Levene's test (F distribution) or Barlett's test (chi-square distribution) to see if variances are equal

Randomised complete block design (RBD)

- Suitable when we have homogenous groups

Model	$y_{ij} = \mu + \tau_i + p_j + e_{ij}$ i=1,2,…t, j= 1, 2,…r
Assumption Parameter constraint	• $e_{ij} \sim NID(0, \sigma^2) independently$ • $\sum_{i=1}^{t}\tau_i = \sum_{j=1}^{r}\rho_j = 0.$
Variance	$s_2^2 = \frac{SS_{Block} + r(t-1) \cdot MS_{Error}}{rt-1}$
Relative efficiency	$RE = \frac{(f_1+1)(f_2+3)s_2^2}{(f_1+3)(f_2+1)s_1^2} \frac{CRD}{RBD}$ If RE=2 => RBD is twice efficient. CRD need a sample size 2 times greater to achieve the same precision.

If assumptions fail:

- Transformation
- Non-parametric (Friedman's test)
  Assumptions:
  - Each block contains t random variables or ranking
  - The blocks are independent
  - Within each block, observations can be arranged in increasing order (not too many ties)

  H_0: Each ranking of the random variables within a block is equally likely

  H_1: At least one treatment has larger observed values.

$R(y_{ij})$: rank from 1 to t assigned to y_{ij} within block j

$$R_i = \sum_{j=1}^{r} R(y_{ij}) \qquad i = 1, 2, \ldots, t$$

Friedman test statistics S (chi-square distribution)

$$= \frac{12}{rt(t+1)} \sum_{i=1}^{t} R_i^2 - 3r(t+1)$$

Factorial experiment

- Examine multiple factors at the same time
- Examine interaction first
  - ⇨ (1) Significant: Need to carry out multiple comparisons on the levels of one factor at each level of the other factor
  - (2) Insignificant: Remove interaction term.

<table>
<tr><td>Model</td><td>$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$
i=1,2,…a, j= 1, 2,…b, k= 1, 2,..r</td></tr>
<tr><td>Assumptions</td><td>• $e_{ij} \sim NID(0, \sigma^2) independently$
• $\Sigma_i\alpha_i=0,\ \Sigma_j\beta_j=0.$
• $\sum_{i=1}^{a}(\alpha\beta)_{ij} = 0$ for all j, $\sum_{j=1}^{b}(\alpha\beta)_{ij} = 0$ for all i.</td></tr>
<tr><td>Main effect</td><td>Main effect of A
$= ½ \{[\overline{a_2b_2 + a_2b_1}] - [\overline{a_1b_2 + a_1b_1}]\}$
treat.'s with A at level 2   treat.'s with A at level 1</td></tr>
<tr><td>Two-factor interaction effect</td><td>$(a_2b_2 - a_1b_2) - (a_2b_1 - a_1b_1)$</td></tr>
<tr><td>Estimated parameters</td><td>
<table>
<tr><td>Source</td><td>SS</td><td>df</td><td>MS</td><td>F</td></tr>
<tr><td>Factor A</td><td>SS_A</td><td>a-1</td><td>$SS_A/(a-1)$</td><td>MS_A/EMS</td></tr>
<tr><td>Factor B</td><td>SS_B</td><td>b-1</td><td>$SS_B/(b-1)$</td><td>MS_B/EMS</td></tr>
<tr><td>Inter[n] AB</td><td>SS_{AB}</td><td>(a-1)(b-1)</td><td>$SS_{AB}/(a-1)(b-1)$</td><td>MS_{AB}/EMS</td></tr>
<tr><td>Error</td><td>ESS</td><td>ab(r-1)</td><td>ESS/ ab(r-1)</td><td></td></tr>
<tr><td>Total</td><td>Total SS</td><td>abr-1</td><td></td><td></td></tr>
</table>
</td></tr>
</table>

$$SS_{Total} = \sum_i \sum_j \sum_k \left(y_{ijk} - \bar{y}_{...}\right)^2$$
$$SS_A = br \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2$$
$$SS_B = ar \sum_j \left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2$$
$$SS_{AB} = r \sum_i \sum_j \left(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}\right)^2$$
$$SS_{Error} = \sum_i \sum_j \sum_k \left(y_{ijk} - \bar{y}_{ij.}\right)^2$$

