STAT 40001/STAT 50001 Statistical Computing

Lecture 19

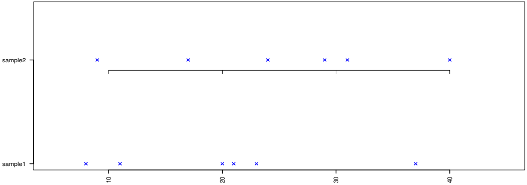
Department of Mathematics and Statistics
Purdue University Northwest

Logic Behind ANOVA

The reason for the word variance in the analysis of variance is that the procedure for comparing the means analyzes the variation in the sample data. To examine how it works let's suppose that independent random samples are taken from two populations say population 1 and population 2 with means μ_1 and μ_2 respectively. Suppose following sample is chosen from each population.

Sample from Population 1	21	37	11	20	8	23
Sample from Population 2	24	31	29	40	9	17

Observe that $\overline{x_1} = 20$ and $\overline{x_2} = 25$. Can we reasonably conclude from these statistics that $\mu_1 \neq \mu_2$?



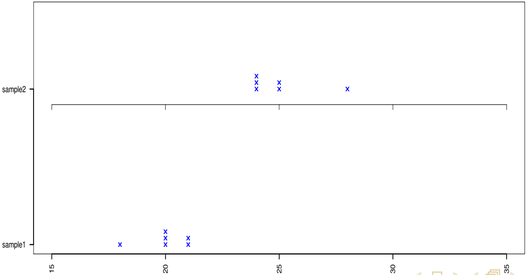
We need to consider the variation within the samples.

Logic Behind ANOVA

Suppose that we have a new data again from population 1 and population 2 as below

Sample from Population 1	21	21	20	18	20	20
Sample from Population 2	25	28	25	24	24	24

Note that $\overline{x_1} = 20$ and $\overline{x_2} = 25$. But this time, we can infer that $\mu_1 \neq \mu_2$ because it seems clear that the difference between the sample means is due to difference between the population means, not the variation within the populations.



Analysis of Variance

Intuitively speaking, in the first case because the variation between the sample means is not large relative to the variation within the samples, we cannot conclude that $\mu_1 \neq \mu_2$ whereas in the second case because the variation between the sample means is large relative to the variation within the samples, we can conclude that $\mu_1 \neq \mu_2$.
We can use the R code below to create the strip charts.

```
> x=c(21, 21, 20, 18, 20, 20)
> y=c(25, 28, 25, 24, 24, 24)
> stripchart(list(sample1=x,sample2=y),method="stack",pch=4,
   offset=1/2,col="blue",lwd=2,las=2,xlim=c(15,35) )
> axis(1,pos=1.9,labels=FALSE)
> axis(1,pos=2.9,labels=FALSE)
```

Elements of Design of Experiment

Design of Experiment (DOE) is a structured, organized method that is used to determine the relationship between the different factors (Xs) affecting a process and the output of that process (Y).
Analysis of variance is the technique we use when all the explanatory variables are categorical. The explanatory variables are called factors and each factor has two ore more levels. When there is a single factor with three or more levels we use one-way ANOVA. But when a single factor has only two levels we can use two sample t-test.

- Randomization
- Blocking
- Replication
- Balanced Design
- Use of Sequential Experimentation
- Adjustment for Covariates
- Use of multiple sizes of experimental units

Experiments with a Single Factor: ANOVA

Suppose we have m treatments or different levels of a single factor that we wish to compare. The observed response from each of the m treatments is a random variable.

The data appears as below

Treatment level	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	y_{m1}	y_{m2}	$\cdots$	y_{mn}	$y_{m.}$	$\bar{y}_{m.}$
					$y_{..}$	$\bar{y}_{..}$

Here y_{ij} denotes the response of j^{th} observation with i^{th} treatment.

Notation:

$$\begin{aligned} \bar{y}_{i.} &= \frac{1}{n}\sum_{j}^{n} y_{ij} = \frac{1}{n} y_{i.} \qquad \text{Treatment mean} \\ \bar{y}_{..} &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j}^{n} y_{ij} = \frac{1}{mn} y_{..} \qquad \text{Grand mean} \end{aligned}$$

ANOVA

A single factor experimental design is said to be a balanced design if the number of observations taken within each treatment is equal, otherwise it is said to be unbalanced.

