# Lab Assignment No. 3: Answer Key

1) Provide descriptive statistics of the dataset.

When considering two independent groups it is important to review not only the pooled information both groups provide, but also the individual univariate measures of central distribution for each group. The pooled mean for the three variables was 124.7083 for the turtle's *Length*, 95.4375 for the *Width*, and 8.3656 for the *Width*. Of the three, *Length* has the largest standard error (2.9581) with the largest range of possible values (93 to 177). *Proc means* also provides the simple percentile confidence intervals for the three variables (the first type of confidence intervals mentioned in class), where the lower and upper bound correspond to the .025 and the .0975 percentile.

Table 1. Pooled percentile confidence intervals

| Variable | Mean | *SD* | *SE* | Lower | Upper |
|---|---|---|---|---|---|
| Length | 124.71 | 20.49 | 2.96 | 118.76 | 130.66 |
| Width | 95.44 | 12.68 | 1.83 | 91.76 | 99.12 |
| Height | 46.38 | 8.36 | 1.21 | 43.94 | 48.80 |

Next each individual group was reviewed. Since the two independent samples Hotelling's $T^2$ considers both samples to have been collected separately (as opposed to a paired samples design) each group should be reviewed by its own.

Table 2. Measures of central distribution for male turtles

| Variable | Mean | *SD* | *SE* | Lower | Upper |
|---|---|---|---|---|---|
| Length | 113.38 | 11.78 | 2.40 | 108.40 | 118.35 |
| Width | 88.29 | 7.07 | 1.44 | 85.30 | 91.28 |
| Height | 40.71 | 3.36 | .68 | 39.29 | 42.12 |

Table 3. Measures of central distribution for female turtles

| Variable | Mean | *SD* | *SE* | Lower | Upper |
|---|---|---|---|---|---|
| Length | 136.04 | 21.25 | 4.34 | 127.07 | 145.01 |
| Width | 102.58 | 13.10 | 2.67 | 97.05 | 108.12 |
| Height | 52.04 | 8.04 | 1.64 | 48.64 | 55.44 |

From Tables 2 and 3 it follows that the female turtle groups resembles the pooled data much more closely than the male turtle group. Moreover, the female group also seems to have much more variability in particular when regarding their lengths (*SD* = 21.25, which is almost double that of the male group). A bivariate investigation of variable pairings for both the male and female groups suggests that multivariate normality may be present (see Figure 1 and Figure 2).

Figure 1. Scatterplot matrix for male turtle data (N=24)

