Lecture 7: Fixed and Random Effects

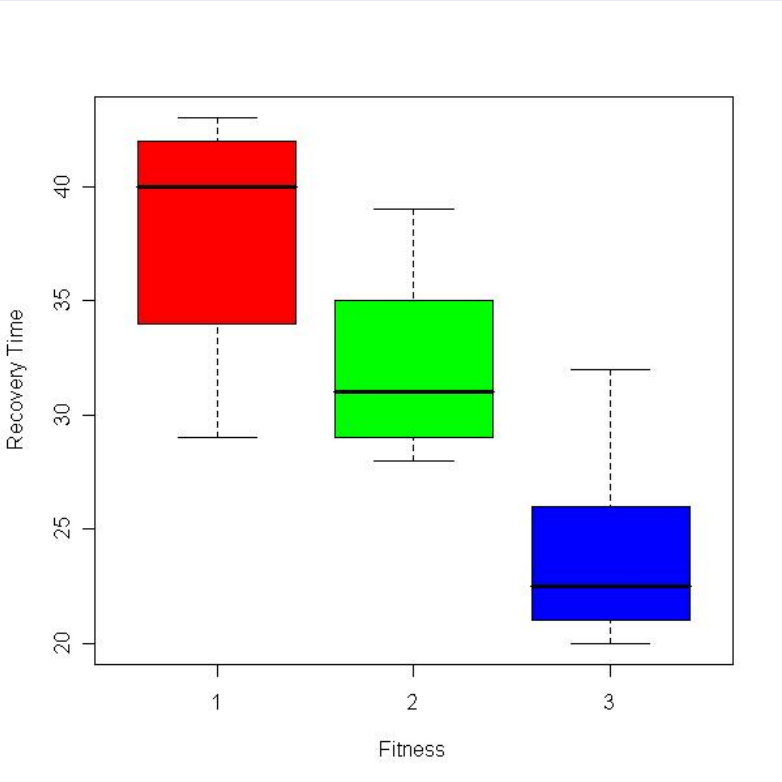
Nancy R. Zhang

Statistics 203, Stanford University

January 28, 2010

One-way ANOVA

Example: rehab surgery



How does prior fitness affect recovery from surgery?
Observations: 24 subjects' recovery time.

Three fitness levels: below average (8), average (10), above average (6).

Can be viewed in two different ways:

1. Extension of "two-sample" t-test to more than two groups.
2. Extension of simple linear regression to case where X is qualitative.

One-way ANOVA model

$$Y_{ij}, 1 \leq i \leq r, 1 \leq j \leq n_i : \quad r \text{ groups and } n_i \text{ samples in } i\text{-th group}$$

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2).$$

Constraint $\sum_{i=1}^{r} \alpha_i = 0$ needed for "identifiability"

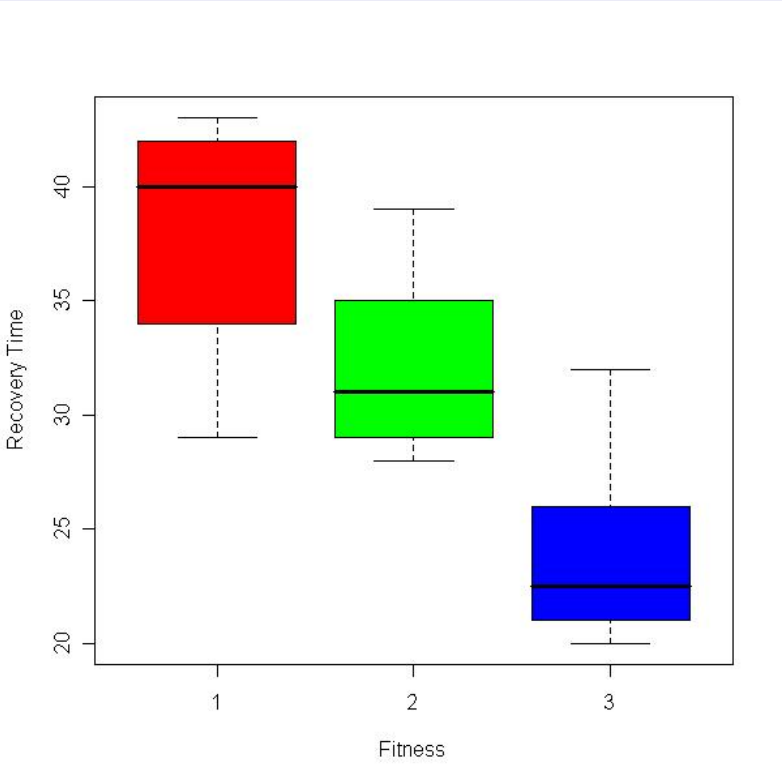
This is equivalent to:

