Lecture 12

Analysis of Variance

One-Way ANOVA

- The response variable Y is continuous
- The explanatory variable is categorical
  - We call it a factor
  - The possible values are called levels
- This is a generalization of the two-sample t-test

Data for one-way ANOVA

- Y, the response variable
- A, the factor
  - I is the number of levels
  - We sometimes refer to these as groups or treatments
- Y_{ij} is the j^{th} observation in the i^{th} group

KNNL Example

- KNNL p 685
- Y is the number of cases of cereal sold
- A is the design of the cereal package
  - There are 4 levels for A because there are 4 different package designs
- i =1 to 4 levels
- j =1 to J_i stores with design i (5,5,4,5)
- Use J if it does not depend on i

Data for one-way ANOVA

```
data a1;
infile '…/ch16ta01.txt';
 input cases design store;
proc print data=a1;
run;
```

The data

Obs	cases	design	store
1	11	1	1
2	17	1	2
3	16	1	3
4	14	1	4
5	15	1	5
6	12	2	1
7	10	2	2

Notation

- For Y_{ij} we use
  - i to denote the level of the factor
  - j to denote the j^{th} observation at factor level i
- i = 1, . . . , I levels of factor A
- j = 1, . . . , J_i observations for level i of factor A

Model

- We assume that the response variable observations are
  - Normally distributed
  - With a mean that may depend on the level of the factor
  - And a variance that does not
  - Independent

Model (2)

- $Y_{ij} = \mu_i + \xi_{ij}$
  - where μ_i is the theoretical mean or expected value of all observations at level i and
  - the ξ_{ij} are iid $N(0, \sigma^2)$
  - $Y_{ij} \sim N(\mu_i, \sigma^2)$, independent
  - This is called the cell means model

Parameters

- The parameters of the model are
  - $\mu_1, \mu_2, \ldots, \mu_I$
  - σ^2

Question – Does our explanatory variable influence Y ? i.e.

Does μ_i depend on i ?

H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$

H_a: not all μ's are the same

Estimates

- Estimate μ_i by the mean of the observations at level i, $\overline{Y}_i$
- $\overline{Y}_i = (\Sigma Y_{ij})/(J_i)$
- For each level we can get an estimate of the variance
- $s_i^2 = (\Sigma(Y_{ij} - \overline{Y}_i)^2)/(J_i - 1)$
- We need to combine these to get an estimate of σ^2

Pooled estimate of σ^2

- If the J_i are all the same we would average the s_i^2
  - We would *not* average the s_i
- In general we pool the s_i^2, giving weights proportional to the df, J_i -1
- The pooled estimate is
- $s^2 = (\Sigma (J_i-1)s_i^2) / (\Sigma (J_i-1))$
- $= (\Sigma (J_i-1)s_i^2)/(n-I)$

Run proc glm

```
proc glm data=a1;
   class design;
   model cases=design;
   means design;
run;
```

Output

```
The GLM Procedure
Class Level Information

Class        Levels    Values
design            4    1 2 3 4
Number of observations     19
```

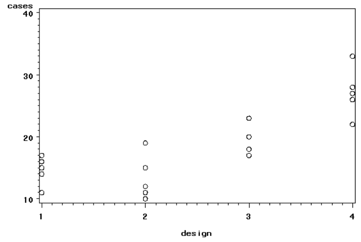
Means statement output

Level of design	N	Mean	Std Dev
1	5	14.6	2.3
2	5	13.4	3.6
3	4	19.5	2.6
4	5	27.2	3.9

Plot the data

```
symbol1 v=circle i=none;
proc gplot data=a1;
   plot cases*design;
run;
```

The plot



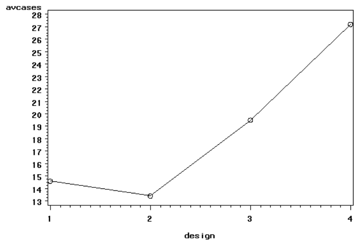
Plot the means

```
proc means data=a1;
   var cases; by design;
   output out=a2 mean=avcases;
proc print data=a2;
symbol1 v=circle i=join;
proc gplot data=a2;
   plot avcases*design;
run;
```

Output Data Set

Obs	design	_FREQ_	avcases
1	1	5	14.6
2	2	5	13.4
3	3	4	19.5
4	4	5	27.2

Plot of the means



Notation

- $Y_{i.} = (\Sigma_j Y_{ij}) / J_i$
- $Y_{..} = (\Sigma_{ij} Y_{ij}) / n$
- n is the total number of observations
- $n = \Sigma_j J_i$

ANOVA Table

Source	df	SS	MS
Model	I-1	$\Sigma_{ij}(Y_{i.}-Y_{..})^2$	SSM/dfM
Error	n-I	$\Sigma_{ij}(Y_{ij}-Y_{i.})^2$	SSE/dfE
Total	n-1	$\Sigma_{ij}(Y_{ij}-Y_{..})^2$	SST/dfT

Anova output

Source	DF	SS	MS	F	P
Model	3	588	196	18.59	<.0001
Error	15	158	10		
Total	18	746			

F test

- F = MSM/MSE
- H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$
- H_1: not all of the μ_i are equal
- Under H_0, F ~ F(I-1, n-I)
- Reject H_0 when F is large
- Report the P-value

