10. Issues on the determination of trial size

§10.1. The general theory

- **Review of hypothesis testing**
  - Null and alternative hypotheses.
  - Simple and composite hypotheses.
  - Test statistic and critical value.
  - Type I and type II errors.
  - Size of a test.
  - Power
  - Indifference region

- **Noncentral χ^2 and F distributions**

  χ^2 **distribution**:

A central χ^2 statistic with df m has a general structure

$$\chi^2 = \boldsymbol{X}' A \boldsymbol{X} / \sigma^2,$$

where $\boldsymbol{X}$ is a vector of i.i.d. $N(0, \sigma^2)$ random variables, and A is an idempotent matrix (i.e., $A^2 = A$) with rank m. In particular, when $A = I$, the χ^2 statistic reduces to

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{m} X_i^2.$$

If some of the components of $\boldsymbol{X}$ does not have zero mean, the statistic is said to have non-central χ^2 distribution with noncentrality parameter $\delta = \boldsymbol{\mu}' A \boldsymbol{\mu} / \sigma^2$, where $\boldsymbol{\mu} = E\boldsymbol{X}$. (Note: the definition of noncentrality parameter is consistent with R but differ from the text book where $\sqrt{\delta}$ is defined as noncentrality parameter.)

F distribution

A central F statistic with df ν_1, ν_2 has a structure

$$F = \frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2},$$

where $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ are two independent random variables with central χ^2 distributions having df ν_1 and ν_2 respectively.

If $\chi^2_{\nu_1}$ has a noncentral χ^2 distribution with noncentrality parameter δ, then F is said to have a noncentral F-distribution with noncentrality parameter δ.

- **General method for the determination of sample size**

The sample size is determined such that the alternatives beyond an indifference region can be detected with a specified power. For any test, while the type I error is controlled at a given level, the power of the test at any particular alternative is an increasing function of the sample size. Sample size can be determined by trial and error through power calculation. The general procedure is as follows:

1. Determine the indifference region;
2. Specify the test size and the minimum power for detecting alternatives beyond the indifference region;
3. Search for the sample size n required by trial and error through the calculation of power at different n's.

Example of F test: For a F test, the test statistic follows a central F distribution with df ν_1 and ν_2, say, under the null hypothesis, but noncentral F distributions under the alternative hypothesis. The df (usually ν_2) and the noncentrality parameter δ are functions of the sample size n. If the size of the test is α and the specified power is $1-\beta$. Let $F_{\nu_1,\nu_2,\delta}$ denote the random variable with the noncentral F distribution. Then, n is determined from the equation

$$P(F_{\nu_1,\nu_2,\delta} > F_{\nu_1,\nu_2,\alpha}) = 1-\beta,$$

where $F_{\nu_1,\nu_2,\alpha}$ is the upper α quantile of the central F distribution with df ν_1 and ν_2.

§10.2. Determination of trial size for the comparison of two treatments

- **Two parallel groups**

The test statistic for the comparison of two treatment in two parallel groups of same size

n is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s}\sqrt{\frac{n}{2}},$$

where s is an estimate of the standard deviation σ. Under normality assumption and the null hypothesis $H_0 : \mu_1 = \mu_2$, t follows a t distribution with df $2(n-1)$.

If s is replaced by σ then t follows a standard normal distribution. In sample size calculation, one can consider, instead of t, the statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma}\sqrt{\frac{n}{2}},$$

taking σ as known (to be determined).

At level α, the test rejects H_0 if $|Z| > \alpha/2$.

If $\Delta = \mu_1 - \mu_2 \neq 0$, the power of the test is

$$
\begin{aligned}
p &= P\left(\left|\frac{\bar{X}_1 - \bar{X}_2}{\sigma}\sqrt{\frac{n}{2}}\right| > z_{\alpha/2}\right) \\
&= P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sigma}\sqrt{\frac{n}{2}} > z_{\alpha/2}\right) \\
&\quad + P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sigma}\sqrt{\frac{n}{2}} < -z_{\alpha/2}\right) \\
&= P\left(\frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\sigma}\sqrt{\frac{n}{2}} > z_{\alpha/2} - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right) \\
&\quad + P\left(\frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\sigma}\sqrt{\frac{n}{2}} < -z_{\alpha/2} - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right)
\end{aligned}
$$

That is,

$$
\begin{aligned}
p &= P\left(Z > z_{\alpha/2} - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right) \\
&\quad + P\left(Z < -z_{\alpha/2} - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right) \\
&\approx P\left(Z > z_{\alpha/2} - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right).
\end{aligned}
$$

where Z denotes the standard normal variable.

