Chap3 - Design Experiments and Variance Analysis

Introduction

ANalysis of VAriance (ANOVA) is a commonly used technique for investigating data by comparing means of subsets of the data. the simplest case is an extension of two-sample t-test for comparing two groups covering situations where there are more than two groups being compared.

Example:

Imagine we suspect that the use of a certain fertilizer will improve the production of a certain plant. In order to develop an experiment to confirm this hypothesis we treat a field with the new fertilizer (treatment 1) and another one under the same or very similar conditions (e.g. sun hours, etc...) in the usual way (control)

Denoting: Y_1 = 'Dried weight of plants treated with new fertilizer (treat. 1)'

$E(Y_1) = E(Y_2)$ $\quad Y_2$ = 'Dried weight of plants treated with the usual way (control)'

$\mu_1 \quad \mu_2$

Hypotheses: (two sided test)

1. $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

Assuming that $Y_1 \sim N(\mu_1, \sigma^2)$ $Y_1 \perp\!\!\!\perp Y_2$
$Y_2 \sim N(\mu_2, \sigma^2)$

having $(Y_{11}, \dots, Y_{1n_1})$ Random sample of Y_1 → $\bar{Y}_1$
$(Y_{21}, \dots, Y_{2n_2})$ Random sample of Y_2 independent from the first one → $\bar{Y}_2$

$\bar{Y}_1 \sim N(\mu_1, \frac{\sigma^2}{n_1})$
$\bar{Y}_2 \sim N(\mu_2, \frac{\sigma^2}{n_2})$

2. Pivotal Quantity

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

where $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$, and ($S_p^2$ pooled)

$$S_i^2 = \sum_{k=1}^{n_i} \frac{(Y_{ik} - \bar{Y}_i)^2}{n_i - 1}, \quad i = 1, 2$$

3. test statistics

Under the validity of H_0:

$$T \mid H_0 = T_0 = \frac{\bar{Y}_1 - \bar{Y}_2 - 0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

4. Observed value of the test statistics

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Let suppose that $n_1 = n_2 = 10$

$\bar{y}_1 = 5.032$ $\quad$ $\bar{y}_2 = 4.661$

$s_1^2 = 0.340$ $\quad$ $s_2^2 = 0.630$

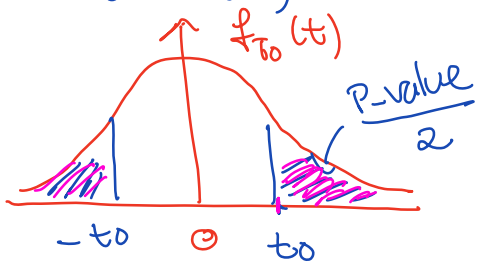
$n_1 + n_2 - 2 = 18$

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2 = \frac{1}{n_1-1}\left(\sum_{i=1}^{n_1} y_{1i}^2 - n_1\bar{y}_1^2\right)$$

$$t_0 = \frac{5.032 - 4.661}{\sqrt{\frac{2}{10}\left(\frac{9\times 0.34 + 9\times 0.63}{10+10-2}\right)}} = 1.1913$$

5. P-value

$T_0 \sim t_{(18)}$



$$\text{P-value} = 2\,P(T_0 \geq 1.1913) = 2\,(1 - F_{t_{(18)}}(1.1913)) =$$

$$= 2\,(1 - 0.8758) = 0.249$$

$\Rightarrow$ do not reject H_0 at the usual α's.

iff H_0 $\forall\, \alpha \geq 0.249$

1%, 5%, 10%

//

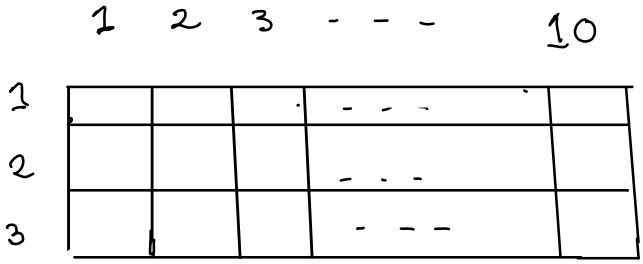
Let us imagine that instead of one new fertilizer we want to compare two fertilizer with the situation where we do not use fertilizer. Only in this way we can prove that the use of a certain fertilizer is worthwhile (or not).

the answer to this problem is the goal of ANOVA. But to Apply the ANOVA methodology we have to be carefull with the Design of the experiment. If we want to collect data from 10 fields, they should have the same caracthenistics and the Association of a field with one treatment should be done in a random way in order to avoid bias. Also, in advanced the field have to be consider with uniform coracteristics.

