One-Factor Designs

Dr. John Mellor-Crummey

Department of Computer Science
Rice University

johnmc@cs.rice.edu

Goals for Today

Understand One-factor Designs

- Motivation
- Properties
- Computing effects
- Estimating experimental errors
- Allocating variation
- Analyzing variance
- Confidence intervals

One-factor Designs

- **Why? Compare several alternatives of one categorical variable**
  - —several processors
  - —several caching schemes
  - —several garbage collection strategies
- **No limits on number of levels the factor can take**
- **Model:** $\boxed{y_{ij} = \mu + \alpha_j + e_{ij}}$
  - —y_{ij} - i^{th} response with factor at level j
  - —μ - mean response
  - —α_j - effect of alternative j

    effects are computed so that their sum is 0: $\sum_j \alpha_j = 0$
  - —e_{ij} - error term

Computing Effects for a One-factor Design

- **Measured data = r observations for each of a alternatives**
  - — **r x a matrix**
  - — **r observations for an alternative arranged in a column**
- **Substituting responses into our model $y_{ij} = \mu + \alpha_j + e_{ij}$ yields ar equations.**
- **Summing them, we get**

$$\sum_{i=1}^{r}\sum_{j=1}^{a} y_{ij} = ar\mu + r\underbrace{\sum_{j=1}^{a}\alpha_j}_{\text{0 by design}} + \underbrace{\sum_{i=1}^{r}\sum_{j=1}^{a} e_{ij}}_{\text{assume mean error = 0}}$$

- **Thus,**

$$\sum_{i=1}^{r}\sum_{j=1}^{a} y_{ij} = ar\mu + 0 + 0$$

and

$$\boxed{\mu = \frac{1}{ar}\sum_{i=1}^{r}\sum_{j=1}^{a} y_{ij}}$$

grand mean, also denoted $\bar{y}_{..}$

Mean Effect for an Alternative

- **Mean effect for an alternative is a column mean**

$$\bar{y}_{.j} = \frac{1}{r}\sum_{i=1}^{r} y_{ij}$$

- **Substituting for y_{ij}**

assume mean error = 0

$$\bar{y}_{.j} = \frac{1}{r}\sum_{i=1}^{r}(\mu + \alpha_j + e_{ij}) = \mu + \alpha_j + \frac{1}{r}\sum_{i=1}^{r} e_{ij} = \mu + \alpha_j$$

- **Therefore, we can estimate mean effect for an alternative α_j**

$$\alpha_j = \bar{y}_{.j} - \mu = \bar{y}_{.j} - \bar{y}_{..}$$

Tabular Computation of Mean Effect

- **Code size comparison study**
  - —three processors R,V,Z
  - —measured number of bytes to code a workload
    - – 5 independent coders for each machine (1 for each of ar entries)
    - – if row entries not independent, need 2 factor analysis (next lecture)

	R	V	Z	
	144.0	101.0	130.0	r replications
	120.0	144.0	180.0	
	176.0	211.0	141.0	
	288.0	288.0	374.0	
	144.0	72.0	302.0	
Column Sum	872.0	816.0	1127.0	2815.0 Grand Sum
Column Mean	174.4	163.2	225.4	187.7 Grand Mean
Column Effect	-13.3	-24.5	37.7	
	α_1	α_2	α_3	

$$\underbrace{\alpha_1 \quad \alpha_2 \quad \alpha_3}_{\text{a alternatives}}$$

- **Interpretation**
  - —avg processor requires 187.7B storage
  - —R uses 13.3B < avg; V uses 24.5B < avg; Z uses 37.7B > avg

Estimating Experimental Errors

- **Predicted response of j[th] alternative**

$$\hat{y}_j = \mu + \alpha_j$$

- **Prediction error** $e_{ij} = y_{ij} - \hat{y}_j = y_{ij} - \mu - \alpha_j$

—**mean error for column and grand mean all will be 0**

e_{ij} = y_{ij} - μ - α_j

R	V	Z
-30.4	-62.2	-95.4
-54.4	-19.2	-45.4
1.6	47.8	-84.4
113.6	124.8	148.6
-30.4	-91.2	76.6

=

R	V	Z
144.0	101.0	130.0
120.0	144.0	180.0
176.0	211.0	141.0
288.0	288.0	374.0
144.0	72.0	302.0

-

R	V	Z
187.7	187.7	187.7
187.7	187.7	187.7
187.7	187.7	187.7
187.7	187.7	187.7
187.7	187.7	187.7

-

R	V	Z
-13.3	-24.5	37.7
-13.3	-24.5	37.7
-13.3	-24.5	37.7
-13.3	-24.5	37.7
-13.3	-24.5	37.7

- **Estimate variance of errors from sum of squared errors**

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{a} e_{ij}^2$$

$$SSE = (-30.4)^2 + (-54.4)^2 + \ldots + (76.6)^2 = 94{,}365.2$$

Allocating Variation

- **Total variation of y can be allocated to the factor and errors**
- **First, square model equation**

