Performance Evaluation:

One-factor Experiments

Hongwei Zhang

http://www.cs.wayne.edu/~hzhang

Experiments should be reproducible: they should all fail in the same way.

--- Finagle's Rule

Acknowledgement: this lecture is partially based on the slides of Dr. Raj Jain.

Outline

- Model & Computation of Effects
- Estimating Experimental Errors
- Allocation of Variation
- ANOVA Table and F-Test
- Confidence Intervals for Effects
- Unequal Sample Sizes
- Visual Diagnostic Tests

Outline

- Model & Computation of Effects
- Estimating Experimental Errors
- Allocation of Variation
- ANOVA Table and F-Test
- Confidence Intervals for Effects
- Unequal Sample Sizes
- Visual Diagnostic Tests

One-factor experiments: Model

- Used to compare alternatives of a single categorical variable.

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

For example, several processors, several caching schemes

r	=	Number of replications
y_{ij}	=	ith response with jth alternative
μ	=	mean response
α_j	=	Effect of alternative j
e_{ij}	=	Error term

$$\sum \alpha_j = 0$$

Computation of effects

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

$$= ar\mu + 0 + 0$$

Assuming error sum is 0

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

Computation of effects (contd.)

$$
\begin{aligned}
\bar{y}_{.j} &= \frac{1}{r}\sum_{i=1}^{r} y_{ij} \\
&= \frac{1}{r}\sum_{i=1}^{r} (\mu + \alpha_j + e_{ij}) \\
&= \frac{1}{r}\left(r\mu + r\alpha_j + \sum_{i=1}^{r} e_{ij}\right) \\
&= \mu + \alpha_j + 0
\end{aligned}
$$

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

Example:
code size on three diff. processors R, V, Z

R	V	Z
144	101	130
120	144	180
176	211	141
288	288	374
144	72	302

Entries in a row are unrelated.

(Otherwise, need a two factor analysis.)

Example (contd.)

	R	V	Z	
	144	101	130	
	120	144	180	
	176	211	141	
	288	288	374	
	144	72	302	
Col Sum	$\Sigma\ y_{.1}$ =872	$\Sigma\ y_{.2}$ =816	$\Sigma\ y_{.3}$ =1127	$\Sigma\ y_{..}$ = 2815
Col Mean	$\bar{y}_{.1}$ =174.4	$\bar{y}_{.2}$ =163.2	$\bar{y}_{.3}$ =225.4	$\mu=\bar{y}_{..}$=187.7
Col Effect	$\alpha_1=\bar{y}_{.1}-\bar{y}_{..}$ =-13.3	$\alpha_2=\bar{y}_{.2}-\bar{y}_{..}$ =-24.5	$\alpha_3=\bar{y}_{.3}-\bar{y}_{..}$ =37.7	

- Average processor requires 187.7 bytes of storage.
- The effects of the processors R, V, and Z are -13.3, -24.5, and 37.7, respectively. That is,
  - R requires 13.3 bytes less than an average processor
  - V requires 24.5 bytes less than an average processor, and
  - Z requires 37.7 bytes more than an average processor.

Outline

- Model & Computation of Effects
- Estimating Experimental Errors
- Allocation of Variation
- ANOVA Table and F-Test
- Confidence Intervals for Effects
- Unequal Sample Sizes
- Visual Diagnostic Tests

Estimating experimental errors

- Estimated response for jth alternative:

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

- Error:

$$e_{ij} = y_{ij} - \hat{y}_j$$

- Sum of squared errors (SSE):

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

Example: for the processor-example

$$\begin{bmatrix} 144 & 101 & 130 \\ 120 & 144 & 180 \\ 176 & 211 & 141 \\ 288 & 288 & 374 \\ 144 & 72 & 302 \end{bmatrix} = \begin{bmatrix} 187.7 & 187.7 & 187.7 \\ 187.7 & 187.7 & 187.7 \\ 187.7 & 187.7 & 187.7 \\ 187.7 & 187.7 & 187.7 \\ 187.7 & 187.7 & 187.7 \end{bmatrix}$$

$$+ \begin{bmatrix} -13.3 & -24.5 & 37.7 \\ -13.3 & -24.5 & 37.7 \\ -13.3 & -24.5 & 37.7 \\ -13.3 & -24.5 & 37.7 \\ -13.3 & -24.5 & 37.7 \end{bmatrix}$$

$$+ \begin{bmatrix} -30.4 & -62.2 & -95.4 \\ -54.4 & -19.2 & -45.4 \\ 1.6 & 47.8 & -84.4 \\ 113.6 & 124.8 & 148.6 \\ -30.4 & -91.2 & 76.6 \end{bmatrix}$$

e_{ij}

$$\text{SSE} = (-30.4)^2 + (-54.4)^2 + \cdots + (76.6)^2 = 94365.20$$

Outline

- Model & Computation of Effects
- Estimating Experimental Errors
- Allocation of Variation
- ANOVA Table and F-Test
- Confidence Intervals for Effects
- Unequal Sample Sizes
- Visual Diagnostic Tests