Notation: we will call factor to a categorical variable:

$$X = \begin{cases} 1, \text{ control} \\ 2, \text{ treatment 1} \\ 3, \text{ treatment 2} \end{cases}$$

and to its Levels we call treatments (or levels), even if they do not have any relation with treatments



we should assign a treatment to each field randomly

this method is often in scientific or medical experiments, engineering design, etc..., when treatments, process, material or products are being compared.

Examples:

- An experiment to study the effect of 4 differents training methods of one course on the scores of the students;
- An experiment to study the effect of 3 different colour of a tea box in the sells;
- Determination of key product design parameters that affect product performance
- selection of design parameters so that the product will work well under a wide variety of fied conditions.

As in regression we have a response variable (Y)

continous

Y - response variable (dependent)

We want to find out how the mean of Y change across a set of conditions that have all been tested within the same experiment. the various conditions being compared in the experiment are defined in terms of one (one way ANOVA) or more (two way ANOVA) categorical variable(s) - Independent variable(s) or explanatory variable(s) call factor(s)

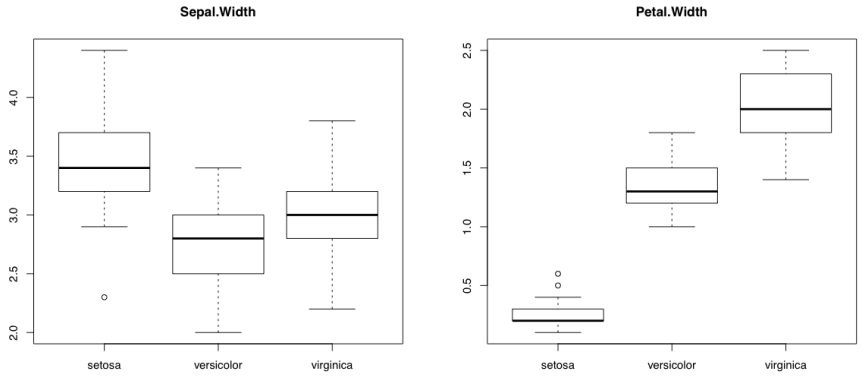
Y - response

x - explantory variable (Factor)

} one way ANOVA

No Assumptions (Linear relationship) is made about the relationship between Y and x. We attempt to check out how the mean of Y differ significantly at different Levels of x (call treatments)

- The ANOVA methodology was proposed by Sir Ronald Fisher in the Rothamstead Experimental Station.
- The aim is to see if there is any difference between groups on some variable. As an example let us see the IRIS data set with the variables Sepal Width and Petal Width by the groups Setosa, Versicolor and Verginica.



- The variable Petal Width looks like different for the three groups.

Single-factor Analysis of variance (one-way ANOVA)

terminology:

Experiment unit: Is the object on which the response and factor are observed

treatment: What we do to the experiment unit

Factor: A controllable experimental variable that is thought to influence the response

a = # treatments (levels of the factor)

y_{ij} = observation of the j-th unit with the treatment i

n_i = # experimental units with treatment i

N = # observations

the response on each of the a treatments is a random variable. the observed data can be summarized in the following table:

treatments	(i,j) observations			totals	Averages
1	y_{11}	y_{12} - - -	y_{1n_1}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22} - - -	y_{2n_2}	$y_{2.}$	$\bar{y}_{2.}$
⋮	⋮			⋮	⋮
a	y_{a1}	y_{a1} - - -	y_{an_a}	$y_{a.}$	$\bar{y}_{a.}$
				$y_{..}$	$\bar{y}_{..} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}}{N}$