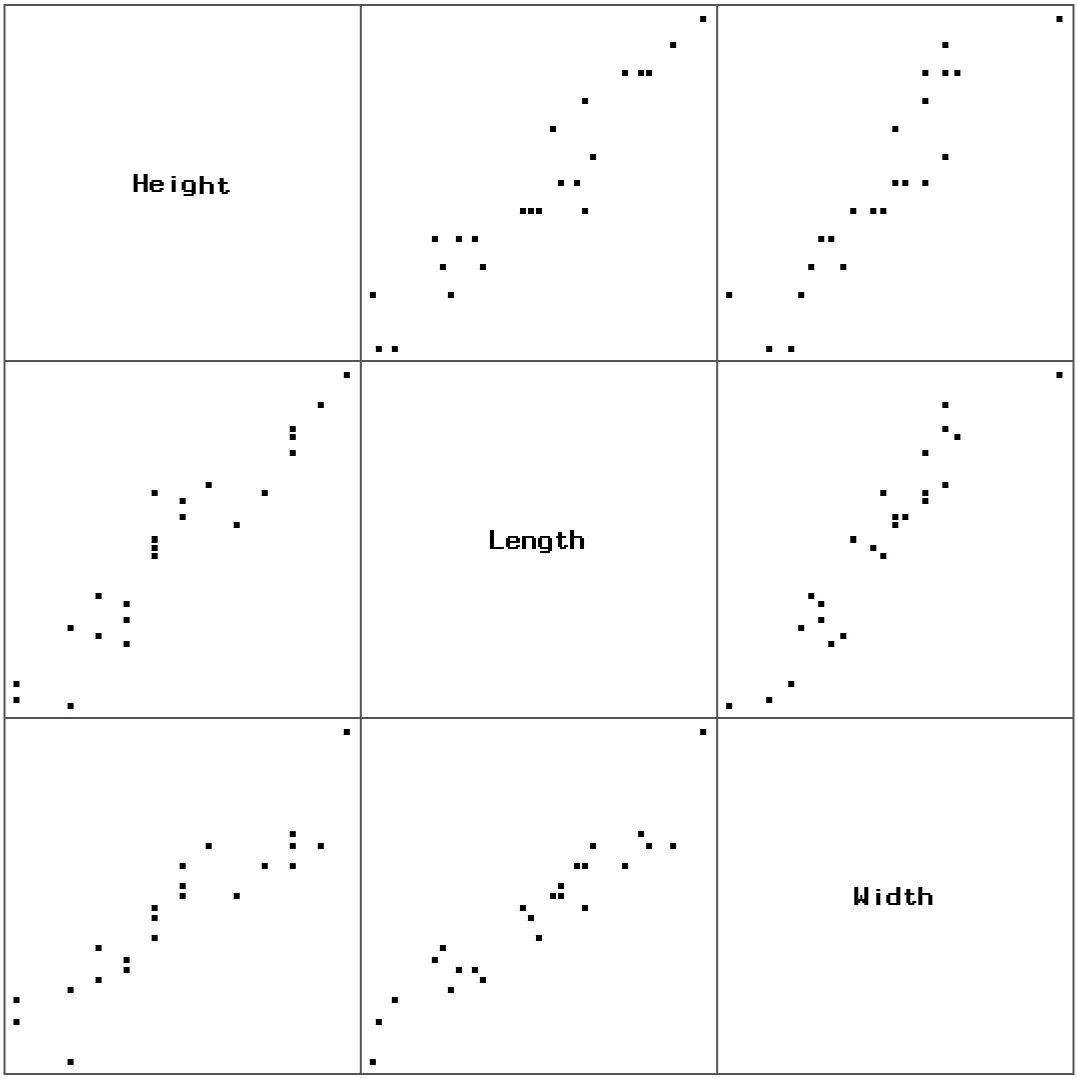
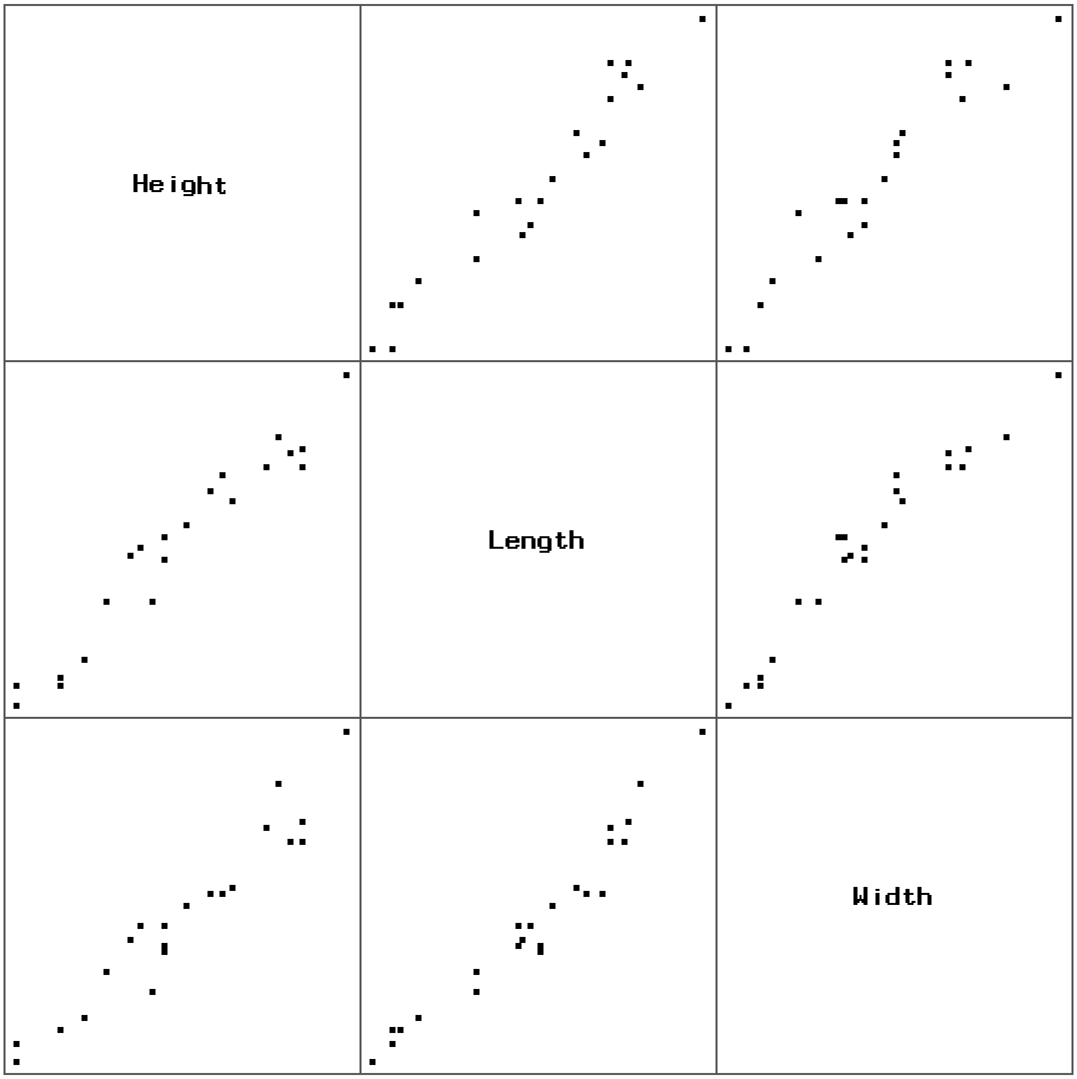


Figure 2. Scatterplot matrix for female turtle data (N=24)



2) Check for the assumption of multivariate normality and equal variance/covariance matrices.