More output

```
R-Square Root MSE
0.788055 3.247563
```

Factor Effects Model

- $Y_{ij} = \mu + \alpha_i + \xi_{ij}$
  - the ξ_{ij} are iid $N(0, \sigma^2)$

Parameters

- The parameters of the model are
  - $\mu, \alpha_1, \alpha_2, \ldots, \alpha_I$
  - σ^2

Hypotheses

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$
- H_1: not all of the μ_i are equal

are translated into

- H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$
- H_1: at least one α_i is not 0

Confidence intervals for means

- $Y_{i\cdot} \sim N(\mu_i, \sigma^2/J_i)$
- CI for μ_i is $Y_{i\cdot} \pm t^* s/\sqrt{J_i}$
- t^* is computed from the t(n-I) distribution

Proc Means

```
proc means data=a1
   mean std stderr clm
   maxdec=2;
   class design;
   var cases;
run;
```

Output

des	N Obs	Mean	Std Dev	Std Error
1	5	14.60	2.30	1.03
2	5	13.40	3.65	1.63
3	4	19.50	2.65	1.32
4	5	27.20	3.96	1.77

Confidence Intervals

des	Lower 95% CL for Mean	Upper 95% CL for Mean
1	11.74	17.46
2	8.87	17.93
3	15.29	23.71
4	22.28	32.12

PROC GLM MEANS STATEMENT

```
proc glm data=a1;
   class design;
   model cases=design;
   means design/t clm;
run;
```

Output

The GLM Procedure

t Confidence Intervals for cases

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	10.54667
Critical Value of t	2.1314

CI Output

des	N	Mean	95% Confidence Limits	
4	5	27.200	24.104	30.296
3	4	19.500	16.039	22.961
1	5	14.600	11.504	17.696
2	5	13.400	10.304	16.496

Multiplicity Problem

- **We have constructed 4 (in general, I) 95% confidence intervals**
- **The overall confidence level is less that 95%**
- **Many different kinds of adjustments have been proposed**
- **We have seen the Bonferroni (use α/I)**

BON option

```
proc glm data=a1;
   class design;
   model cases=design;
   means design/bon clm;
run;
```

Output

```
Bonferroni t Confidence
   Intervals for cases

Alpha                          0.05
Error Degrees of Freedom         15
Error Mean Square          10.54667
Critical Value of t         2.83663
```

Bonferroni CIs

```
                Simultaneous 95%
des N   Mean Confidence Limits

4    5 27.200   23.080 31.320
3    4 19.500   14.894 24.106
1    5 14.600   10.480 18.720
2    5 13.400    9.280 17.520
```

Differences in means

- Distribution of $Y_{i.}$-$Y_{k.}$ is
- $N(\mu_i-\mu_k, (\sigma^2/J_i)+(\sigma^2/J_k))$
- CI for μ_i-μ_k is $Y_{i.}$-$Y_{k.}$ ± $t^*s(Y_{i.}-Y_{k.})$
- where $s(Y_{i.}-Y_{k.}) = s(\sqrt{\frac{1}{J_i}+\frac{1}{J_k}})$

t^*

- We deal with the multiplicity problem by adjusting t^*
- Many different choices are available

LSD

- Least Significant Difference (LSD)
- Ignore multiplicity
- Use t(n-I)
- Also called T in SAS

Bonferroni

- Use the error budget idea
- There are I(I-1)/2 comparisons among I means
- So, replace α by α/(I(I-1)/2) and use t(n-I)

Tukey

- Based on the studentized range distribution (max minus min divided by the standard deviation)
- $t^* = q^*/\sqrt{2}$
- Details are in KNNL Section 17.5

Scheffe

- Based on the F distribution
- $t^* = \sqrt{(I-1)F(1-\alpha; I-1, N-I)}$
- Takes care of multiplicity for all linear combinations of means

Multiple Comparisons

- LSD is too liberal
- Scheffe is too conservative
- Bonferroni is ok for small I
- Tukey (HSD) is recommended

Example

```
proc glm data=a1;
   class design;
   model cases=design;
   means design/
     lsd tukey bon scheffe;
run;
```

LSD

```
t Tests (LSD) for cases
NOTE: This test controls the
Type I comparisonwise error rate,
not the experimentwise error rate.
Alpha                           0.05
Error Degrees of Freedom          15
Error Mean Square           10.54667
Critical Value of t          2.13145
```

Tukey

```
Tukey's Studentized Range (HSD)
   Test for cases
NOTE: This test controls the
Type I experimentwise error rate.
Alpha                          0.05
Error Degrees of Freedom         15
Error Mean Square          10.54667
Critical Value of Studentized
                  Range   4.07597
```

4.07/sqrt(2)= 2.88

Tukey Intervals (CLDIFF option for equal cell sizes)

design Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
4 - 3	7.700	1.421	13.979	***
4 - 1	12.600	6.680	18.520	***
4 - 2	13.800	7.880	19.720	***
3 - 4	-7.700	-13.979	-1.421	***
3 - 1	4.900	-1.379	11.179	
3 - 2	6.100	-0.179	12.379	
1 - 4	-12.600	-18.520	-6.680	***
1 - 3	-4.900	-11.179	1.379	
1 - 2	1.200	-4.720	7.120	
2 - 4	-13.800	-19.720	-7.880	***
2 - 3	-6.100	-12.379	0.179	
2 - 1	-1.200	-7.120	4.720	

Output (option lines)

	Mean	N	design
A	27.200	5	4
B	19.500	4	3
B			
B	14.600	5	1
B			
B	13.400	5	2