SERIK SAGITOV, Chalmers Tekniska Högskola , March 6, 2014

Chapter 12. Analysis of variance

Chapter 11:	$I = 2$ samples	independent samples	paired samples
Chapter 12:	I samples	one-way layout	two-way layout

1 One-way layout

Single main factor (factor A) with I levels (I treatments). Data consist of I independent IID samples $(Y_{i1}, \ldots, Y_{iJ})$, $i = 1, \ldots, I$. The goal is to test

H_0: all I treatments have the same effect, vs H_1: there are systematic differences.

Example: seven labs with $I = 7$, $J = 10$. P. 444: data and boxplots.

Normal theory model

Normally distributed observations $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$ with equal variances.

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

obs = overall mean + differential effect + noise

Maximum likelihood estimates

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

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

$\mathrm{SS_{TOT}} = \sum\sum(Y_{ij} - \bar{Y}_{..})^2$ total sum of squares with the total df $= IJ - 1$,

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

$\mathrm{SS_E} = \sum\sum\hat{\epsilon}_{ij}^2$ within samples (error) sum of squares, where $\hat{\epsilon}_{ij} = Y_{ij} - \bar{Y}_{i.}$ are residuals.

Degrees of freedom and mean squares:

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

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

Pooled sample variance $s_p^2 = \mathrm{MS_E} = \frac{1}{I(J-1)}\sum_i\sum_j(Y_{ij} - \bar{Y}_{i.})^2$ is an unbiased estimate of σ^2.

F-test for $H_0 : \alpha_1 = \ldots = \alpha_I = 0$ against $H_1 : \alpha_u \neq \alpha_v$ for some (u, v).

Reject H_0 for large values of $F = \frac{\mathrm{MS_A}}{\mathrm{MS_E}}$ with the null distribution $F_{I-1,I(J-1)}$.

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

Example: seven labs. The 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. Which of them are significant?

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, the number of significant results in k tests $X \sim \text{Bin}(k, \frac{\alpha}{k})$. Thus, due to independence, $\text{P}(X \geq 1|H_0) = 1 - (1 - \frac{\alpha}{k})^k \approx \alpha$.

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

Simultaneuos $100(1-\alpha)\%$ CI for $\binom{I}{2}$ pairwise differences $(\alpha_u - \alpha_v)$

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

Flexibility: works for different sample sizes as well after replacing $\sqrt{\frac{2}{J}}$ by $\sqrt{\frac{1}{J_u} + \frac{1}{J_v}}$.

Example: seven labs
95% CI for one difference $(\alpha_u - \alpha_v)$ using $s_p = \sqrt{0.0037} = 0.061$ and $t_{63}(0.025) = 2.00$:

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

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

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

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 are between labs (1,4), (1,5), (1,6).

Tukey method
If I independent samples $(Y_{i1}, \ldots, Y_{iJ})$ taken from $\text{N}(\mu_i, \sigma^2)$ have the same size J, then the sample means $\bar{Y}_{i.} \sim \text{N}(\mu_i, \frac{\sigma^2}{J})$ are independent and

$$\tfrac{\sqrt{J}}{s_p} \max_{u,v} |\bar{Y}_{u.} - \bar{Y}_{v.} - (\mu_u - \mu_v)| \sim \text{SR}(I, I(J-1))$$

Studentized range distribution $\text{SR}(t, \nu)$ has two parameters: t = number of samples, ν is the number of degrees of freedom used in the variance estimate s_p^2. Table 6, p. A14–19 gives

$$q_{t,\nu}(\alpha) = 100(1-\alpha)\%\text{-percentiles of } \text{SR}(t, \nu).$$

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

Example: seven labs. Using $q_{7,60}(0.05) = 4.31$ we find four significant pairwise differences: (1,4), (1,5), (1,6), (3,4), since $(\bar{Y}_{u.} - \bar{Y}_{v.}) \pm q_{7,63}(0.05) \cdot \frac{0.061}{\sqrt{10}} = (\bar{Y}_{u.} - \bar{Y}_{v.}) \pm 0.083$.

Kruskal-Wallis test
Nonparametric test for H_0: all observations are equal in distribution, no treatment effects. No assumption of normality.
Pooled sample size $N = J_1 + \ldots + J_I$. Pooled sample ranking: R_{ij} = ranks of Y_{ij} with $\sum_{i,j} R_{ij} = \frac{N(N+1)}{2}$ and $\bar{R}_{..} = \frac{N+1}{2}$.

Kruskal-Wallis test statistic $K = \frac{12}{N \cdot (N+1)} \sum_{i=1}^{I} J_i \cdot (\bar{R}_{i.} - \frac{N+1}{2})^2$

Reject H_0 for large K using the approximate null distribution $K \overset{a}{\sim} \chi^2_{I-1}$.