Allocation of variation

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

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

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

0

$$\text{SSY} = \text{SS0} + \text{SSA} + \text{SSE}$$

$$\text{SS0} = \sum_{i=1}^{r} \sum_{j=1}^{a} \mu^2 = ar\mu^2$$

Allocation of variation (contd.)

$$\text{SSA} = \sum_{i=1}^{r}\sum_{j=1}^{a}\alpha_j^2 = r\sum_{j=1}^{a}\alpha_j^2$$

- Total variation of y (SST):

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

Example: for the processor-example

$$
\begin{aligned}
\text{SSY} &= 144^2 + 120^2 + \cdots + 302^2 = 633639 \\
\text{SS0} &= ar\mu^2 \\
&= 3 \times 5 \times (187.7)^2 = 528281.7 \\
\text{SSA} &= r \sum_j \alpha_j^2 \\
&= 5[(-13.3)^2 + (-24.5)^2 + (37.6)^2] \\
&= 10992.1 \\
\text{SST} &= \text{SSY} - \text{SS0} \\
&= 633639.0 - 528281.7 = 105357.3 \\
\text{SSE} &= \text{SST} - \text{SSA} \\
&= 105357.3 - 10992.1 = 94365.2
\end{aligned}
$$

Example (contd.)

Percent variation explained by processors $= 100 \times \frac{10992.13}{105357.3} = 10.4\%$

89.6% of variation in code size is due to experimental errors (programmer differences).

Is 10.4% statistically significant?

Outline

- Model & Computation of Effects
- Estimating Experimental Errors
- Allocation of Variation
- ANOVA Table and F-Test
- Confidence Intervals for Effects
- Unequal Sample Sizes
- Visual Diagnostic Tests

Analysis of variance (ANOVA)

- Importance ≠ Significance
- Important ⇒ Explains a high percent of variation
- Significance
  ⇒ High contribution to the variation compared to that by errors.
- Degree of freedom
  = Number of independent values required to compute

SSY	=	SS0	+	SSA	+	SSE
ar	=	1	+	(a-1)	+	a(r-1)

Note that the degrees of freedom also add up.

F-Test

- Purpose: To check if SSA is *significantly* greater than SSE.
- Errors are normally distributed $\Rightarrow$ SSE and SSA have chi-square distributions.

  The ratio $(SSA/\nu_A)/(SSE/\nu_e)$ has an F distribution.

  where ν_A=a-1 = degrees of freedom for SSA

  ν_e=a(r-1) = degrees of freedom for SSE

  Computed ratio > $F_{[1-\alpha;\, \nu_A,\, \nu_e]}$ $\Rightarrow$
  SSA is significantly higher than SSE.

  SSA/ν_A is called mean square of A or (MSA).

  Similary, MSE=SSE/ν_e

ANOVA table for one-factor experiments

Compo-nent	Sum of Squares	%Variation	DF	Mean Square	F-Comp.	F-Table
y	SSY=$\sum y_{ij}^2$		ar			
$\bar{y}_{..}$	SS0=$ar\mu^2$		1			
y-$\bar{y}_{..}$	SST=SSY-SS0	100	ar-1			
A	SSA $= r\Sigma\ \alpha_i^2$	$100\left(\frac{\text{SSA}}{\text{SST}}\right)$	a-1	$\text{MSA} = \frac{\text{SSA}}{a-1}$	$\frac{\text{MSA}}{\text{MSE}}$	$F\ [1-\alpha; a-1, a(r-1)]$
e	SSE=SST- SSA	$100\left(\frac{\text{SSE}}{\text{SST}}\right)$	$a(r-1)$	$\text{MSE} = \frac{\text{SSE}}{a(r-1)}$		

Example: for the processor-example

Compo-nent	Sum of Squares	%Variation	DF	Mean Square	F-Comp.	F-Table
y	633639.00					
$y_{..}$	528281.69					
y-$y_{..}$	105357.31	100.0%	14			
A	10992.13	10.4%	2	5496.1	0.7	2.8
Errors	94365.20	89.6%	12	7863.8		

$$s_e = \sqrt{\text{MSE}} = \sqrt{7863.77} = 88.68$$

- Computed F-value < F from Table

  $\Rightarrow$ The variation in the code sizes is mostly due to experimental errors and not because of any significant difference among the processors.

