STAT 5000

Statistical Methods I

Week 7

Fall 2024

Dr. Danica Ommen

Unit 2

False Discovery Rate

False Discovery Rate (FDR)

- In genomic studies, we are simultaneously testing a huge number of hypotheses, each relating to a feature.
  - Gene expression experiments: tens of thousands of genes are measured for each sample
  - GWAS studies: millions of SNPs are tested simultaneously.
  - Microbiome studies: depending on the precision, there could be thousands of OTUs measured for each sample.

False Discovery Rate (FDR)

- Suppose one test of interest has been conducted for each of the m genes in an RNA-seq experiment to study gene expression.
- Often, the test of interest is that for each gene, whether the mean expression levels change (i.e., the gene is differentially expressed) between different treatments or not.
- Let $H_{o1}, H_{o2}, \cdots, H_{om}$ denote the null hypotheses (interpreted as non-differential expression) corresponding to the m tests (genes).
- If we set level 5% for each test, the number of type I error (false positives) is expected to be 5% of m_o, the total number of genes that are not differentially expressed.

False Discovery Rate (FDR)

- Considering the high dimensionality of RNA-seq data, the number of errors is expected to be really big if we only control error at the level of individual test.
- Considering the high dimensionality of RNA-seq data, multiple comparison adjustments such as Bonferroni's method is impractical.

False Discovery Rate (FDR)

- FDR (or pFDR = positive FDR) is an alternative error rate that can be useful for RNA-seq experiments or other genomic studies.
- Table of Outcomes for m Tests

Hypothesis	Accept Null	Reject Null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

- FDR (Benjamini and Hochberg, 1995)

$$FDR = E\left(\frac{V}{R}\middle| R > 0\right) Pr(R > 0)$$

False Discovery Rate (FDR)

A Conceptual Description of FDR

- Suppose a scientist conducts 100 independent RNA-seq experiments.
- For each experiment, the scientist produces a list of genes declared to be differentially expressed by testing a null hypothesis for each gene.
- For each list consider the ratio of the number of false positive results to the total number of genes on the list (set this ratio to 0 if the list contains no genes).
- The FDR is approximated by the average of the ratios described above.

False Discovery Rate (FDR)

Is Experiment-Wise Error Rate Too Conservative for Genomic Studies?

- Suppose that one of the 100 gene lists consists of 500 genes declared to be differentially expressed.
- Suppose that 1 of those 500 genes is not truly differentially expressed but that the other 499 are.
- This list is considered to be in error and such lists are allowed to make up only a small proportion of the total number of lists if experiment-wise error rate is to be controlled.
- However such a list seems quite useful from the scientific viewpoint.

False Discovery Rate (FDR)

FDR: The Appropriate Error Rate for Genomic Studies?

- The hypothetical gene list discussed previously with 1 false positive and 499 true positives would be a good list that would help to keep the FDR down.
- Some of the gene lists may contain a high proportion of false positive results and yet the method we are using may still control FDR at a given level because it is the average performance across repeated experiments that matters. (This comment applies to the control of experiment-wise error rate as well.)

Unit 2

Intro to Blocking

Key to Statistical Significance

Differences on response variable between groups are larger than the differences within groups.

- t-test statistic

$$t = \frac{(\bar{Y}_i - \bar{Y}_j) - 0}{S_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

- F-test statistic

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

Variation within Groups: Problem

- When σ^2 is large compared to differences between means
  - Fail to reject H_o of equal means even when differences between means exist.
- Why would σ^2 be large?
  - Response variable has large amount of variation.
  - Experimental units are not homogeneous with respect to response variable.

Variation within Groups: Solution

- Choose more homogeneous experimental units.
  - Reduces variation in response variable - more likely to produce significant result.
  - Reduces generalizability of experimental results.
- Use more heterogeneous experimental units.
  - Increases variation in response variable - less likely to produce significant result.
  - Increases generalizability of experimental results.

Reduce Variation through Blocking

Block

A group of experimental units that, prior to treatment, are expected to be more like one another (with respect to response variables) than experimental units in general.

In simple words, blocks are groups of similar experimental units.

- Group experimental units into blocks before assigning treatments
- Randomly assign treatments to experimental units within each block separately

Types of Blocking

- Sorting
- Subdividing
- Reusing
- Matching

Types of Blocking

Sorting

- You are interested in the effect of two different instructional methods on achievement in mathematics of 8th graders.
- Sort students by their Iowa Test math scores from 7th grade.
- Students within each block will have similar Iowa Test math scores.

Subdividing

Reusing

Matching

Types of Blocking

Sorting

Subdividing

- You are interested in the yield of three varieties of soy beans.
- You have 12 fields across Iowa that you can use.
- Divide each field into 3 sections and plant one variety on each section.

Reusing

Matching

Types of Blocking

Sorting

Subdividing

Reusing

- You are interested in the determining which of two brands of golf balls travels the furthest when hit with a five iron.
- Have each person hit both types of golf ball (reuse each person).

Matching

Types of Blocking

Sorting

Subdividing

Reusing

Matching

- You are interested in the determining which of two brands of golf balls travels the furthest when hit with a five iron.
- Pair two golfers with same skill level.
- Have one person hit one brand of golf ball and other person hit the other brand of golf ball.

