Serik Sagitov, Chalmers Tekniska Högskola , May 24, 2004

Chapter 12. Analysis of variance

Ch 11:	I=2	indep. samples	paired samples
Ch 12:	$I\geq 2$	one-way layout	two-way layout

1. One-way layout

One factor (factor A) with I levels (I treatments)

I independent IID samples $(Y_{i1}, \ldots, Y_{iJ})$, $i = 1, \ldots, I$

H_0: all I treatments have the same effect

H_1: there are systematic differences

Ex 1: seven labs

p. 444: data and boxplots, $I = 7$, $J = 10$

Normal theory model

Normally distributed observation $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$

$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $\Sigma\,\alpha_i = 0$, $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$

obs = overall mean + differential effect + error

Maximum likelihood estimates

$\hat{\mu} = \bar{Y}_{..}$ pooled sample mean $\bar{Y}_{..}$

$\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..}$ sample means $\bar{Y}_{1.}, \ldots, \bar{Y}_{I.}$

Sums of squares: $\mathrm{SS_{TOT}} = \mathrm{SS_A} + \mathrm{SS_E}$

$\mathrm{SS_{TOT}} = \Sigma\,\Sigma(Y_{ij} - \bar{Y}_{..})^2$ total sum of squares

$\mathrm{SS_A} = J\,\Sigma\,\hat{\alpha}_i^2$ between samples (factor A) sum of sq

$\mathrm{SS_E} = \Sigma\,\Sigma\,\hat{\epsilon}_{ij}^2$ within samples (error) sum of squares

residuals $\hat{\epsilon}_{ij} = Y_{ij} - \bar{Y}_{i.}$

Degrees of freedom and mean squares:

$\mathrm{df_A} = I-1$, $\mathrm{MS_A} = \frac{\mathrm{SS_A}}{\mathrm{df_A}}$, $\mathrm{E(MS_A)} = \sigma^2 + \frac{J}{I-1}\sum \alpha_i^2$

$\mathrm{df_E} = I(J-1)$, $\mathrm{MS_E} = \frac{\mathrm{SS_E}}{\mathrm{df_E}}$, $\mathrm{E(MS_E)} = \sigma^2$

total df $= IJ-1$

Pooled sample variance $s_p^2 = \mathrm{MS_E}$
an unbiased estimate of σ^2

F-test

$H_0 : \alpha_1 = \ldots = \alpha_I = 0$ against

$H_1 : \alpha_u \neq \alpha_v$ for some (u, v)

Reject H_0 for large values of $F = \frac{\mathrm{MS_A}}{\mathrm{MS_E}}$

null distribution of F is $F_{I-1,I(J-1)}$

If $Z_i \sim$ N(0,1) indep., then $\frac{(Z_1^2+\ldots+Z_m^2)/m}{(Z_{m+1}^2+\ldots+Z_{m+n}^2)/n} \sim F_{m,n}$

Ex 1: seven labs

normal probability plot of residuals $\hat{\epsilon}_{ij}$, p. 450

Anova-1 table

Source	df	SS	MS	F	P-value
Labs	6	.125	.0210	5.66	.0001
Error	63	.231	.0037		
Total	69	.356			

Multiple comparisons: $\binom{7}{2} = 21$ pairwise comparisons

Lab	1	3	7	2	5	6	4
Mean	4.062	4.003	3.998	3.997	3.957	3.955	3.920

Bonferroni method

Take α as an overall level in k independent tests

if each done at significance level α/k

Proof: given that H_0 is true

number of significant results in k tests $X \sim \text{Bin}(k, \frac{\alpha}{k})$

$\text{P}(X \geq 1|H_0) = 1 - (1 - \frac{\alpha}{k})^k \approx \alpha$

> Warning: $k = \binom{I}{2}$ pairwise comparisons are not independent as required by the Bonferroni method

Simultaneuos $100(1-\alpha)\%$ CI

for $\binom{I}{2}$ pairwise differences $(\alpha_u - \alpha_v)$

$(\bar{Y}_{u.} - \bar{Y}_{v.}) \pm t_{I(J-1)}(\frac{\alpha}{I(I-1)}) \cdot s_p\sqrt{\frac{2}{J}}$

