Problem 1 (*adapted from J&W Exercise 6.23*)
The data can be found in "iris.dat". Columns correspond to sepal length, sepal width, petal length, petal width, and type (1,2,3 for setosa, versicolor, and virginica, respectively). *Note that this problem asks for analysis only on the two width variables.*
The MANOVA null hypothesis of interest is $H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \boldsymbol{\mu_3}$, where $\boldsymbol{\mu_i}$ is the vector of mean widths for the ith iris type. The sum of squares and cross products (SSCP) matrices for this hypothesis and error, respectively, are

$$\mathbf{H} = \begin{bmatrix} 11.3449 & -22.9327 \\ -22.9327 & 80.4133 \end{bmatrix} \quad \text{and} \quad \mathbf{E} = \begin{bmatrix} 16.9620 & 4.8084 \\ 4.8084 & 6.1566 \end{bmatrix}$$

With $\mathbf{T} = \mathbf{H} + \mathbf{E}$, we can collect these in a MANOVA table:

Source	SSCP	df
Treatments	**H**	2
Error	**E**	147
Total	**T**	149

Wilk's Lambda value is .0383 and corresponds to an F statistic of 299.94 with 4 and 292 degrees of freedom. This is significant evidence (p-value ≈ 0) that at least two of the iris types differ in at least one mean width measurement. Simultaneous 95% confidence limits are given below for each pair of iris types and width measurements. A Bonferroni adjustment of .05/6 is made to account for the multiplicity (individual confidence level is 99.17%).

Obs	Dependent	i	j	LowerCL	UpperCL
1	swidth	1	2	0.476296	0.839704
2	swidth	1	3	0.272296	0.635704
3	swidth	2	3	-0.385704	-0.022296
4	pwidth	1	2	-1.189470	-0.970530
5	pwidth	1	3	-1.889470	-1.670530
6	pwidth	2	3	-0.809470	-0.590530

The sample covariance matrices seem to be quite different, particularly the variances of the petal width, so the assumption of equal population covariance matrices is doubtful. SAS code for this problem is

```
proc glm data=iris;
  class type;
  model swidth pwidth = type / clparm alpha=.00833;
  estimate '1 vs 2' type 1 -1 0;
  estimate '1 vs 3' type 1 0 -1;
  estimate '2 vs 3' type 0 1 -1;
  manova h=type / printe printh;
  run; quit;
```

Problem 2

Twenty drivers participate in a repeated measures experiment done to investigate how talking on a cell phone affects driving skill. Each driver responds to driving situations on a simulator while experiencing three different conditions. The first condition is complete silence. In the second condition, the driver and a researcher have a conversation. In the third condition, the driver talks on a cell phone with the researcher. For each condition, the percentage of correct and timely responses to various driving situations is recorded. In the "skill.dat" dataset, these percentages are given in order of the conditions just described, with the first column corresponding to the driver's sex.

a) Briefly explain why the usual ANOVA model is not appropriate for modeling the relationships among the variables here.

The usual ANOVA model would assume that responses under different conditions are independent. This is not appropriate here because the same people experience all three conditions, and there would likely be correlation among their responses.

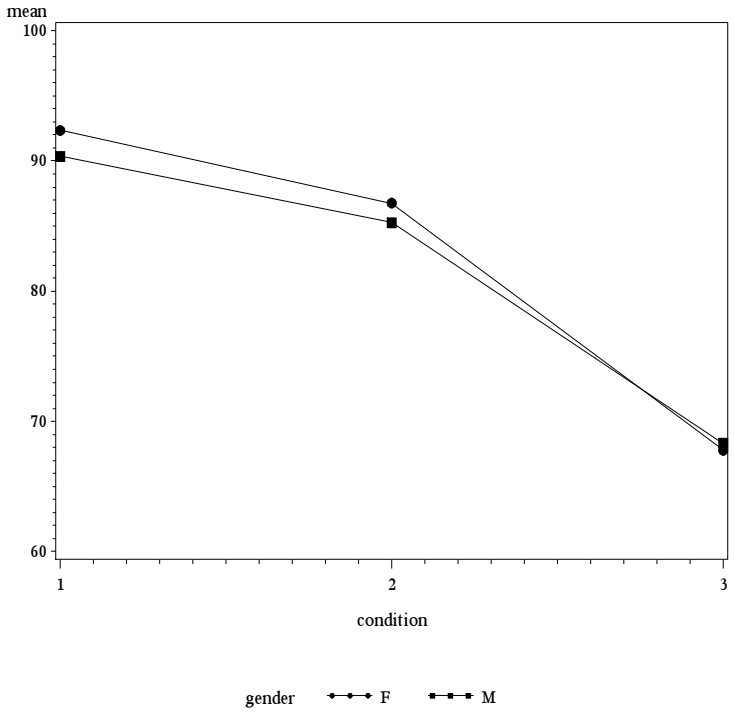
b) Explain what interaction between sex and driving condition would mean in this situation. Conduct the appropriate test for interaction between sex and driving condition.