E.g., programmer difference

Outline

- Model & Computation of Effects
- Estimating Experimental Errors
- Allocation of Variation
- ANOVA Table and F-Test
- Confidence Intervals for Effects
- Unequal Sample Sizes
- Visual Diagnostic Tests

Confidence intervals for effects

- Estimates are random variables

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	$\frac{\sum e_{ij}^2}{a(r-1)}$	

Degrees of freedom for errors = a(r-1)

- For the confidence intervals, use t values at r(a-1) degrees of freedom.
- Mean responses: $\hat{y}_j = \mu + \alpha_j$
- Contrasts $\sum h_j\ \alpha_j$: Use for α_1-α_2

Example: for the processor-example

$$\text{Error variance } s_e^2 = \frac{94365.2}{12} = 7863.8$$

$$\begin{aligned} \text{Std Dev of errors} &= \sqrt{(\text{Var. of errors})} \\ &= 88.7 \end{aligned}$$

$$\text{Std Dev of } \mu = s_e/\sqrt{ar} = 88.7/\sqrt{15} = 22.9$$

$$\begin{aligned} \text{Std Dev of } \alpha_j &= s_e/\sqrt{\{(a-1)/(ar)\}} \\ &= 88.7/\sqrt{(2/15)} = 32.4 \end{aligned}$$

Example (contd.)

- For 90% confidence, $t_{[0.95;\ 12]} = 1.782$.
- 90% confidence intervals:

$$\begin{aligned}
\mu &= 197.7 \mp (1.782)(22.9) = (146.9, 228.5) \\
\alpha_1 &= -13.3 \mp (1.782)(32.4) = (-71.0, 44.4) \\
\alpha_2 &= -24.5 \mp (1.782)(32.4) = (-82.2, 33.2) \\
\alpha_3 &= 37.6 \mp (1.782)(32.4) = (-20.0, 95.4)
\end{aligned}$$

The code size on an average processor is significantly different from zero.

Processor effects are not significant.

Example (contd.)

- Using h_1=1, h_2=-1, h_3=0, ($\sum h_j$=0):

$$\text{Mean } \alpha_1 - \alpha_2 = \bar{y}_{.1} - \bar{y}_{.2} = 174.4 - 163.2 = 11.2$$

$$\text{Std dev of } \alpha_1 - \alpha_2 = \frac{s_e}{\sqrt{(\sum h_j^2/ar)}}$$

$$= \frac{88.7}{\sqrt{(2/15)}} = 56.1$$

$$90\% \text{ CI for } \alpha_1 - \alpha_2 = 11.2 \mp (1.782)(56.1)$$

$$= (-88.7,\ 111.1)$$

CI includes zero $\Rightarrow$ one isn't superior to other.

Example (contd.)

- Similarly,

$$90\%\ \text{CI for } \alpha_1 - \alpha_3$$
$$= (174.4 - 225.4) \mp (1.782)(56.1)$$
$$= (-150.9,\ 48.9)$$

$$90\%\ \text{CI for } \alpha_2 - \alpha_3$$
$$= (163.2 - 225.4) \mp (1.782)(56.1)$$
$$= (-162.1,\ 37.7)$$

Any one processor is not superior to another.

Outline

- Model & Computation of Effects
- Estimating Experimental Errors
- Allocation of Variation
- ANOVA Table and F-Test
- Confidence Intervals for Effects
- Unequal Sample Sizes
- Visual Diagnostic Tests

Unequal sample sizes

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

By definition:

$$\sum_{j=1}^{a} r_j \alpha_j = 0$$

Here, r_j is the number of observations at *j*th level.