If the power required is at least $1 - \beta$ for the difference beyond Δ, the required n should satisfy

$$z_{\alpha/2} - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}} = -z_\beta.$$

Hence

$$n = \frac{2\sigma^2(z_{\alpha/2} + z_\beta)^2}{\Delta^2}.$$

For a particular trial, both Δ and σ^2 must be specified. The σ^2 is the variance of the measurements within each group.

- **Two treatment comparison with other designs**

Matched pair experiments

For the matched pair experiments, the test statistic is

$$t = \frac{\sqrt{n}\bar{d}}{s_d},$$

where s_d^2 is the estimate of $\text{Var}(X_{i1} - X_{i2}) = 2\sigma^2(1 - \rho)$, where ρ is the inter correlation within each pair, σ^2 denotes the same variance as in the parallel groups design.

The sample size required to achieve power $1 - \beta$ at mean difference Δ is then

$$\begin{aligned} n^* &= \frac{2\sigma^2(1-\rho)(z_{\alpha/2} + z_\beta)^2}{\Delta^2} \\ &= n/\text{RE}, \end{aligned}$$

where $\text{RE} = 1/(1-\rho)$ is the relative efficiency of the matched pair experiments with respect to the parallel groups design, and n is the sample size required by the parallel groups design to achieve the same power.

In general, if one design (A) has a relative efficiency RE to another design (B), if the sample size required by design B is n, then the sample size required for design A is

$$n^* = n/\text{RE}.$$

The RE is essentially the ratio of two variances. In general, σ^2 can be decomposed into

$$\sigma^2 = \sigma^2_{\text{CF}} + \sigma^2_e,$$

where σ^2_e is the variance of random error and σ^2_{CF} is the variance of prognostic factors. If a design can control the prognostic factors, the σ^2_e can be estimated and used in the test statistics. It then results in a relative efficiency

$$\text{RE} = \frac{\sigma^2_{\text{CF}} + \sigma^2_e}{\sigma^2_e}.$$

Covariate control

If a covariate is controlled, and the correlation of the covariate with the response is ρ, then the relative efficiency of the trial with covariate controlled to the parallel groups design is

$$\text{RE} = \frac{1}{1 - \rho^2}.$$

Note that, in the regression model

$$X = \beta_0 + \beta Z + \epsilon,$$

we have

$$\sigma^2 = \sigma_e^2 + \beta^2 \sigma_Z^2 = \sigma_e^2 + \rho^2 \sigma^2.$$

Hence

$$\frac{\sigma^2}{\sigma_e^2} = \frac{\sigma^2}{\sigma^2(1 - \rho^2)} = 1/(1 - \rho^2).$$

The required sample size for the trial is then

$$n^* = n/\text{RE} = n(1 - \rho^2).$$

Two-period crossover study

The relative efficiency is

$$\mathrm{RE} = \frac{\sigma_s^2 + \sigma_e^2}{\sigma_e^2},$$

where σ_s^2 is the variance due to subject.

If n denotes the number of measurements for each group in a parallel groups study (in this case it is the same as the sample size for each group) required to achieve the desired power, the number of measurements required for the crossover study is then

$$n^* = n(\frac{\sigma_e^2}{\sigma_s^2 + \sigma_e^2}) = n(1 - R),$$

where $R = \frac{\sigma_s^2}{\sigma_s^2+\sigma_e^2}$. Since each sample will contribute to two measurements in the crossover study, the required sample size is then

$$\frac{n^*}{2} = \frac{n(1 - R)}{2}.$$

§10.3. Determination of trial size for the comparison of more than two treatments

- **Parallel groups**

  Detection of overall effect

  If the purpose of study is to detect any difference among the treatments, the sample size calculation should be based on the F-test for the total treatment effects. The test statistic is given by

$$F = \frac{\text{BMS}}{\text{WMS}},$$

where WMS is an unbiased estimate of σ^2 and BMS is given by

$$(g-1)\text{BMS} = \sum_{i=1}^{g} n(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2,$$

which is a quadratic form of

$$(\sqrt{n}\bar{X}_{1\cdot}, \dots, \sqrt{n}\bar{X}_{g\cdot}).$$

Under $H_0 : \mu_1 = \cdots = \mu_g$, $\sqrt{n}\bar{X}_{i\cdot}$'s are i.i.d. $N(0, \sigma^2)$ variables, and $(g-1)\mathrm{BMS}/\sigma^2$ follows a central χ^2-distribution. If H_0 is not true, it follows a noncentral χ^2 distribution with noncentrality parameter

$$\delta = \frac{n \sum_{i=1}^{g} (\mu_i - \bar{\mu})^2}{\sigma^2}.$$

The sample size is then determined by solving

$$P(F_{\nu_1,\nu_2,\delta} > F_{\nu_1,\nu_2,\alpha}) = 1 - \beta,$$

with $\nu_1 = g - 1$, $\nu_2 = n(g - 1)$ and δ given above.

