Statistics 210 – Part 3
Statistical Methods

Hal S. Stern
Department of Statistics
University of California, Irvine
sternh@uci.edu

- Thus far:
  - design of experiments
  - two sample methods
  - one factor ANOVA
  - pairing/blocking
  - simple regression/correlation
  - multiple regression
- To come:
  - regression and ANOVA
  - factorial ANOVA
  - random/mixed effects

Multiple regression

Relationship between regression and ANOVA

- My main objection to the text is that it teaches regression first and then treats ANOVA as a special case of regression
- This is true to some extent (as we shall now see)
- But this doesn't help one understand statistics
- Recall the one factor ANOVA model with r groups

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- We can rewrite as a regression model
  - introduce $r-1$ indicator variables
  - $x_{i1} = 1$ if obs i is in group 1 and 0 o.w.
    $x_{i2} = 1$ if obs i is in group 2 and 0 o.w.
    ....
    $x_{i,r-1} = 1$ if obs i is in group $r-1$ and 0 o.w.
  - omitted group (group r) becomes reference group
  - the regression model is

$$Y_i = \beta_o + \beta_1 x_{i1} + \cdots + \beta_{r-1} x_{i,r-1} + \epsilon_i$$

Multiple regression

Regression and ANOVA (cont'd)

- Recall the regression model for ANOVA is

$$Y_i = \beta_o + \beta_1 x_{i1} + \cdots + \beta_{r-1} x_{i,r-1} + \epsilon_i$$

- Interpretation
  - $E(Y|\mathbf{x}) = \beta_o$ for group r
  - $E(Y|\mathbf{x}) = \beta_o + \beta_j$ for group j $(1 \leq j \leq r-1)$
  - β_j is difference in means between group j and group r
- ANOVA F-test of equal means is the same as the regression F-test of $\beta_1 = \cdots = \beta_{r-1} = 0$
- There is great flexibility in defining the indicator variables

Multiple regression

Regression and ANOVA: coding/contrasts

- Previous definition of indicator variables doesn't yield estimates of group means
- Suppose we want estimates of group means
  - use r indicator variables and omit intercept

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_r x_{ir} + \epsilon_i$$

    so that $\beta_j = E(Y|\text{group } j)$

- Suppose we want factor effects version (grand mean + treatment effects)
  - new definition for $r - 1$ indicators

$$x_{ij} = \begin{cases} 1 & \text{obs } i \text{ in group } j \\ -1 & \text{obs } i \text{ in group } r \\ 0 & \text{otherwise} \end{cases}$$

    for $j = 1, \ldots, r - 1$
  - can show $\beta_o = \sum_{i=1}^{r} \mu_i / r$ and $\beta_j = \mu_j - \beta_o$

Multiple regression

Regression and ANOVA: coding/contrasts

- Can also build in contrasts, e.g., suppose there are three treatments (lo,med,hi)
  - define two indicators as follows

$$x_i = \begin{cases} -1 & \text{obs i in low grp} \\ 0 & \text{obs i in med grp} \\ 1 & \text{obs i in hi grp} \end{cases} \qquad z_i = \begin{cases} -1 & \text{obs i in low grp} \\ 2 & \text{obs i in med grp} \\ -1 & \text{obs i in hi grp} \end{cases}$$

  - fit $Y_i = \beta_o + \beta_x x_i + \beta_z z_i + \epsilon_i$
  - reconcile two models as follows

group	ANOVA mean	regr mean
low	μ_1	$\beta_o - \beta_x - \beta_z$
med	μ_2	$\beta_o + 2\beta_z$
hi	μ_3	$\beta_o + \beta_x - \beta_z$

  - equating last two columns yields

$$\begin{aligned} \beta_o &= (\mu_1 + \mu_2 + \mu_3)/3 \quad \text{(overall mean)} \\ \beta_x &= (\mu_3 - \mu_1)/2 \quad \text{(linear contr)} \\ \beta_z &= (\mu_1 - 2\mu_2 + \mu_3)/6 \quad \text{(quadr. contr)} \end{aligned}$$

Multiple regression

Analysis of covariance (ANCOVA)

- Two views of the analysis of covariance
  - ANOVA view (doing an ANOVA of Y with part of error explained by covariate/predictor x)
  - regression view (compare regr. lines across groups)
- ANOVA view (Ch. 25 of text)
  - model written two equiv ways

$$\begin{aligned} Y_{ij} &= \mu_* + \tau_i + \gamma x_{ij} + \epsilon_{ij} \quad \text{or} \\ &= \mu + \tau_i + \gamma(x_{ij} - \bar{x}) + \epsilon_{ij} \end{aligned}$$

    with second version preferred because μ is interpreted as grand mean
  - ANOVA table looks like

source of variation	degrees of freedom
groups	$r-1$
covariate	1
error	$N-r-1$
total	$N-1$

  - equivalent to $ANOVA$ of Y and $ANOVA$ of X with regression of Y-residuals on X-residuals
  - can compare estimated responses at $\bar{x}$ (known as the adjusted treatment mean – adjusted for x) which is equal to $= \mu + \tau_j$

Multiple regression

Analysis of covariance (cont'd)

- Regression view (Ch 11.4, pg 1018 of Ch 25)
  - can allow separate regression lines of Y on x for different groups
  - define $z_{ij} = 1$ for obs i in group j and 0 o.w. $(j = 1, 2, \ldots, r-1)$
  - could use other codings as well
  - model

$$Y_i = \beta_o + \beta_1 x_i + \sum_{j=1}^{r-1} (\alpha_{oj} z_{ij} + \alpha_{1j} z_{ij} x_{ij}) + \epsilon_i$$

  - interpretation:
    * for group r: $Y_i = \beta_o + \beta_1 x_i + \epsilon_i$
    * for group j $(j < r)$: intercept is $\beta_o + \alpha_{oj}$ and slope is $\beta_1 + \alpha_{1j}$
  - if $\alpha_{1j} = 0$ for all j, then we have r parallel lines (this is known as the ANCOVA model)
  - we can test this hypothesis (F-test)
  - note if $\alpha_{oj} = \alpha_{1j} = 0$ for all j, then all r groups are described by the same regression line

Multiple regression

Analysis of covariance (cont'd)

- Comments on ANCOVA
  - regression view seems quite natural to me here
  - can easily generalize to several covariates
  - can easily generalize to nonlinear models (e.g., polynomial)
  - ANCOVA versus blocking
    - blocking is another way to "control" for x
    - ANCOVA is better than blocking on x if $\text{corr}(x, y)$ is high (say $> .7$)
    - blocking on x is better if $\text{corr}(x, y)$ is moderate
  - ANCOVA versus differencing
    - frequently x is an earlier version of the same variable as Y (e.g., pretest, birthweight)
    - then can do ordinary ANOVA on $Y - x$
    - simpler model - implicitly assumes slope of x is one
  - beware extrapolation - ANCOVA automatically extrapolates, i.e., it will compare two groups with non-overlapping x distns