Effect of Blocking

- Reduce σ^2 = variation within experimental units
- Variation due to block variable is removed from σ^2 (and its estimate)
- Easier to detect difference between treatment groups

Unit 2

Blocking: Matched Pairs

Matched Pairs

- Experiments with two treatments
- Blocks have one or two experimental units
  - One unit (reuse)
    - Receives both treatments
    - Order of treatments is random
  - Two units (match)
    - Two treatments randomly assigned to pair
    - One unit receives one treatment
    - Other unit receives other treatment

Data

- Y_{i1} = response to treatment 1 in the ith block
- Y_{i2} = response to treatment 2 in the ith block
- n blocks, assumed independent

	Treatment 1	Treatment 2
Block 1	Y_{11}	Y_{12}
Block 2	Y_{21}	Y_{22}
⋮	⋮	⋮
⋮	⋮	⋮
⋮	⋮	⋮
Block n	Y_{n1}	Y_{n2}

Data

- Responses within a block are not independent; usually Y_{i1} has a positive correlation with Y_{i2}
- $\mu_1 =$ mean response for treatment 1
- $\mu_2 =$ mean response for treatment 2
- $\mu_d = \mu_1 - \mu_2 =$ mean difference in responses between treatments

Analysis: Expected Value

- $D_i = Y_{i1} - Y_{i2} =$ difference in response between treatments for each of the n blocks
- Estimate μ_d with

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(Y_{i1} - Y_{i2}) = \bar{Y}_1 - \bar{Y}_2$$

Analysis: Variance

$$
\begin{aligned}
Var(\bar{D}) &= Var(\frac{1}{n}\sum_{i=1}^{n} D_i) \\
&= \frac{1}{n^2} Var(\sum_{i=1}^{n} D_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(D_i) \quad \text{(assumes independent blocks)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_{i1} - Y_{i2}) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} [Var(Y_{i1}) + Var(Y_{i2}) - 2Cov(Y_{i1}, Y_{i2})]
\end{aligned}
$$

Analysis: Variance

$$
\begin{aligned}
Var(\bar{D}) &= \frac{1}{n^2}\sum_{i=1}^{n}[Var(Y_{i1}) + Var(Y_{i2}) - 2Cov(Y_{i1}, Y_{i2})] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right] \\
&= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\frac{\rho\sigma_1\sigma_2}{n} \\
&= Var(\bar{Y}_1) + Var(\bar{Y}_2) - 2Cov(\bar{Y}_1, \bar{Y}_2)
\end{aligned}
$$

- ρ is *within* block correlation between Y_{i1} and Y_{i2}
- ρ is positive, making overall $Var(\bar{D})$ smaller

Analysis: Variance

Estimate $Var(\bar{D})$ using observed differences D_i:

- Unbiased estimate of $Var(D_i)$

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

- Standard Error of $\bar{D}$

$$\text{SE}(\bar{D}) = \frac{s_d}{\sqrt{n}}$$

Analysis

Hypothesis Test

- $H_0 : \mu_d = 0$ vs. $H_a : \mu_d \neq 0$
- Test Statistic:

$$t = \frac{\bar{D}}{(s_d/\sqrt{n})}$$

- p-value:

$$2 \times P(t_{n-1} > |t|)$$

Analysis

Confidence Interval

100$(1-\alpha)$% Confidence Interval for μ_d:

$$\bar{D} \pm t_{n-1,1-\alpha/2} \frac{s_d}{\sqrt{n}}$$

Example: Monkey Nerve Cell Study

- Goal: Explore creatine phosphate (CP) concentrations in rhesus monkey nerve cells.
- n=8 monkeys
- Nerves extending from one side of the spinal cord were severed
- CP concentrations (mg per 100g of tissue) were measured during the regeneration process
- Nerves extending from the other side of the spinal cord were kept intact (control).
- CP concentrations were also measured on the control side.