$$y_{ij}^2 = \mu^2 + \alpha_j^2 + e_{ij}^2 + 2\mu\alpha_j + 2\mu e_{ij} + 2\alpha_j e_{ij}$$

- **Adding corresponding terms of ar equations**

$$\sum_{i,j} y_{ij}^2 = \sum_{i,j} \mu^2 + \sum_{i,j} \alpha_j^2 + \sum_{i,j} e_{ij}^2 + \text{cross product terms}$$

SSY = SS0 + SSA + SSE

all add to 0 because

$$\sum_j \alpha_j = 0, \quad \sum_{i,j} e_{ij} = 0$$

- **Total variation**

$$SST = \sum_{i,j} (y_{ij} - \bar{y}_{..})^2 = \sum_{i,j} y_{ij}^2 - ar\bar{y}_{..}^2 = SSY - SS0 = SSA + SSE$$

Allocating Variation for Code Size Study

	R	V	Z	
	144.0	101.0	130.0	
	120.0	144.0	180.0	
	176.0	211.0	141.0	
	288.0	288.0	374.0	
	144.0	72.0	302.0	
Column Sum	872.0	816.0	1127.0	2815.0 Grand Sum
Column Mean	174.4	163.2	225.4	187.7 Grand Mean
Column Effect	-13.3	-24.5	37.7	
	α_1	α_2	α_3	

$$SSY = 144^2 + 120^2 + \ldots + 302^2 = 633{,}639$$

$$SS0 = ar\mu^2 = (3)(5)(187.7)^2 = 528{,}281.7$$

$$SSA = r\sum_j \alpha_j^2 = 5((-13.3)^2 + (-24.5)^2 + (37.6)^2) = 10{,}992.1$$

$$SST = SSY - SS0 = 633{,}639.0 - 528{,}281.7 = 105{,}357.3$$

$$SSE = SST - SSA = 105{,}357.3 - 10{,}992.1 = 94{,}365.2$$

% variation explained by processors = 100 * SSA/SST
= 100*(10,922.1/105,357.3) = 10.4%

Analyzing Variance I

- To determine if a factor is *statistically significant*
  - —compare its contribution with that of errors
- Unexplained variation is high →
  variation due to factor may be statistically insignificant
- Analysis of Variance (ANOVA)
- Degrees of freedom

$$SSY = SS0 + SSA + SSE$$

$$\underbrace{ar}_{\text{sum of ar terms: all independent}} = \underbrace{1}_{\text{single term } \mu^2\text{: repeated ar times}} + \underbrace{(a-1)}_{\text{sum of } (\alpha_j)^2\text{: a-1 independent } (\sum\alpha_j = 0)} + \underbrace{a(r-1)}_{\text{sum of } (e_{ij})^2\text{: a(r-1) independent since } e_{.j} \text{ across all replications sum to 0}}$$

Analyzing Variance II

- **Degrees of freedom**

  SSY = SS0 + SSA + SSE

  ar = 1 + (a-1) + a(r-1)

- **Mean square values**

  MSA = SSA/(a-1)

  MSE = SSE/(a(r-1))

  $$s_e = \sqrt{MSE}$$

- **Is effect of alternative significant?**
  - —if errors are normally distributed, then SSE & SSA have chi-square distributions
  - —ratio of [SSA/(a-1)]/[SSE/(a(r-1))] = MSA/MSE has F distribution
    - – numerator has ν_A = (a-1) degrees of freedom
    - – denominator has ν_E = a(r-1) degrees of freedom
  - —Significant if $[SSA/\nu_A]/[SSE/\nu_E] > F_{[1-\alpha;\ \nu A;\ \nu E]} = F_{[1-\alpha;\ a-1;\ a(r-1)]}$