Multifactor studies

Introduction

- Now consider more ANOVA-type designs with regression tools available to help with analysis
- Experiments/observational studies with more than one factor
- Examples:
  - vary price (3 levels) and advertising media (2 levels) to explore effect on sales
  - model family purchases using income (4 levels) and family stage (4 levels) as factors
- Why use multifactor studies?
  - efficient (can learn about more than one factor with same set of subjects)
  - added info (can learn about factor interactions)
  - but ... too many factors can be costly, hard to analyze

Factorial designs

Introduction

- Factorial design - takes all possible combinations of levels of the factors as separate treatments
- Example: 3 levels of factor A (a_1, a_2, a_3) and 2 levels of factor B (b_1, b_2) yields 6 treatments ($a_1b_1, a_1b_2, a_2b_1, a_2b_2, a_3b_1, a_3b_2$)
- Terminology:
  - complete factorial (all combinations used) vs fractional factorial (only a subset used)
  - fractional factorials done later
- Outline
  - factorial design with two factors
  - factorial designs with blocking
  - factorial designs with more than two factors
  - factorials with no replication (including fractional factorials)

Factorial designs

Crossed vs nested factors

- Factorials refer to ”crossing” two factors (all combinations of one factor are matched with all combinations of the other)
- Alternative structure is nesting
- Example: schools and instructors are factors
  - want two instructors at each school
  - crossed design requires same two instructors teach at each school
  - alternative: use diff’t instructors in each school
  - then instructors are **nested** within schools
- Crossed:

	Instructor	
School	1	2
A		
B		
C		

- Nested:

	Instructor					
School	1	2	3	4	5	6
A			x	x	x	x
B	x	x			x	x
C	x	x	x	x		

or

	Instructor	
School	1	2
A		
B		
C		

- Nested designs are discussed in Chapter 26 of text and later in this course

Factorial designs

Two factor study - model

- The cell means version of the model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \qquad \epsilon_{ijk} \text{ iid } N(0, \sigma^2)$$

  where $i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, n_{ij}$

  - μ_{ij} = mean of all units given level i of factor A and level j of factor B
  - $\mu_{i\cdot} = \frac{1}{b} \sum_j \mu_{ij}$ is mean response at level i of factor A
  - $\mu_{\cdot j} = \frac{1}{a} \sum_i \mu_{ij}$ is mean response at level j of factor B
  - $\mu_{\cdot\cdot} = \frac{1}{ab} \sum_i \sum_j \mu_{ij}$ is overall mean response

- We begin with assumption of equal sample size $n_{ij} = n$. It is an important assumption that will be reconsidered later. (Note that $N = abn$)

Factorial designs

Two factor study - model

- The factor effects version of the model

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

  where $\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$
  and $i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n_{ij}$

  - μ = overall mean response ($= \mu_{\cdot\cdot}$)
  - α_i = effect of level i of factor A ($= \mu_{i\cdot} - \mu_{\cdot\cdot}$)
  - β_j = effect of level j of factor B ($= \mu_{\cdot j} - \mu_{\cdot\cdot}$)
  - $(\alpha\beta)_{ij}$ = interaction ($= \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}$)
  - more on interactions shortly

- We begin with assumption of equal sample size $n_{ij} = n$. It is an important assumption that will be reconsidered later. (Note that $N = abn$)

Factorial designs

Cell means vs factor effects

- Two models are equivalent
- To see this note

  $$\mu_{ij} = \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{.j} - \mu_{..})$$
  $$+(\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..})$$

  or

  $$\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

- Factor effects model is much more popular than the cell means model

Factorial designs

Two factor study - interactions

- Factor effects model: $(\alpha\beta)_{ij}$ is interaction effect
  - $(\alpha\beta)_{ij} = 0$ for all i, j means an additive model
  - $(\alpha\beta)_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)$ is the difference between the mean for factor levels i, j and what would be expected under an additive model
  - the effect of factor A is not the same at every level of factor B
  - the effect of factor B is not the same at every level of factor A
  - can see interactions as we did in randomized block ANOVA (i.e., plot response vs factor A and connect points at the same level of factor B)
  - can also see an interaction in tables of means

Factorial designs

Two factor study - relation to regression

- Can write either cell means or factor effects model as a regression model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$

- Illustrate with factor effects model
- Example: $a = 2, b = 3, n = 2$

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{13} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \\ (\alpha\beta)_{23} \end{pmatrix} + \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{131} \\ \epsilon_{132} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{231} \\ \epsilon_{232} \end{pmatrix}$$

- But this model is overparameterized (the X matrix is not of full rank)

- E.g., col 2 + col 3 = col 1, col 7 + col 8 + col 9 = col 2, etc.

- Need to recode

Factorial designs

Two factor study - relation to regression

- Rewrite factor effects model ($a = 2, b = 3, n = 2$) as full rank regression model

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \end{pmatrix} + \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{131} \\ \epsilon_{132} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{231} \\ \epsilon_{232} \end{pmatrix}$$

- Other parameters determined by model restrictions, e.g.,
  $\alpha_2 = -\alpha_1, \beta_3 = -\beta_1 - \beta_2, (\alpha\beta)_{21} = -(\alpha\beta)_{11}$

- Note the cell means model can also be written in regression form

Factorial designs

Two factor study - ANOVA table

source of variation	degrees of freedom	sums of squares
factor A	$a-1$	$nb\sum_i(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot\cdot\cdot})^2$
factor B	$b-1$	$na\sum_j(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdot\cdot\cdot})^2$
interaction	$(a-1)(b-1)$	$n\sum_i\sum_j(\bar{Y}_{ij\cdot}-\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot j\cdot}+\bar{Y}_{\cdot\cdot\cdot})^2$
error	$ab(n-1)$	$\sum_i\sum_j\sum_k(Y_{ijk}-\bar{Y}_{ij\cdot})^2$
total	$abn-1$	$\sum_i\sum_j\sum_k(\bar{Y}_{ijk}-\bar{Y}_{\cdot\cdot\cdot})^2$