Example: seven labs
Actual measurements replaced by their ranks $1 \div 70$. With the observed test statistic $K = 28.17$ and df $= 6$, we get a P-value ≈ 0.0001.

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

2 Two-way layout

Two factors: factor A with I levels (levels = rows) and factor B with J levels (levels = columns). Data $\{Y_{ijk}, 1 \le i \le I, 1 \le j \le J, 1 \le k \le K\}$ is represented by $I \cdot J$ cells with K observations per cell. The total number of observations is $= I \cdot J \cdot K$.

Normal theory model

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

Grand mean + main effect for factor A +main effect for factor B + interaction + noise

Key assumption: $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$ are independent and have the same variance

Parameter constraints and degrees of freedom

$\sum \alpha_i = 0$, $\mathrm{df_A} = I - 1$; $\sum \beta_j = 0$, $\mathrm{df_B} = J - 1$;

$\sum \delta_{i1} = 0, \ldots, \sum \delta_{iJ} = 0$, $\sum \delta_{1j} = 0, \ldots, \sum \delta_{Ij} = 0$, $\mathrm{df_{AB}} = IJ - I - (J-1) = (I-1)(J-1)$

MLE

$\hat\mu = \bar Y_{...}$ $\quad \hat\alpha_i = \bar Y_{i..} - \bar Y_{...}$ $\quad \hat\beta_j = \bar Y_{.j.} - \bar Y_{...}$ $\quad \hat\delta_{ij} = \bar Y_{ij.} - \bar Y_{...} - \hat\alpha_i - \hat\beta_j = \bar Y_{ij.} - \bar Y_{i..} - \bar Y_{.j.} + \bar Y_{...}$

Example: iron retention
Raw data, p. 396, X_{ijk} = percentage of iron retained is not normally distributed. Factor A: different iron forms $I = 2$, factor B: dosage levels $J = 3$, $K = 18$ observations per cell.
Transformed data $Y_{ijk} = \ln(X_{ijk})$, p. 462-463: boxplots and plots of cell standard deviations vs cell means. MLEs for the transformed data

$$\bar Y_{...} = 1.92,\ \|\bar Y_{ij.}\| = \begin{pmatrix} 1.16 & 1.90 & 2.28 \\ 1.68 & 2.09 & 2.40 \end{pmatrix},\ \hat\alpha_1 = -0.14,\ \hat\alpha_2 = 0.14$$

$$\hat\beta_1 = -0.50,\ \hat\beta_2 = 0.08,\ \hat\beta_3 = 0.42,\ \|\hat\delta_{ij}\| = \begin{pmatrix} -0.12 & 0.04 & 0.08 \\ 0.12 & -0.04 & -0.08 \end{pmatrix}$$

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

$\mathrm{SS_{TOT}} = \sum_i \sum_j \sum_k (Y_{ijk} - \bar Y_{...})^2$ $\qquad \mathrm{df_{TOT}} = IJK - 1$

$\mathrm{SS_A} = JK \sum_i (\bar Y_{i..} - \bar Y_{...})^2$ $\qquad \mathrm{df_A} = (I - 1)$

$\mathrm{SS_B} = IK \sum_j (\bar Y_{.j.} - \bar Y_{...})^2$ $\qquad \mathrm{df_B} = (J - 1)$

$SS_{AB} = K\sum_i\sum_j(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$ $\quad df_{AB} = (I-1)(J-1)$

$SS_E = \sum_i\sum_j\sum_k(Y_{ijk} - \bar{Y}_{ij.})^2$ $\quad df_E = IJ(K-1)$

Mean squares

$MS_A = \frac{SS_A}{df_A}$ $\quad E(MS_A) = \sigma^2 + \frac{JK}{I-1}\sum_i \alpha_i^2$

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

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

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

Pooled sample variance $s_p^2 = MS_E = \frac{1}{IJ(K-1)}\sum_i\sum_j\sum_k(Y_{ijk} - \bar{Y}_{ij.})^2$ unbiased estimate of σ^2.

Three F-tests

Null hypothesis	Property	Test statistics and null distribution
H_A: $\alpha_1 = \ldots = \alpha_I = 0$	$E(MS_A) = \sigma^2$	$F_A = \frac{MS_A}{MS_E} \sim F_{df_A, df_E}$
H_B: $\beta_1 = \ldots = \beta_J = 0$	$E(MS_B) = \sigma^2$	$F_B = \frac{MS_B}{MS_E} \sim F_{df_B, df_E}$
H_{AB}: all $\delta_{ij} = 0$	$E(MS_{AB}) = \sigma^2$	$F_{AB} = \frac{MS_{AB}}{MS_E} \sim F_{df_{AB}, df_E}$

Reject null hypothesis for large values of the respective test statistic F.
Inspect normal probability plot for residuals $\hat{\epsilon}_{ijk} = Y_{ijk} - \bar{Y}_{ij.}$.