$$Y_{ij} = \mu + \alpha_1 I_{fitness=1} + \alpha_2 I_{fitness=2} + \alpha_3 I_{fitness=3} + \varepsilon_{ij},$$

Can always phrase an ANOVA problem as a multiple linear regression problem using indicator variables.

$$Y_{ij} = \mu + \alpha_1 I_{fitness=1} + \alpha_2 I_{fitness=2} + \alpha_3 I_{fitness=3} + \varepsilon_{ij},$$

Example: rehab surgery



Design matrix:

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \vdots & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \vdots \\ 0 & 0 & 1 \end{pmatrix}$$

F-test for one-way ANOVA

Source	*SS*	*df*	MS	*E*(*MS*)
Treatments	$SSTR = \sum_{i=1}^{r} n_i \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$	$r-1$	SSTR/(r-1)	$\sigma^2 + \frac{\sum_{i=1}^{r} n_i \alpha_i^2}{r-1}$
Error	$SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i.})^2$	$n-r$	SSE/(n-r)	σ^2

$$H_0 : \alpha_1 = \cdots = \alpha_r = 0$$

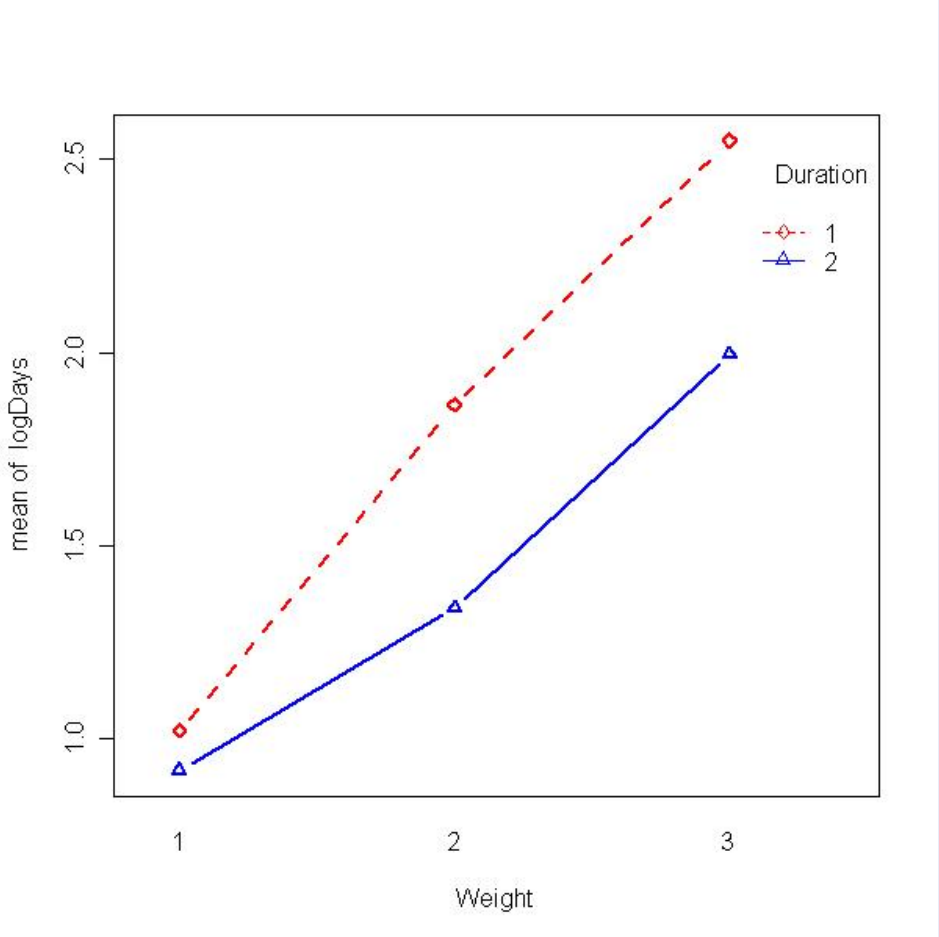
Under H_0,

$$F = \frac{MSTR}{MSE} = \frac{\frac{SSTR}{df_{TR}}}{\frac{SSE}{df_E}} \sim F_{df_{TR}, df_E}$$