Analyzing Variance for Code Size Study

	R	V	Z	
	144.0	101.0	130.0	
	120.0	144.0	180.0	
	176.0	211.0	141.0	
	288.0	288.0	374.0	
	144.0	72.0	302.0	
Column Sum	872.0	816.0	1127.0	2815.0 Grand Sum
Column Mean	174.4	163.2	225.4	187.7 Grand Mean
Column Effect	-13.3	-24.5	37.7	
	α_1	α_2	α_3	

observed difference mostly due to experimental error not significant difference among processors

$$SSY = 144^2 + 120^2 + \ldots + 302^2 = 633{,}639$$

$$SS0 = ar\mu^2 = (3)(5)(187.7)^2 = 528{,}281.7$$

$$SSA = r\sum_j \alpha_j^2 = 5((-13.3)^2 + (-24.5)^2 + (37.6)^2) = 10{,}992.1$$

$$SST = SSY - SS0 = 633{,}639.0 - 528{,}281.7 = 105{,}357.3$$

$$SSE = SST - SSA = 105{,}357.3 - 10{,}992.1 = 94{,}365.2$$

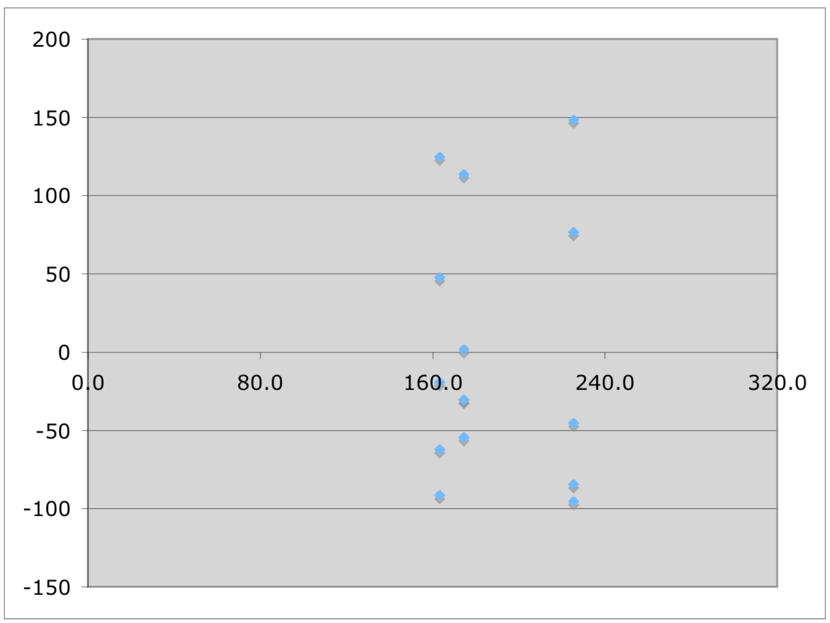
MSA = SSA/2 = 5496.1; MSE = SSE/(3(5-1)) = 7863.8
MSA/MSE = .7; $F_{[.90;\ 2;\ 12]}$ = 2.81; (MSA/MSE)/ $F_{[.90;\ 2;\ 12]}$ < 1

Assumptions of One-factor Experiments

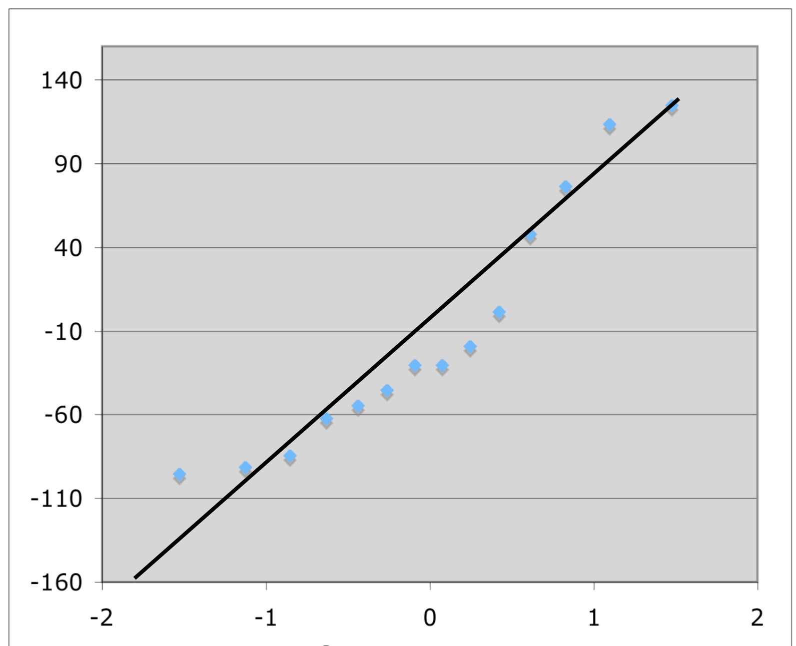
- **Effects of various factors are additive**
- **Errors are additive**
- **Errors are independent of factor levels**
- **Errors are normally distributed**
- **Errors have the same variance for all factor levels**