Example: Monkey Nerve Cell Study

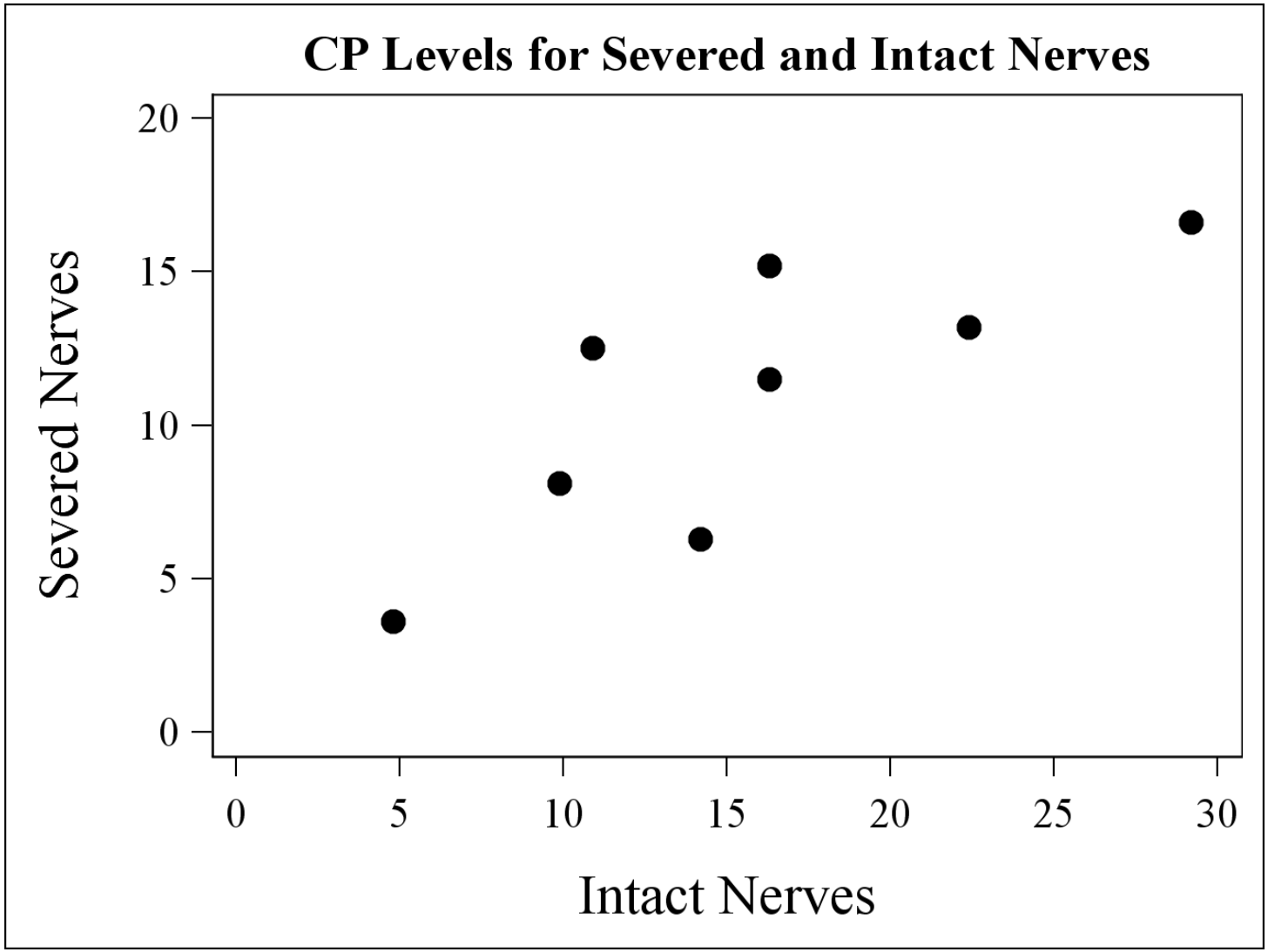
- Randomization
  - Randomly select four monkeys to have nerves severed on the right side
  - The other four monkeys will have nerves severed on the left side

Example: Monkey Nerve Cell Study

CP concentrations in rhesus monkey nerve cells

Monkey	Severed nerves	Intact nerves
1	11.5	16.3
2	3.6	4.8
3	12.5	10.9
4	6.3	14.2
5	15.2	16.3
6	8.1	9.9
7	16.6	29.2
8	13.1	22.4

Example: Monkey Nerve Cell Study



Example: Monkey Nerve Cell Study

- Experimental units: monkeys
- Treatments: severed nerve, control
- Paired data (repeated measurements)
  - One side (severed nerves)
  - Other side (control)
- Used a paired t-test
  - Test $H_0 : \mu_{\text{severed}} = \mu_{\text{control}}$
  - against $H_a : \mu_{\text{severed}} \neq \mu_{\text{control}}$

Example: Monkey Nerve Cell Study

Pair	Difference (severed-control)
1	$d_1 = Y_{11} - Y_{12} = -4.8$
2	$d_2 = Y_{21} - Y_{22} = -1.0$
3	$d_3 = Y_{31} - Y_{32} = 1.6$
4	$d_4 = Y_{41} - Y_{42} = -7.9$
5	$d_5 = Y_{51} - Y_{52} = -1.1$
6	$d_6 = Y_{61} - Y_{62} = -1.8$
7	$d_7 = Y_{71} - Y_{72} = -12.6$
8	$d_8 = Y_{81} - Y_{82} = -9.3$

- Sample mean: $\bar{d} = -4.64$
- Sample variance: $S_d^2 = 23.87$
- Standard error for mean difference: $S_{\bar{d}} = \sqrt{\frac{S_d^2}{n}} = 1.73$

Example: Monkey Nerve Cell Study

- t-statistic:

$$t = \frac{\bar{d} - 0}{S_{\bar{d}}} = -2.685 \text{ on } n - 1 = 7 \text{ df}$$

- Two-sided p-value=0.0313
- Using a type I error level of $\alpha = 0.05$ we can reject the null hypothesis and conclude that CP concentration is different during nerve cell regeneration from normal tissue.
- A 95% confidence interval for the difference in mean CP concentrations for severed and intact nerve cells

$$\bar{d} \pm t_{(n-1),1-\alpha/2}\sqrt{\frac{S_d^2}{n}}$$

Alternative Computation for Sample Variance of Differences:

$$Var(\bar{D}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\frac{\rho\sigma_1\sigma_2}{n}$$

is also estimated as

$$\begin{aligned} S_{\bar{D}}^2 &= \frac{S_1^2}{n} + \frac{S_2^2}{n} - 2\frac{\hat{\rho}S_1\,S_2}{n} \\ &= \frac{20.21}{8} + \frac{57.90}{8} - 2\frac{(.79469)(4.4956)(7.609)}{8} \\ &= 2.97 \end{aligned}$$

$$S_{\bar{d}} = \sqrt{2.97} = 1.72$$

Consequence of Doing the Wrong Test

If the pairing (correlation) is incorrectly ignored we would compute

$$\begin{aligned} S_{\bar{Y}_1-\bar{Y}_2} &= \sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{n}} \\ &= \sqrt{\frac{20.2107}{8}+\frac{57.8971}{8}} = 3.06 \end{aligned}$$

and

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{\bar{Y}_1-\bar{Y}_2}} = -1.48 \text{ on } 11.4 \text{ df}$$

with two-sided p-value=0.1660

Diagnosing Assumptions:

- Model assumptions are:
  - Blocks are independent $\rightarrow$ differences are independent
  - D_i are i.i.d. $N(\mu_d = \mu_1 - \mu_2, \sigma_d^2)$

    where $\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$

Diagnosing Assumptions:

- Independence of differences
  - Examine study to determine if responses from one block could affect responses from any other block.
  - Critical problem if this fails
  - Observations from same block will usually be positively correlated

Diagnosing Assumptions:

- Normal distribution for differences
  - Normal probability plot for differences
  - Effects of non-normality
    - t-test sensitive to outliers
    - t-test sensitive to skewness of the distribution of possible differences
    - If sample size is large, t-test is fairly robust to these problems

Nonparametric Tests

- If smaller sample sizes with non-normal differences
  - Wilcoxon signed rank test
  - Sign test

Wilcoxon Signed Rank Test

- Order the *absolute values* of the differences
- Assign ranks to the differences
- Test statistic: T = sum of ranks with positive differences
- T is approximately normal (for large n) with
  - mean $n(n+1)/4$
  - variance $n(n+1)(2n+1)/24$
- Corrections for tied values (see any text on nonparametric statistics).
- Not quite as good as a t-test for normal data, but much better than a t-test on skewed distributions

Wilcoxon Signed Rank Test: Monkey Example

Pair	|Difference|	rank
1	$\|d_1\| = \|-4.8\| = 4.8$	5
2	$\|d_2\| = \|-1.0\| = 1.0$	1
3	$\|d_3\| = \|1.6\| = 1.6$	3
4	$\|d_4\| = \|-7.9\| = 7.9$	6
5	$\|d_5\| = \|-1.1\| = 1.1$	2
6	$\|d_6\| = \|-1.8\| = 1.8$	4
7	$\|d_7\| = \|-12.6\| = 12.6$	8
8	$\|d_8\| = \|-9.3\| = 9.3$	7

- T = 3 and n=8
- $|Z| = \frac{|T-n(n+1)/4|-0.5}{\sqrt{n(n+1)(2n+1)/24}} = 2.11$ with p-value=0.035
- Exact p-value=0.039

Sign Test

- Based on numbers of positive and negative differences.
- Ties (zero differences) are ignored
- Ignores the sizes of the observed differences
- Not as powerful as Wilcoxon signed rank
- Can be useful for censored data
- p-value obtained from the binomial distribution
  (use normal approximation to binomial for large samples)

Sign Test: Monkey Example

Pair	Difference	sign
1	$d_1 = -4.8$	-
2	$d_2 = -1.0$	-
3	$d_3 = 1.6$	+
4	$d_4 = -7.9$	-
5	$d_5 = -1.1$	-
6	$d_6 = -1.8$	-
7	$d_7 = -12.6$	-
8	$d_8 = -9.3$	-

- S = 1 positive difference out of n=8 pairs
- $H_0 : Pr(+) = Pr(-) = 0.5 \quad \text{vs} \quad H_a : Pr(+) \neq 0.5$
- p-value = $2\left[\binom{8}{0}(0.5)^8 + \binom{8}{1}(0.5)^8\right] = 0.070$

Unit 2

Blocking: RCBD

Block

A group of experimental units that, prior to treatment, are expected to be more like one another (with respect to one or more response variables) than experimental units in general. (In simple words, groups of similar experimental units.)

Randomized Complete Block Design (RCBD)

Experimental design in which separate and completely randomized treatment assignments are made for each of multiple blocks in such a way that all treatments have at least one experimental unit in each block.

Typical RCBD Set-up

- J treatments
- *n* blocks with *J* units in each block
  - Units within each block are similar
  - Within each block, randomly assign *J* treatments to the units so that one experimental unit for each treatment
  - Each block is essentially a repetition of the experiment

NOTE: If blocks have many units (some multiple of J), then we can apply each treatment more than once

Model: (experiments with one unit per treatment per block)

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

- $i = 1, \ldots, n$ indexes blocks
- $j = 1, \ldots, J$ indexes treatments
- τ_j are fixed treatment effects (with $\tau_J = 0$)
- β_i are block effects
  - Could be fixed effects with $\beta_n = 0$
  - Could be random effects with $\beta_i \sim N(0, \sigma_B^2)$
- Additive model (same treatment effects in each block)
- $\epsilon_{ij} \sim N(0, \sigma_e^2)$

ANOVA Table

source of variation	degrees of freedom	sums of squares
blocks	$n-1$	$J\sum_{i=1}^{n}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$
treatments	$J-1$	$n\sum_{j=1}^{J}(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})^2$
error	$(n-1)(J-1)$	$\sum_{i=1}^{n}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i\cdot}-\bar{Y}_{\cdot j}+\bar{Y}_{\cdot\cdot})^2$
total	$nJ-1$	$\sum_{i=1}^{n}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$