$a \geq 2$

$i = 1, \ldots, a$

$j = 1, \ldots n_i$

where : $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij} \Rightarrow \bar{y}_{i\cdot} = \frac{y_{i\cdot}}{n_i}$

$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}$

$N = \sum_{i=1}^{a} n_i \qquad \bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{N}$

We will apply a Linear model to describe the observations :

ONE-WAY ANOVA Model

Model : $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, $i = 1, \dots, a$
$j = 1, \dots, n_i$

where :

Y_{ij} - random variable - response associated with the j-th replication under the i-th treatment ;

μ - overall mean common to all treatments;

τ_i - i-th treatment effect ;

ε_{ij} - random error

a - number of treatments

n_i - number of replications of treatment i

this model can be written as

$Y_{ij} = \mu_i + \varepsilon_{ij}$ where $\mu_i = \mu + \tau_i$

$$Y_{ij} = \underbrace{\mu + \tau_i}_{\mu_i} + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}$$

$i = 1, \dots, a$

$j = 1, \dots, n_i$

Assumptions: $\varepsilon_{ij} \underset{\text{n.i.d.}}{\sim} N(0, \sigma^2)$

$\Rightarrow Y_{ij} \underset{\text{Indep.}}{\sim} N(\mu_i, \sigma^2)$

- $E(Y_{ij}) = E(\mu_i + \varepsilon_{ij}) = \mu_i$
- $VAR(Y_{ij}) = VAR(\mu_i + \varepsilon_{ij}) = VAR(\varepsilon_{ij}) = \sigma^2$

to be able to use this model we will require that the observations are taken in random order and that the environment in which the treatments are used is as uniform as possible. this experimental design is called Completely Random design

the levels of the factor are fixed in advanced $\Rightarrow$ fixed effect model

In this model, the treatment effect τ_i are defined as deviation from the overall mean μ

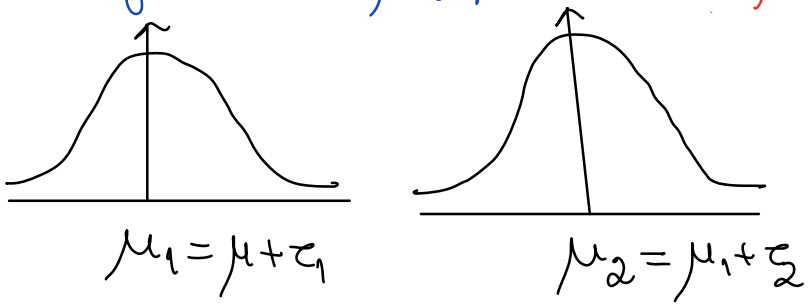
We will assume that:

$$\sum_{i=1}^{a} \tau_i = 0$$

What we want to check out?

If the mean of Y is the same for all treatments

$Y_{ij} \sim N(\mu_i, \sigma^2)$ if $\mu_1 = \mu_2 = \cdots = \mu_a$



same variance but different Locations?

Test:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$ vs $H_1: \mu_i \neq \mu_j$ for some (i,j)

$\Leftrightarrow$

$H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0$ vs $H_1: \exists_i : \tau_i \neq 0$

If H_0 is true than $E(Y_{ij}) = \mu$ (Same Location)

i.e, all observations are taken from the same normal with mean μ and variance σ^2. therefore, if H_0 is true changing the Levels of the factor has no effect on the mean response

- In each case the errors ε_{ij} are independent and identically distributed $N(0, \sigma^2)$. In the second parametrization μ is an overall mean and τ_i is a treatment effect. In order not to have too many parameters we apply the constraint $\sum_{i=1}^{a} \tau_i = 0$. Alternatively we may make μ represent the mean for the first population and set $\tau_1 = 0$. In this case $\tau_2, \ldots, \tau_a$ represent differences of the other population means from the first.

- Note that we thus have three key assumptions assumed in all populations:
  1. normality,
  2. independence,
  3. equal variances.

  Our primary interest is to test if the treatment means are all the same, i.e. to test:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ vs $H_1 : \mu_i \neq \mu_j$ for some (i, j),

or in the equivalent formulation

$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$ vs $H_1 : \tau_i \neq 0$ for some i.