- Expected mean squares
  $E(MSA)=\sigma^2+nb\sum_i(\mu_{i\cdot}-\mu_{\cdot\cdot})^2/(a-1)$
  $E(MSB)=\sigma^2+na\sum_j(\mu_{\cdot j}-\mu_{\cdot\cdot})^2/(b-1)$
  $E(MSAB)=\sigma^2+n\frac{\sum_i\sum_j(\mu_{ij}-\mu_{i\cdot}-\mu_{\cdot j}+\mu_{\cdot\cdot})^2}{(a-1)(b-1)}$
  $E(MSE)=\sigma^2$
- All MS are independent
- Can test hypotheses about A, B, AB using F-tests, e.g., test $H_o:(\alpha\beta)_{ij}=0$ for all i,j
  using $F=MSAB/MSE$ vs $F_{(a-1)(b-1),ab(n-1)}$ distn
  - same as regression tests (for equal n case)
  - if we do three .05 tests than experimentwise error rate is larger
  - if no significant interactions, then can combine $SSAB$ with SSE (more df for estimating error)

Factorial designs

Two factor study - contrasts/factor effects

- Key point: the two factor factorial design can be viewed as a one factor design with ab treatments
- Inference for a single mean μ_{ij}, a factor level mean $\mu_{i\cdot}$, or linear combinations etc. are straightforward
- We illustrate with confidence interval expressions (assuming equal sample size in each cell)
- $100(1-\alpha)\%$ confidence interval for
  μ_{ij} is $\bar{Y}_{ij\cdot} \pm t_{df(error)}\sqrt{MSE/n}$
  $\mu_{i\cdot}$ is $\bar{Y}_{i\cdot\cdot} \pm t_{df(error)}\sqrt{MSE/(nb)}$
  $\mu_{\cdot j}$ is $\bar{Y}_{\cdot j\cdot} \pm t_{df(error)}\sqrt{MSE/(na)}$
  $\sum_i \sum_j c_{ij}\mu_{ij}$ is $\sum_{ij} c_{ij}\bar{Y}_{ij\cdot} \pm t_{df(error)}\sqrt{\frac{MSE}{n}\sum_i \sum_j c_{ij}^2}$
  $\sum_i c_i\mu_{i\cdot}$ is $\sum_i c_i\bar{Y}_{i\cdot\cdot} \pm t_{df(error)}\sqrt{\frac{MSE}{nb}\sum_i c_i^2}$
  $\sum_j c_j\mu_{\cdot j}$ is $\sum_j c_j\bar{Y}_{\cdot j\cdot} \pm t_{df(error)}\sqrt{\frac{MSE}{na}\sum_j c_j^2}$

Factorial designs

Two factor study - multiple comparisons

- Can handle the multiple comparisons issue with the usual procedures
  - Bonferroni
    - specify number of comparisons
    - adjust t critical value
  - Scheffe
    - replace t critical value with $\sqrt{kF_{k,df_{error}}}$
    - value of k depends on specific interest, e.g., use $k = a$ to protect all linear combos of factor A means, use $k = a - 1$ to protect all contrasts of factor A levels, etc.
  - Tukey (honest pairwise comparisons)
    - number of treatments we use for the table depends on which pairwise comparisons are being made (A, B, or AB)

Factorial designs

Two factor study - contrasts/factor effects

- Contrasts: linear combinations with $\sum_{ij} c_{ij} = 0$ (or $\sum_i c_i = 0$, etc.)
- From this point on take contrasts as $\sum_{ij} c_{ij}\mu_{ij}$ (others like $\sum_i c_i\mu_{i\cdot}$ are special cases)
- Great flexibility for testing hypotheses
- Consider an example with $a = 2$ and $b = 2$
  - true means are $\mu_{11} = \mu_{12} = \mu_{21} = 0, \mu_{22} = 4$ (first subscript is level of A, second is for B)
  - no response unless A and B both at level 2
  - usual ANOVA F-test for A is equiv to contrast with $c = (1, 1, -1, -1)$ which has value -4
  - usual ANOVA F-test for B is equiv to contrast with $c = (1, -1, 1, -1)$ which has value -4
  - usual ANOVA F-test for AB is equiv to contrast with $c = (1, -1, -1, 1)$ which has value 4
  - the contrast $c = (-1, -1, -1, 3)/\sqrt{3}$ corresponds exactly with the true pattern, has same $\sum_{ij} c_{ij}^2$ as others, and has value 6.93

Factorial designs

Two factor study - contrasts/factor effects

- It is possible to design contrasts that operate solely on A, solely on B, solely on the interaction, or on all three
- Illustrate with examples in $a = 2, b = 3$ case
  - For each example $c = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix}$
  - Contrast in A: $c = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \end{pmatrix}$

    all cols are the same, wts sum to zero in cols
  - Contrast in B: $c = \begin{pmatrix} -1 & 2 & -1 \\ -1 & 2 & -1 \end{pmatrix}$

    all rows are the same, wts sum to zero in rows
  - Contrast in AB: $c = \begin{pmatrix} -1 & 2 & -1 \\ 1 & -2 & 1 \end{pmatrix}$

    wts sum to zero in each row and in each column
  - Contrast out of all three: $c = \begin{pmatrix} -1 & 5 & -1 \\ -1 & -1 & -1 \end{pmatrix}$

    wts sum to zero overall, but not in rows/cols

Factorial designs

Two factor study - contrasts/factor effects

- Definition of orthogonal contrasts is same as in one-factor study
- As in one-way ANOVA it is possible to decompose any SS into orthogonal contrasts
- For example we can decompose SS(A) into SS for $a-1$ orthogonal contrasts
- Multiplying a vector of contrast weights for A, and a vector of weights for B yields a contrast for the interaction

Factorial designs

Two factor study - model diagnostics

- Same as in earlier ANOVA/regression models
- Residuals are $e_{ijk} = Y_{ijk} - \bar{Y}_{ij.}$
- Check for normal errors
- Check for constant variance (plot/test resids vs A, vs B, vs AB)
- Check for independence
- Normality is least important
- Remedies - transformation, weighted least squares, etc.
- New idea - interactions can be hard to interpret, may use transformations to avoid interactions

Factorial designs

Two factor study - unequal sample sizes

- What happens if n_{ij} is not the same for each treatment (i.e., combination of factors)
- Experiment: happens if there are missing values
  Observ study: this will generally be the case
- Three main ideas to know about
  - contrast analysis still works
  - regression analysis approach
  - unweighted means approach
- Also briefly consider what happens if $n_{ij} = 0$ for some treatment