Expectations for Mean Squares

- Residual (error) Mean Square: $E(MS_{error}) = \sigma_e^2$
- Fixed Treatment Effects (with $\bar{\tau} = \sum_{j=1}^{J} \tau_j$)

$$E(MS_{treatments}) = \sigma_e^2 + \frac{n}{J-1} \sum_{j=1}^{J} (\tau_j - \bar{\tau})^2$$

- Fixed Blocks (with $\bar{\beta} = \sum_{i=1}^{n} \beta_i$)

$$E(MS_{blocks}) = \sigma_e^2 + \frac{J}{n-1} \sum_{i=1}^{n} (\beta_i - \bar{\beta})^2$$

- Random Blocks

$$E(MS_{blocks}) = \sigma_e^2 + J\sigma_\beta^2$$

Tests for Treatment Effects

- Test the null hypothesis of no treatment effects

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_J$$

against alternative that at least one mean is different.

- Reject H_0 if

$$F = \frac{MS_{treatments}}{MS_{error}} \geq F_{(J-1,(n-1)(J-1)),1-\alpha}$$

Note that

$$\frac{E(MS_{treatments})}{E(MS_{error})} = \frac{\sigma_e^2 + \frac{n}{J-1}\sum_{j=1}^{J}(\tau_j - \bar{\tau})^2}{\sigma_e^2}$$

Test for Block Effects

- Test the null hypothesis of no block effects:
  - For fixed block effects, $H_0 : \beta_1 = \beta_2 = \ldots = \beta_n$
  - For random block effects, $H_0 : \sigma^2_\beta = 0$
- Reject H_o if

$$F = \frac{MS_{blocks}}{MS_{error}} \geq F_{(n-1,(n-1)(J-1)),1-\alpha}$$

- Usually this test is of little interest, because variation among blocks is anticipated.

Example: Penicillin Experiment

$$\left.\begin{array}{l} \textit{Process A} \\ \textit{Process B} \\ \textit{Process C} \\ \textit{Process D} \end{array}\right\} \text{Levels of a “fixed” treatment factor}$$

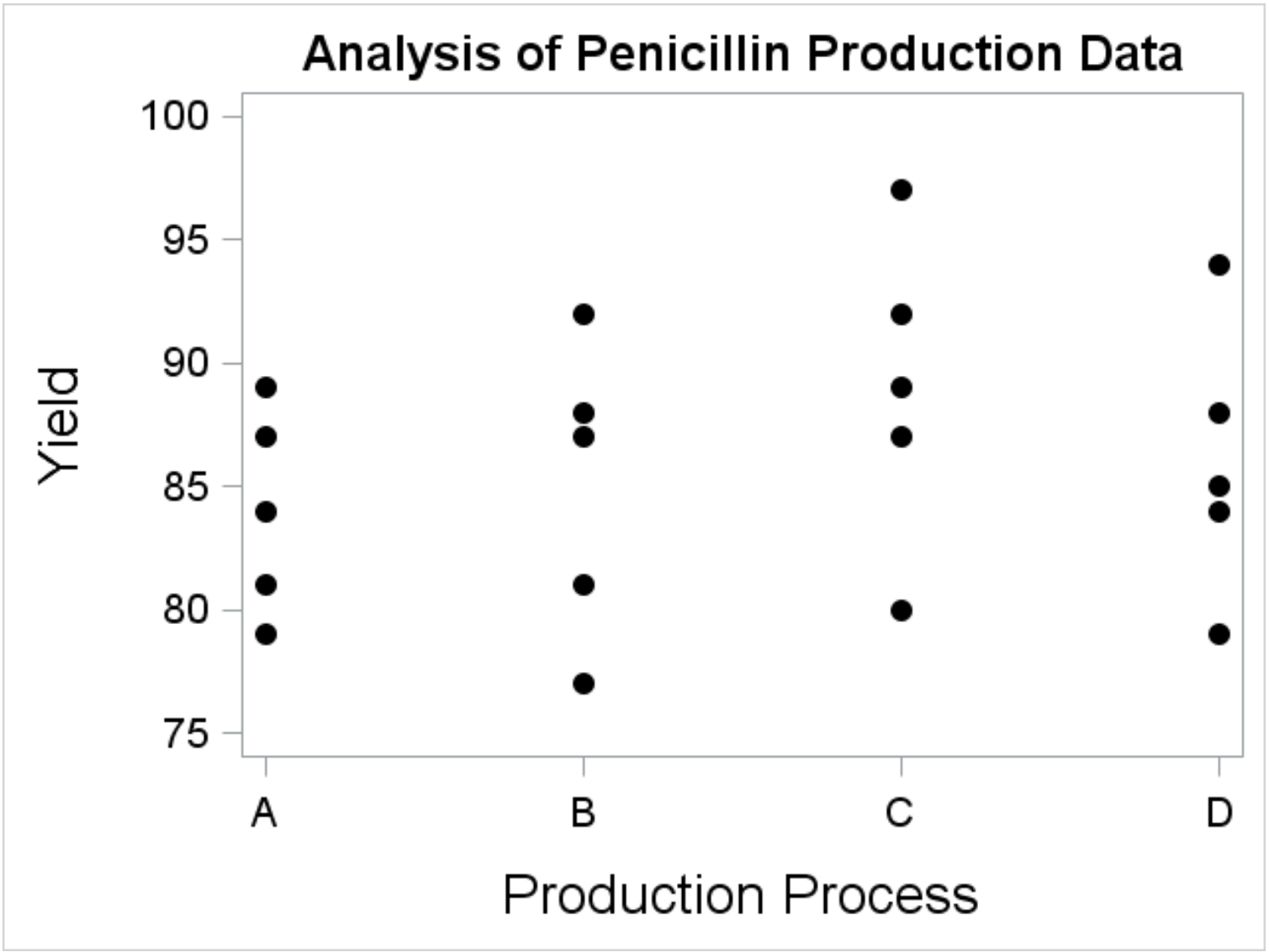
- Blocks correspond to different batches of raw material
- Random sample of five batches from all possible batches
- Split each batch into four parts:
  - run each process on one randomly selected part
  - randomize the order in which the processes are run within each batch
- To repeat this experiment, need to use a different set of batches of raw material