Pairwise comparisons

If only certain pairwise comparisons are of interest, the sample size can be determined as if under consideration is a two parallel groups study. The test size must be adjusted by the Bonferroni criterion.

- **Trial size determination for designs controlling prognostic factors**

  The sample size calculation for other designs is done with some modification on the degrees of freedom and the noncentrality parameter:

  – The second df in the F statistic ν_2 is adjusted to the df of the residual sum of squares.

  – The noncentrality parameter is adjusted to δRE, where RE is the relative efficiency of the design under consideration to the parallel groups design.

Randomized block design

ν_2: $(n-1)(g-1)$.

RE: s^2_{est}/RMS.

The RMS is the residuals mean sum of squares resulting from the randomized block design, and s^2_{est} is an estimate of the variance within each treatment groups if the experiment has not been in random blocks which is given by

$$s^2_{est} = \frac{(n-1)\mathrm{BMS} + n(g-1)\mathrm{RMS}}{ng-1}.$$

In the sample size calculation, the RE can be taken from other similar studies.

Stratified study

ν_2: $g(n-S)$, S being number of strata.

RE: s^2_{est}/RMS.

RMS: from the stratified experiment.

$$s_{est}^2 = \frac{1}{n_{\cdot\cdot} - g}[(n_{\cdot\cdot} - gS)s^2 + \sum_{i=1}^{g}\sum_{a=1}^{S} n_{ai}(\bar{X}_{ai} - \bar{X}_{\cdot i})^2],$$

where s^2 is the pooled variance across treatments and strata.

Covariate control

ν_2: $g(n-1)-1$.

RE: $1/(1-\rho^2)$, ρ being the correlation between the response variable and the covariate.

Latin square design

ν_2: $(qg-2)(g-1)$. $n = qg$

RE: $\text{RMS}_{rb}/\text{RMS}$, where

$$\text{RMS}_{rb} = \frac{\text{CMS} + (g-1)\text{RMS}}{g}.$$

The RMS and CMS are the residual and column mean sum of squares for the Latin square design. The relative efficiency is relative to a complete randomized block design.

Crossover study

ν_2: $(qg-3)(g-1)$. $n = qg$

RE: $\frac{\sigma_s^2+\sigma_e^2}{\sigma_e^2}$, where σ_s^2 is the variance due to subject.

Note that the n refers to the number of measurements for each treatment. The resulted n^* from the sample size calculation is also the number of measurements for each treatment. While a parallel groups design requires a total gn samples, the crossover study only needs n^* samples.

Balanced incomplete block design

ν_2: $n(k-1)-g+1$.

The RE is taken as the efficiency factor EFF. The relative efficiency is relative to a randomized block design.

§10.4. R function for trial size calculation

```
# R function for computing sample size of parallel groups design

parallel.sample.size = function(g,sigma,mu.v,alpha, p ,n.start) {
 # Input:
 # g: number of groups;
 # sigma: error standard deviation;
 # mu.v:  \sum ( mu_i - \bar{\mu })^2;
 # alpha: size of the test;
 # p: power of the test at assumed sigma and mu.v;
```

```
# n.start: a starting value for the group sample size.
#
# Output:
# n:  required group sample size.

#determination of the bounds of n
 flag1 = 0
 flag2 = 0
 while (flag1 == 0 | flag2 == 0) {
   n = n.start
   df1 = g-1
   df2 = g*(n-1)
   delta = (n * mu.v)/sigma^2
   F1 = qf(1-alpha, df1, df2)
   power = 1 - pf( F1, df1, df2, delta)
   if (power >= p) {
     n.u = n
     flag1 = 1
     n.start = floor(n*0.618)
   } else {
     n.l = n
     flag2 = 1
     n.start = floor(n/0.618)
   }
}

# search for the required n
 flag3 = 0
 while (flag3 == 0) {
   n.tmp = floor( n.l + (n.u-n.l)*0.618 )
   n = n.tmp
```

```
    df1 = g-1
    df2 = g*(n-1)
    delta =(n * mu.v)/sigma^2
    F1 = qf(1-alpha, df1, df2)
    power = 1 - pf( F1, df1, df2, delta)
    if (power >= p) {
      n.u = n.tmp
    } else {
      n.l = n.tmp
    }
    if (n.u-n.l <= 1 ) flag3 = 1
  }

  n = n.tmp
  list(n=n, alpha = alpha, beta = 1-p)
}

# Example
 g=4
 sigma = 3
 mu.v =  9.901
 alpha = 0.05
 p = 0.8
 n.start = 8
 parallel.sample.size(g,sigma,mu.v,alpha, p ,n.start)
```