N =total number of observations:

$$N = \sum_{j=1}^{a} r_j$$

Parameter estimation

Parameter	Estimate	Variance
μ	$\bar{y}_{..}$	s_e^2/N
α_j	$\bar{y}_{.j}$-$\bar{y}_{..}$	$s_e^2(N-r_j)/(Nr_j)$
$\mu+\alpha_j$	$\bar{y}_{.j}$	s_e^2/r_j
$\Sigma\ h_j\alpha_j,\ \Sigma\ h_j=0$	$h_j\ \bar{y}_{.j}$	$s_e^2 \sum_{j=1}^{a}\ (h_j^2/r_j)$
s_e^2	$\sum e_{ij}^2/\{N-a\}$	

Degrees of freedom for errors = N-a

Analysis of variance

Component	Sum of Squares	%Variation	DF	Mean Square	F-Comp.	F-Table
y	SSY=$\sum y_{ij}^2$		N			
$\bar{y}_{..}$	SS0=$N\mu^2$		1			
y-$\bar{y}_{..}$	SST=SSY-SS0	100	N-1			
A	SSA = $\sum_{j=1}^{a} r_j \alpha_j^2$	$100\left(\frac{\text{SSA}}{\text{SST}}\right)$	a-1	MSA=$\frac{\text{SSA}}{a-1}$	$\frac{\text{MSA}}{\text{MSE}}$	$F_{[1-\alpha;a-1,N-a]}$
e	SSE=SST- SSA	$100\left(\frac{\text{SSE}}{\text{SST}}\right)$	N-a	MSE=$\frac{\text{SSE}}{N-a}$		

Example: for the processor-example

	R	V	Z		
	144	101	130		
	120	144	180		
	176	211	141		
	288	288			
	144				
Column Sum	872	744	451	2067	
Column Mean	174.40	186.00	150.33		172.25
Column effect	2.15	13.75	-21.92		

- All means are obtained by dividing by the number of observations added.
- The column effects are 2.15, 13.75, and -21.92.

Example (contd.)

$$\begin{bmatrix} 144 & 101 & 130 \\ 120 & 144 & 180 \\ 176 & 211 & 141 \\ 288 & 288 & \\ 144 & & \end{bmatrix} = \begin{bmatrix} 172.25 & 172.25 & 172.25 \\ 172.25 & 172.25 & 172.25 \\ 172.25 & 172.25 & 172.25 \\ 172.25 & 172.25 & \\ 172.25 & & \end{bmatrix} + \begin{bmatrix} 2.15 & 13.75 & -21.92 \\ 2.15 & 13.75 & -21.92 \\ 2.15 & 13.75 & -21.92 \\ 2.15 & 13.75 & \\ 2.15 & & \end{bmatrix}$$

$$+ \begin{bmatrix} -30.40 & -85.00 & -20.33 \\ -54.40 & -42.00 & 29.67 \\ 1.60 & 25.00 & -9.33 \\ 113.60 & 102.00 & \\ -30.40 & & \end{bmatrix}$$

Example (contd.)

- Sums of Squares:

$$\begin{aligned}
\text{SSY} &= \sum y_{ij}^2 = 397375 \\
\text{SS0} &= N\mu^2 = 356040.75 \\
\text{SSA} &= 5\alpha_1^2 + 4\alpha_2^2 \\
&\quad +3\alpha_3^2 = 2220.38 \\
\text{SSE} &= (-30.40)^2 + (-54.40)^2 + \cdots \\
&\quad +(-9.33)^2 = 39113.87 \\
\text{SST} &= \text{SSY} - \text{SS0} = 41334.25
\end{aligned}$$

- Degrees of Freedom:

$$\begin{array}{rcccccc}
\text{SSY} & = & \text{SS0} & + & \text{SSA} & + & \text{SSE} \\
\text{N} & = & 1 & + & \text{(a-1)} & + & \text{N-a} \\
12 & = & 1 & + & 2 & + & 9
\end{array}$$

Example (contd.)

Compo-nent	Sum of Squares	%Variation	DF	Mean Square	F-Comp.	F-Table
y	397375.00					
$y_{..}$	356040.75					
y-$y_{..}$	41334.25	100.00%	11			
A	2220.38	5.37%	2	1110.19	0.26	3.01
Errors	39113.87	94.63%	9	4345.99		

$$s_e = \sqrt{\text{MSE}} = \sqrt{4345.99} = 65.92$$

Conclusion: Variation due to processors is insignificant as compared to that due to modeling errors.

Outline

- Model & Computation of Effects
- Estimating Experimental Errors
- Allocation of Variation
- ANOVA Table and F-Test
- Confidence Intervals for Effects
- Unequal Sample Sizes
- Visual Diagnostic Tests

Visual diagnostic tests for the one-factor experimental analysis

Assumptions:

1. Factor effects are additive.
2. Errors are additive.
3. Errors are independent of factor levels.
4. Errors are normally distributed.
5. Errors have the same variance for all factor levels.