NOTE: $Var(nX)=n^2Var(X)$
$Var(X+Y)= Var(X) + Var(Y)$

2^n Factorial Design

$Var(\hat{A}) = \frac{\sigma^2}{r2^{n-2}}$ SST= $\frac{Y(.)^2}{2^n r}$

Yates' Algorithm

	Col 1	Col 2	Col 3 = n
$\sum(1)$	$a+1$	$ab+b+a+1$	T
$\sum a$	$ab+b$	$ac+c+abc+bc$	$Y(A)$
$\sum b$	$ac+c$	$a-1+ab-b$	$Y(B)$
$\sum ab$	$abc+bc$	$ac-c+abc-bc$	$Y(AB)$
$\sum c$	$a-1$	$ab+b-a-1$	$Y(C)$
$\sum ac$	$ab-b$	$abc+bc-ac-c$	$Y(AC)$
$\sum bc$	$ac-c$	$ab-b-a+1$	$Y(BC)$
$\sum abc$	$abc-bc$	$abc-bc-ac+c$	$Y(ABC)$

Partial confounding**:** Confounding different effects in each rep

Fractional replication**:** (1) find identity relation, (2) find the effect subgroup, (3) Decide fractional replicate and its aliases

Random effects model

- Treatments which are drawn at random from a population of treatments

Model	$y_{ij} = \mu + \alpha_i + e_{ij}$ i=1,2,…t, j= 1, 2,…r
Assumption	• $e_{ij} \sim NID(0, \sigma^2) independently$ • $\alpha_i \sim NID(0, \sigma^2) independently$
Hypothesis	H_0: $\sigma_a^2 = 0$ (no treatment effects) H_1: $\sigma_a^2 > 0$ (there is treatment difference)
Estimated parameters	$MSA = \hat{\sigma}_e^2 + r\hat{\sigma}_a^2$ $MSW = \hat{\sigma}_e^2$ The variance among X accounts for x% of the variation and the variance within X accounts for the other (1-x)%.
CI	A 100(1-α)% CI for σ_e^2 is given by $\Pr\left\{\frac{SSW}{A} < \sigma_e^2 < \frac{SSW}{B}\right\} = 1 - \alpha$ where A = $\chi^2_{N-t,\alpha/2}$ (*upper* $\alpha/2$ point of χ^2_{N-t}) and $B = \chi^2_{N-t,1-\alpha/2}$ (*lower* $\alpha/2$ point of χ^2_{N-t}; $B < A$) are values that separate the upper tail area (higher side) of $\alpha/2$ and $(1 - \alpha/2)$, respectively. An *approximate* 100(1-2α)% CI for σ_a^2 is given by $\Pr\left\{\frac{SSA(1-F_u/F_0)}{rC} < \sigma_a^2 < \frac{SSA(1-F_l/F_0)}{rD}\right\} \approx 1-2\alpha$ where $C = \chi^2_{t-1,\alpha/2}$ (upper tail) $D = \chi^2_{t-1,1-\alpha/2}$ (lower tail; $D < C$) $F_0 = MSA/MSW$ $F_u = F^{\alpha/2}_{t-1,N-t}$ (upper) $F_l = F^{1-\alpha/2}_{t-1,N-t}$ (lower; $F_l < F_u$)
Intraclass correlation coefficient	$\rho_I = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$ ρ_I is a measure of how similar (or dissimilar) units *between* groups are, compared with similarity *within* groups. It is the proportion of variability attributable to groups. In the extreme case of no difference between groups, ρ_I would be zero.

The analysis of covariance

- Examine one factor, but also take into account extraneous (continuous) variables
- We can only measure covariate during the experiment
- The influence of covariate on the response is unknown

Model	$y_{ij} = \mu + \tau_i + \beta\left(x_{ij} - \bar{x}_{..}\right) + e_{ij}$ y_{ij}: $(X) for\ jth\ subject\ in\ ith\ treatment$ μ: $overall\ mean\ (X)$ τ_i: $effect\ of\ the\ ith\ treatment\ on\ (X)$ β: coefficient for the linear regression of y_{ij} on x_{ij} x_{ij}: covariate for jth subject in ith treatment $\bar{x}_{..}$: overall covariate mean
Assumption Parameter constraint	• $e_{ij} \sim NID(0, \sigma^2) independently$ • $\sum_{i=1}^{t} \tau_i = 0$ • Common slope β (not significant)

Adjusted mean of Y	$\overline{y_{i_{adj}}} = \overline{y_{i\cdot}} + \hat{\beta}(\overline{x_{\cdot\cdot}} - \overline{x_{i\cdot}})$

Survey design

Sampling frame: a list of sampling units
Sampling unit: non-overlapping units for sampling
Unit: A group of elements
Element: an object on which measures are taken
NOTE: unit can be element.