Reject H_0 at level α if $F > F_{1-\alpha, df_{TR}, df_E}$.

Two-Way ANOVA

Example: rehab time from kidney failure



Recovery time depends on weight gain between treatments and duration of treatment.

Two levels of duration, three levels of weight gain.

Two-way ANOVA model: observations:

$$(Y_{ijk}), 1 \leq i \leq a, 1 \leq j \leq b, 1 \leq k \leq n_{ij}$$

a groups in first grouping variable (A),
b groups in second grouping variable (B),
n_{ij} samples in (i, j)-"cell".

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim N(0, \sigma^2).$$

In kidney example, $a = 3$ (weight gain), $b = 2$ (duration of treatment), $n_{ij} = 10$ for all (i, j).

Using indicator variables, this is still a multiple regression problem.

Two-way ANOVA: main questions of interest

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim N(0, \sigma^2).$$

Constraints needed for identifiability

- $\sum_{i=1}^{a} \alpha_i = 0$
- $\sum_{j=1}^{b} \beta_j = 0$
- $\sum_{j=1}^{b} (\alpha\beta)_{ij} = 0, 1 \leq i \leq a$
- $\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, 1 \leq j \leq b.$

- Are there main effects for the grouping variables?

$$H_0 : \alpha_1 = \cdots = \alpha_a = 0, \qquad H_0 : \beta_1 = \cdots = \beta_b = 0.$$

- Are there interaction effects?

$$H_0 : (\alpha\beta)_{ij} = 0, 1 \leq i \leq a, 1 \leq j \leq b.$$

Decomposition of Variance

$$SST = SSA + SSB + SSAB + SSE$$

Term	$SS\ (n_{ij} = n)$
A	$SSA = nb\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2$
B	$SSB = na\sum_{j=1}^{b}\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2$
AB	$SSAB = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)^2$
Error	$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2$

Two-way ANOVA table ($n_{ij} = n$)

SS	df	$E(MS)$
SSA	$a-1$	$\sigma^2 + nb\frac{\sum_{i=1}^{a}\alpha_i^2}{a-1}$
SSB	$b-1$	$\sigma^2 + na\frac{\sum_{j=1}^{b}\beta_j^2}{b-1}$
$SSAB$	$(a-1)(b-1)$	$\sigma^2 + n\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2}{(a-1)(b-1)}$
SSE	$(n-1)ab$	σ^2

F-tests for two-way ANOVA

SS	df	$E(MS)$
SSA	$a-1$	$\sigma^2 + nb\frac{\sum_{i=1}^{a}\alpha_i^2}{a-1}$
SSB	$b-1$	$\sigma^2 + na\frac{\sum_{j=1}^{b}\beta_j^2}{b-1}$
$SSAB$	$(a-1)(b-1)$	$\sigma^2 + n\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2}{(a-1)(b-1)}$
SSE	$(n-1)ab$	σ^2