Model for the data

It is usually useful to describe the observations from an experiment with a model. One way to write this model is

$$y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{Means Model}$$

where y_{ij} is the response of the j^{th} observation with the i^{th} treatment$(i = 1, 2..., m, \quad j = 1, 2..., n)$, μ_i is the mean of the i^{th} factor level or treatment and ϵ_{ij} is a random error component that arises due to all other source of variability.

We assume that $E(\epsilon_{ij}) = 0$. Hence $E(y_{ij}) = \mu_i$.

Another way of writing the model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad \text{Effects Model}$$

where μ is the overall mean and τ_i is a parameter unique to the i^{th} treatment called the i^{th} treatment effect.

Analysis of Fixed effect Models (Single Factor ANOVA)

In single factor ANOVA we consider the statistical model

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where $i = 1, 2, ..., m, \quad j = 1, 2, 3..., n$. We are interested in testing the equality of the m treatment means. Hence, the appropriate hypotheses are

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = ... = \mu_m \\ H_1 &: \quad \mu_i \neq \mu_j \quad \text{for at least one pair } (i, j) \end{aligned}$$

We may rewrite this as

$$\begin{aligned} H_0 &: \quad \tau_1 = \tau_2 = ... = \tau_m = 0 \\ H_1 &: \quad \tau_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

as we can write $\mu_i = \mu + \tau_i$ which yields

$$\mu = \frac{1}{m} \sum_{i=1}^{m} \mu_i$$

so that

$$\sum_{i=1}^{m} \tau_i = 0$$

Hence, testing the equality of treatment means is equivalent to testing the treatment effects are zero.

Decomposition of the Total Sum of Squares

Recall that the $y_{i.}$ represent the total of the observations under i^{th} treatment and $y_{..}$ represent the grand sum of all the observations.

Subject	Tret1	Tret2	$\cdots$	Tret m
1	y_{11}	y_{21}	$\cdots$	y_{m1}
2	y_{12}	y_{22}	$\cdots$	y_{m2}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
n	y_{1n}	y_{2n}	$\cdots$	y_{mn}

Here y_{ij} denotes the response of the j^{th} subject with the i^{th} treatment. This is an example of crossover design.
Notations:

$$\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} = \frac{1}{n} y_{i.} \qquad \text{Treatment mean}$$

$$\bar{y}_{..} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij} = \frac{1}{mn} y_{..} = \frac{1}{N} y_{..} \qquad \text{Grand mean}$$

Decomposition of Sum of Squares

Consider the partitioning of the total variability in the data into its components. Let us denote by TSS the total corrected sum of squares and it is given by

$$TSS = \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$$

which may be written as

$$\begin{aligned} TSS &= \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}[(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})]^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 + n\sum_{i=1}^{m}(\bar{y}_{i.} - \bar{y}_{..})^2 \end{aligned}$$

This is called the fundamental ANOVA identity.

$$TSS = SS_{Treatments} + SSE$$

Also note that we can partition the degrees of freedom as

$$N - 1 = (m - 1) + (N - m)$$

Statistical Inference

Since we assume that ϵ_{ij} are independent and normally distributed with mean zero and variance σ^2 the observations y_{ij} are independent and normally distributed with mean $\mu_i = \mu + \tau_i$ and variance σ^2 therefore, TSS/σ^2 has chi-square distribution with $N-1$ degrees of freedom. Also note that under null hypothesis $(\tau_i = 0)$ SSE/σ^2 has chi-square distribution with $N-m$ degrees of freedom and $SS_{Treatments}/\sigma^2$ has chi-square distribution with $m-1$ degrees of freedom and we have

ANOVA Table

Source of variation	Sum of Squares	DF	MS	F_0
Between Treatments	$n\sum_{i=1}^{m}(\bar{y}_{i.} - \bar{y}_{..})^2$	$m-1$	$MS_{Treatments}$	$\frac{MS_{Treatments}}{MSE}$
Within Treatments	$TSS - SS_{Treatments}$	$N-m$	MSE	
Total	$\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	$N-1$		