Probability concepts

$$E(Y) = \mu$$
$$= \sum_{j=1}^{k} y_j p_j$$

$Var(Y) = \sum_{j=1}^{k} (y_j - \mu)^2 p_j = E(Y^2) - [E(Y)]^2$

Simple random sampling

- Sampling is done without replacement.
- Simpliest, appropriate with no prior information.

$$SE(\bar{y}) = \sqrt{Var(\bar{y})} = \sigma \sqrt{\frac{(1-f)}{n}}$$

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{L} (N_i - 1)\sigma_i^2 + \frac{1}{N-1} \sum_{i=1}^{L} N_i (\mu_i - \mu)^2$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\mathrm{Var}(\bar{y}) = (1-f)\frac{\sigma^2}{n}$$

$$Y_T = N\bar{y};\ \mathrm{Var}(Y_T) = \mathrm{Var}(N\bar{y}) = N^2(1-f)\frac{\sigma^2}{n}$$

Use sample proportion $\hat{p}$ to estimate population proportion p with

$$\left(\begin{aligned} Var(\hat{p}) &= \frac{N}{N-1} \cdot pq \cdot \frac{(1-f)}{n} \\ &\cong (1-f) \cdot \frac{pq}{n} \qquad \text{since } \frac{N}{N\text{-}1} \cong 1 \text{ for large } N \end{aligned} \right)$$

Where q = 1 – p.

Sampling size

$$n \geq N \left\{ 1 + N \left(\frac{B}{z_{\alpha/2}\sigma} \right)^2 \right\}^{-1}$$

To estimate pop. total within D of true value

$$n \geq N \left\{ 1 + N^{-1} \left(\frac{D}{z_{\alpha/2}\sigma} \right)^2 \right\}^{-1}$$

To estimate pop. proportion with a specified margin of error B at siginifcance level α

Ignoring the fpc and taking $(N/(N-1)) = 1$, we get

$$n \geq \frac{p(1-p)\, z_{\alpha/2}^2}{B^2}$$

Not ignoring the fpc: After some manipulation

$$n \geq N \left[1 + \frac{N-1}{p(1-p)} \left(\frac{B}{z_{\alpha/2}} \right)^2 \right]^{-1}$$

If no p is given, use p=0.5 (conversative + large sample size)

Stratified random sampling

- Use to reduce variance
- Obtain best results when within-stratum differences is small, and large differences between stratum means.

$$\bar{y}_{ST} = \sum_{1}^{L} \frac{N_i}{N} \bar{y}_i$$

$$\text{with } \mathrm{Var}(\bar{y}_{ST}) = \mathrm{Var}\left(\sum_{1}^{L} \frac{N_i}{N} \bar{y}_i \right) = \mathrm{Var}\left(\sum_{i=1}^{L} W_i \bar{y}_i \right) = \sum_{i=1}^{L} W_i^2 \mathrm{Var}(\bar{y}_i)$$

$$= \sum_{i=1}^{L} W_i^2 (1-f_i) \frac{\sigma_i^2}{n_i};$$

$$y_{T,ST} = N \cdot \bar{y}_{ST} = \sum_{i=1}^{L} N_i \bar{y}_i$$

$$\text{with } \mathrm{Var}(y_{T,ST}) = \mathrm{Var}(N \cdot \bar{y}_{ST}) = N^2 \cdot \mathrm{Var}(\bar{y}_{ST})$$

$$= N^2 \cdot \sum_{i=1}^{L} W_i^2 (1-f_i) \frac{\sigma_i^2}{n_i} = \sum_{i=1}^{L} N_i^2 (1-f_i) \frac{\sigma_i^2}{n_i}$$

$$\hat{\mathrm{Var}}(y_{T,ST}) = N^2 \hat{\mathrm{Var}}(\bar{y}_{ST}) = \sum_{i=1}^{L} N_i^2 (1-f_i) \cdot \frac{s_i^2}{n_i};$$

$$\hat{p}_{ST} = \sum_{i=1}^{L} \left(\frac{N_i}{N} \right) \cdot \hat{p}_i$$

$$\hat{\mathrm{Var}}(\hat{p}_{ST}) = \sum_{i=1}^{L} \left(\frac{N_i}{N} \right)^2 \cdot \hat{\mathrm{Var}}(\hat{p}_i)$$

$$= \sum_{i=1}^{L} \left(\frac{N_i}{N} \right)^2 \cdot (1-f_i) \cdot \left(\frac{\hat{p}_i \hat{q}_i}{n_i - 1} \right)$$

Sample size

The sample size for stratum i according to ***proportional allocation***

$$n_i = \frac{n \cdot N_i}{N} = f \cdot N_i.$$

The optimal sample size in stratum *i* is,

$$n_i = n \cdot \left(\frac{\frac{N_i \sigma_i}{\sqrt{c_i}}}{\sum_{j=1}^{L} \frac{N_j \sigma_j}{\sqrt{c_j}}} \right)$$

When sampling costs over the strata are equal,
i.e. $c_1 = c_2 = \ldots = c_L$, this reduces to

$$n_i = n \cdot \frac{N_i \sigma_i}{\sum_{j=1}^{L} N_j \sigma_j}$$

This is known as *Neyman allocation*.