It is important that when evaluating multivariate normality it is done separately within both groups. Since each group is compared against the other the pooled assessment would be inappropriate because the pooled scenario is never considered.

In order to evaluate if the data contain multivariate outliers two specific aspects should be investigated: Leverage values (or Mahalonobis Distances) and Mardia's skewness and kurtosis estimates. For the simplicity in the computations we shall consider leverage values in this example for indicators of potential multivariate outliers. Again, three possible cut-off criterions could be considered when evaluating individual leverage values. These three would be:

[1] $$\frac{2 \times p}{N} = \frac{2 \times 3}{24} = .2500$$

[2] $$\frac{2 \times (p+1)}{N} = \frac{2 \times 4}{24} = .3333$$

[3] $$\frac{\chi_{df=3,\ \alpha=.01}}{N-1} + \frac{1}{N} = \frac{11.3449}{24-1} + \frac{1}{24} = .4932 + .0417 = .5349$$

Given the persistence of the authors of our text book (Raykov & Marcoulides, 2008) that we ought to consider more liberal cut-off criteria when debating the removal of multivariate outliers we would continue with the third criterion (which is suggested in the book). This decision is further substantiated when later evaluating the multivariate skewness and kurtosis. In the case of both groups, the maximum leverage values were .4445 for the female turtle sample and .3148 for the male turtles. Since neither of those values exceeded the .5349 cut-off value no observations were flagged as multivariate outliers. Individual group Mardia's skewness and kurtosis estimates were computed (Table 4).

Table 4. Mardia's multivariate normal distribution estimates

| Group | Skewness | *p* | Kurtosis | *p* |
|---|---|---|---|---|
| Male | 1.4567 | .7279 | 14.5870 | .8535 |
| Female | 2.8099 | .1995 | 12.7463 | .3135 |

Both groups had non-significant skewness and kurtosis estimates suggesting that the data are multivariate normally distributed. Further, this also supports the usage of the earlier decided on cut-off criterion for the leverage values. If we would have missed an influential observation in the data it would have been most likely been detected by Mardia's coefficients.

In the case of a two independent samples Hotelling's $T^2$ the assumption of equal covariance matrices ought to be evaluated. Much like in the case of Levene's univariate two-samples *t*-test evaluation of equal variances, Bartlett's Test assesses the equality of equal variance/covariance matrices. In this example Bartlett's Test was requested through the *proc glm* statement. The chi-square approximation of Bartlett's statistic was $\chi^2$(df=6) = 23.40, *p* < .001, which suggests that the two groups have unequal variance/covariances. However, for the purpose of this exercise we shall continue as if the test would have been non-significant. Technically, at this point we ought to apply the appropriately corrected formulas in order to make informed decisions regarding the two independent samples.

3) Assuming that $\Sigma_1 = \Sigma_2$, test $H_o : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ at the $\alpha = 0.05$ level using Hotelling's $T^2$ test. Discuss.

The two independent samples Hotelling's $T^2$ provided positive evidence against the null hypothesis. With a $T^2 = 72.38$, $F\ (3, 44) = 23.08$, $p < .001$, we can conclude that there is at least one pairing of means in the three measured turtle dimensions on which the two genders differ significantly. However, in order to further investigate these relationships we ought to consider the simultaneous confidence intervals for all three variables.

4) What are the simultaneous 95% confidence intervals?

The simultaneous 95% confidence intervals can be computed for each individual variable *i* by using the following relational formula:

$$95\%\ C.I. = (\bar{x}_{1i} - \bar{x}_{2i}) \pm \sqrt{\frac{p \times (n_1 + n_2 - 2)}{\frac{1}{n_1} + \frac{1}{n_2}} \times F_{p,\ n_1+n_2-p-1} \times \frac{S_{pooled}}{n_1 + n_2 - p - 1}}\ ,$$

where

$$S_{pooled} = \frac{(n_1 - 1) \times S_1 + (n_2 - 1) \times S_2}{n_1 + n_2 - 2}\ .$$