Factorial designs

Two factor study - unequal sample sizes

- Contrast analysis
  - only info from ANOVA table required for contrasts is $MSE = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2 / (\sum_{ij}(n_{ij} - 1))$
  - if $L = \sum_{ij} c_{ij}\mu_{ij}$ with $\sum_{ij} c_{ij} = 0$, then $\hat{L} = \sum_{ij} c_{ij}\bar{Y}_{ij.}$ with $s.e.(\hat{L}) = \sqrt{MSE \sum_{ij} \frac{c_{ij}^2}{n_{ij}}}$
  - $100(1-\alpha)\%$ CI for L is $\hat{L} \pm t_{df,1-\alpha/2} s.e.(\hat{L})$ and $df = \sum_{ij}(n_{ij} - 1)$
  - contrasts in a single factor, $L = \sum_i c_i \mu_{i.}$
    - can be written as general contrasts ($L = \sum_i c_i \mu_{i.} = \sum_i \sum_j c_i \mu_{ij}/b$)
    - derive CI directly from $\hat{L} = \sum_i c_i \hat{\mu}_{i.} = \sum_i c_i \left(\frac{1}{b}\sum_j \bar{Y}_{ij.}\right)$
  - this is another nice feature of contrasts (don't require equal sample size)

Factorial designs

Two factor study - unequal sample sizes

- Regression analysis
  - regression approach does not assume $n_{ij} = n$
  - can test factor A (or B or AB) using regression methods, i.e., compare reduced model (B, AB) to full model (A, B, AB)
  - difficulties
    - tests of A, B, and AB are not independent
    - result of test for B depends on whether AB is in the model or not
    - $SS(A), SS(B), SS(AB), SS(Error)$ won't sum to $SS(Total)$
  - source of problems is that regression analysis automatically weights by n_{ij}
  - in ANOVA usually want each μ_{ij} to have equal weight
  - regression is appropriate if n_{ij} are meaningful

Factorial designs

Two factor study - unequal sample sizes

- Unweighted means analysis
  - a nice approach (in concept) at least
  - MSE as usual
  - obtain $MSA, MSB, MSAB$ from two factor ANOVA on cell means $\bar{Y}_{ij}$
  - problem: the MSE is on a different scale than the other MS (larger by a factor related to the size of the trt groups)
  - compute the harmonic mean, $n_h = \left(\frac{1}{ab}\sum_{ij}\frac{1}{n_{ij}}\right)^{-1}$
  - multiply $MSA, MSB, MSAB$ by n_h (or divide MSE by n_h)
  - now can carry out (independent) F-tests
  - this is similar to the contrast approach
  - described in Snedecor and Cochran

Factorial designs

Two factor study - unequal sample sizes

- What if one treatment combination is never observed, i.e., $n_{ij} = 0$?
- No problem if we assume no interactions
- There is a problem if we believe interactions exist
  - can't fit full interaction model
  - can't even compare main effects because of lack of balance
  - text describes a "partial analysis" approach: use a portion of the table, omitting the row/col with the empty cell

Factorial designs

Two factor study - randomized block design

- Can perform the two factor study in blocks
- Note block is a third factor!
- Repeat full factorial experiment ($r = ab$ treatments) in each block
- Assume no block and treatment interactions
- ANOVA table (see Sect 27.6 for more details)

source of variation	degrees of freedom	
block	$n-1$	
treatments	$ab-1$	
factor A		$a-1$
factor B		$b-1$
interaction AB		$(a-1)(b-1)$
error	$(ab-1)(n-1)$	
total	$abn-1$	

- Can make less restrictive assumptions - allow for block$*A$ and block$*B$ interactions (still no block$*A*B$ interactions) but this reduces df_{error}

Factorial designs

More than two factors

- No conceptual difficulties in increasing the number of factors
- Introduce factor C with c levels
- ANOVA table

source of variation	degrees of freedom	sums of squares
factor A	$a-1$	$nbc\sum_i(\bar{Y}_{i\cdots}-\bar{Y}_{\cdots\cdot})^2$
factor B	$b-1$	$nac\sum_j(\bar{Y}_{\cdot j\cdot\cdot}-\bar{Y}_{\cdots\cdot})^2$
factor C	$c-1$	$nab\sum_k(\bar{Y}_{\cdot\cdot k\cdot}-\bar{Y}_{\cdots\cdot})^2$
interaction AB	$(a-1)(b-1)$	$nc\sum_i\sum_j(\bar{Y}_{ij\cdot\cdot}-\bar{Y}_{i\cdots}-\bar{Y}_{\cdot j\cdot\cdot}+\bar{Y}_{\cdots\cdot})^2$
interaction AC	$(a-1)(c-1)$	$nb\sum_i\sum_k(\bar{Y}_{i\cdot k\cdot}-\bar{Y}_{i\cdots}-\bar{Y}_{\cdot\cdot k\cdot}+\bar{Y}_{\cdots\cdot})^2$
interaction BC	$(b-1)(c-1)$	$na\sum_j\sum_k(\bar{Y}_{\cdot jk\cdot}-\bar{Y}_{\cdot j\cdot\cdot}-\bar{Y}_{\cdot\cdot k\cdot}+\bar{Y}_{\cdots\cdot})^2$
interaction ABC	$(a-1)(b-1)(c-1)$	$SS(ABC)$
error	$abc(n-1)$	$\sum_i\sum_j\sum_k\sum_l(Y_{ijkl}-\bar{Y}_{ijk\cdot})^2$
total	$abcn-1$	$\sum_i\sum_j\sum_k\sum_l(Y_{ijkl}-\bar{Y}_{\cdots\cdot})^2$

$$SS(ABC)=n\sum_i\sum_j\sum_k(\bar{Y}_{ijk\cdot}-\bar{Y}_{ij\cdot\cdot}-\bar{Y}_{i\cdot k\cdot}-\bar{Y}_{\cdot jk\cdot}+\bar{Y}_{i\cdots}+\bar{Y}_{\cdot j\cdot\cdot}+\bar{Y}_{\cdot\cdot k\cdot}-\bar{Y}_{\cdots\cdot})^2$$

- Contrasts also are straightforward extensions
- Usually hope for nonsiginficant ABC interaction
- AB interaction: effect of B depends on level of A
  ABC interaction:
    effect of AB interaction depends on level of C
    or effect of C depends on levels of A,B

Factorial designs

Studies with no replications

- Suppose we have two factors (A with a levels and B with b levels) but only ab experimental units
- This might happen because of cost considerations or practical constraints
- Randomized block design is an example
- Can fit two factor factorial model with interactions but then no error estimate
- Resolution: hope for no interactions and use $MS(AB)$ to estimate σ^2
- ANOVA table

source of variation	degrees of freedom	sums of squares
factor A	$a-1$	$b\sum_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$
factor B	$b-1$	$a\sum_j(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$
error	$(a-1)(b-1)$	$\sum_i\sum_j(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$
total	$ab-1$	$\sum_i\sum_j(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$