Like in the regression model we can define:

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

$\bar{y}_{..}$ → total mean of observations

$$\bar{y}_{..} = \frac{y_{..}}{N}$$

$$N = \sum_{i=1}^{a} n_i$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - N\bar{y}_{..}^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - N\left(\frac{y_{..}}{N}\right)^2$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$$

and as in regression:

$$SST = SSTR + SSE$$

SSTR → treatments, SSE → error

$$SSTR = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$$

Between-treatment

$$= \sum_{i=1}^{a} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{a} n_i (\bar{y}_{i.}^2 - 2\bar{y}_{i.}\bar{y}_{..} + \bar{y}_{..}^2)$$

$$= \sum_{i=1}^{a} n_i \bar{y}_{i.}^2 - 2\bar{y}_{..}\sum_{i=1}^{a} n_i \bar{y}_{i.} + N\bar{y}_{..}^2$$

$n_1\bar{y}_{1.} + \cdots + n_a\bar{y}_{a.} = y_{..} = N\bar{y}_{..}$

$$= \sum_{i=1}^{a} n_i \bar{y}_{i.}^2 - N\bar{y}_{..}^2 = \sum_{i=1}^{a} n_i\left(\frac{y_{i.}}{n_i}\right)^2 - N\bar{y}_{..}^2$$

$$= \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

within-treatment

$$MSTR = \frac{SSTR}{a-1} \; ; \; MSE = \frac{SSE}{N-a} = \hat{\sigma}^2$$

$$MST = \frac{SST}{N-1}$$

IT can be proved that

$$E(SSTR) = (a-1)\sigma^2 + \sum_{i=1}^{a} n_i \tau_i$$

If H_0 is true ($\tau_1 = \tau_2 = \cdots = \tau_a = 0$) then

$E(SSTR) = (a-1)\sigma^2$ and $E(MSTR) = E\left[\frac{SSTR}{a-1}\right] = \sigma^2$

ie, under H_0 MSTR is an unbiased estimator of σ^2

IT can also be proved that

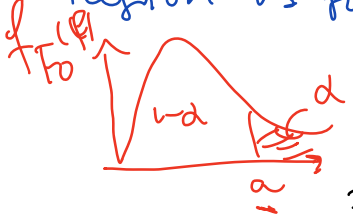
$E(MSE) = \sigma^2$ if H_0 is true or not and

MSE $\perp\!\!\!\perp$ MSTR

Under H_0, we have that

$$F_0 = \frac{MSTR}{MSE} = \frac{\frac{SSTR}{a-1}}{\frac{SSE}{N-a}} \overset{H_0}{\sim} F(a-1, N-a)$$

If H_0 is false $E(MSTR) > \sigma^2$ thus the rejection region is for large values of F_0



reject H_0 if $F_0 > F^{-1}_{F(a-1, N-a)}(1-\alpha)$

IF H_0 is false MSTR will be large. and

$F_0 >> 1$

ANOVA table (one-way - fixed model)

Source of variation	Sum of squares	Degrees of Freedom	Mean Square	F_0
treatments	SSTR	$a-1$	MSTR	$\frac{MSTR}{MSE}$
Error	SSE	$N-a$	MSE	
total	SST	$N-1$		

Pontual estimators of model parameters

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, \dots, a \quad j = 1, \dots, n_i$$

$$E[y_{ij}] = \mu_i = \mu + \tau_i \quad ; \quad \hat{E}[y_{ij}] = \hat{y}_{ij}$$

$$\hat{\mu}_i = \bar{y}_{i.} \quad \text{and} \quad \hat{\mu} = \bar{y}_{..}, \text{ so } \hat{y}_{ij} = \hat{\mu}_i = \bar{y}_{i.}$$

$$\tau_i = \mu_i - \mu \quad ; \quad \hat{\tau}_i = \hat{\mu}_i - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..}$$

residuals estimators

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij}$$

$$r_{ij} = e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.}$$

Example:

Let y = Percentage of smokers in a population

Factor = income with 5 levels

Income	y_{ij}			$y_{i\cdot}$
1	38	42	14	94
2	41	41	16	98
3	36	39	18	93
4	32	36	15	83
5	28	33	17	78