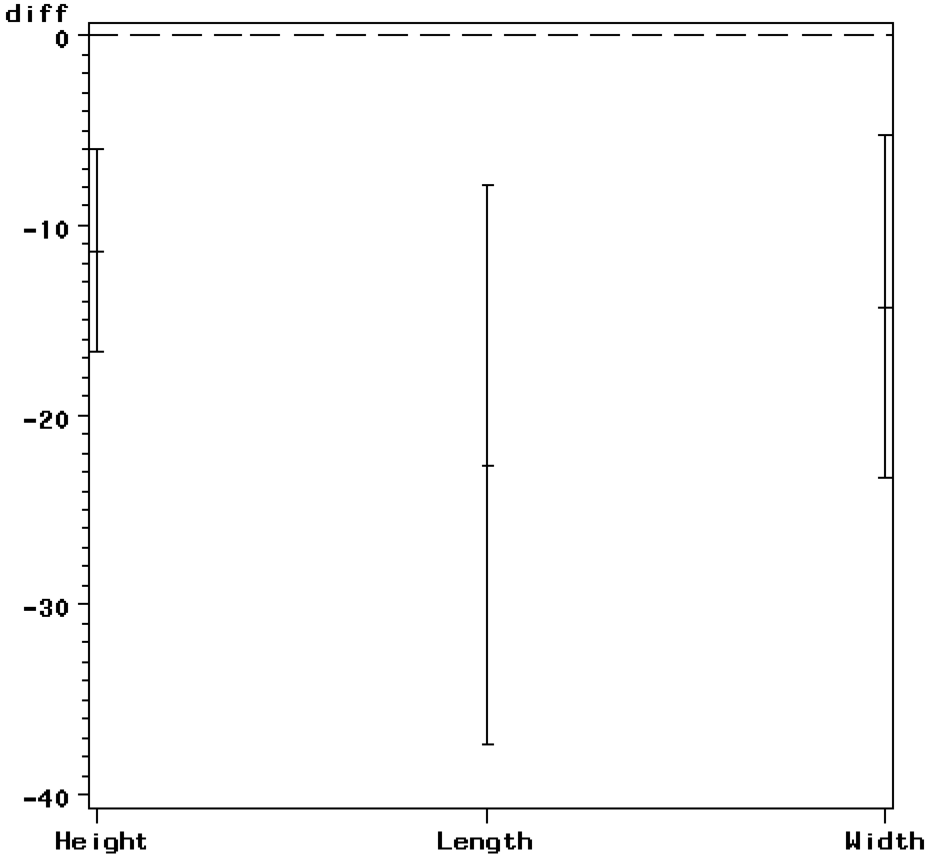
Table 5 shows the corresponding set of 95% simultaneous confidence intervals. Also, these confidence intervals can be used to produce a confidence interval plot that can assist in the interpretation of the values (Figure 3).

Table 5. 95% simultaneous confidence intervals

| Variable | *F* | $S_{\text{pooled}}$ | $\bar{x}_{1i} - \bar{x}_{i2}$ | Lower | *Upper* |
|---|---|---|---|---|---|
| Height | 2.82 | 38.00 | -11.33 | -16.62 | -6.04 |
| Length | 2.82 | 295.14 | -22.67 | -37.41 | -7.93 |
| Width | 2.82 | 110.89 | -14.29 | -23.33 | -5.26 |

In both Table 5 as well as Figure 3 it is apparent from the negative mean differences that the female group ($\bar{x}_2$) is the larger of the two. Since none of the three confidence intervals includes zero in their interval it can be concluded that they all vary significantly from one another.

Figure 3. 95% simultaneous confidence intervals



5) What specific conclusions can your draw about the three variables in regard to the two turtle genders?

Form these confidence intervals it can be deduced that all three variables vary significantly between the two genders. Since we are subtracting the female variable means from the male ones and the resulting mean difference is negative we can infer that the female turtles are on average larger than the male ones. Moreover, since none of the three simultaneous confidence intervals include zero we can further state that the two turtle genders differ significantly along all three of the biological measures. Further, from the confidence intervals the prior information about the turtle length is substantiated that it is the one measurement with the largest standard error.

6) What are the Bonferroni corrected 95% confidence intervals?

The corresponding Bonferroni corrected 95% confidence intervals (from class known as type II confidence intervals) are also produced by the syntax provided in lab, and are here summarized in Table 6.

Table 6. 95% Bonferroni corrected confidence intervals

| Variable | *t* | $S_{\text{pooled}}$ | $\bar{x}_{1i} - \bar{x}_{i2}$ | Lower | *Upper* |
|---|---|---|---|---|---|
| Height | 2.48 | 38.00 | -11.33 | -15.75 | -6.91 |
| Length | 2.48 | 295.14 | -22.67 | -34.99 | -10.34 |
| Width | 2.48 | 110.89 | -14.29 | -21.84 | -6.74 |

The Bonferroni corrected 95% confidence intervals are not as conservative as the previously reported simultaneous confidence intervals. Nonetheless, it follows that again all of the three variables differ significantly between the two gender groups.

7) How do these last confidence intervals differ from the previous? Which of the two sets would you utilize and why?