Factorial designs

Studies with no replications - 2 factors

- ANOVA table on previous slide
- Expected mean squares
  $E(MSA) = \sigma^2 + b\sum_i(\mu_{i.} - \mu_{..})^2/(a-1)$
  $E(MSB) = \sigma^2 + a\sum_j(\mu_{.j} - \mu_{..})^2/(b-1)$
  $E(MSE) = \sigma^2$
- Usual tests
- Note that we have two plausible estimates for μ_{ij} under model
  - Y_{ij}
  - $\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}$
  - second is preferred (less variance)

Factorial designs

Studies with no replications - 2 factors

- Model checking - as usual except now also very concerned about nonadditivity
- Tukey's test for non-additivity
  - mentioned earlier in randomized block context
  - tests a specific form of nonadditivity
  - $H_o : Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$
    $H_a : Y_{ij} = \mu + \alpha_i + \beta_j + D\alpha_i\beta_j + \epsilon_{ij}$
  - compute $SS(AB)$ (see motivation in text)

$$SS(AB) = \frac{\left(\sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})Y_{ij}\right)^2}{\sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \sum_j (\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2}$$

  - compute adjusted error sum of squares
    $SSE.AB = SS(Error) - SS(AB)$
  - test $F = SS(AB)/MSE.AB$ with 1 and $(a-1)(b-1)-1$ d.f.
- If we find non-additivity, perhaps can transform to remove it

Factorial designs

Studies with no replication - 2^K studies

- Assume K factors, each at two levels
- Known as 2^K factorial
- One application: factors are really continuous and we want to explore response to factors .... leads to response surface methods (Ch. 32)
- Some special features
  - all main effects, interactions are 1 d.f.
  - the regression approach works nicely
    - 1 column of X for each main effect (with +1/-1 coding)
    - interaction columns by multiplication
    - all columns are orthogonal
- With replication, no new problems here
- With no replication same problem as discussed previously but with some new solutions

Factorial designs

Studies with no replication - 2^K studies

- Estimating σ^2 in 2^K study without replication
  - pool SS from nonsignif factors/interactions to estimate σ^2; if we pool p terms, then

    $$\sigma^2 = (2^K \sum_q b_q^2)/p$$

    where b_q is regression coefficient
  - normal probability plot - coefficients b_q with no real effect are "normal" with constant variance
    * estimate slope from those points (slope $= \sigma/2^K$)
    * half-normal plot (ordered $|b_j|$ vs $\Phi^{-1}(\frac{j+n-.125}{2n+.5})$) works a bit better
  - center point replication
    * construct one new treatment at intermediate levels of each factor - called a center point
    * take n_o observations at this new center point
    * these observations add n_o d.f. to error d.f.
    * split extra SS into $n_o - 1$ d.f. for pure error and 1 d.f. for lack-of-fit

Factorial designs

Studies with no reps - fractional factorials

- Assume K factors, each at two levels
- Sometimes 2^K is too many observations
- Can run 2^{K-J} fractional factorial (fraction 2^{-J} of a full factorial)
- Can't estimate all 2^K effects
- Introduce confounding by carefully selecting those treatments to use
- Note still have problem estimating σ^2 unless there is some replication
- Example of fractional factorial on next slide

Factorial designs

Studies with no reps - fractional factorials

- 2^K study with $K = 3$ (call factors A, B, C)
- Design matrix for full factorial regression (no rep)

obs	μ	A	B	C	AB	AC	BC	ABC
1	1	1	1	1	1	1	1	1
2	1	1	1	-1	1	-1	-1	-1
3	1	1	-1	1	-1	1	-1	-1
4	1	1	-1	-1	-1	-1	1	1
5	1	-1	1	1	-1	-1	1	-1
6	1	-1	1	-1	-1	1	-1	1
7	1	-1	-1	1	1	-1	-1	1
8	1	-1	-1	-1	1	1	1	-1

- Consider $J = 1$ - half-factorial

obs	μ	A	B	C	AB	AC	BC	ABC
1	1	1	1	1	1	1	1	1
4	1	1	-1	-1	-1	-1	1	1
6	1	-1	1	-1	-1	1	-1	1
7	1	-1	-1	1	1	-1	-1	1

  - note that can't distinguish between μ and ABC, they are said to be confounded
  - so are A and BC, B and AC, C and AB
  - signif A effect may actually be BC effect
  - use only main effects in analysis
  - very useful if no interactions
  - other half-factorials are possible - each will confound different effects

Random effects

One factor studies (Section 25.1)

- Recall distinction btwn fixed and random effects
  - fixed - factor levels used in study are of intrinsic interest
  - random - factor levels are not of intrinsic interest but rather constitute a sample from a population
- One-factor random effects model

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \mu_i \text{ iid } N(\mu, \sigma_\mu^2) \quad \epsilon_{ij} \text{ iid } N(0, \sigma^2)$$

  for $i = 1, \ldots, r$; and $j = 1, \ldots, n_i$
  - can also write as factor effects model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ with $\tau_i \sim N(0, \sigma_\mu^2)$
  - marginal distn of Y_{ij} has mean μ, variance $\sigma_\mu^2 + \sigma^2$, and $\text{cov}(Y_{ij}, Y_{ij'}) = \sigma_\mu^2$
  - distn of $\mathbf{Y}$ with $r = 2, n_1 = n_2 = 2$

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \\ \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2+\sigma_\mu^2 & \sigma_\mu^2 & 0 & 0 \\ \sigma_\mu^2 & \sigma^2+\sigma_\mu^2 & 0 & 0 \\ 0 & 0 & \sigma^2+\sigma_\mu^2 & \sigma_\mu^2 \\ 0 & 0 & \sigma_\mu^2 & \sigma^2+\sigma_\mu^2 \end{pmatrix}\right)$$

  - note: if we consider μ_i fixed (condition on it), then Y_{ij} are indep with mean μ_i and variance σ^2 (the fixed effects model)

Random effects

One factor studies

- From this point on assume equal sample sizes $n_i = n$ for all i
- Unequal sample sizes is a nuisance but can be handled
- ANOVA table is same as for fixed effects model

source of variation	degrees of freedom	sums of squares
treatments	$r - 1$	$\sum_i n(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$
error	$r(n-1) = N - r$	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$
total	$rn - 1 = N - 1$	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$

- Expected mean squares slightly different than for fixed effects case
  $E(MSTr) = \sigma^2 + n\sigma_\mu^2$ and $E(MSE) = \sigma^2$
- F-test now used to test $H_o : \sigma_\mu^2 = 0$