$y_{\cdot\cdot} = 446$

$n_i = 3, \forall i$

$N = a \times n = 5 \times 3 = 15$

$\sum_{i=1}^{5} \sum_{j=1}^{3} y_{ij}^2 = 14870$

We can use analysis of variance to test the hypothesis that different income levels do not affect the mean percentage of smokers in the population, ie:

Model: $y_{ij} = \mu + \tau_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}$; $i = 1, \dots, 5$ (a) (income levels)

$j = 1, \dots, 3$ (n) (replications)

$a = 5$

$n_1 = n_2 = \dots = n_5 = 3$

Hypotheses:

$H_0: \mu_1 = \mu_2 = \dots = \mu_5$ vs $H_1: \mu_i \neq \mu_j$ for some (i, j)

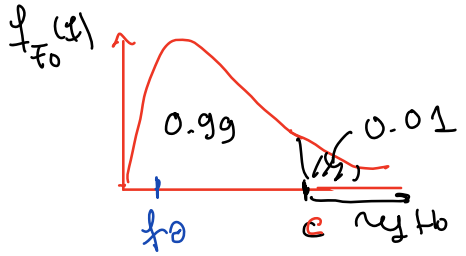
$\tau_1 = \tau_2 = \dots = \tau_5 = 0$ vs $\tau_i \neq 0$ for at least one i

$N = 5 \times 3 = 15$

test statistic :

$$F_0 = \frac{MSTR}{MSE} \sim F(a-1, N-a) \equiv F(4, 10)$$

For $\alpha = 0.01$, reject H_0 if $F_0 > F^{-1}_{F(4,10)}(0.99) = c$



$$F^{-1}_{F(4,10)}(0.99) = 5.9943 \quad (\downarrow \text{table})$$

Critical region = $]\, 5.9943; +\infty\, [$

Decision : (ANOVA table)

$$SSTR = \sum_{i=1}^{a} \frac{y_{i\cdot}^2}{n_i} - \frac{y_{\cdot\cdot}^2}{N} = \sum_{i=1}^{5} \frac{y_{i\cdot}^2}{3} - \frac{y_{\cdot\cdot}^2}{15}$$

$$= \frac{94^2 + 98^2 + 93^2 + 83^2 + 78^2}{3} - \frac{(446)^2}{15} = 92.9333$$

$$SST = \sum_{i=1}^{5} \sum_{j=1}^{3} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{N} = 14870 - \frac{(446)^2}{15} = 1608.933$$

$$SSE = SST - SSTR = 1608.933 - 92.9333$$

$$= 1516$$

ANOVA table

Source of variation	SS	df	MS	f_0
treatments	92.9333	$4 = (a-1)$	23.2333	0.1532
Error	1516	10	151.6	
total	1608.9333	$14 = N-1$		

thus, $f_0 = 0.1532 \notin$ critical Region, thus we do not reject H_0 at 1% significant level

---

(Level)

If the number of observation under each treatment is equal we say the design is balanced

Balanced Design $n_1 = n_2 = \cdots = n_a = n$

observation : Choosing a balanced design has two important advantages:

- ANOVA is relatively insensitive to small departures from the assumption of equality of variances if the sample size are equal.
- the power of the test is maximized if the samples are of the same size.

Interval Estimation for the mean response of each treatment

- The ANOVA is a powerful procedure for test the homogeneity of a set of means. However, if we reject the null hypothesis and accept the stated alternative that the means are not all equal we still do not know which of the population means are equal and which are different. We can analyse how the treatments differers.

Inferences for a single mean value: μ_i.

punctual estimator $\hat{\mu}_i = \bar{Y}_{i\cdot}$

IF $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ then

$Y_{ij} \sim N(\mu_i, \sigma^2)$ and

$\bar{Y}_{i\cdot} = \left(\sum_{i=1}^{n_i} \frac{Y_{ij}}{n_i} \right) \sim N\left(\mu_i, \frac{\sigma^2}{n_i}\right)$, so

$T = \dfrac{\bar{Y}_{i\cdot} - \mu_i}{\sqrt{\frac{\sigma^2}{n_i}}} \sim N(0,1)$, but σ^2 is unknown and $\hat{\sigma}^2 = MSE$