Decision Process:
We know that $F_0 = \frac{SS_{Treatments}/(m-1)}{SSE/(N-m)} = \frac{MS_{Treatments}}{MSE}$ is the test statistic for the hypothesis of no difference in treatment means. We compute the value of F_0 and observe the value of $F_{\alpha,m-1,N-m}$ using the F-table. We should reject the null hypothesis(H_0) and conclude that there are differences in the treatment means if

$$F_0 > F_{\alpha,m-1,N-m}$$

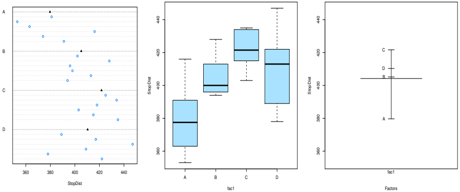
Example-Tire Data

A tire manufacturer is interested in investigating the handling properties for different tread patterns. The data frame Tire in PASWR package has the stopping distance measured to the nearest foot for a standard size car to come to the complete stop from a speed of 60 miles per hour. There are six measurements of the stopping distance for four different tread patterns labeled A, B, C and D.(Note that same car and same driver were assigned). The order of treatments was assigned at random.

```
> library(PASWR)
> data(Tire)
> Tire
   StopDist tire
1       391    A
2       374    A
3       416    A
4       363    A
5       353    A
6       381    A
7       394    B
8       413    B
9       398    B
10      396    B
11      428    B
12      402    B
13      435    C
14      415    C
15      403    C
16      418    C
17      434    C
18      425    C
19      422    D
20      378    D
21      409    D
22      447    D
23      417    D
24      389    D
```

Example-Tire Data

The `oneway.plots(StopDist,tire)` function creates the plots below.



We can perform the analysis using `aov` or `oneway.test` (available in PASWR)

```
> summary(aov(StopDist~tire))
            Df Sum Sq Mean Sq F value  Pr(>F)
tire         3   5673  1891.0   5.328 0.00732 **
Residuals   20   7099   354.9
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> oneway.test(StopDist~tire, var.equal=T)
        One-way analysis of means
data:  StopDist and tire
F = 5.3278, num df = 3, denom df = 20, p-value = 0.007316
```

Decision: No tire tread effect has been rejected as $p < \alpha$.

Checkout

model.tables(aov(StopDist∼tire), type= "means")

Example- Soil Type data

An experiment has been conducted to measure the effect of soil on crop yield.Crop yields per unit area were measured from 10 randomly selected fields on each of the three soil types namely: sand, clay and loam. All fields were sown with the same variety of seed, fertilizer etc. The dataset is available in http://user.mendelu.cz/drapela/Forest_Biometry/Data/yields.txt
We use the R code below to perform the analysis of the data

```
> data<-read.table("http://user.mendelu.cz/drapela/Forest_Biometry/Data/yields.txt", header=T)
> attach(data)
> y<-c(sand,clay,loam)
> soil<-factor(rep(1:3,c(10,10,10)))
> plot(soil,y, names=c("sand","clay","loam"), ylab="yield")
> summary(aov(y˜soil))
> par(mfrow=c(2,2))
> plot(aov(y˜soil))
> plot.design(y˜soil)
```

Observing the effects in each soil type

```
> summary(lm(y˜soil))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.900      1.081   9.158 9.04e-10 ***
soil2          1.600      1.529   1.047  0.30456
soil3          4.400      1.529   2.878  0.00773 **
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that Soil 1 (sand) is the reference level and has a mean of 9.9. Note that we can fit a the model without intercept term using

```
> lm(y˜soil-1)# The R-squared value is not quite correct.
```

Example- Income

The data set female.inc included in the UsingR package contains income data for females age 15 or over in the United States for the year 2001 broken down by race.