Flexibility: works for different sample sizes as well

replace $\sqrt{\frac{2}{J}}$ by $\sqrt{\frac{1}{J_u} + \frac{1}{J_v}}$

Ex 1: seven labs

95% CI for one difference $(\alpha_u - \alpha_v)$

$(\bar{Y}_{u.} - \bar{Y}_{v.}) \pm t_{63}(0.025) \cdot \frac{s_p}{\sqrt{5}} = (\bar{Y}_{u.} - \bar{Y}_{v.}) \pm 0.055$

where $t_{63}(0.025) = 2.00$, $s_p = \sqrt{0.0037} = 0.061$

Simultaneuos 95% CI for $(\alpha_u - \alpha_v)$ by Bonferroni method

$(\bar{Y}_{u.} - \bar{Y}_{v.}) \pm t_{63}(\frac{.05}{42}) \cdot \frac{s_p}{\sqrt{5}} = (\bar{Y}_{u.} - \bar{Y}_{v.}) \pm 0.086$

Labs	1–4	1–6	1–5	3–4	7–4	2–4	1–2
Diff	0.142	0.107	0.105	0.083	0.078	0.077	0.065

significant differences bewteen labs (1,4), (1,5), (1,6)

Tukey method

If the sample sizes are equal J, then

$\bar{Y}_{i.} \sim \mathrm{N}(\mu + \alpha_i, \frac{\sigma^2}{J})$ are independent and

$\frac{\sqrt{J}}{s_p} \max_{u,v} |\bar{Y}_{u.} - \bar{Y}_{v.} - (\alpha_u - \alpha_v)| \sim \mathrm{SR}(I, I(J-1))$

Studentized range distribution $\mathrm{SR}(t, \nu)$

t = number of samples, ν = df (the variance estimate)

Table 6, p. A14-A19

$q_{t,\nu}(\alpha) = 100(1-\alpha)\%$-percentile of $\mathrm{SR}(t, \nu)$

$$\text{Simultaneuos CI} = (\bar{Y}_{u.} - \bar{Y}_{v.}) \pm q_{I,I(J-1)}(\alpha) \cdot \frac{s_p}{\sqrt{J}}$$

Ex 1: seven labs

$(\bar{Y}_{u.} - \bar{Y}_{v.}) \pm q_{7,63}(0.05) \cdot \frac{0.061}{\sqrt{10}} = (\bar{Y}_{u.} - \bar{Y}_{v.}) \pm 0.083$

where $q_{7,60}(0.05) = 4.31$

significant differences (1,4), (1,5), (1,6), (3,4)

Kruskal-Wallis test

Nonparametric test for

H_0: all observations are equal in distribution

when ϵ_{ij} are non-normal

Pooled sample size $N = J_1 + \ldots + J_I$

pooled sample ranking: R_{ij} = ranks of Y_{ij}

$\Sigma_{i,j} R_{ij} = \frac{N(N+1)}{2}$, $\bar{R}_{..} = \frac{N+1}{2}$

$$\text{Test statistic } K = \frac{12}{N \cdot (N+1)} \Sigma_{i=1}^{I} J_i \cdot (\bar{R}_{i.} - \tfrac{N+1}{2})^2$$

Reject H_0 for large K

approximate null distribution $K \overset{a}{\sim} \chi^2_{I-1}$

Ex 1: seven labs

Actual measurements replaced by their ranks $1 \div 70$

Labs	1	2	3	4	5	6	7
	70	4	35	6	46	48	38
	63	3	45	7	21	5	50
	53	65	40	13	47	22	52
	64	69	41	20	8	28	58
	59	66	57	16	14	37	68
	54	39	32	26	42	2	1
	43	44	51	17	9	31	15
	61	56	25	11	10	34	23
	67	24	29	27	33	49	60
	55	19	30	12	36	18	62
Means	58.9	38.9	38.5	15.5	26.6	27.4	42.7

$K = 28.17$, df $= 6$, P-value ≈ 0.0001

2. Two-way layout

Two factors: factor A with I levels (levels = rows)
factor B with J levels (levels = columns)

Data $\{Y_{ijk}, 1 \le i \le I, 1 \le j \le J, 1 \le k \le K\}$

$I \cdot J$ cells with K observations per cell

total number of observations $= I \cdot J \cdot K$

Normal theory model

$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}$

grand mean + main effects + interaction

independent random errors $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$

Parameter constraints

$\Sigma\, \alpha_i = 0$, $\mathrm{df}_{\mathrm{A}} = I - 1$ $\quad$ $\Sigma\, \beta_j = 0$, $\mathrm{df}_{\mathrm{B}} = J - 1$

$\Sigma\, \delta_{i1} = 0, \ldots, \Sigma\, \delta_{iJ} = 0$ $\quad$ $\Sigma\, \delta_{1j} = 0, \ldots, \Sigma\, \delta_{Ij} = 0$