Visual Diagnostic Tests

- **Same as discussed in earlier lectures**
  - **—quantile-quantile plot of residuals vs. predicted response**
    - **if plot approximately linear, can assume normality**
  - **—scatter plot of residuals vs. predicted response**
    - **confirm no trend in residuals or their spread**
    - **magnitude of errors smaller than response by order of magnitude → ignore trend**



no trend, but residuals not small!



slightly S shaped -> short tails

Variance for Effects

- **Estimated model parameters are random variables**
  - —**based on one sample**
  - —**estimates from another sample would be different**
- **Variance of model parameters can be estimated from errors**

parameter	estimate	variance
μ	$\bar{y}_{..}$	s_e^2 / ar
α_j	$\bar{y}_{.j} - \bar{y}_{..}$	$s_e^2 (a-1)/ar$
$\mu+\alpha_j$	$\bar{y}_{.j}$	s_e^2 / r
$\sum_{j=1}^{a} h_j \alpha_j$, $\sum_{j=1}^{a} h_j = 0$	$\sum_{j=1}^{a} h_j \bar{y}_{.j}$	$\sum_{j=1}^{a} s_e^2 h_j^2 / ar$
s_e^2	$\sum e_{ij}^2 /(a(r-1))$	

Variance for μ

- **Expression for μ in terms of random variables y_{ij}'s**

$$\mu = \frac{1}{ar}\sum_{i=1}^{a}\sum_{j=1}^{r} y_{ij}$$

- **What is the coefficient a_{ij} for each y_{ij}?**

$$a_{ij} = \left\{ \frac{1}{ar} \right.$$

$$\mathrm{Var}(\mu) = Var\left(\frac{1}{ar}\sum_{i=1}^{a}\sum_{j=1}^{r} y_{ij}\right)$$

Assuming
errors normally distributed
zero mean, variance $(\sigma_e)^2$
What is the variance for μ?

$$\sigma_\mu^2 = \left(ar\left(\frac{1}{ar}\right)^2\right)\sigma_e^2$$

$$= \frac{1}{ar}\sigma_e^2$$

Variance for α_j

- **Expression for α_j in terms of random variables y_{ij}'s**

$$\alpha_j = \bar{y}_{.j} - \mu = \bar{y}_{.j} - \bar{y}_{..} = \frac{1}{r}\sum_{i=1}^{r} y_{ij} - \frac{1}{ar}\sum_{i=1}^{r}\sum_{j=1}^{a} y_{ij}$$

- **What is the coefficient a_{ikj} for each y_{ik} for α_j?**

$$a_{ikj} = \begin{cases} \dfrac{1}{r} - \dfrac{1}{ar} & k = j \\ -\dfrac{1}{ar} & \text{otherwise} \end{cases}$$

Assuming
errors normally distributed
zero mean, variance $(\sigma_e)^2$
What is the variance for α_j?

- **Var of error $e_{\alpha j}$ for α_j**

$$Var\left(e_{\alpha_j}\right) = Var\left(\sum_{i=1}^{r}\sum_{k=1}^{a} a_{ikj} y_{ik}\right)$$

$$\sigma^2_{e_{\alpha_j}} = \left(r\left(\frac{1}{r} - \frac{1}{ar}\right)^2 + (ar - r)\left(\frac{1}{ar}\right)^2 \right)\sigma_e^2$$

$$= \frac{(a-1)}{ar}\sigma_e^2$$

Linear Contrasts

- **Linear combinations of effects**

$$\sum_{j=1}^{a} h_j \alpha_j \qquad \text{where} \qquad \sum_{j=1}^{a} h_j = 0$$

$$\text{mean} = \sum_{j=1}^{a} h_j \bar{y}_{.j} \qquad \text{variance} = \sum_{j=1}^{a} h_j^2 s_e^2 / ar$$

Confidence Intervals for Effects

- Compute confidence intervals with t values read out of table at $a(r-1)$ degrees of freedom (DOF of errors)