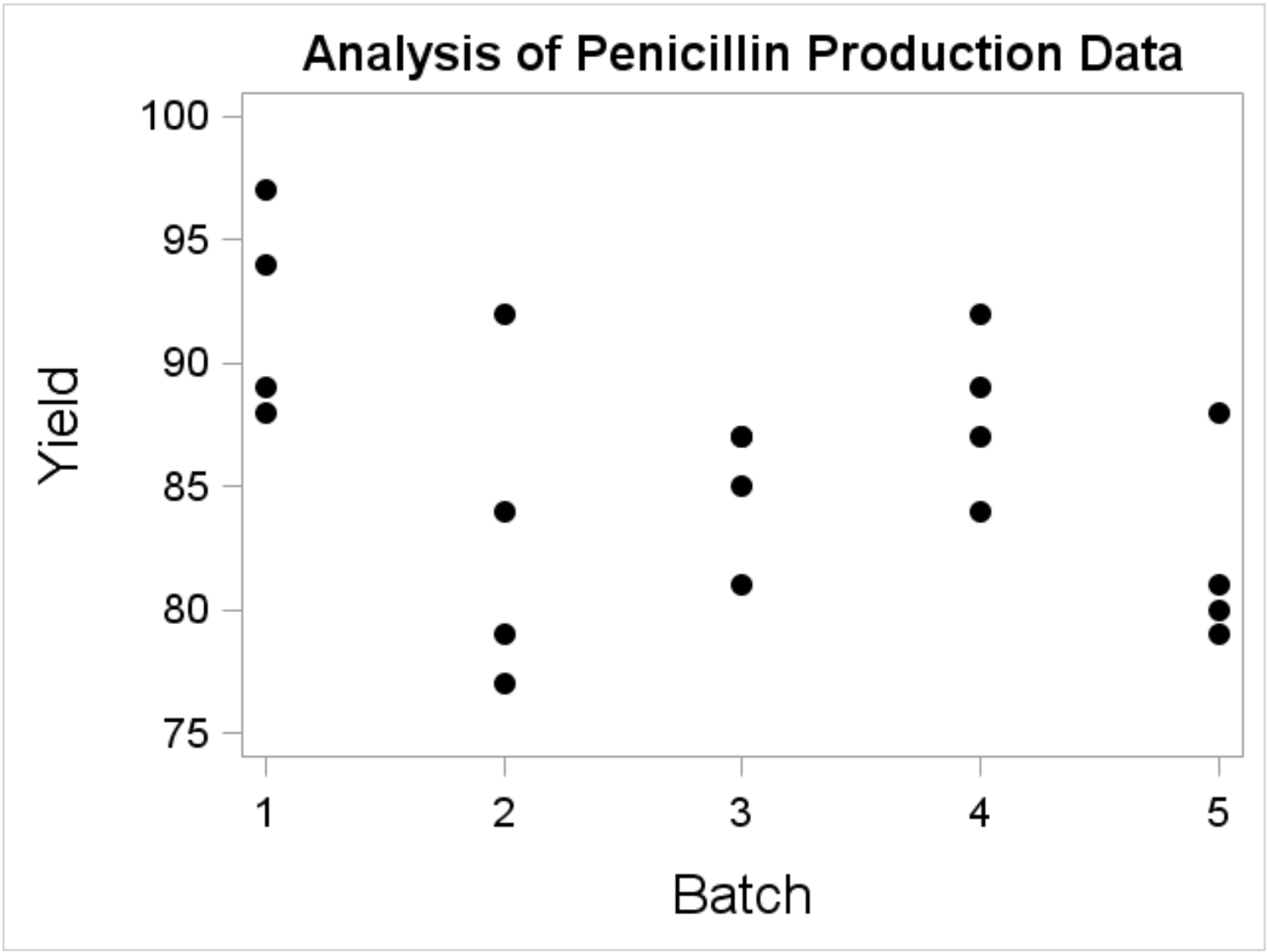
Example: Penicillin Experiment

Data: Yields

	Processes			
Batch	A	B	C	D
1	89	88	97	94
2	84	77	92	79
3	81	87	87	85
4	87	92	89	84
5	79	81	80	88
Means	84	85	86	89

Data Source: Box, Hunter & Hunter (1978), Statistics for Experimenters, Wiley & Sons, New York.