Random effects

One factor studies

- Inference for quantities of interest
  - overall mean μ
    - estimate is $\hat{\mu} = \bar{Y}_{..}$
    - $\text{Var}(\bar{Y}_{..}) = (n\sigma_\mu^2 + \sigma^2)/(rn)$
    - $s.e.(\bar{Y}_{..}) = \sqrt{MSTr/(rn)}$ with d.f. $r - 1$
    - different CI than for fixed effects model
    - why? effective sample size for μ is r not rn
    - note: $\text{Var}(\bar{Y}_{i.}) = (n\sigma_\mu^2 + \sigma^2)/n$ would again be estimated using $MSTr$
  - error variance σ^2
    - same results as for fixed effects, i.e.,
    - MSE is unbiased point estimate
    - $\frac{r(n-1)MSE}{\sigma^2} \sim \chi^2_{r(n-1)}$ can be used for confidence intervals or tests

Random effects

One factor studies

- Inference for quantities of interest (cont'd)
  - random effects variance σ_μ^2
    - already seen that we can test $H_o : \sigma_\mu^2 = 0$ using F-test
    - can estimate σ_μ^2 in several ways
    - method of moments (equating MS's and their expectation) yields $\hat{\sigma}_\mu^2 = (MSTr - MSE)/n$
    - unbiased but can be negative
    - CIs - Satterthwaite approx in book
    - maximum likelihood estimate is a bit harder, but by definition can't be negative

Random effects

One factor studies

- Inference for quantities of interest (cont'd)
  - intraclass correlation $\rho_I = \sigma_\mu^2/(\sigma_\mu^2 + \sigma^2)$
    - measures correlation of two obs in same treatment group
    - key result for inference is that $(\frac{MSTr}{n\sigma_\mu^2+\sigma^2})/(\frac{MSE}{\sigma^2}) \sim F_{r-1,r(n-1)}$
    - (under $H_o\!:\!\sigma_\mu^2 = 0$ the variance factors cancel)
    - can rewrite key result as $\left(1 + \frac{n\rho_I}{1-\rho_I}\right)^{-1} \frac{MSTr}{MSE} \sim F_{r-1,r(n-1)}$
    - this form can be used to get tests/CIs for ρ
    - CI : Define $L = \frac{1}{n}\left(\frac{MSTr}{MSE}\left(\frac{1}{F_{1-\alpha/2}} - 1\right)\right)$ and $U = \frac{1}{n}\left(\frac{MSTr}{MSE}\left(\frac{1}{F_{\alpha/2}} - 1\right)\right)$, then $(\frac{L}{1+L}, \frac{U}{1+U})$ is a $100(1-\alpha)\% CI$ for ρ_I

Random effects

Two crossed factors (Section 25.2-25.4)

- Usual two factor factorial design with both factors considered random
- Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$
  with $\alpha_i \sim N(0, \sigma^2_\alpha), \quad \beta_j \sim N(0, \sigma^2_\beta),$
  $(\alpha\beta)_{ij} \sim N(0, \sigma^2_{\alpha\beta}), \quad \epsilon_{ijk} \sim N(0, \sigma^2)$
  and everything independent of everything else
- Interpretation:
  $E(Y_{ijk}) = \mu$ and $\text{Var}(Y_{ijk}) = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta} + \sigma^2$
  with $\text{Cov}(Y_{ijk}, Y_{ij'k'}) = \sigma^2_\alpha$
  and $\text{Cov}(Y_{ijk}, Y_{i'jk'}) = \sigma^2_\beta$
  and $\text{Cov}(Y_{ijk}, Y_{ijk'}) = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta}$
- Example: reproducibility and repeatability studies in engineering (A = parts, B = operators)

Random effects

Two crossed factors

- ANOVA table is same as fixed effects 2-factor factorial except for expected mean squares
  $E(MSA) = \sigma^2 + nb\sigma_\alpha^2 + n\sigma_{\alpha\beta}^2$
  $E(MSB) = \sigma^2 + na\sigma_\beta^2 + n\sigma_{\alpha\beta}^2$
  $E(MSAB) = \sigma^2 + n\sigma_{\alpha\beta}^2$
  $E(MSE) = \sigma^2$
- This means that to test
  $H_o : \sigma_\alpha^2 = 0$ use $F = MSA/MSAB$
  $H_o : \sigma_\beta^2 = 0$ use $F = MSB/MSAB$
  $H_o : \sigma_{\alpha\beta}^2 = 0$ use $F = MSAB/MSE$
- Estimates of variance components can be obtained by method of moments or maximum likelihood

Mixed models

Models with random and fixed effects (Section 25.2-25.4)

- Extremely useful in practice (SAS's PROC MIXED)
- Quick summary of ideas here (Chapter 25)
- Use two-factor crossed design to illustrate
- Model

  $$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

  with $i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n$
  - i subscript is fixed effect ($\sum_i \alpha_i = 0$)
  - j subscript is random (β_j iid $N(0, \sigma^2_\beta)$)
  - $(\alpha\beta)_{ij}$ iid $N(0, \sigma^2_{\alpha\beta})$ (the unrestricted model)
  - ϵ_{ijk} iid $N(0, \sigma^2)$
  - $\beta_j, (\alpha\beta)_{ij}, \epsilon_{ijk}$ are pairwise independent
- Comments
  - randomized block (with random block effects) is an example with $\sigma^2_{\alpha\beta} = 0$
  - random effects are useful way to introduce correlation (spatial, time, genetics, cluster)

Mixed models

Models with random and fixed effects

- There is a more general version of the two-factor crossed design mixed linear model
- The restricted model assumes
  - $(\alpha\beta)_{ij} \sim N(0, \frac{a-1}{a}\sigma^2_{\alpha\beta})$
  - $\sum_i (\alpha\beta)_{ij} = 0$ for each j
  - $\text{Cov}((\alpha\beta)_{ij}, (\alpha\beta)_{i'j}) = -\frac{1}{a}\sigma^2_{\alpha\beta}$
  - use of $(a-1)/a$ makes expected mean squares turn out nicer
- The restricted model is more popular. Why?
  - more general (allows positive or negative correlation within level of j)

	restricted	unrestricted
$\text{Cov}(Y_{ijk}, Y_{ijk'})$	$\sigma^2_\beta + \frac{a-1}{a}\sigma^2_{\alpha\beta}$	$\sigma^2_\beta + \sigma^2_{\alpha\beta}$
$\text{Cov}(Y_{ijk}, Y_{i'jk'})$	$\sigma^2_\beta - \frac{1}{a}\sigma^2_{\alpha\beta}$	σ^2_β
$\text{Cov}(Y_{ijk}, Y_{i'j'k'})$	0	0

  - matches finite population results (exp mean squares)

Random effects

Nested or hierarchical designs (Chapter 26)

- Recall discussion of nesting and crossing
- Not always possible for every level of A and B to be paired
- Example: suppose factor A is schools and factor B is instructors
  - if we want two instructors at each school
  - crossed design requires same two instructors teach at each school
  - alternative: use diff't instructors in each school
  - then instructors are **nested** within schools
- Crossed:

	Instructor	
School	1	2
X		
Y		
Z		

- Nested:

	Instructor					
School	1	2	3	4	5	6
X			x	x	x	x
Y	x	x			x	x
Z	x	x	x	x		

or

	Instructor	
School	1	2
X		
Y		
Z		

Random effects

Nested designs

- It is possible that both factors are fixed,
  but more common for nested factors to be random
- Consequently, need only skim Chapter 26
- Assume equal sample sizes
- Also allow for possibility of repeated measurements at level i of A and level j of B (for example students within instructors within schools)
- Model

$$Y_{ijk} = \mu + \tau_i + \epsilon_{j(i)} + \eta_{ijk}$$

  with $i = 1, \ldots, r;\ j = 1, \ldots, n;\ k = 1, \ldots, m$
  - μ = overall mean
  - τ_i = trt group i effect (fixed w/ $\sum_i \tau_i = 0$)
  - $\epsilon_{j(i)}$ = jth unit nested within treatment i
  - $\epsilon_{j(i)}$'s are iid $N(0, \sigma^2)$
  - usually not feasible to block
    (can't put every trt in nested unit)
  - η_{ijk} = measurement error
  - η_{ijk}'s are iid $N(0, \sigma_\eta^2)$

Random effects

Nested designs

- ANOVA table

source of variation	degrees of freedom	sums of squares
treatments (A)	$r-1$	$\sum_i nm(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$
nested units (B)	$r(n-1)$	$\sum_i \sum_j m(Y_{ij\cdot} - \bar{Y}_{i\cdot\cdot})^2$
error	$rn(m-1)$	$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij\cdot})^2$
total	$rnm-1$	$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{\cdot\cdot\cdot})^2$

- Expected value of mean squares:
  $E(MSTr) = \sigma_\eta^2 + m\sigma^2 + nm \sum_i \tau_i^2/(r-1)$
  $E(MSB) = \sigma_\eta^2 + m\sigma^2$
  $E(MSE) = \sigma_\eta^2$

- F-test for treatment effects uses MSB as error

- In fact, general principle is ... test sources of variation vs the first nested (or crossed) random factor

- MSB is sometimes known as experimental error

- MSE is sometimes known as observational error

Random effects

Nested designs

- Inference for treatment effects
  - $\bar{Y}_{i\cdot\cdot}$ is sample mean for trt i
  - $\text{Var}(\bar{Y}_{i\cdot\cdot}) = (\sigma_\eta^2 + m\sigma^2)/(mn)$
  - estimate above using $MSB/(mn)$
  - if we know σ^2 and σ_η^2 then we can use the variance expression to design a study (pick m and n to minimize variance for fixed cost)
- Inference for variance components
  - MSE estimates σ_η^2
  - $s^2 = (MSB - MSE)/m$ estimates σ^2 (method of moments)
  - usual difficulty with s^2 (can be negative)
  - use maximum likelihood to avoid

Random effects

Nested designs

- Random treatment effects
  - if the treatment effect is viewed as random, then the nested model is known as a variance components model
  - often used for studying variation at different levels of a measurement process (sampling trees, then leaves, and then measurements on leaves)
  - $\text{Var}(\bar{Y}_{...}) = \frac{\sigma_\tau^2}{r} + \frac{\sigma^2}{rn} + \frac{\sigma_\eta^2}{rnm}$
  - estimates of variance components can be used to design sampling scheme (how many trees? how many leaves per tree? how many measures per leaf?)
  - method of moments estimates are naive
  - maximum likelihood, restricted ML (REML) are better

Repeated measures

Models with repeated measurements on units (Chapter 27)

- Key point: sometimes one experimental unit is measured more than once (more than one response)
  - ex: 3 diff't treatments on each of n units (see below)
  - ex: 1 treatment applied to each unit, response measured at a number of time points (later)
  - other variations are possible
- Example: consider n subjects with 3 treatments

	treatment		
subject	trt 1	trt 2	trt 3
1	Y_{11}	Y_{12}	Y_{13}
2	Y_{21}	Y_{22}	Y_{23}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_{n1}	Y_{n2}	Y_{n3}

- Advantage: precision (each subject is a block)
- Disadvantages:
  interference (carryover effect, order effects),
  strong assumpt. about correl. structure (more later)
- Randomization of treatment sequence avoids some of the disadvantages

Repeated measures

Models with repeated measurements on units

- Model:

$$Y_{ij} = \mu + \rho_i + \tau_j + \epsilon_{ij}$$

  $i = 1, \ldots, n;\ j = 1, \ldots, r$

  - μ is overall mean
  - ρ_i iid $N(0, \sigma_\rho^2)$ are subject effects (random)
  - τ_j are trt effects (fixed w/ $\sum_j \tau_j = 0$)
  - ϵ_{ij} iid $N(0, \sigma^2)$ errors
  - ρ_i, ϵ_{ij} independent

- Comments
  - this is randomized block design (subjects as blocks) with random block effects
  - no replications
  - the r treatments can be one factor or $r = ab$ trts from a factorial
  - use mixed model (with $n = 1$) if interactions
  - implies $\text{Corr}(Y_{ij}, Y_{ij'}) = \sigma_\rho^2/(\sigma_\rho^2 + \sigma^2)$ for all treatments(responses)

Repeated measures

Models with repeated measurements on units

- Recall model assumes $\mathrm{Corr}(Y_{ij}, Y_{ij'}) = \frac{\sigma_\rho^2}{(\sigma_\rho^2+\sigma^2)}$ for all measurements
  - this is a strong assumption
  - will generally be false if responses are measurements taken at diff't times (nearby measurements will be more highly correlated)
  - if false, one needs to specify a mixed model that accounts for the correlation structure

Repeated measures

Models with repeated measurements on units

- Inference - ANOVA table

source	d.f.	SS
subjects	$n-1$	$r\sum_{i=1}^{n}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$
treatments	$r-1$	$n\sum_{j=1}^{r}(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})^2$
error	$(n-1)(r-1)$	$\sum_{i=1}^{n}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{i\cdot}-\bar{Y}_{\cdot j}+\bar{Y}_{\cdot\cdot})^2$
total	$IJ-1$	$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$

  where last row is actually interaction used as error

- Expected mean squares
  $E(MSSub)=\sigma^2+r\sigma_\rho^2$
  $E(MSTr)=\sigma^2+\frac{n}{r-1}\sum_{j=1}^{r}\tau_j^2$
  $E(MSE)=\sigma^2$

- Test for treatments uses $F=MSTr/MSE$
  (MSE - first term w/ trt crossed by random factor)