$\mathrm{df}_{\mathrm{AB}} = IJ - I - (J-1) = (I-1)(J-1)$

MLE

$\hat{\mu} = \bar{Y}_{...}$ $\quad$ $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}$ $\quad$ $\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}$

$\hat{\delta}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{...} - \hat{\alpha}_i - \hat{\beta}_j$

Ex 2: iron retention

factor A: $I = 2$ different iron forms

factor B: $J = 3$ dosage levels, $K = 18$ obs per cell

Raw data, p. 396: $X_{ijk} = \%$ of iron retained

transformed data $Y_{ijk} = \ln(X_{ijk})$

p. 462-463: boxplots and plots of cell SDs vs cell means

MLE for the transformed data

$\bar{Y}_{...} = 1.92$ $\quad$ $\|\bar{Y}_{ij.}\| = \begin{pmatrix} 1.16 & 1.90 & 2.28 \\ 1.68 & 2.09 & 2.40 \end{pmatrix}$

$\hat{\alpha}_1 = -0.14$, $\hat{\alpha}_2 = 0.14$

$\hat{\beta}_1 = -0.50$, $\hat{\beta}_2 = 0.08$, $\hat{\beta}_3 = 0.42$

$$\|\hat{\delta}_{ij}\| = \begin{pmatrix} -0.12 & 0.04 & 0.08 \\ 0.12 & -0.04 & -0.08 \end{pmatrix}$$

Sums of squares

$\mathrm{SS_{TOT}} = \mathrm{SS_A} + \mathrm{SS_B} + \mathrm{SS_{AB}} + \mathrm{SS_E}$

$\mathrm{SS_{TOT}} = \Sigma_i \Sigma_j \Sigma_k (Y_{ijk} - \bar{Y}_{...})^2$

$\mathrm{SS_A} = JK \, \Sigma_i \, \hat{\alpha}_i^2$

$\mathrm{SS_B} = IK \, \Sigma_j \, \hat{\beta}_j^2$

$\mathrm{SS_{AB}} = K \, \Sigma_i \Sigma_j \, \hat{\delta}_{ij}^2$

$\mathrm{SS_E} = \Sigma_i \Sigma_j \Sigma_k (Y_{ijk} - \bar{Y}_{ij.})^2$, $\mathrm{df_E} = IJ(K-1)$

Mean squares

$\mathrm{MS_A} = \frac{\mathrm{SS}_A}{\mathrm{df_A}}$ $\quad$ $\mathrm{E(MS_A)} = \sigma^2 + \frac{JK}{I-1} \Sigma_i \, \alpha_i^2$

$\mathrm{MS_B} = \frac{\mathrm{SS_B}}{\mathrm{df_B}}$ $\quad$ $\mathrm{E(MS_B)} = \sigma^2 + \frac{IK}{J-1} \Sigma_j \, \beta_j^2$

$\mathrm{MS_{AB}} = \frac{\mathrm{SS_{AB}}}{\mathrm{df_{AB}}}$ $\quad$ $\mathrm{E(MS_{AB})} = \sigma^2 + \frac{K}{(I-1)(J-1)} \Sigma_i \Sigma_j \, \delta_{ij}^2$

$\mathrm{MS_E} = \frac{\mathrm{SS_E}}{\mathrm{df_E}}$ $\quad$ $\mathrm{E(MS_E)} = \sigma^2$

Three F-tests

H_{A}: $\alpha_1 = \ldots = \alpha_I = 0$ $\quad$ $F_{\mathrm{A}} = \frac{\mathrm{MS_A}}{\mathrm{MS_E}} \sim F_{\mathrm{df_A},\mathrm{df_E}}$

H_{B}: $\beta_1 = \ldots = \beta_J = 0$ $\quad$ $F_{\mathrm{B}} = \frac{\mathrm{MS_B}}{\mathrm{MS_E}} \sim F_{\mathrm{df_B},\mathrm{df_E}}$

H_{AB}: all $\delta_{ij} = 0$ $\quad$ $F_{\mathrm{AB}} = \frac{\mathrm{MS_{AB}}}{\mathrm{MS_E}} \sim F_{\mathrm{df_{AB}},\mathrm{df_E}}$

