22S:30/105, Statistical Methods and Computing
Instructor: Cowles
Lab 7, April 19, 2006
One-way ANOVA
Inference in Regression

1 Downloading datasets

Download the following data files from the course web page:

```
autistic.dat
hotdogs.dat
OECD.dat
```

2 Immunology

(example taken from Daniel *Biostatistics: A Foundation for Analysis in the Health Sciences*

Research by Singh et al. (1999) as reported in the journal *Clinical Immunology and Immunopathology* is concerned with immune abnormalities in autistic chilcren. As part of their research, they took measurements on the serum concentration of an antigen in three samples of children, autistic children, normal children, and mentally-handicapp0ed children (non-Down's-syndrome). All children were 10 years old or younger.

This dataset contains two variables:

```
concentration of the antigen (in units per milliliter of serum)
group,    coded A for autistic
               N for normal
               M for mentally handicapped
```

1. What population(s) do the researchers wish to study?

2. If the researchers believed that the distributions of serum concentrations of this antigen were normal in each of the populations of interest, what additional assumption would be needed to justify the use of one-way ANOVA to analyze these data?

3. What is the null hypothesis?

4. What is the alternative hypothesis?

5. Read the dataset into SAS:

```
options linesize = 75 ;

data autistic ;
infile 'c:\temp\autistic.dat' ;
input conc group $ ;
run ;
p
```

6. Use SAS to check the assumptions of one-way ANOVA.

```
proc sort data = autistic ;
by group ;
run ;

proc univariate plot data = autistic ;
var conc ;
by group ;
run ;

proc means data = autistic ;
var conc ;
by group ;
run ;
```

   (a) Do the distributions of the sample data appear to be roughly normal?

   (b) Is the largest sample standard deviation no more than twice as large as the smallest sample standard deviation?

3 Food

The "hotdogs" dataset contains data on the sodium and calories contained in each of 54 major hotdog brands. The variables are:

```
type -- Beef, Meat, or Poultry
calories per hotdog
sodium per hotdog
```

There are many other brands of hotdogs on the market besides those included in this dataset. We are interested in determining whether the mean number of calories per hotdog is the same in all of the three types of hotdogs.

1. What population(s) do we researchers wish to study?

2. What is the null hypothesis?

3. What is the alternative hypothesis?

4. Read the dataset into SAS:

```
data hotdogs ;
infile 'c:\temp\hotdogs.dat' ;
input type $ calories sodium ;
run ;
```

5. Use SAS to check the assumptions of one-way ANOVA.

```
proc sort data = hotdogs ;
by type ;
run ;
```

```
proc univariate plot data = hotdogs ;
var calories ;
by type;
run ;
```

```
proc means data = hotdogs ;
var calories ;
by type ;
run ;
```

   (a) Do the distributions of the sample data appear to be roughly normal?

   (b) Is the largest sample standard deviation no more than twice as large as the smallest sample standard deviation?

6. Use SAS to test your hypotheses at the $\alpha = .05$

```
proc anova data = hotdogs ;
class type ;
model calories = type ;
run ;
```

The SAS output is as follows:

```
                    The ANOVA Procedure

                 Class Level Information

    Class          Levels    Values
    type               3      Beef Meat Poultry


           Number of observations     54
```

Dependent Variable: calories

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	17692.19510	8846.09755	16.07	<.0001
Error	51	28067.13824	550.33604		
Corrected Total	53	45759.33333			

R-Square	Coeff Var	Root MSE	calories Mean
0.386636	16.12935	23.45924	145.4444

Source	DF	Anova SS	Mean Square	F Value	Pr > F
type	2	17692.19510	8846.09755	16.07	<.0001

7. Can we reject the overall hypothesis of equality of means?

8. Are we justified in carrying out pairwise t-tests to look for significant differences between individual pairs of means?

9. Add a "means" statement to carry out the pairwise t-tests with Bonferroni correction.

```
proc anova data = hotdogs ;
class type ;
model calories = type ;
means type / bon alpha = .05 ;
run ;
```

The new part of the output is:

```
                    The ANOVA Procedure

              Bonferroni (Dunn) t Tests for calories

  NOTE: This test controls the Type I experimentwise error rate, but it
  generally has a higher Type II error rate than Tukey's for all pairwise
                            comparisons.

              Alpha                        0.05
              Error Degrees of Freedom       51
              Error Mean Square         550.336
              Critical Value of t       2.47551


       Comparisons significant at the 0.05 level are indicated by ***.

                          Difference
              type        Between      Simultaneous 95%
           Comparison      Means       Confidence Limits

     Meat    - Beef         1.856      -17.302   21.013
     Meat    - Poultry     39.941       20.022   59.860   ***
     Beef    - Meat        -1.856      -21.013   17.302
     Beef    - Poultry     38.085       18.928   57.243   ***
     Poultry - Meat       -39.941      -59.860  -20.022   ***
     Poultry - Beef       -38.085      -57.243  -18.928   ***
```

10. Which population means are significantly different at the .05 level?

4 International health economics

Reference: http://www.oecd.org/publications/figures/

The OECD dataset is collated from the above web page of the Organization for
Economic Cooperation and Development (OECD). It provides summary statistics
for the 29 member nations. The variables are as follows:

```
name:    name of country
pcgdp:   per capita gross domestic product (1998)
         reported in US dollars converted using Purchasing Power Parities to
            adjust for differences in price levels between countries
pch:     per capita health care expenditures (1996)
         reported in US dollars converted using Purchasing Power Parities
beds:    in-patient hospital beds per 1000 population (1996)
```

```
los:      average length of stay in days for hospital patients (1996)
docs:     doctors per 1000 population (1996)
infmort:  infant mortality (1996)
          number of deaths of infants < 1 yr of age per 1000 live births
region:   region of the world
```

Suppose we want to get predicted values of pch if we know pcgdp. We want predicted
values for the countries that are in the dataset, as well as for a hypothetical new
country with pcgdp = $20,000.

We will put a dummy record in the dataset with a missing value for pch and the
desired value of the explanatory variable. This record will not be included in SAS's
calculation of the regression coefficients, but SAS will give us predicted values, as
well as a confidence interval and prediction interval, for it. SAS's symbol for a
missing value is a period.

```
***************
* Reading in  *
* the dataset *
*************** ;

data OECD ;
infile 'c:\temp\OECD.dat' ;
input name $13. pcgdp pch beds los doc infmort region $ ;
run ;


**************
*  Checking  *
************** ;

* Note the extra record with missing value for pch ;

proc print data = OECD ;
run ;


**************
* Regression *
**************  ;

proc reg data = OECD ;
model pch = pcgdp / clb ;
run ;


********************
```

```
* Predicted values *
* and residuals    *
******************** ;

proc reg data = OECD ;
model pch = pcgdp / p ;
id name ;
run ;


*********************
* Confidence limits *
* for means of sub- *
* populations       *
******************** ;

proc reg data = OECD ;
model pch = pcgdp / clm ; /* clm gets conf limits for the mean */
id name ;
run ;


*********************
* Prediction limits *
* for individual    *
* predictions       *
******************** ;

proc reg data = OECD ;
model pch = pcgdp / cli ; /* cli gets prediction interval for
                            new individual */
id name ;
run ;


*********************
* Scatterplots and *
* Residual plots   *
******************** ;

proc reg data = OECD ;
model pch = pcgdp / p ;
plot pch * pcgdp / symbol =  '.' ;
run ;
plot residual. * predicted. / symbol = '.' ;
run ;
```

Scatterplots and residual plots may also be obtained easily using the automated feature in SAS called "Insight."