Analysis of Penicillin Production Data
100
95
90
85
80
75
Yield
A
B
C
D
Production Process


Analysis of Penicillin Production Data
Yield
100
95
90
85
80
75
1
2
3
4
5
Batch

Example: Penicillin Experiment

Model:

$Y_{ij} =$	$\mu + \tau_j$	$+\beta_i$	$+\epsilon_{ij}$
$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
Yield for the j^{th} process applied to the i^{th} batch	mean yield for the j^{th} process, averaging across the entire population of possible batches	random batch effect	random error

where $\beta_i \sim N(0, \sigma^2_\beta)$, $\epsilon_{ij} \sim N(0, \sigma^2_e)$, all ϵ_{ij} & β_i are independent

ANOVA

Source	d.f.	SS	MS	F	p-value
Blocks	4	264	66.000	3.50	0.0407
Processes	3	70	23.333	1.24	0.3387
Error	12	226	18.833		
Total	19	560			

- $\hat{\sigma}_e^2 = MS_{error} = 18.833$
- $S_{\bar{Y}_{.j} - \bar{Y}_{.k}} = \sqrt{MS_{error} \frac{2}{n}} = \sqrt{(18.833)\frac{2}{5}} = 2.745$
- MS_{blocks} is an estimate of $E(MS_{blocks}) = \sigma_e^2 + 4\sigma_\beta^2$
- An estimate of σ_β^2 is

$$\frac{MS_{blocks} - MS_{error}}{4} = \frac{66 - 18.833}{4} = 11.792$$

Difference in Estimated Treatment Means

- $Y_{ij} - Y_{ik} = (\mu + \beta_i + \tau_j + \epsilon_{ij}) - (\mu + \beta_i + \tau_k + \epsilon_{ik})$
  $= \tau_j - \tau_k + \epsilon_{ij} - \epsilon_{ik}$
- $Var(Y_{ij} - Y_{ik}) = Var(\epsilon_{ij} - \epsilon_{ik}) = 2\sigma_e^2$
- Similarly,

$$
\begin{aligned}
Var(\bar{Y}_{.j} - \bar{Y}_{.k}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_{ij} - \frac{1}{n}\sum_{i=1}^{n} Y_{ik}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n}(Y_{ij} - Y_{ik})\right) \\
&= Var\left(\frac{1}{n}\sum_{i=1}^{n}(\tau_j - \tau_k + \epsilon_{ij} - \epsilon_{ik})\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(\tau_j - \tau_k + \epsilon_{ij} - \epsilon_{ik})\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} 2\sigma_e^2\right) = \frac{2\sigma_e^2}{n}
\end{aligned}
$$

Example: Penicillin Experiment

Model:
$$\begin{aligned} \mu_j = E(Y_{ij}) &= E(\mu + \tau_j + \beta_i + e_{ij}) \\ &= \mu + \tau_j + E(\beta_i) + E(e_{ij}) \\ &= \mu + \tau_j \qquad i = 1, 2, 3, 4 \end{aligned}$$

= mean yield for j^{th} process, averaging across all possible batches

PROC GLM and PROC MIXED in SAS fit a model with $\tau_4 = 0$. Then

- $\mu = \mu_4$ is the mean yield for process D
- $\tau_j = \mu_j - \mu_4 \qquad j = 1, 2, 3, 4.$

Contrasts

- Variances for contrasts among the treatment means are computed in a similar manner:
  - The design is balanced (each treatment occurs the same number of times in each block)
  - Block effects cancel out because $\sum_j c_j = 0$
  - variance for contrast estimate: $Var(\sum_j c_j \bar{Y}_{\cdot j}) = \sigma_e^2 \sum_j c_j^2/n$, estimated by $MS_{error} \sum_j c_j^2/n$
- Can use orthogonal contrasts to partition treatment sums of squares

Variance-Covariance Structure:

$$\begin{aligned} Var(Y_{ij}) &= Var(\mu + \tau_j + \beta_i + e_{ij}) \\ &= Var(\beta_i + e_{ij}) \\ &= Var(\beta_i) + Var(e_{ij}) \\ &= \sigma^2_\beta + \sigma^2_e \qquad \text{for all } (i,j) \end{aligned}$$

is the variance of a measurement of yield for one run of a process with a random sample of a batch

Variance-Covariance Structure:

For runs of different processes on the same batch:

$$
\begin{aligned}
&Cov(Y_{ij}, Y_{ik}) \\
&= Cov(\mu + \tau_j + \beta_i + e_{ij}, \mu + \tau_k + \beta_i + e_{ik}) \\
&= Cov(\beta_i + e_{ij},\ \beta_i + e_{ik}) \\
&= Cov(\beta_i, \beta_i) + Cov(\beta_i,\ e_{ij}) + Cov(e_{ik}, \beta_i) + Cov(e_{ij},\ e_{ik}) \\
&= Var(\beta_i) = \sigma^2_\beta \qquad \text{for all } j \neq k
\end{aligned}
$$

Variance-Covariance Structure:

Correlation Among Yields for Runs on the Same Batch:

$$\rho = \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{Var(Y_{ij})Var(Y_{ik})}}$$

$$= \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_e} \text{ for } j \neq k$$