$$MS = SS/df$$

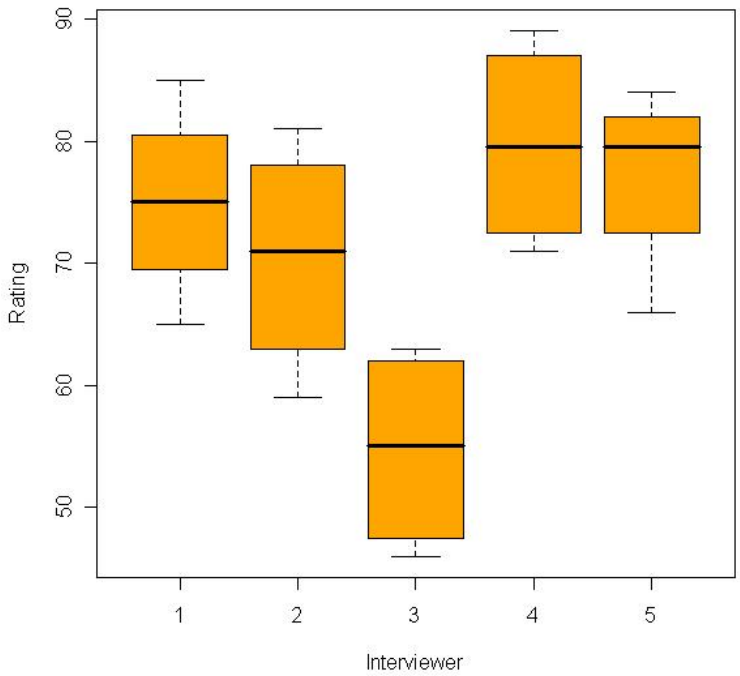
F-tests:

$$F_{AB} = MSAB/MSE \sim F((a-1)(b-1),(n-1)ab)$$

$$F_A = MSA/MSE \sim F(a-1,(n-1)ab)$$

$$F_B = MSB/MSE \sim F(b-1,(n-1)ab)$$

Example



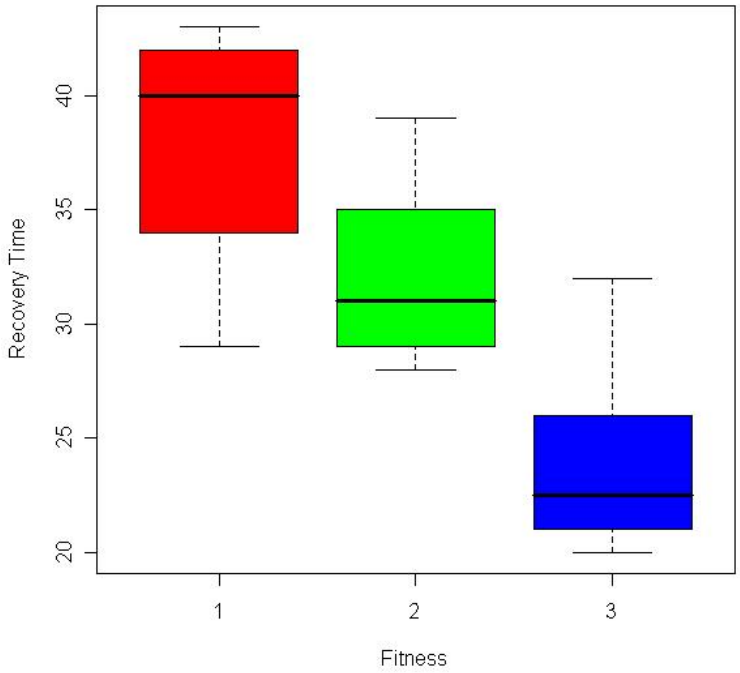
Setting: Personnel management in a large enterprise.

Question: Does the interviewer have an effect on the rating of job candidates?

Data: 5 interviewers selected at random, each interviews 4 candidates selected at random.

What is different about this data set?