So, we can obtain the following:

Pivotal Quantity

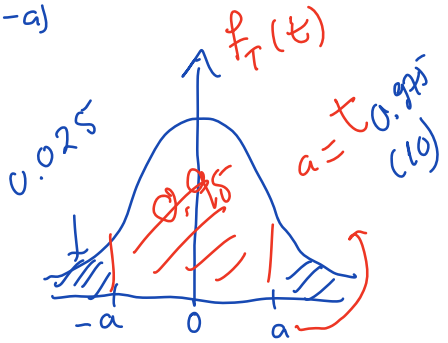
$$T = \frac{\bar{Y}_{i\cdot} - \mu_i}{\sqrt{\frac{MSE}{n_i}}} \sim t_{(N-a)}$$

$$CI_{(1-\alpha)\times 100\%}(\mu_i) = \left[\bar{Y}_{i\cdot} \pm t_{1-\alpha/2}(N-a) \sqrt{\frac{MSE}{n_i}} \right]$$

where $t_{1-\alpha/2}(N-a) = F^{-1}_{t_{(N-a)}}(1-\alpha/2)$

Example: obtain $CI_{95\%}(\mu_1)$

Pivotal Quantity: $T = \frac{\bar{Y}_{1\cdot} - \mu_1}{\sqrt{\frac{MSE}{n_1}}} \sim t_{(10)}$



Choosing the symmetrical C.I.

$$P(-a \le T \le a) = 0.95 \iff P\left(-t_{0.975(10)} \le \frac{\bar{Y}_{1\cdot} - \mu_1}{\sqrt{\frac{MSE}{n_1}}} \le t_{0.975(10)}\right) = 0.95$$

$$\iff P\left(-t_{0.975(10)}\sqrt{\frac{MSE}{n_1}} \le \bar{Y}_{1\cdot} - \mu_1 \le t_{0.975}\sqrt{\frac{MSE}{n_1}}\right) = 0.95 \iff$$

$$\iff P\left(-\bar{Y}_{1\cdot} - t_{0.975(10)}\sqrt{\frac{MSE}{n_1}} \le -\mu_1 \le -\bar{Y}_{1\cdot} + t_{0.975}\sqrt{\frac{MSE}{n_1}}\right) = 0.95$$

$\iff$ (multiply by (-1))

$$P\left(\underbrace{\bar{Y}_{1\cdot}}_{\text{R.V}} - t_{0.975}\sqrt{\underbrace{\frac{MSE}{n_1}}_{\text{R.V.}}} \le \mu_1 \le \underbrace{\bar{Y}_{1\cdot}}_{\text{R.V}} + t_{0.975}\sqrt{\underbrace{\frac{MSE}{n_1}}_{\text{R.V.}}}\right) = 0.95$$

Random confidence Interval

Conclusion :

$y_{1\cdot} = 94$; $\bar{y}_{1\cdot} = \frac{94}{3}$

$MSE = \hat{\sigma}^2 = 151.6$; $t_{0.975(10)} \simeq 2.228$

$$CI_{95\%}(\mu_1) = \left[\frac{94}{3} \pm 2.228 \times \sqrt{\frac{151.6}{3}} \right] \simeq [15.50 ; 47.17]$$

$$CI_{95\%}(\mu_2) = \left[\frac{98}{3} \pm 2.228 \sqrt{\frac{151.6}{3}} \right] \simeq [16.83 ; 48.51]$$

$$CI_{95\%}(\mu_3) = \left[\frac{93}{3} \pm 2.228 \sqrt{\frac{151.6}{3}} \right] \simeq [15.16 ; 46.84]$$

$$CI_{95\%}(\mu_4) = \left[\frac{83}{3} \pm 2.228 \sqrt{\frac{151.6}{3}} \right] \simeq [11.83 ; 43.51]$$

$$CI_{95\%}(\mu_5) = \left[\frac{78}{3} \pm 2.228 \sqrt{\frac{151.6}{3}} \right] \simeq [10.16 ; 41.84]$$