Design effect

$$deff = \frac{Var(\text{estimate under current sampling plan})}{Var(\text{estimate under SRS with same sample size})}$$

If deff <<1, then stratified random sampling is better than SRS.

Cluster sampling

- Easy to implement
- Clusters are generally geographical entites
- Useful when there is a large within-cluster variation but small between-cluster variation.

a) When all cluster sizes are equal to L:

$$\bar{y}_{CL} = \frac{1}{n}\sum_{1}^{L}\bar{y}_i = \frac{1}{n}\sum_{i=1}^{L}\left(\frac{1}{L}\sum_{j=1}^{L} y_{ij}\right)$$

$$\text{with } \hat{Var}(\bar{y}_{CL}) = (1-\frac{n}{N})\frac{s_t^2}{nM^2}$$

$$\text{where } s_t^2 = \frac{1}{n-1}\sum_{i=1}^{L}\left(\sum_{j=1}^{L} y_{ij} - \frac{\frac{N}{n}\sum_{i=1}^{L}\sum_{j=1}^{L} y_{ij}}{N}\right)$$

b) When cluster sizes are not all equal, we estimate the mean through ratio:

$$\bar{y}_{CL} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}, \text{ with } \hat{Var}(\bar{y}_{CL}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)s_r^2$$

$$\text{where } s_r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{CL} m_i)^2}{n-1}$$

Systematic sampling

- Simple, save time and effort
- Useful when we do not have a list of population
- If the population is period, DO NOT USE.

Method A: When N/k is an integer, choose a unit at random from the first k^th unit. Take every k^th unit from the starting unit.
Method B: Choose a unit at random from the population. Starting point depends on remainder.
When N/k is not an integer, use Method B to ensure an unbiased estimator of μ

$\bar{y}_{sys}$ = mean of a systematic sample, with

$$Var(\bar{y}_{sys}) = \tfrac{N-1}{N}\sigma^2 - \tfrac{k(m-1)}{N} S_W^2$$

$$where \quad \sigma = population\ SD,$$

$$S_W^2 = \tfrac{1}{k(m-1)}\sum_{i=1}^{k}\sum_{j=1}^{m}(u_{ij} - \bar{u}_{i.})^2 \ and$$

$$\bar{u}_{i.} = \tfrac{1}{m}\sum_{j=1}^{m} u_{ij}$$

Ratio estimator

$r = \frac{\bar{y}}{\bar{x}}$, and

$E(r) \cong \frac{\mu_y}{\mu_x}$ (approximation is good for large samples)

$$Var(r) \cong \frac{1-f}{n}\cdot\frac{1}{\mu_x^2}\cdot\sum_{i=1}^{N}\frac{(Y_i - RX_i)^2}{N-1}$$

We estimate

$\sum_{i=1}^{N}\frac{(Y_i - RX_i)^2}{N-1}$ by $\sum_{i=1}^{n}\frac{(y_i - rx_i)^2}{n-1}$

and μ_x^2 by $\bar{x}^2$, giving

$$\hat{Var}(r) = \frac{1-f}{n}\cdot\frac{1}{\bar{x}^2}\cdot\sum_{i=1}^{n}\frac{(y_i - rx_i)^2}{n-1} = \frac{1-f}{n}\cdot\frac{1}{\bar{x}^2}\cdot s_e^2$$

$$\hat{\tau}_y = Y_{T,r} = \tau_x\cdot\left(\frac{\bar{y}}{\bar{x}}\right)$$

$$\text{with } Var(\hat{\tau}_y) = Var(\tau_x\cdot\left(\frac{\bar{y}}{\bar{x}}\right)) = \tau_x^2 Var(r)$$

$$\hat{Var}(Y_{Tr}) = \tau_x^2 \hat{Var}(r)$$

$$= (N\mu_x)^2 \frac{1-f}{n} \frac{1}{\bar{x}^2} \frac{1}{n-1} \sum_1^n (y_i - rx_i)^2$$

$$\cong \frac{N(N-n)}{n(n-1)} \sum_1^n (y_i^2 - 2rx_i y_i + r^2 x_i^2)$$

$$= \frac{N(N-n)}{n} S_e^2.$$

Regression estimator

$$\bar{y}_{LR} = \bar{y} - \hat{\beta}(\bar{x} - \mu_X)$$

$$Var(\bar{y}_{LR}) \approx \tfrac{1-f}{n}\left[\tfrac{n-1}{n-2}\right][s_y^2 - \hat{\beta}^2 s_x^2]$$

where

$$s_y^2 = \tfrac{1}{n-1}\sum_1^n (y_i - \bar{y})^2 \quad and \quad s_x^2 = \tfrac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$$

MSE= $\left[\frac{n-1}{n-2}\right][s_y^2 - \hat{\beta}^2 s_x^2]$