```
> library(UsingR)
> data(female.inc)
> attach(female.inc)
> boxplot(income ~ race, data=female.inc)
> boxplot(log(income,10) ~ race, data=female.inc)
> sapply(with(female.inc,split(income,race)),median)
> aov(income ~ race, data=female.inc)
> summary(aov(income ~ race, data=female.inc))
```

Example- Baseball Data

The data set hall.fame in the UsingR package contains statistics for several major league baseball players. We would like to perform a one-way test to see whether the mean batting average, BA, is the same for Hall of Fame members (Hall.Fame.Membership) as for the other players.

```
> library(UsingR)
> data(hall.fame)
> attach(hall.fame)
> plot(Hall.Fame.Membership,BA)
> aov(BA~Hall.Fame.Membership)
> sapply(with(hall.fame,split(BA,Hall.Fame.Membership)),mean)
```

Example-Child Birth weight

The data set babies in the UsingR package contains information on birth weight of a child and the mother's smoking status. The birth weight, wt, is coded in ounces and smoke is a numeric value: 0 for never, 1 for smokes now, 2 for smoked until current pregnancy, 3 for smoked previously but not now, and 9 if unknown.

‘

```
> library(UsingR)
> attach(babies)
> data<- subset(babies, select=c("wt","smoke"))
> plot(wt~factor(smoke))
> sapply(with(babies,split(wt,factor(smoke))),mean)
> summary(aov(wt~factor(smoke)))
```

Model Assumptions

The values in the ANOVA table and the subsequent inferences made from those values are based on the assumption that the data follow the model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad \text{Effects Model}$$

where μ is the overall mean and τ_i is a parameter unique to the i^{th} so are fixed but unknown numbers called the i^{th} treatment effect. ϵ_{ij} are independent normals with a mean zero and constant variance. Consequently, the three basic assumptions corresponding the errors:

- independentce
- normal distribution
- constant variance

We can use `checking.plots(model)` from `PASWR` library to check the status of the above assumptions

Testing for equality of variance-Bartlett's test

Bartlett's test is a statistical test that is used to test homoscedasticity, that is, if multiple samples are from populations with equal variances.

$$\begin{aligned} H_0 &: \quad \sigma_1^2 = \sigma_2^2 =, \cdots, = \sigma_m^2 \\ H_1 &: \quad \text{at least one } \sigma_i^2 \text{ is different} \end{aligned}$$

The test statistic for this test is given by

$$\begin{aligned} \chi_0^2 &= \frac{q}{c} \\ \text{where} \quad q &= (N-m)\ln s_p^2 - \sum_{i=1}^{m}(n_i - 1)\ln s_i^2 \\ c &= 1 + \frac{1}{3(m-1)}\left[\sum_{i=1}^{m}\frac{1}{n_i-1} - \frac{1}{N-m}\right] \\ s_p^2 &= \frac{1}{N-m}\sum_{i=1}^{m}(n_i-1)s_i^2 \\ s_i^2 &= \frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2 \end{aligned}$$

The quantity q is large when the sample variances s_i^2 differ greatly and is equal to zero when all s_i^2 are equal.
Decision Criteria Reject H_0 when

$$\chi_0^2 > \chi^2_{\alpha, m-1}$$

To determine if three different studying techniques lead to different exam scores, a professor randomly assigns 10 students to use each technique (Technique A, B, or C) for one week and then makes each student take an exam of equal difficulty. The exam scores of the 30 students are shown below:

```
df <-data.frame(group = rep(c("A","B", "C"), each=10),
     score = c(85, 86, 88, 75, 78, 94, 98, 79, 71, 80,
                91, 92, 93, 85, 87, 84, 82, 88, 95, 96,
                79, 78, 88, 94, 92, 85, 83, 85, 82, 81))