Reject null hypothesis

for large values of the respective test statistic F

Inspect normal probability plot

for residuals $\hat{\epsilon}_{ijk} = Y_{ijk} - \bar{Y}_{ij.}$

Ex 2: iron retention

Anova-2 table for the transformed iron retention data

Source	df	SS	MS	F	P
Iron form	1	2.074	2.074	5.99	0.017
Dosage	2	15.588	7.794	22.53	0.000
Interaction	2	0.810	0.405	1.17	0.315
Error	102	35.296	0.346		
Total	107	53.768			

Significant effect due to iron form

log scale difference $\hat{\alpha}_2 - \hat{\alpha}_1 = \bar{Y}_{2..} - \bar{Y}_{1..} = 0.28$

multiplicative effect of $e^{0.28} = 1.32$ on a linear scale

interaction is not significant

Additive model

If $K = 1$ we cannot estimate interaction

$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$

$\hat{\mu} = \bar{Y}_{..}$ $\quad \hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..}$ $\quad \hat{\beta}_i = \bar{Y}_{.j} - \bar{Y}_{..}$

$\hat{\epsilon}_{ij} = Y_{ij} - \bar{Y}_{..} - \hat{\alpha}_i - \hat{\beta}_i$

Sums of squares

$\mathrm{SS_{TOT}} = \mathrm{SS_A} + \mathrm{SS_B} + \mathrm{SS_E}$

$\mathrm{SS_A} = J \sum_i \hat{\alpha}_i^2$ $\quad \mathrm{df_A} = I - 1$

$\mathrm{SS_B} = I \sum_j \hat{\beta}_J^2$ $\quad \mathrm{df_B} = J - 1$

$\mathrm{SS_E} = \sum_i \sum_j \hat{\epsilon}_{ij}^2$ $\quad \mathrm{df_E} = (I-1)(J-1)$

If $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$, apply F-tests to test H_A and H_B

Randomized block design

Experimental design with I treatments
randomly assigned within each of J blocks

To test H_0: $\alpha_1 = \ldots = \alpha_I = 0$, no treatment effects
use two-way layout ANOVA

The block effect is anticipated and is not of major interest

Block	Treatments	Observation
Homogen. plot of land divided into I subplots	I fertilizers	The yield on the subplot (i, j)
A four-wheel car	4 tire types	tire's life-length
A litter of I animals	I diets	the weight gain

Ex. 3: experiment on itching

Data, p. 467

$I = 7$ treatments to relieve itching

$J = 10$ blocks (male volunteers aged 20-30)

$K = 1$ observation per cell

Y_{ij} = the duration of the itching in seconds

Boxplots and normal prob. plot of residuals, p. 468-469
placebo cell variance: different response to placebo

Anova-2 table:

Source	df	SS	MS	F	P
Drugs	6	53013	8835	2.85	0.018
Subjects	9	103280	11476	3.71	0.001
Error	54	167130	3096		
Total	69	323422			

$$\boxed{\text{Simultaneuos CI} = (\bar{Y}_{u.} - \bar{Y}_{v.}) \pm q_{I,(I-1)(J-1)}(\alpha) \cdot \frac{s_p}{\sqrt{J}}}$$

Tukey's method of multiple comparison reveals

$q_{I,(I-1)(J-1)}(\alpha) \cdot \frac{s_p}{\sqrt{J}} = q_{7,54}(0.05) \cdot \sqrt{\frac{3096}{10}} = 75.8$

only one significant difference: papaverine vs placebo

Treat	2	1	6	7	4	5	3
Mean	208.4	191.0	176.5	167.2	148.0	144.3	118.2

Friedman's test

Nonparametric test, when ϵ_{ij} are non-normal, to test

H_0: no treatment effects

Ranking within the block number j

$(R_{1j}, \ldots, R_{Ij})$ = ranks of $(Y_{1j}, \ldots, Y_{Ij})$

$R_{1j} + \ldots + R_{Ij} = \frac{I(I+1)}{2}$

$\frac{1}{I}(R_{1j} + \ldots + R_{Ij}) = \frac{I+1}{2}$ and $\bar{R}_{..} = \frac{I+1}{2}$

$$\boxed{\begin{array}{c}\text{Test statistic } Q = \frac{12J}{I(I+1)} \Sigma_{i=1}^{I} (\bar{R}_{i.} - \frac{I+1}{2})^2 \\ \text{approximate null distribution } Q \overset{a}{\sim} \chi^2_{I-1}\end{array}}$$

Q is a measure of agreement between J rankings

reject H_0 for large values of Q

Ex. 3: experiment on itching

R_{ij} and $\bar{R}_{i.}$ are given on p. 470

$\frac{I+1}{2} = 4$, $Q = 14.86$, df = 6, P-value ≈ 0.0214