The Bonferroni corrected 95% confidence intervals are estimated using a slightly different formula, which relies on a *t* critical value, rather than the *F* statistic seen earlier.

$$Bonferroni\ 95\%\ C.I. = (\bar{x}_{i1} - \bar{x}_{i2}) \pm t_{CRIT} \times \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times S_{pooled}}\ ,$$

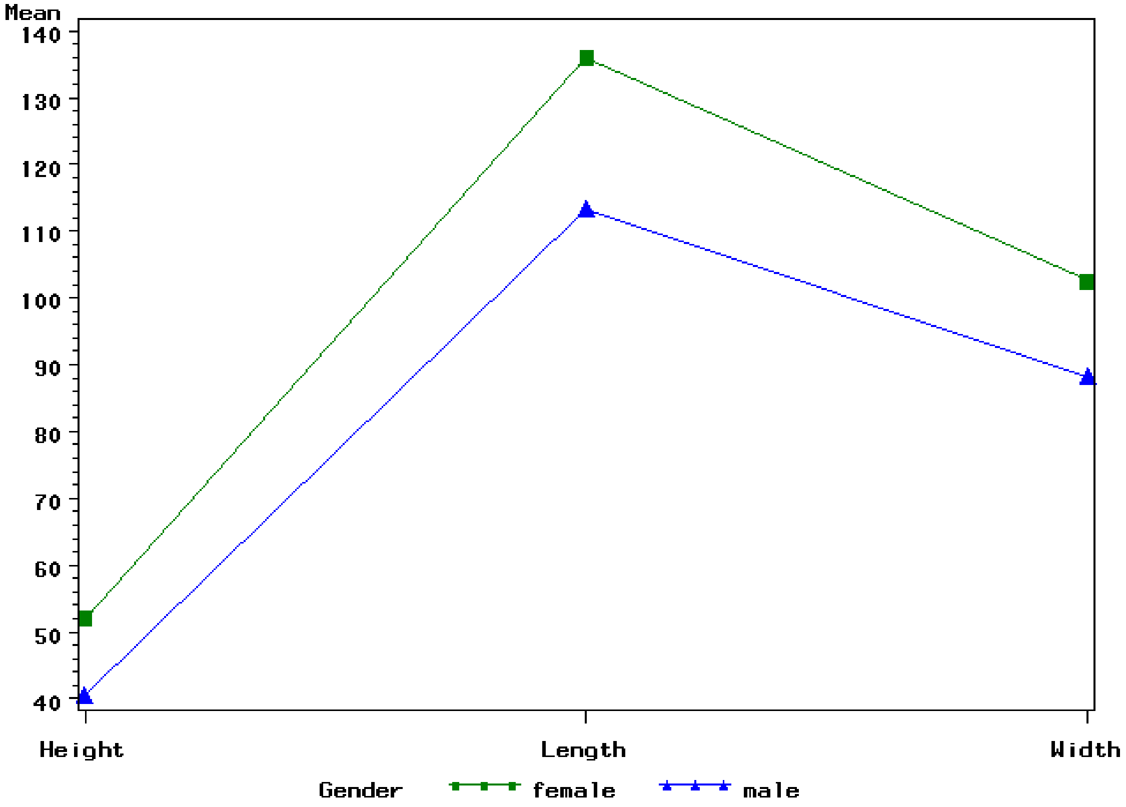
whereby the corresponding $\alpha$ value used in the determination of the critical *t* value is divided by the number of variables for which confidence intervals are being estimated. Bonferroni corrected confidence intervals are also more appropriate when no linear combinations between the three variables are sought. In our specific example all three measurements are those of biological dimensions of turtles. Therefore it is safe to assume that linear combinations of the three variables are highly plausible. This makes the Bonferroni corrected 95% confidence intervals a second choice to the simultaneously estimated confidence intervals.

8) Provide a traditional profile analysis means plot.

The traditional profile plot shows the means of each variable separately for the groups of interest. This is only appropriate for variables which are all measured on the same scale. If variables were not on the same scaled they would have to be transformed in order to be compatible for this type of visual inspection. Figure

4 shows the three variables seem to vary systematically for the two groups in that male dimensions are always smaller than their female counterparts.