bartlett.test(score ~ group, data = df)
```

The test returns the following results:
Test statistic B: 3.3024
P-value: 0.1918

Comparisons Among Treatment Means

Once the null hypothesis (treatment means are equal) is rejected we usually want more information. Since the null hypothesis is rejected, there is a difference between treatment means, but exactly which means differ is not specified by ANOVA. If we want to know this information we need further comparisons and analysis among groups of treatments. The procedure is called multiple comparison methods. One method to examine treatment effects is called a contrast.

Suppose we are interested in comparing only pairs of means. This means we are interested in the contrasts of the form $\Gamma = \mu_i - \mu_j$ for all $i \neq j$. There are several different procedures among them two popular methods are

a. Tukey's Honest Significant Difference (HSD) Test

b. The Fisher's Least Significant Difference(LSD) Method.

Tukey's honest significant difference(HSD) Method

Suppose after conducting the ANOVA procedure we have rejected the null hypothesis of equal treatment means, we wish to test all pairwise mean comparisons:

$$\begin{aligned} H_0 &: \mu_i = \mu_j \\ H_1 &: \mu_i \neq \mu_j \end{aligned}$$

for all $i \neq j$. In order to have the overall significance level exactly α when the sample sizes are equal and the level is at most α when the sample sizes are not equal we use Tukey's test.
Tukey's procedure uses the studentized range statistic. The studentized range statistic for equal sample sizes is given by

$$q = \frac{\overline{y}_{max} - \overline{y}_{min}}{\sqrt{\frac{MSE}{n}}}$$

where $\overline{y}_{max}$ and $\overline{y}_{min}$ are the largest and smallest sample means out of a group of m sample means.
Decision process:
Two means are significantly different at level α if

$$|\overline{y}_{i.} - \overline{y}_{j.}| \geq T_\alpha$$

Example

In the manufacturing of clothing a wear testing machine is used to measure the resistance to abrasion of different fabrics. The data below gives the loss of weight of the material in grams after a specified number of cycles for 4 different types of fabrics: A, B, C and D

A	B	C	D
1.93	2.55	2.40	2.33
2.38	2.72	2.68	2.40
2.20	2.75	2.31	2.28
2.25	2.70	2.28	2.25

```
> A<-c(1.93,2.38,2.20,2.25)
> B<-c(2.55,2.72,2.75,2.70)
> C<-c(2.40,2.68,2.31,2.28)
> D<-c(2.33,2.40,2.28,2.25)
> y<-c(A,B,C,D)
> type<-factor(rep(c("A","B","C","D"),c(4,4,4,4)))
> plot(type,y, names=c("A","B","C","D"), ylab="weightloss")
> summary(aov(y˜type))
            Df Sum Sq Mean Sq F value  Pr(>F)
type         3 0.5201 0.17337   8.534 0.00264 **
Residuals   12 0.2438 0.02031
```

Example

```
> TukeyHSD(aov(y~type))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = y ~ type)
$type
       diff         lwr         upr     p adj
B-A  0.4900  0.19078408  0.78921592 0.0019081
C-A  0.2275 -0.07171592  0.52671592 0.1631572
D-A  0.1250 -0.17421592  0.42421592 0.6148291
C-B -0.2625 -0.56171592  0.03671592 0.0928894
D-B -0.3650 -0.66421592 -0.06578408 0.0159938
D-C -0.1025 -0.40171592  0.19671592 0.7429566
```

It can be observed that the Tukey's procedure indicates that at $\alpha = 0.05$ there is a significant difference between fabric type A and type B and fabric type B and type D.

Checkout:

`> plot(TukeyHSD(aov(y~type)))`: Use option `las=2`

Nonparametric Methods in the ANOVA

The nonparametric counterpart of ANOVA is the Kruskal-Wallis test. This test is used to test the null hypothesis that m treatment means are identical against the alternative hypothesis that some of the treatments generate observations that are different from the others.

Test scores in three form (Form A, B and C) are listed below. Is there there difference in the means?

```
> A<-c(63,64,95,64,60,85)
> B<-c(58,56,51,84,77)
> C<-c(85,79,59,89,80,71,43)
> Test=stack(list("Test A"=A, "Test B"=B, "Test C"=C))
> plot(values~ind, data=Test, xlab="Test", ylab="Scores")
> kruskal.test(values~ind,data=Test)

        Kruskal-Wallis rank sum test

data:  values by ind
Kruskal-Wallis chi-squared = 1.7753, df = 2, p-value = 0.4116
```

Decision: The large p-value indicates no reason to doubt the null hypothesis of equal mean