STATISTICS 571 **Discussion** #11

I. Review – Analysis of Variance

1. ANOVA:
   (a) Compare means three or more populations $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
   (b) Mathematical Model: $Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij}$. Y_{ij} is the jth observation of ith sample (group) , μ_i is the mean for group i, μ is the grandmean, and $\alpha_i = \mu_i - \mu$. ϵ_{ij} is the error.
   (c) Variation Among Samples (Groups):
   $$\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$$
   This is also called group sum square SS_{group}. The d.f. of SS_{group} is $k-1$
   (d) Variation Within Samples :
   $$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{k}(n_i - 1)s_i^2$$
   where $s_i^2 = \frac{\sum_{i=1}^{k}(Y_{ij}-\bar{Y}_i)^2}{n_i-1}$ is the ith sample covaraince. This is also called error sum square SS_{error}. The d.f. of SS_{error} is $n-k$
2. F-test and ANOVA table
   (a) Let group mean square $MS_{group} = \frac{SS_{group}}{k-1}$ and error mean square $MS_{error} = \frac{SS_{error}}{n-k}$
   (b) Let $F = \frac{MS_{group}}{MS_{error}}$, then under H_0, $F \sim F_{k-1,n-k}$, F distribution with $k-1$ and $n-k$ degrees of freedom
   (c) ANOVA table

Source	df	SS	MS	F	p-value
groups	k-1	$\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$	$\frac{SS_{group}}{(k-1)}$	$F = \frac{MS_{group}}{MS_{error}}$	$p = P(F_{k-1,n-k} > F)$
error	n-k	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$	$\frac{SS_{error}}{(n-k)}$		
Total	n-1	SS_{total}			

3. Confidence Interval and Tukey's HSD simultaneously confidence interval
   (a) $R^2 = \frac{SS_{group}}{SS_{total}}$, and the estimator of common variance $\hat{\sigma^2} = MS_{error}$
   (b) $SE(\bar{Y}_i - \bar{Y}_j) = \sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$, t^* value is obtained from the t distribution with d.f n-k. The C.I is
   $$\bar{Y}_i - \bar{Y}_j - t^* \times SE < \mu_i - \mu_j < \bar{Y}_i - \bar{Y}_j + t^* \times SE$$
   (c) If we want to be 95% confident that all population mean differences are contained in their intervals,(which is called simultaneous confidence interval) we need to increase the size of multiplier (t^*). A commonly used method is called Tukey's honestly significant difference (HSD)

II.Practice Problems

1. The bright yellow head of the adult Egyptian vulture requires carotenoid pigments. These pigments cannot be synthesized by the vultures,though, so they must be obtained through their diet. Unfortunately, carotenoids are scarce in rotten flesh and bones, but they are readily available in the dung of ungulates. Perhaps for this reason, Egyptian vultures are frequently seen eating the droppings of cows, goats, and sheep in Spain, where they have been studied. Ungulates are common in some areas but not in others. A study was performed to measure plasma carotenoids in wild-caught vultures at four locations in Spain:

Site	mean concentration	standard deviation	n
1	1.86	1.22	22
2	5.75	2.46	72
3	6.44	3.42	77
4	11.37	1.96	11

   (a) Complete the ANOVA table for the problem

   (b) Compuate R^2 for this analysis

   (c) Compute a 95% confidence interval for the mean difference between site 1 and site 2

   (d) Use ANOVA to test the equality of means among the four strains. Interpret the results within the context of the problem

2. A researcher is studying the effectiveness of three methods of reducing smoking. He wants to determine whether the mean reduction in the number of cigarettes smoked daily differs from one method to another among men patients. Each smoked about 60 cigarettes per day before treatment. Four randomly chosen members of the group pursue method I; four pursue method II ; and so on. The reductions in the number of cigarettes smoked daily are as follows:

Method		
I	II	III
10	19	11
9	20	13
9	21	15
8	20	13

   (a) Make a dotplot of the number of reductions versus the methods. Do the samples look like they were drawn from populations with equal mean? Do the samples have low skewness and similar standard deviation?

   (b) Use lm() and anova() in R to find the ANOVA table. From this table, what is the numerical estimate of the common standard deviation in all populations?

   (c) Use ANOVA to test the equality of mean reductions among three methods. Interpret your results in the context of the problem

   (d) Use Tukey's honestly significant difference (HSD) to compute simultaneous 95% confidence intervals for all pairwise differences

III Solutions of the Practice problems

1. (a) $n = 22 + 72 + 77 + 11 = 182$,The grand mean (overall mean)

$$\begin{aligned}\bar{Y} &= \sum_{i=1}^{k} \frac{n_i \bar{Y_i}}{n} \\ &= \frac{22 * 1.86 + 72 * 5.75 + 77 * 6.44 + 11 * 11.37}{182} = 5.91\end{aligned}$$

$$\begin{aligned}SS_{group} &= \sum_{i=1}^{k} n_i (\bar{Y_i} - \bar{Y})^2 \\ &= 22 * (1.86 - 5.91)^2 + 72 * (5.75 - 5.91)^2 + 77 * (6.44 - 5.91)^2 + 11 * (11.37 - 5.91)^2 \\ &= 712.25\end{aligned}$$

$$\begin{aligned}SS_{error} &= \sum_{i=1}^{k} (n_i - 1) s_i^2 \\ &= 21 * 1.22^2 + 71 * 2.46^2 + 76 * 3.42^2 + 10 * 1.96^2 \\ &= 1388.26\end{aligned}$$

ANOVA table

Source	df	SS	MS	F	p-value
groups(sites)	3	712.25	237.42	30.44	<0.0001
error	178	1388.26	7.80		
Total	181	2100.51			

(b) the $R^2 = \frac{SS_{group}}{SS_{total}} = 712.25/2100.51 = 0.34$

(c) By the formula $SE(\bar{Y_i} - \bar{Y_j}) = \sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ we have $SE(\bar{Y_1} - \bar{Y_2}) = \sqrt{7.8} * \sqrt{1/22 + 1/72} = 0.68$. $\bar{Y_1} - \bar{Y_2} = -3.89$, $t^* = 1.97$ (quantile of t-distribution with d.f of 178). So the 95% C.I. is

$$-3.89 - 1.97 * 0.68 < \mu_1 - \mu_2 < -3.89 + 1.97 * 0.68$$

that is $(-5.23, -2.55)$

(d) From the ANOVA table we find that the p-value ¡0.0001, we reject the null hypothesis that all 4 sites have the same mean concentration.

2. R code

```
method1=c(10,9,9,8)
method2=c(19,20,21,20)
method3=c(11,13,15,13)
alldata=c(method1,method2,method3)
trt=c(rep(1,4), rep(2,4), rep(3,4))
trt=factor(trt); trt
library(lattice)
dotplot(alldata~trt)
fit1=lm(alldata~trt)
anova(fit1)
```

```
Analysis of Variance Table

Response: alldata
          Df  Sum Sq Mean Sq F value    Pr(>F)
trt        2 248.000 124.000      93 9.748e-07 ***
Residuals  9  12.000   1.333

aov1 = aov(alldata~trt)
TukeyHSD(aov1)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = alldata ~ trt)
$trt
    diff       lwr       upr     p adj
2-1   11  8.720337 13.279663 0.0000008
3-1    4  1.720337  6.279663 0.0021919
3-2   -7 -9.279663 -4.720337 0.0000338
```

Note that the estimator of common standard deviation is $\sqrt{1.333} = 1.154$