- Inference for treatment means
  - single treatment means aren't very useful
    (as in randomized block)
  - difference in treatment means $\bar{Y}_{\cdot j}-\bar{Y}_{\cdot j'}$
    - estimates $\tau_j-\tau_{j'}$
    - $\text{Var}(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot j'})=\sigma^2\left(\frac{1}{n}+\frac{1}{n}\right)$
    - estimate σ^2 using MSE and obtain t intervals
    - works for any contrast of trt means

Repeated measures

Models with repeated measurements on units

- Model checking
  - typical checks using residuals (constant error variance, normality of errors)
  - need to check for response by subject interaction (nonadditivity)
  - normal probability plot of subject effects $\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$
  - can check equi-correlated response assumption (compound symmetry) by computing pooled within subject covariance matrix W

$$W_{jj'} = \sum_i (Y_{ij} - \bar{Y}_{\cdot j})(Y_{ij'} - \bar{Y}_{\cdot j'})/(n-1)$$

Repeated measures

Models w/ between and within subject factors

- Sometimes subjects can't be used for all factors in a study
- Examples:
  individual can't be used if gender is a factor,
  subplot can't be used to test diff't irrigation
- Then, some factors are "between subjects" and some factors are "within subjects"
- Agricultural expts of this type are split plot designs (between = "whole plot", within = "subplot")
- Example of layout of data
  - one between subject factor A (two levels a_1, a_2)
  - n subjects at each level (nested)
  - one within subject factor B (three levels b_1, b_2, b_3)

factor A	subject	factor B b_1	b_2	b_3
a_1	1	Y_{111}	Y_{112}	Y_{113}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n	Y_{n11}	Y_{n12}	Y_{n13}
a_2	$n+1$	$Y_{n+1,2,1}$	$Y_{n+1,2,2}$	$Y_{n+1,2,3}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$2n$	$Y_{2n,1,1}$	$Y_{2n,1,2}$	$Y_{2n,2,3}$

Repeated measures

Models w/ between and within subject factors

- Model:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \rho_{i(j)} + \epsilon_{ijk}$$

  $i = 1, \ldots, n;\ j = 1, \ldots, a;\ k = 1, \ldots, b$

  - μ is overall mean
  - α_j - between subject trt effects (fixed w/ $\sum_j \alpha_j = 0$)
  - β_k - within subject trt effects (fixed w/ $\sum_k \beta_k = 0$)
  - $(\alpha\beta)_{jk}$ are interactions (with usual assumpt.)
  - $\rho_{i(j)}$ iid $N(0, \sigma^2_\rho)$ are subject effects (random, nested within factor A)
  - ϵ_{ijk} iid $N(0, \sigma^2)$ errors
  - $\rho_{i(j)}, \epsilon_{ijk}$ independent

Repeated measures

Models w/ between and within subject factors

- Interpretation of the model
  Y_{ijk} has mean $\mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$
  Y_{ijk} has variance $\sigma_\rho^2 + \sigma^2$
  $\text{Cov}(Y_{ijk}, Y_{ijk'}) = \sigma_\rho^2$
  other covariances are zero
- Comments
  - A, B could themselves be factorial designs on more than one factor
  - could also accommodate blocks for between subject factor (e.g., siblings)
  - some/all factors could be random effects
  - same correlation structure (compound symmetry model) as before

Repeated measures

Models w/ between and within subject factors

- Inference - ANOVA table

source	d.f.	SS
A	$a-1$	$bn\sum_{j=1}^{a}(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots})^2$
subjects	$a(n-1)$	$b\sum_{i=1}^{n}\sum_{j=1}^{a}(\bar{Y}_{ij\cdot}-\bar{Y}_{\cdot j\cdot})^2$
B	$b-1$	$an\sum_{k=1}^{b}(\bar{Y}_{\cdot\cdot k}-\bar{Y}_{\cdots})^2$
AB	$(a-1)(b-1)$	$n\sum_{j=1}^{a}\sum_{k=1}^{b}(\bar{Y}_{\cdot jk}-\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdot\cdot k}+\bar{Y}_{\cdots})^2$
error (Bxsubj)	$a(n-1)(b-1)$	$\sum_{i=1}^{n}\sum_{j=1}^{a}\sum_{k=1}^{b}(Y_{ijk}-\bar{Y}_{\cdot jk}-\bar{Y}_{ij\cdot}+\bar{Y}_{\cdot j\cdot})^2$
total	$nab-1$	$\sum_{i=1}^{n}\sum_{j=1}^{a}\sum_{k=1}^{b}(Y_{ijk}-\bar{Y}_{\cdots})^2$

- Notes:
  - organization of table is diff't than in book
  - first two rows are between subjects (total d.f. $= an-1$)
  - this is d.f. if we had one measurement per subject
  - last three rows are within subjects (total d.f. $= an(b-1)$)

Repeated measures

Models w/ between and within subject factors

- Expected mean squares
  $E(MSA) = \sigma^2 + b\sigma_\rho^2 + \frac{bn}{a-1}\sum_j \alpha_j^2$
  $E(MSSub) = \sigma^2 + b\sigma_\rho^2$
  $E(MSB) = \sigma^2 + \frac{an}{b-1}\sum_k \beta_k^2$
  $E(MSAB) = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{j,k}(\alpha\beta)_{jk}^2$
  $E(MSE) = \sigma^2$

- Tests of hypotheses
  - test vs nearest random factor
  - test for A uses $MSA/MSSub$
    (same error for all between subjects factors)
  - test for B uses $MSAB/MSE$ (similar for AB)
  - if there were more within factors, each would have its own error because the error here is actually $B \times$ Subject interaction

Repeated measures

Models w/ between and within subject factors

- Inference for contrasts
  - in factor A we have $\hat{L} = \sum_j c_j \bar{Y}_{\cdot j \cdot}$
    with s.e. $\sqrt{MSSub\left(\frac{1}{nb}\sum_j c_j^2\right)}$
  - in factor B we have $\hat{L} = \sum_k c_k \bar{Y}_{\cdot\cdot k}$
    with s.e. $\sqrt{MSE\left(\frac{1}{na}\sum_k c_k^2\right)}$
  - in factor AB we have $\hat{L} = \sum_{j,k} c_{jk} \bar{Y}_{\cdot jk}$
    with s.e. $\sqrt{MSE\left(\frac{1}{n}\sum_{j,k} c_{jk}^2\right)}$
- Model checking
  - as in previous repeated measures model
  - check constant variance σ_ρ^2 within levels of A using plot of $(\bar{Y}_{ij\cdot} - \bar{Y}_{\cdot j \cdot})$ vs j