STAT:2010/4200
Lab 8.
One-way ANOVA

1 Datasets

We will need the following dataset

```
autistic.dat
hotdogs.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 children. 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-handicapped 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:

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

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?

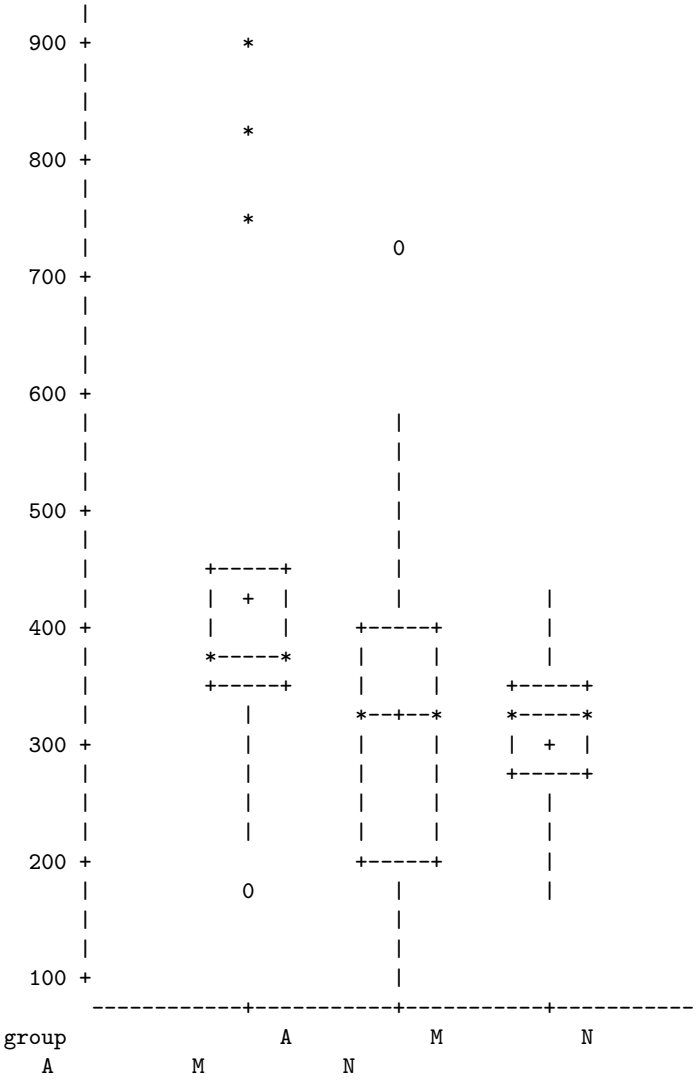
   (b) Is the largest sample standard deviation no more than twice as large as the smallest sample standard deviation?

In summary, it is not safe to use ANOVA to perform a test for H_0 and H_a specified above. We can resort to a nonparametric test instead.

Part of the SAS output:

The UNIVARIATE Procedure
Variable: conc

Schematic Plots



The MEANS Procedure

---------- group=A ----------

Analysis Variable : conc

N	Mean	Std Dev	Minimum	Maximum
23	419.9130435	178.0262782	170.0000000	900.0000000

```
-------------------------------- group=M --------------------------------

                        Analysis Variable : conc

     N            Mean         Std Dev         Minimum         Maximum

    15     329.3333333     170.9518172     105.0000000     715.0000000


-------------------------------- group=N --------------------------------

                        Analysis Variable : conc

     N            Mean         Std Dev         Minimum         Maximum

    33     305.0000000      63.8112451     165.0000000     435.0000000
```

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:

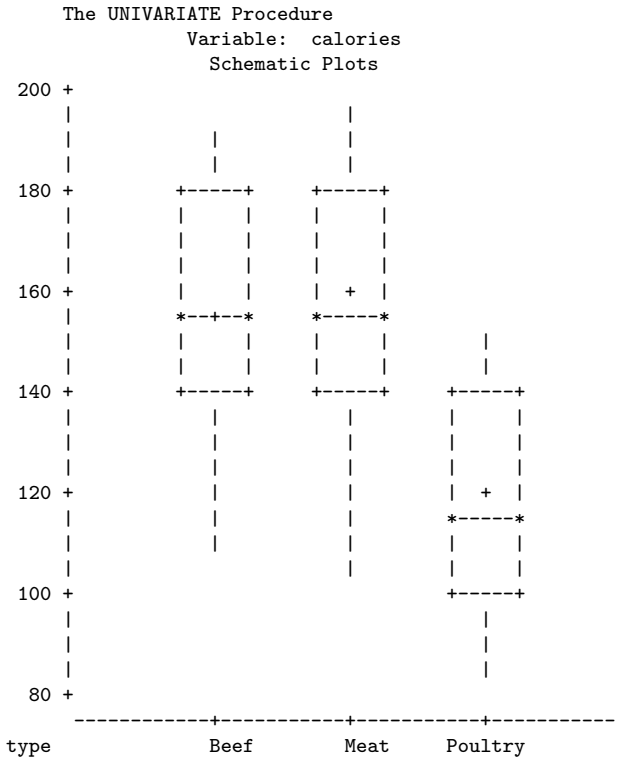
```
data hotdogs ;
infile 'c:\temp\hotdogs.dat' ;
input type $ calories sodium ;
run ;
```

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 ;
```



```
                          The MEANS Procedure
-------------------------------- type=Beef --------------------------------

                      Analysis Variable : calories

     N            Mean         Std Dev         Minimum         Maximum

    20     156.8500000      22.6420080     111.0000000     190.0000000


-------------------------------- type=Meat --------------------------------

                      Analysis Variable : calories

     N            Mean         Std Dev         Minimum         Maximum

    17     158.7058824      25.2357997     107.0000000     195.0000000


------------------------------ type=Poultry -------------------------------

                      Analysis Variable : calories

     N            Mean         Std Dev         Minimum         Maximum

    17     118.7647059      22.5514119      86.0000000     152.0000000
```

(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

                                      Sum of
 Source                     DF       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?
Remark: The above SAS output looks a bit different from what we see in chap25 notes p.20 (where A, B, C ... are used to indicate which pairs are not significantly different). Actually, for this follow-up test of ANOVA, SAS might produce either one of the two types of output, and you need to know how to read either one (explained in more details in class.) For the one above: we are doing three pairs of comparisons (SAS listed six, but three of them are redundant because, say, they looked at the difference between meat-beef and then that of beef-meat).