Compare to previous cases

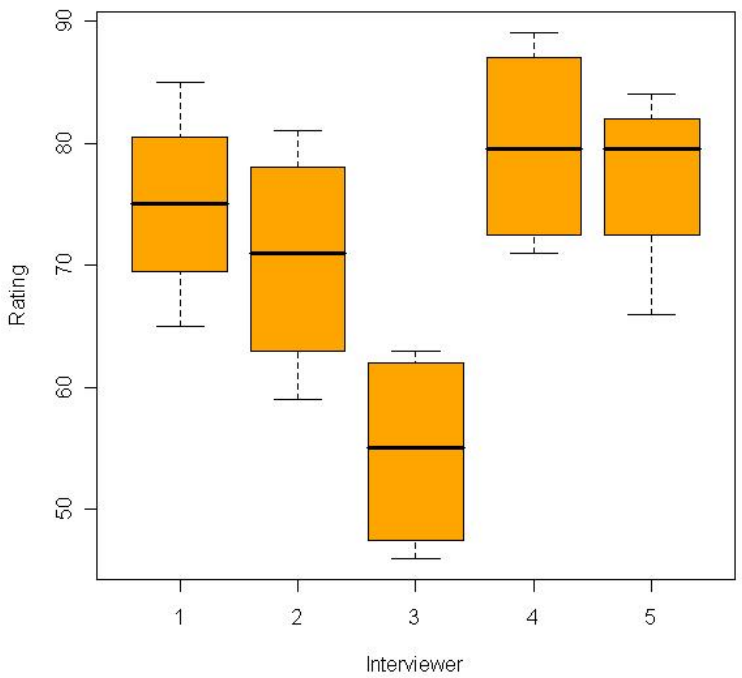


How does prior fitness affect recovery from surgery? Observations: 24 subjects' recovery time.

Three fitness levels: below average (8), average (10), above average (6).

Here, fitness level is of intrinsic interest. They are not random.

Example



Setting: Personnel management in a large enterprise.

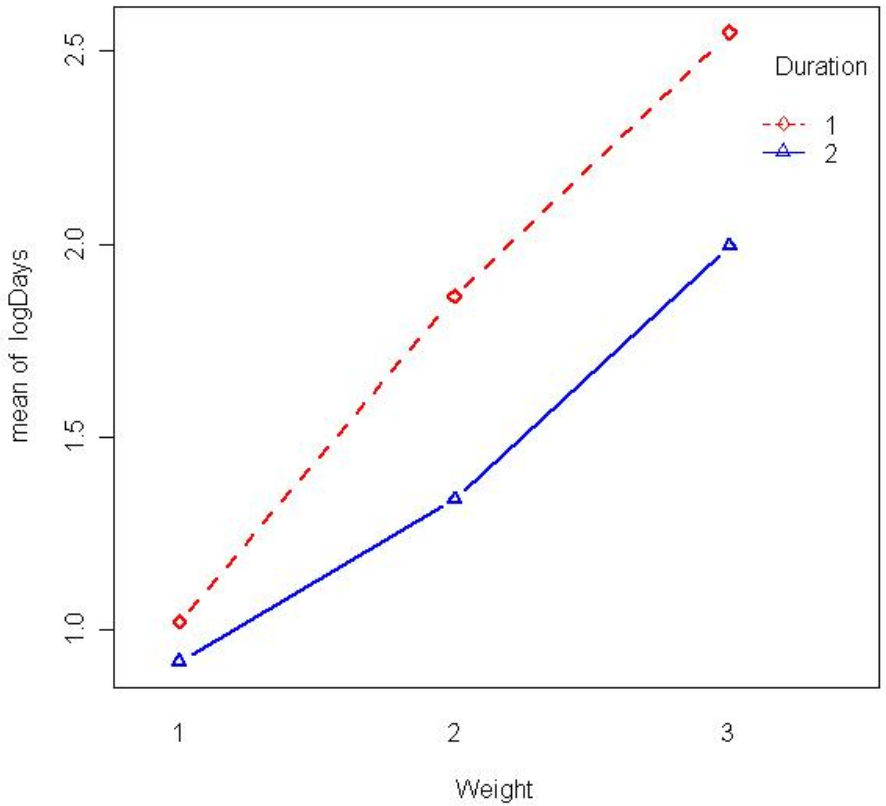
Question: Does the interviewer have an effect on the rating of job candidates?

Data: 5 interviewers selected at random, each interviews 4 candidates selected at random.

The interviewers are *random draws* from a larger population.

We are interested in the larger population and not these 5 specific interviewers.