Figure 4. Traditional means profile plot

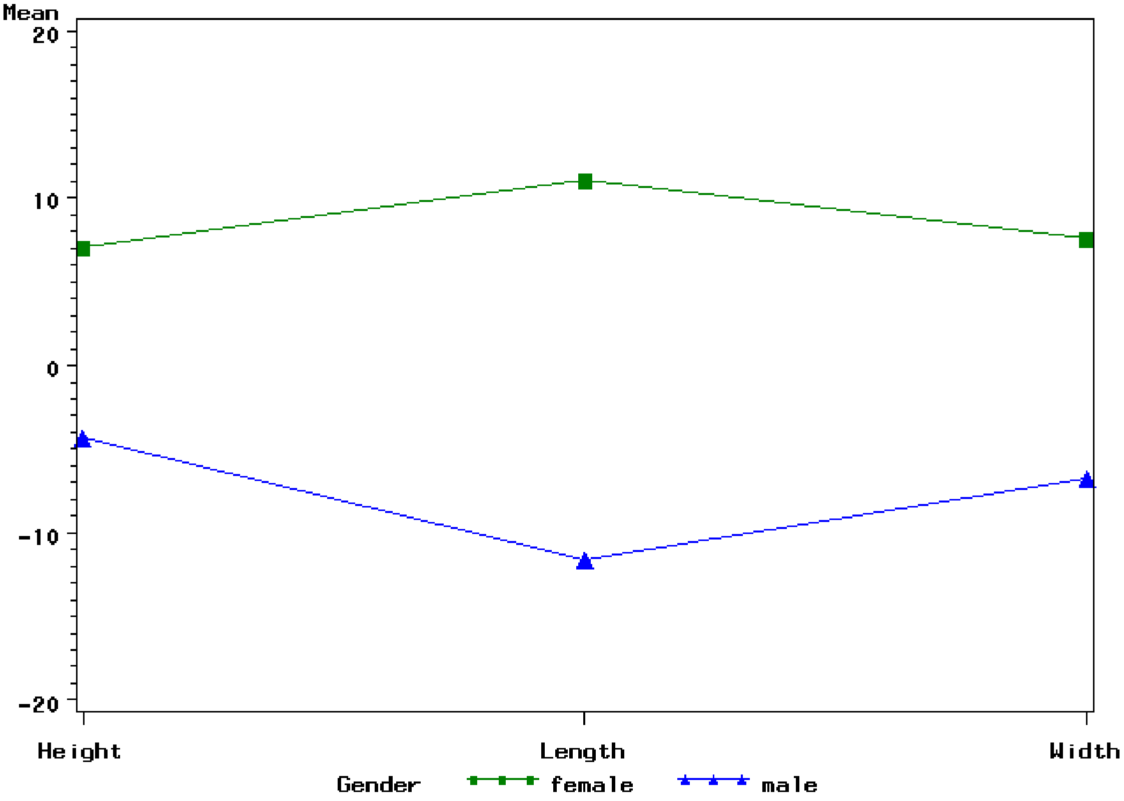


9) Given that a biologist would expect an average length of 125cm, width 95cm and a height of 45cm in general, give the deviation means plot for the two turtle genders.

In most instances variables that are being compared may have an expected standard against we wish to compare final results. This can often act as a sampling/quality check were sampled means can be compared against an expected value. In our example the average turtle dimensions, across both genders, were reported by the biologists as 125cm long, 95 cm wide and 45 cm tall. In order to see how much our two groups deviate specifically from these expected pooled means these means ought to be subtracted from each group respectively and the replotted.

Figure 5 shows the deviation means profile plot for the two groups. In this particular instance it appears that given this hypothesized pooled turtle size average (for the three variables) the male and female groups differed most significantly along the length of the turtle's body. In fact the traditional means plot would suggest that the two group profiles are parallel. The deviation means plot on the other hand is nearly parallel with the exception of the *Length* variable.

Figure 5. Deviation means profile plot



10) Are the two profiles parallel? Support your answer with the appropriate test. What final conclusions can you draw about the two turtle groups?

Judging the traditional means profile plot one would think the two profiles to be parallel. However, a more detailed investigation demonstrates that although the two groups do appear to vary by a constant (were females are reliably always above the expected pooled mean and males below) they seem to diverge non-parallel on the Length variable. In order to investigate parallel profiles lag deviation scores were computed between variables and their immediate predecessors. The new deviation scores were submitted to a follow up Hotelling's $T^2$ in order to test for the parallel nature of the two profiles. The resulting $T^2$ rejected the null hypothesis in favor of two non parallel profiles, $T^2$ = 15.42, $F$ (2, 45) = 7.54, $p$ < .01.