$\mu_1 - \mu_2$; $\mu_2 - \mu_3$ $\mu_4 - \mu_5$

$\mu_1 - \mu_3$ $\mu_2 - \mu_4$

$\mu_1 - \mu_4$ $\mu_2 - \mu_5$

$\mu_1 - \mu_4$ $\mu_3 - \mu_4$

$\mu_3 - \mu_4$

$$\binom{5}{2} = \frac{5!}{2!\,3!}$$

Paired Comparison-inference for $(\mu_i - \mu_j)$

$\bar{Y}_{i\cdot} \perp\!\!\!\perp \bar{Y}_{j\cdot}$

From

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \overset{\text{indep}}{\sim} N(\mu_i - \mu_j, \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_j}),$$

we have the pivotal variable:

$$\boxed{\frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - (\mu_i - \mu_j)}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t_{(N-a)}}$$

- The $(1-\alpha) \times 100\%$ confidence interval for $(\mu_i - \mu_j)$ is given by:

$$CI_{(1-\alpha)\times 100\%}(\mu_i - \mu_j) = \left((\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}) \pm t_{1-\frac{\alpha}{2}(N-a)} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \right).$$

Example: $CI_{95\%}(\mu_1 - \mu_2) = ?$

$$T = \frac{(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{MSE \times \frac{2}{3}}} \sim t(10)$$

$$CI_{95\%}(\mu_1-\mu_2) = \left[\frac{94-98}{5} \pm 2.228 \times \sqrt{\frac{151.6 \times 2}{5}}\right]$$

$$= [-23.73 ; 21.065]$$

- As a general rule, if the CI does not contain zero, then these two means can considered statistically different (with confidence level $(1-\alpha)$). $\left(\frac{\alpha}{2}\right)$
- With a treatment there are $g = a(a-1)/2$ pairs of means to be compared and we want the overall confidence level for all intervals to be "correct" $(1-\alpha) \times$ % of the times.
- For example, if we construct many 95% confidence intervals, the chance that they all contain the true values of the parameters that they estimate will be lower than 95%. For g independent confidence intervals we have $P(\text{all confidence intervals cover their parameters}) = 0.95^g$.

g	1	2	3	...	10	...	100
0.95^g	0.9500	0.9025	0.8574	...	0.5987	...	0.0059

- One possible correction for the two by two investigate differences is the follow: the par of means μ_i and μ_j are declared significantly different if

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > LSD,$$

where $LSD = z\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ and $z = t_{1-\frac{\alpha}{2g}(N-a)}$ where g is the total number of comparisons under study. This is called Bonferroni correction.

Example: Multiple comparision

$$g = \binom{a}{2} = \binom{5}{2} = \frac{5!}{2!3!} = \frac{5\times4\times3!}{2!3!} = \frac{5\times4}{2}$$

$$= 10$$

$$(1-\alpha) = 0.95 \Rightarrow \alpha = 0.05$$

$$C.I_{95\%}(\mu_1-\mu_2) = \dots$$

Quantile: $t_{1-\frac{0.05}{2\times10}}(10) = t_{0.9975}(10)$ (g)

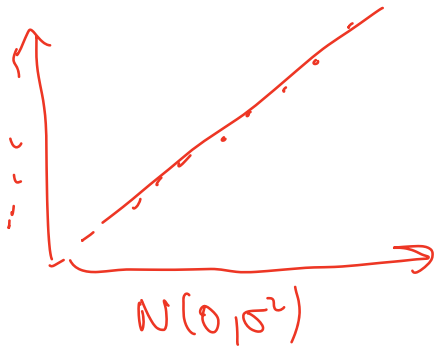
$$= 3.581$$

$$CI_{95\%}(\mu_1-\mu_2) = \left[\frac{94-98}{3} \pm 3.581\sqrt{\frac{MSE\times2}{3}}\right]$$

$$= [-37.334 ; 34.667]$$

with Bonferroni connection

- Residuals are $e_{ij} = y_{ij} - \bar{y}_{i\cdot}$. The residuals analysis and model checking can be performed as we did in multiple regression, e.g., qq-plots, plots of e_{ij} vs $\bar{y}_{i\cdot}$ and e_{ij} vs factor levels.



$\varepsilon_{ij} \sim N(0,\sigma^2)$
i.i.d.