Tukey's simultaneuos CI for $\mu_u - \mu_v = (\bar{Y}_{u..} - \bar{Y}_{v..}) \pm q_{I,\nu}(\alpha)\cdot\frac{s_p}{\sqrt{J}}$ with $\nu = IJ(K-1)$.

Example: 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. Estimated log scale difference $\hat{\alpha}_2 - \hat{\alpha}_1 = \bar{Y}_{2..} - \bar{Y}_{1..} = 0.28$ yields the 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. The additive model without interaction $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$.
Maximum likelihood estimates

$\hat{\mu} = \bar{Y}_{..}$, $\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..}$, $\hat{\beta}_i = \bar{Y}_{.j} - \bar{Y}_{..}$, $\hat{\epsilon}_{ij} = Y_{ij} - \bar{Y}_{..} - \hat{\alpha}_i - \hat{\beta}_i = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$

Sums of squares $SS_{TOT} = SS_A + SS_B + SS_E$

$SS_{TOT} = \sum_i\sum_j(\bar{Y}_{ij} - \bar{Y}_{..})^2$ $\quad df_{TOT} = IJ - 1$

$SS_A = J\sum_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$ $\quad df_A = I - 1$ $\quad MS_A = \frac{SS_A}{df_A}$ $\quad F_A = \frac{MS_A}{MS_E} \sim F_{df_A, df_E}$

$SS_B = I\sum_j(\bar{Y}_{.j} - \bar{Y}_{..})^2$ $\quad df_B = J - 1$ $\quad MS_B = \frac{SS_B}{df_B}$ $\quad F_B = \frac{MS_B}{MS_E} \sim F_{df_B, df_E}$

$SS_E = \sum_i\sum_j(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$ $\quad df_E = (I-1)(J-1)$

Pooled sample variance $s_p^2 = MS_E = \frac{1}{(I-1)(J-1)}\sum_i\sum_j(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$ unbiased estimate of σ^2

Tukey's simultaneuos CI for $\mu_u - \mu_v = (\bar{Y}_{u.} - \bar{Y}_{v.}) \pm q_{I,\nu}(\alpha)\cdot\frac{s_p}{\sqrt{J}}$ with $\nu = (I-1)(J-1)$.

3 Randomized block design

Blocking is used to remove the effects of a few of the most important nuisance variables. Randomization is then used to reduce the contaminating effects of the remaining nuisance variables.

"Block what you can, randomize what you cannot."

Experimental design with I treatments randomly assigned within each of J blocks.
To test H_0: $\alpha_1 = \ldots = \alpha_I = 0$ of no treatment effects use the two-way layout ANOVA.
The block effect is anticipated and is not of major interest. Examples:

Block	Treatments	Observation
A homogeneous plot of land divided into I subplots	I fertilizers each applied to a randomly chosen subplot	The yield on the subplot (i, j)
A four-wheel car	4 types of tires tested on the same car	tire's life-length
A litter of I animals	I diets randomly assigned to I sinlings	the weight gain

Example: experiment on itching
Data presented on p. 467: Y_{ij} = the duration of the itching in seconds,
$I = 7$ treatments to relieve itching,
$J = 10$ blocks (male volunteers aged 20-30),
$K = 1$ observation per cell.
Boxplots and a normal probability plot of residuals, p. 468-469. Notice 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			

Tukey's method of multiple comparison $q_{I,(I-1)(J-1)}(\alpha) \cdot \frac{s_p}{\sqrt{J}} = q_{7,54}(0.05) \cdot \sqrt{\frac{3096}{10}} = 75.8$ reveals only one significant difference: papaverine vs placebo with $208.4 - 118.2 = 90.2 > 75.8$.

Treatment	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 j-th block: $(R_{1j}, \ldots, R_{Ij})$ = ranks of $(Y_{1j}, \ldots, Y_{Ij})$ so that $R_{1j} + \ldots + R_{Ij} = \frac{I(I+1)}{2}$, implying $\frac{1}{I}(R_{1j} + \ldots + R_{Ij}) = \frac{I+1}{2}$ and $\bar{R}_{..} = \frac{I+1}{2}$.

Test statistic $Q = \frac{12J}{I(I+1)} \sum_{i=1}^{I} (\bar{R}_{i.} - \frac{I+1}{2})^2$ has an approximate null distribution $Q \stackrel{a}{\sim} \chi^2_{I-1}$.

Since Q is a measure of agreement between J rankings, we reject H_0 for large values of Q.

Example: experiment on itching
From the values R_{ij} and $\bar{R}_{i.}$ are given on p. 470 we find $\frac{I+1}{2} = 4$, $Q = 14.86$, df = 6, P-value ≈ 0.0214.