Tests:

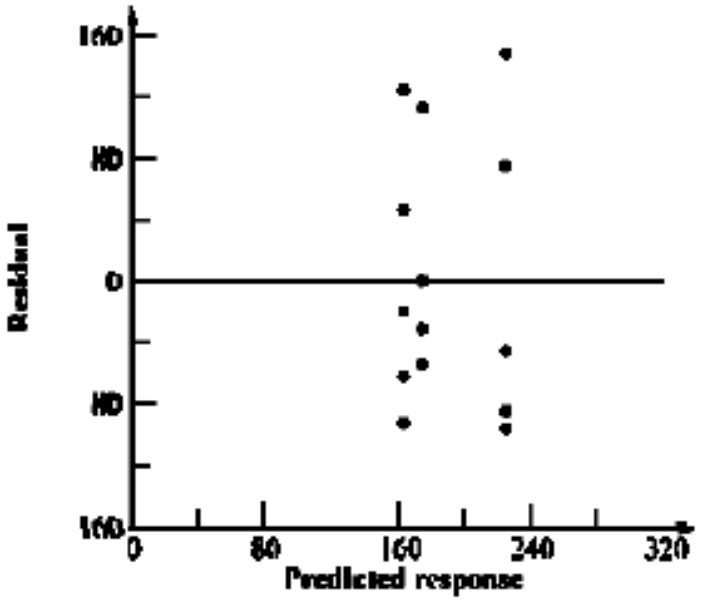
- Residuals versus predicted response:

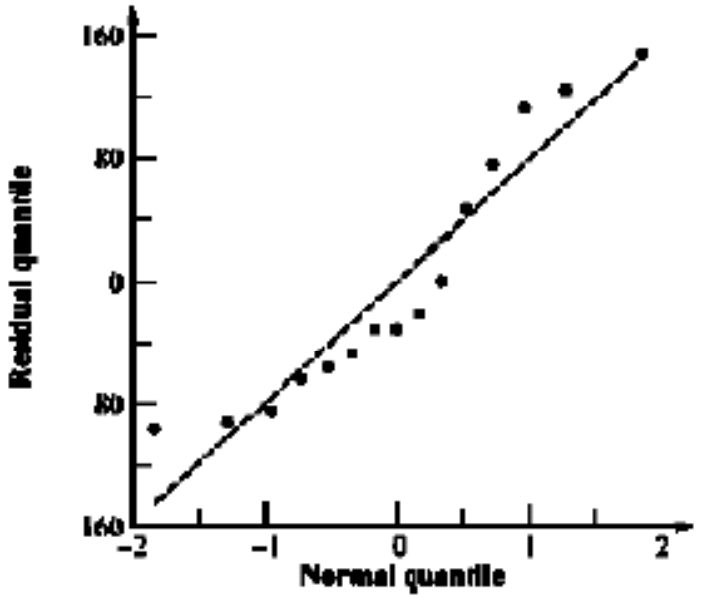
  No trend $\Rightarrow$ Independence

  Scale of errors $<<$ Scale of response $\Rightarrow$ Ignore visible trends.

- Normal quantilte-quantile plot linear $\Rightarrow$ Normality

Example: for the processor-example





- Horizontal and vertical scales similar
  $\Rightarrow$ Residuals are not small $\Rightarrow$ Variation due to factors is small compared to the unexplained variation
- No visible trend in the spread
- Q-Q plot is S-shaped $\Rightarrow$ shorter tails than normal.

Summary

- Model & Computation of Effects
- *Estimating Experimental Errors*
- *Allocation of Variation*
- *ANOVA Table and F-Test*
- Confidence Intervals for Effects
- *Unequal Sample Sizes*
- *Visual Diagnostic Tests*

Further reading

- Chapter 21: two-factor full factorial design without replications
  - Model: $y_{ij} = \mu + \alpha_j + \beta_i + e_{ij}$

y_{ij} = Observation with A at level j and B at level i
μ = mean response
α_j = effect of factor A at level j
β_i = effect of factor B at level i
e_{ij} = error term

- Chapter 22: two-factor full factorial design with replications
  - Model: $y_{ijk} = \mu + \alpha_j + \beta_i + \gamma_{ij} + e_{ijk}$

y_{ijk} = Response in the kth replication with factor A at level j and factor B at level i
μ = mean response
α_j = Effect of factor A at level j
β_i = Effect of Factor B at level i
γ_{ij} = Effect of interaction between factors A and B
e_{ijk} = Experimental error