**Note.** The different *F* critical value results from the different number of degrees of freedom, since we are now submitting two deviation scores rather than three raw variables.

```
/* ---------------------------------------------------------- */
/*Lab Assignment No.3: Independant Samples T-squared (Syntax) */
/*                 Due February 25, 2009                 */
/* ---------------------------------------------------------- */

%let path='C:\ YOUR PATH HERE \';
libname turt &path;

/* QUESTION NO. 1 */

/* Importing the data */
proc import file='E:\Teaching\Mulivariate S09\Lab 6\Homework
3\turtle.xls'
     out = turt.turtle
     dbms = EXCEL2000;
run;

/* Reviewing your data */
proc means data=turt.turtle mean std stderr min max median clm;
run;
proc means data=turt.turtle mean std stderr min max median clm;
     by Gender;
run;

/* QUESTION NO. 2 */

/* When evaluating multivariate normlaity one ought to test each
   group individually rather than the poopled data */
data males;
     set turt.turtle;
     if Gender='male';
data females;
     set turt.turtle;
     if Gender='female';
run;

proc insight data=males;
  scatter Heigth Length Width * Heigth Length Width;
run;
quit;
proc insight data=females;
  scatter Heigth Length Width * Heigth Length Width;
run;
quit;

/* Let us investigate the leverage values
   for each of our turtle groups */
proc reg data=males noprint;
     model Heigth = Length Width / influence;
     output out=hiim H=Leverage;
run;
quit;
proc reg data=females noprint;
```

```
     model Heigth = Length Width / influence;
     output out=hiif H=Leverage;
run;
quit;

/* Critical leverage values we should consider */
proc iml;

     N = 24;
     hiicrit1 = (2*3)/N;
     hiicrit2 = 2*((3+1)/N);
     hiicrit3 = ((2*(gaminv(.99,1.5)))/(N-1))+(1/N);

     print 'Conservative leverage cut-off';
     print hiicrit1;
     print 'Regression leverage cut-off';
     print hiicrit2;
     print 'Raycov and Marcoulides leverage cut-off';
     print hiicrit3;

quit;
proc means data=hiim max;
     var Leverage;
run;
proc means data=hiif max;
     var Leverage;
run;

/* Producing the Mardia's coefficients for the
   two turtle genders */
data hiim;
     set males;
     drop Gender;
data hiif;
     set females;
     drop Gender;
run;
/* RUN THE MULTINORM MACRO FIRST */

%multnorm(data=hiim, var = Heigth Length Width)
%multnorm(data=hiif, var = Heigth Length Width)

/* Let us test for the assumption of homogeneity of
   the variance/covariance matrices */

/* This will produce Bartlett's Test */
proc discrim data=turt.turtle pool=test;
  class Gender;
  var Heigth Length Width;
run;

/* QUESTION NO. 3 */

/* Computing the Hotelling's T-squared Statistic */
```

```
proc iml;

  /* This will start a sub-routing or internal macro in proc iml
*/
  start hotel2;
    n1=nrow(x1);
    n2=nrow(x2);
    k=ncol(x1);
    one1=j(n1,1,1);
    one2=j(n2,1,1);
    ident1=i(n1);
    ident2=i(n2);

     /* Produces the mean vector for the 'male' group */
     ybar1=x1`*one1/n1;
    s1=x1`*(ident1-one1*one1`/n1)*x1/(n1-1.0);
     print 'Male Turtles';
     print n1 ybar1;
     print s1;

     /* Produces the mean vector for the 'female' group */
    ybar2=x2`*one2/n2;
    s2=x2`*(ident2-one2*one2`/n2)*x2/(n2-1.0);
     print 'Female Turtles';
     print n2 ybar2;
    print s2;

     /* Computing the pooled covariance matrix */
    spool=((n1-1.0)*s1+(n2-1.0)*s2)/(n1+n2-2.0);
    print spool;

     /* Computing Hotelling's T-squared */
     pp = spool*(1/n1+1/n2);

     print 'Covariances in computation';
     print pp;

    t2=(ybar1-ybar2)`*inv(spool*(1/n1+1/n2))*(ybar1-ybar2);

     /* Transforming the T-squared into an F statistic */
    f=((n1+n2-k-1)/(k*(n1+n2-2)))*t2;

     /* Producing the criterions */
    df1=k;
    df2=n1+n2-k-1;
    p=1-probf(f,df1,df2);
     fcrit = finv(.95, df1, df2);
    print t2 f df1 df2 p fcrit;
  finish;

  use turt.turtle;
    read all var{Heigth Length Width} where (Gender="female")
into x1;
```

```
    read all var{Heigth Length Width} where (Gender="male") into
x2;
  run hotel2;

quit;

/* QUESTION NO. 4 - 7 */

/* Simultaneous Confidence Intervals */
/* Setting a parameter equal to the variable number */
%let p=3;

/* Making a temporary dataset that is only the real notes
   Also, note that we are flattening the datafile by having all
   the observations in one column, and not organized by variables
*/
data males;
  set turt.turtle;
  if Gender="male";
  variable="Heigth";   x=heigth; output;
  variable="Length";   x=length;   output;
  variable="Width";    x=width;  output;
  keep Gender variable x;
 run;
proc sort data=males;
  by variable;
 run;