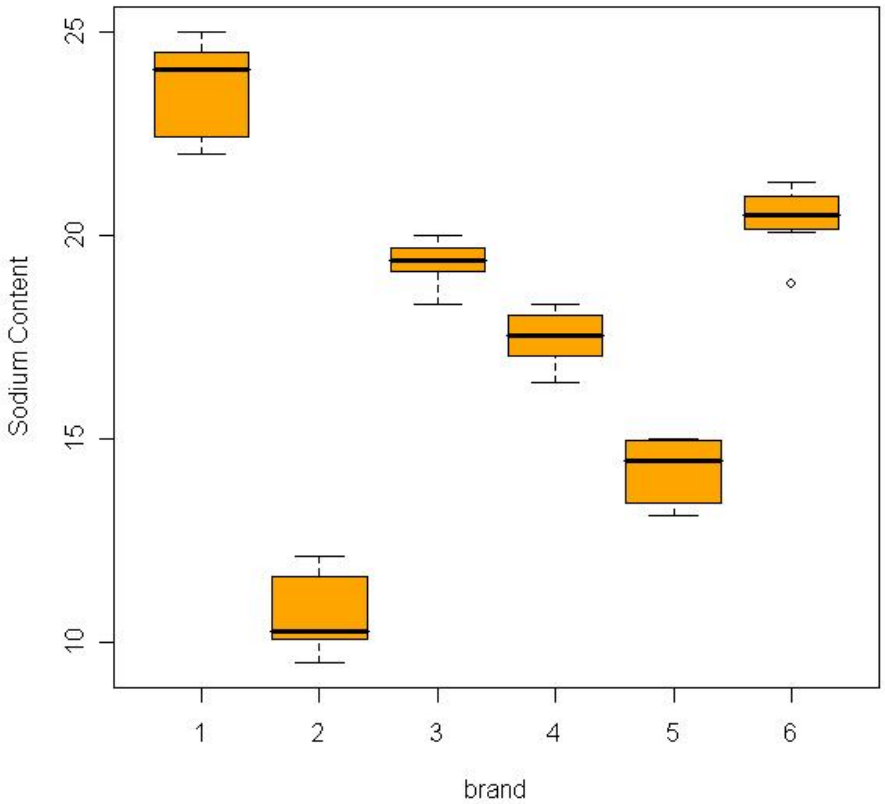
Another Example



Recovery time depends on weight gain between treatments and duration of treatment.

Two levels of duration, three levels of weight gain.

Another Example



How does the sodium in beer differ between brands?

6 randomly chosen brands,
8 bottles tested per brand.

Random Effects Model

Assuming that cell-sizes are the same, i.e. equal observations for each "subject" (brand of beer).

$$Y_{ij} \sim \mu_. + \alpha_i + \varepsilon_{ij}, \quad 1 \leq i \leq r, 1 \leq j \leq n$$

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

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

Parameters:

- μ is the population mean;
- σ^2 is the measurement variance;
- σ_α^2 is the population variance of effect (i.e. variation in sodium content of beer).

Decomposition of Variance and Covariance

$$Var(Y_{ij}) = \sigma_\alpha^2 + \sigma^2$$

But only one parameter in mean function:

$$E(Y_{ij}) = \mu.$$

- The observations are no longer independent:

$$\mathrm{Cov}(Y_{ij}, Y_{i'j'}) = \begin{cases} \sigma_\alpha^2 + \sigma^2, & i = i', \ j = j'; \\ \sigma_\alpha^2, & i = i', \ j \neq j'; \\ 0, & i \neq i', \ j \neq j'. \end{cases}$$

- Random effects models are also called “variance components” models.

When cell sizes are the same (balanced),

$$\widehat{\mu}_{\cdot} = \overline{Y}_{\cdot\cdot} = \frac{1}{nr}\sum_{i,j} Y_{ij\cdot}$$

This also changes estimates of σ^2 – see ANOVA table below. We might guess that $df = nr - 1$ and

$$\widehat{\sigma}^2 = \frac{1}{nr-1}\sum_{i,j}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2.$$

This is *not* the case.

Source	*SS*	*df*	*E*(*MS*)
Treatments	$SSTR = \sum_{i=1}^{r} n\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\sigma^2 + n\sigma_\alpha^2$
Error	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\cdot})^2$	$(n-1)r$	σ^2

One way ANOVA: r groups, n observations in each group.