Results for runs on different batches are independent:

$$Cov(Y_{ij}, Y_{\ell k}) = 0 \quad \text{for } i \neq \ell$$

Variance-Covariance Matrix for the Four Runs on the Same Batch

$$Var\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} = \begin{bmatrix} \sigma^2_\beta + \sigma^2_e & \sigma^2_\beta & \sigma^2_\beta & \sigma^2_\beta \\ \sigma^2_\beta & \sigma^2_\beta + \sigma^2_e & \sigma^2_\beta & \sigma^2_\beta \\ \sigma^2_\beta & \sigma^2_\beta & \sigma^2_\beta + \sigma^2_e & \sigma^2_\beta \\ \sigma^2_\beta & \sigma^2_\beta & \sigma^2_\beta & \sigma^2_\beta + \sigma^2_e \end{bmatrix}$$

$$= \sigma^2_\beta \, J + \sigma^2_e \, I$$

- J is a matrix of all ones
- I is the identity matrix
- This is a *compound symmetric* covariance structure

Model for Random Blocks

Write this model in the form $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$

$$\begin{bmatrix} Y_{11}\\ Y_{12}\\ Y_{13}\\ Y_{14}\\ Y_{21}\\ Y_{22}\\ Y_{23}\\ Y_{24}\\ Y_{31}\\ Y_{32}\\ Y_{33}\\ Y_{34}\\ Y_{41}\\ Y_{42}\\ Y_{43}\\ Y_{44}\\ Y_{51}\\ Y_{52}\\ Y_{53}\\ Y_{54} \end{bmatrix} = \begin{bmatrix} 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\ 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\ 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\ 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\ 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1 \end{bmatrix} \begin{bmatrix} \mu\\ \tau_1\\ \tau_2\\ \tau_3\\ \tau_4 \end{bmatrix} + \begin{bmatrix} 1&0&0&0&0\\ 1&0&0&0&0\\ 1&0&0&0&0\\ 1&0&0&0&0\\ 0&1&0&0&0\\ 0&1&0&0&0\\ 0&1&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&1&0&0\\ 0&0&1&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&1&0\\ 0&0&0&1&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 0&0&0&0&1\\ 0&0&0&0&1\\ 0&0&0&0&1 \end{bmatrix} \begin{bmatrix} \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\\ \beta_5 \end{bmatrix} + \begin{bmatrix} e_{11}\\ e_{12}\\ e_{13}\\ e_{14}\\ e_{21}\\ e_{22}\\ e_{23}\\ e_{24}\\ e_{31}\\ e_{32}\\ e_{33}\\ e_{34}\\ e_{41}\\ e_{42}\\ e_{43}\\ e_{44}\\ e_{51}\\ e_{52}\\ e_{53}\\ e_{54} \end{bmatrix}$$

Model for Random Blocks

Impose the SAS constraint $\tau_4 = 0$

$$
\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{34} \\ Y_{41} \\ Y_{42} \\ Y_{43} \\ Y_{44} \\ Y_{51} \\ Y_{52} \\ Y_{53} \\ Y_{54} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \\ e_{41} \\ e_{42} \\ e_{43} \\ e_{44} \\ e_{51} \\ e_{52} \\ e_{53} \\ e_{54} \end{bmatrix}
$$

ANOVA

For this constrained model, the least squares estimator for

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}$$

is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Multiple Comparisons

Tukey-Kramer HSD (*Honest Significant Difference*)

- Declare a significant difference in treatment means if

$$|\bar{Y}_i - \bar{Y}_j| \geq \frac{1}{\sqrt{2}} q_{(r, df_{error}, 1-\alpha)} \sqrt{MS_{error}\left(\frac{1}{n} + \frac{1}{n}\right)}$$

- For the penicillin study

$$HSD = \frac{1}{\sqrt{2}}(4.199)\sqrt{18.833\left(\frac{1}{5} + \frac{1}{5}\right)} = 8.15$$

- Order sample means from smallest to largest:

Process A	Process B	Process D	Process C
84	85	86	89

Fixed or Random Blocks?

- F-test for treatment effects is the same whether blocks are fixed or random
- $Var(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot k}) = \sigma_e^2(\frac{2}{n})$ is the same whether blocks are fixed or random. Use

$$S_{\bar{Y}_{\cdot j} - \bar{Y}_{\cdot k}} = \sqrt{MS_{error}\frac{2}{n}}$$

Fixed or Random Blocks?

- Standard error of a single treatment mean is different:
  - Fixed blocks: $Var(\bar{Y}_{.j}) = \sigma_e^2/n$

    Use $S_{\bar{Y}_{.j}} = \sqrt{MS_{error}\frac{1}{n}}$ with df_{error}
  - Random blocks: $Var(\bar{Y}_{.j}) = (\sigma_e^2 + \sigma_\beta^2)/n$

    Use

$$S_{\bar{Y}_{.j}} = \sqrt{\frac{(J-1)MS_{error} + MS_{blocks}}{Jn}}$$

    with $df = \frac{[\frac{J-1}{J}MS_{error} + \frac{1}{J}MS_{blocks}]^2}{\frac{[\frac{J-1}{J}MS_{error}]^2}{df_{error}} + \frac{[\frac{1}{J}MS_{blocks}]^2}{df_{blocks}}}$

Questions?

Contact me:

Email: dmommen@iastate.edu

Student Office Hours: Thursdays @ 10-11 am