MATH 4310 **Name**_ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Test #2

Problem 1: A study examines the vital capacity measurements of 60 adult males classified by 4 different types of occupations and three age groups. The file **test21.csv** contains the values of vital capacity (VC) vs age group (AGE) and occupation (OCC).

i) Apply the Anova test to answer the following questions: (a) does the vital capacity differs among individuals with different occupations, (b) does the vital capacity differences among individuals with different age groups, and (c) is there an interaction between age and occupation? Use $\alpha = 0.01$ for all the tests.

ii) Use the Tukey's HSD procedure to test for significant differences among individual pairs of means for age group and occupation, if appropriate (you can ignore the interaction term in the Tukey's HSD procedure). Justify your conclusion.

(i) *We use the Anova to test the null hypothesis that there is no difference among the means of (a) individuals with different occupations, (b) different age and (b) that there is no interaction between age and occupation.*

```
> data21 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test21.csv")
> data21$AGE = factor(data21$AGE,levels=unique(data21$AGE))
> data21$OCC = factor(data21$OCC,levels=unique(data21$OCC))
> data21.model = aov(VC~AGE+OCC+AGE:OCC, data = data21)
> anova(data21.model)

Analysis of Variance Table

Response: VC
             Df  Sum Sq Mean Sq F value    Pr(>F)
AGEGROUP      2 12.3088  6.1544 29.3817 4.652e-09 ***
OCC           3 19.7785  6.5928 31.4750 2.129e-11 ***
AGEGROUP:OCC  6  8.9489  1.4915  7.1205 1.825e-05 ***
Residuals    48 10.0542  0.2095
```

The p-value in the above shows that for each of the 3 cases p-value < 0.01. Thus, we reject the null hypothesis, and we conclude that: (a) vital capacity differs among individuals with different occupations, (b) vital capacity differs among individuals with different age groups, and (c) there is an interaction between age and occupation

(ii) *We run the Tukey's HSD procedure:*

```
> TukeyHSD(data21.model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = VC ~ AGE + OCC + AGE:OCC, data = data21)

$AGE
       diff         lwr       upr     p adj
2-1 -0.7395 -1.08952404 -0.389476 0.0000164
3-1  0.3465 -0.00352404  0.696524 0.0528802
3-2  1.0860  0.73597596  1.436024 0.0000000

$OCC
         diff         lwr       upr     p adj
b-a 0.2073333 -0.23743004 0.6520967 0.6045104
c-a 0.4613333  0.01656996 0.9060967 0.0393618
d-a 1.4940000  1.04923663 1.9387634 0.0000000
c-b 0.2540000 -0.19076337 0.6987634 0.4338300
d-b 1.2866667  0.84190330 1.7314300 0.0000000
d-c 1.0326667  0.58790330 1.4774300 0.0000008
```

For the age factor, we observe p-value < 0.01 only for the comparison 2-1 and 3-2. For the occupation factor, we observe p-value < 0.01 only for the comparison d-a, d-b and d-c. All the other comparisons are not statistically significant at level $\alpha = 0.01$.

Problem 2: An experiment was run on six pregnant women to evaluate the effect of labor on glucose production and utilization. Glucose concentrations were collected on the six subjects during four stages of labor: latent (A1) and active (A2) phases of cervical dilatation, fetal expulsion (B), and placental expulsion (C); data are stored in file **test22.csv**

i) Apply the Anova test (with blocks) to answer the following question: (a) is there an effect of labor on glucose production and utilization? (b) Is the experimental design balanced or not? Use $\alpha = 0.01$ for all these tests. [Hint: the subject variable is the factor block]

ii) Use the Tukey's HSD procedure to test for significant differences among the four stages of labor, if appropriate.

(i) *We apply the two-way anova with blocks. This is an additive model where the first term in the formula is the block factor.*

```
> data22 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test22.csv")
> data22$GROUP = factor(data22$GROUP,levels=unique(data22$GROUP))
> data22$SUBJ = factor(data22$SUBJ,levels=unique(data22$SUBJ))
> str(data22)
'data.frame':   24 obs. of  3 variables:
 $ GC   : num  3.6 3.53 4.02 4.9 4.06 3.97 4.4 3.7 4.8 5.33 ...
 $ GROUP: Factor w/ 4 levels "A1","A2","B",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ SUBJ : Factor w/ 6 levels "1","2","3","4",..: 1 2 3 4 5 6 1 2 3 4 ...
> table(data22$GROUP,data22$SUBJ)

    1 2 3 4 5 6
  A1 1 1 1 1 1 1
  A2 1 1 1 1 1 1
  B  1 1 1 1 1 1
  C  1 1 1 1 1 1
> data22.model = aov(GC~SUBJ+GROUP, data = data22)
> anova(data22.model)
Analysis of Variance Table

Response: GC
          Df Sum Sq Mean Sq F value    Pr(>F)
SUBJ       5 8.7735 1.75470   6.426 0.0022156 **
GROUP      3 8.3409 2.78030  10.182 0.0006583 ***
Residuals 15 4.0960 0.27306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The table shows that the experimental design is balanced.
Since the p-value corresponding to the GROUP variable is less than 0.01, we conclude that there is a statistically significant effect of labor on glucose production and utilization.
NOTE: solution is the same using GC~SUBJ+GROUP or GC~GROUP+SUBJ (due to balanced design)