- Fixed effect model:

Source	SS	df	$E(MS)$
Treatments	$SSTR = \sum_{i=1}^{r} n\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\sigma^2 + n\frac{\sum_{i=1}^{r}\alpha_i^2}{r-1}$
Error	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\cdot})^2$	$(n-1)r$	σ^2

- Random effect model:

Source	SS	df	$E(MS)$
Treatments	$SSTR = \sum_{i=1}^{r} n\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\sigma^2 + n\sigma_\alpha^2$
Error	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\cdot})^2$	$(n-1)r$	σ^2

Inference for population mean: $\mu_.$

- Easy to check that

$$E(\overline{Y}_{..}) = \mu_.$$

$$\text{Var}(\overline{Y}_{..}) = \frac{n\sigma_\alpha^2 + \sigma^2}{rn}.$$

- To come up with a t statistic that we can use for test, CIs, we need to find an estimate of $\text{Var}(\overline{Y}_{..})$. ANOVA table says $E(MSTR) = n\sigma_\alpha^2 + \sigma^2$.
- Therefore,

$$\frac{\overline{Y}_{..} - \mu_.}{\sqrt{\frac{SSTR}{(r-1)rn}}} \sim t_{r-1}$$

Inference for population mean: $\mu_.$

$$\frac{\overline{Y}_{..} - \mu_.}{\sqrt{\frac{SSTR}{(r-1)rn}}} \sim t_{r-1}$$

- Why $r-1$ degrees of freedom? Imagine we could record an infinite number of observations for each group, so that $\overline{Y}_{i.} \to \mu_i$, or that $\sigma^2 = 0$.
- To learn anything about $\mu_.$ we still only have r observations $(\mu_1, \ldots, \mu_r)$.
- Sampling more within an individual cannot narrow the CI for $\mu_.$.

One-way ANOVA (random)

Source	SS	df	$E(MS)$
Treatments	$SSTR = \sum_{i=1}^{r} n\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\sigma^2 + n\sigma_\alpha^2$
Error	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\cdot})^2$	$(n-1)r$	σ^2

- Only change here is the expectation of $MSTR$ which reflects randomness of α_i's.
- ANOVA table is still useful to setup tests: the same F statistics for fixed effect models will work here.
- Test for random effect: $H_0 : \sigma_\alpha^2 = 0$ based on

$$F = \frac{MSTR}{MSE} \sim F_{r-1,(n-1)r} \qquad \text{under } H_0.$$

Estimating σ_α^2

- From the ANOVA table

$$\sigma_\alpha^2 = \frac{E(SSTR/(r-1)) - E(SSE/((n-1)r))}{n}.$$

- Natural estimate:

$$S_\alpha^2 = \frac{SSTR/(r-1) - SSE/((n-1)r)}{n}$$

- Problem: this estimate can be negative. If it is, set to 0.

Two-way ANOVA (random)

Example: productivity study

- Imagine a study on the productivity of employees in a large manufacturing company.
- Company wants to get an idea of daily productivity, and how it depends on which machine an employee uses.
- Study: take m employees and r machines, having each employee work on each machine for a total of n days.
- As these employees are not *all* employees, and these machines are not *all* machines it makes sense to think of both the effects of machine and employees (and interactions) as random.

Two-way ANOVA (random)

Observations, for $1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n$:

$$Y_{ijk} \sim \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij},$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2),$$

$$\alpha_i \sim N(0, \sigma^2_\alpha),$$

$$\beta_j \sim N(0, \sigma^2_\beta),$$

$$(\alpha\beta)_{ij} \sim N(0, \sigma^2_{\alpha\beta}).$$

Sums of squares

Identical to fixed effects model of last class

$$SSA = nb \sum_{i=1}^{r} \left(\overline{Y}_{i..} - \overline{Y}_{...} \right)^2$$

$$SSB = na \sum_{j=1}^{m} \left(\overline{Y}_{.j.} - \overline{Y}_{...} \right)^2$$

$$SSAB = n \sum_{i=1}^{r} \sum_{j=1}^{m} \left(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...} \right)^2$$

ANOVA tables: Two-way (random)

SS	*df*	$E(MS)$
SSA	$r-1$	$\sigma^2 + nm\sigma^2_\alpha + n\sigma^2_{\alpha\beta}$
SSB	$m-1$	$\sigma^2 + nr\sigma^2_\beta + n\sigma^2_{\alpha\beta}$
SSAB	$(m-1)(r-1)$	$\sigma^2 + n\sigma^2_{\alpha\beta}$
SSE	$(n-1)ab$	σ^2

- To test $H_0 : \sigma^2_\alpha = 0$ use *SSA* and *SSAB*.
- To test $H_0 : \sigma^2_{\alpha\beta} = 0$ use *SSAB* and *SSE*.