/* Saving out a dataset that will contain not only the
   specific unique means for each real note variable type
   but also the corresponding variances */
proc means data=males noprint;
  by variable;
  var x;
  id Gender;
  output out=pop1 n=n1 mean=xbar1 var=s21;
run;

/* Same as with the real data we will produce the corresponding
   matrix of means and variances for the fake group */
data females;
  set turt.turtle;
  if Gender="female";
  variable="Heigth";   x=heigth; output;
  variable="Length";   x=length;   output;
  variable="Width";    x=width;  output;
  keep Gender variable x;
proc sort data=females;
  by variable;
proc means data=females noprint;
  by variable;
  id Gender;
  var x;
  output out=pop2 n=n2 mean=xbar2 var=s22;
```

```
run;

/* We shall combine the two produced outputs
   into a permanent dataset */
data turt.combine;
  merge pop1 pop2;
  by variable;
  xdiff=xbar1-xbar2;
  f=finv(0.95,&p,n1+n2-&p-1);
  t=tinv(1-0.025/&p,n1+n2-2);
  sp=((n1-1)*s21+(n2-1)*s22)/(n1+n2-2);
  losim=xbar1-xbar2-sqrt(&p*(n1+n2-2)*f*(1/n1+1/n2)*sp/(n1+n2-&p-
1));
  upsim=xbar1-xbar2+sqrt(&p*(n1+n2-2)*f*(1/n1+1/n2)*sp/(n1+n2-&p-
1));
  lobon=xbar1-xbar2-t*sqrt((1/n1+1/n2)*sp);
  upbon=xbar1-xbar2+t*sqrt((1/n1+1/n2)*sp);
  drop Type _TYPE_ _FREQ_;
run;
proc print;
run;

/* We can now plot these confidence intervals.
   This will produce another temporary dataset with
   repeated means etc. to form the confidence interval */
data scis;
  merge pop1 pop2;
  by variable;
  f=finv(0.95,&p,n1+n2-&p-1);
  sp=((n1-1)*s21+(n2-1)*s22)/(n1+n2-2);
  diff=xbar1-xbar2; output;
  diff=xbar1-xbar2-sqrt(&p*(n1+n2-2)*f*(1/n1+1/n2)*sp/(n1+n2-&p-
1)); output;
  diff=xbar1-xbar2+sqrt(&p*(n1+n2-2)*f*(1/n1+1/n2)*sp/(n1+n2-&p-
1)); output;
  run;

/* This will plot the new temp data showing the confidence
intervals */
proc gplot data=scis;
  title 'Simoultaneous Confidence Intervals';
  axis1 length=4 in;
  axis2 length=6 in;
  plot diff*variable / vaxis=axis1 haxis=axis2 vref=0 lvref=21;
  symbol v=none i=hilot color=black;
run;
quit;

/* QUESTION NO. 8 */

/* Profile Plots */

/* Traditional profile plots for sample means
   need to be created from another flattened file */
```

```
data profile;
  set turt.turtle;
  variable="Heigth";   x=heigth; output;
  variable="Length";   x=length; output;
  variable="Width";    x=width;  output;
proc sort;
  by Gender variable;
proc means;
  by Gender variable;
  var x;
  output out=a mean=xbar;
run;

/* Plotting the means */
proc gplot;
  title 'Traditional Means Profile Plot';
  axis1 length=4 in label=("Mean");
  axis2 length=6 in;
  plot xbar*variable=Gender / vaxis=axis1 haxis=axis2;
  symbol1 v=K f=special h=2 i=join color=green;
  symbol2 v=L f=special h=2 i=join color=blue;
run;
quit;

/* QUESTION NO. 9 */

/* Deviation Means Profile Plot */

data profile2;
  set turt.turtle;
  variable="Heigth";   x=heigth-45; output;
  variable="Length";   x=length-125; output;
  variable="Width";    x=width-95;  output;
proc sort;
  by Gender variable;
proc means;
  by Gender variable;
  var x;
  output out=a2 mean=xbar;
run;

/* Plotting the means */
proc gplot;
  title 'Unique Means Profile Plot';
  axis1 length=4 in label=("Mean");
  axis2 length=6 in;
  plot xbar*variable=Gender / vaxis=axis1 haxis=axis2;
  symbol1 v=K f=special h=2 i=join color=green;
  symbol2 v=L f=special h=2 i=join color=blue;
run;
quit;

/* QUESTION NO. 10 */
/* Let us now test whether the two profiles are parallel */

/* First we are computing the lag difference scores */
data turt.diffset;
  set turt.turtle;
  diff1=Width-Length;
  diff2=Heigth-Width;
run;

/* Next we need to run the T-square proc iml statement this is
   the same as earlier with the difference that we are going to
   submit the difference scores and not the original variables */

proc iml;

title 'Testing for parallel profiles';
  start hotel2;
    n1=nrow(x1);
    n2=nrow(x2);
    k=ncol(x1);
    one1=j(n1,1,1);
    one2=j(n2,1,1);
    ident1=i(n1);
    ident2=i(n2);
    ybar1=x1`*one1/n1;
    s1=x1`*(ident1-one1*one1`/n1)*x1/(n1-1.0);
    print n1 ybar1;
    print s1;
    ybar2=x2`*one2/n2;
    s2=x2`*(ident2-one2*one2`/n2)*x2/(n2-1.0);
    print n2 ybar2;
    print s2;
    spool=((n1-1.0)*s1+(n2-1.0)*s2)/(n1+n2-2.0);
    print spool;
    t2=(ybar1-ybar2)`*inv(spool*(1/n1+1/n2))*(ybar1-ybar2);
    f=(n1+n2-k-1)*t2/k/(n1+n2-2);
    df1=k;
    df2=n1+n2-k-1;
    p=1-probf(f,df1,df2);
     fcrit=finv(.95,df1,df2);
     print t2 f df1 df2 p fcrit;
  finish;

  use turt.diffset;
    read all var{diff1 diff2} where (Gender="male") into x1;
    read all var{diff1 diff2} where (Gender="female") into x2;
  run hotel2;

quit;
```