```
> TukeyHSD(data22.model, which = "GROUP")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GC ~ SUBJ + GROUP, data = data22)
```

```
$GROUP
           diff         lwr      upr     p adj
A2-A1 0.6666667 -0.20287041 1.536204 0.1653989
B-A1  1.3366667  0.46712959 2.206204 0.0024454
C-A1  1.4816667  0.61212959 2.351204 0.0009660
B-A2  0.6700000 -0.19953708 1.539537 0.1624015
C-A2  0.8150000 -0.05453708 1.684537 0.0699315
C-B   0.1450000 -0.72453708 1.014537 0.9622295
```

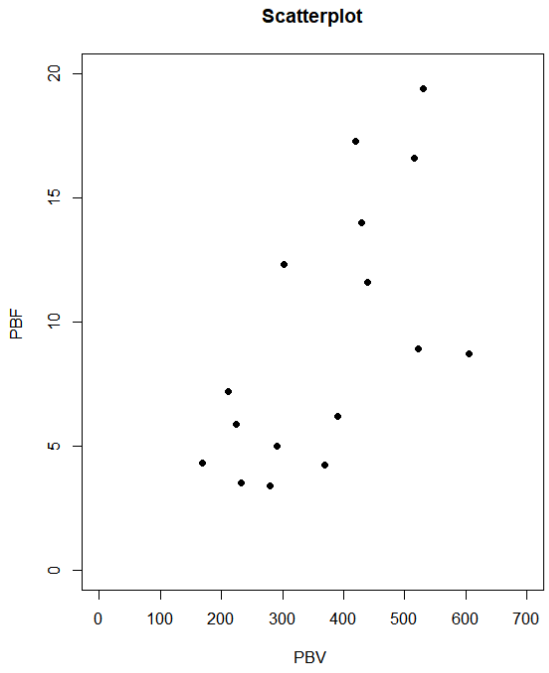
The Tukey test shows that there is a statistically significant difference (p_adj < 0.01) between the stages B-A1 and C-A1.

The differences between the other stages are not statistically significant.

Problem 3: Pulmonary blood flow (PBF) and pulmonary blood volume (PBV) values were recorded for 16 infants and children with congenital heart disease, see file: **test23.csv**

- **(i)** Write the equation of the linear regression equation of the PBF (the response variable) as a function of the PBV (the explanatory variable). (round to 3 decimal digits)
- **(ii)** Test the hypothesis H0: $\beta_1=0$ vs H1: $\beta_1\neq 0$ with significance level $\alpha = 0.05$ and $\alpha = 0.005$
- **(iii)** Compute the approximate 95\% confidence interval of β_1. (round to 3 decimal digits)
- **(iv)** Would we obtain the same value of β_1 (regression coefficient) and r^2 (coefficient of determination) if we interchange x and y in the R formulas? Explain why we obtain or we do not obtain the same quantity.

```
> data23 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test23.csv")
> x <- data23$PBV
> y <- data23$PBF
> plot(x, y, main="Scatterplot",xlab="PBV ", ylab="PBF ", pch=19)
```




```
> relation <- lm(y~x)
> print(summary(relation))
Call:
lm(formula = y ~ x)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-6.4389 -3.5963  0.1949  3.3508  6.7782

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.028332   3.267897  -0.009  0.99320
x            0.025119   0.008331   3.015  0.00927 **
---
Residual standard error: 4.262 on 14 degrees of freedom
Multiple R-squared:  0.3937,   Adjusted R-squared:  0.3504
F-statistic: 9.091 on 1 and 14 DF,  p-value: 0.009269
```

i) We write the equation of the regression line

y = -0.028+ 0.025 x

ii) Test the hypothesis about β_1 at significance levels $\alpha = 0.05$ and $\alpha = 0.005$.

We use the p-value = 0.00927 from the R output and conclude that, at level $\alpha = 0.05$ we do reject H0 since p- value < 0.05 but at level $\alpha = 0.005$ we do not reject H0 since p-value>0.005

iii) We compute the approximate 95\% confidence interval of β_1. Since n=16, the number of degrees of freedom is r=16-2=14. Hence t(alpha/2,r=14) = qt(1-0.05/2,14)=2.145

*C.I. = 0.025 +/- 2.145 *0.008 = (0.008,0.042)*

iv) Would we obtain the same value of β_1 and r^2 if we interchange x and y?

If we interchange x and y, the new value of β_1 would change since the model would become $x = \beta_0 + \beta_1 y$ and in this case β_1 would measure the slope of a different line (new slope is the reciprocal of the prior one).
The value of the correlation coefficient would not change since the model associated with the bivariate normal distribution is symmetric in x and y. The formula of the correlation coefficient is symmetric with respect to x and y. In fact, correlation measures the strength of linear relationship between two variables and this is independent of the order in which variables are taken.

```
> cor(y, x)
[1] 0.6274564
> cor(x,y)
[1] 0.6274564
```