Interaction between sex and driving condition would allow for a different mean response for each combination of sex and condition. This mean response would not necessarily be the sum of each individual factor's effect.

A test of H_0: no interaction versus H_a: interaction yields a test statistic of $F = .27$ with p-value .7653, which is not significant evidence to claim that sex interacts with the driving conditions here.

c) Provide a profile plot with the three conditions along the horizontal axis, the mean percentages given on the vertical axis, and with separate lines for the two sexes. Comment on how this supports your conclusion from part b) above.

The plot below shows nearly parallel lines (nearly one common line) for males and females. This is consistent with our test above, which found insignificant evidence for an interaction between sex and condition, because the effect of changing the driving condition is nearly constant for both sexes.



d) How do males and females compare in this situation? Test for an overall effect due to sex. Comment on the appropriateness of this test in light of your answers from parts b) and c).

The test of H_0 : no gender effect versus H_a gender effect results in a test statistic of $F = .19$ with p-value .6667. This is not significant evidence that the mean response times differ between the sexes. It should be noted that this test averages over the three driving conditions. This is reasonable to do from our test and plot above, which showed that the effect of gender is not significantly affected by the driving conditions.

e) Letting X_1, X_2, and X_3 denote the percentages for the three conditions, define $Y_1 = X_2 - X_1$ and $Y_2 = X_3 - X_2$, and consider Hotelling's T^2 test of $E(\mathbf{Y}) = \mathbf{0}$ versus $E(\mathbf{Y}) \neq \mathbf{0}$. In terms of the variables in this situation, what exactly is being tested with these hypotheses. Conduct this test.

The null hypothesis that $E(\mathbf{Y}) = \mathbf{0}$ means that both $E(X_1) - E(X_2) = 0$ and $E(X_2) - E(X_3) = 0$. This condition is equivalent to $E(X_1) = E(X_2) = E(X_3)$, or in words, that the population mean driving skills are equal for all three conditions.

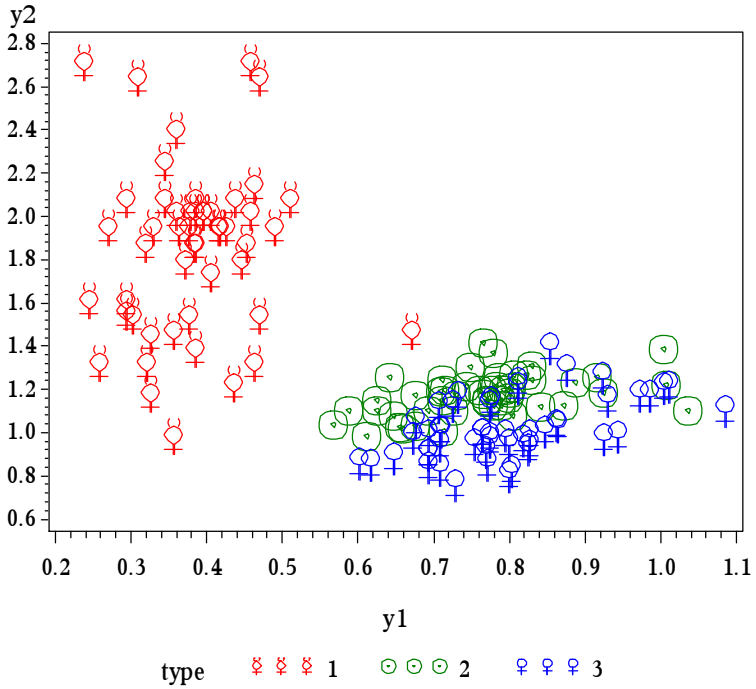
The F test statistic is 101.83 with 2 and 18 degrees of freedom. The p-value is approximately zero, which is significant evidence to reject H_0 and conclude that $E(\mathbf{Y}) \neq \mathbf{0}$, which implies that there is a condition effect on driving skill.

```
data driver;
 infile 'v:\skill.dat';
 input gender $ p1 p2 p3;
 y1=p2-p1;
 y2=p3-p2; run;
proc glm data=driver;
 class gender;
 model p1 p2 p3 = gender;
 manova h=gender m=p3-p2,p2-p1;
 manova h=gender m=p1+p2+p3; run;
data driver2;
  set driver;
  condition=1;  k=p1; output;
  condition=2;  k=p2; output;
  condition=3;  k=p3; output;
  drop p1 p2 p3;  run;
proc sort;
  by gender condition;  run;
proc means;
  by gender condition;
  var k;
  output out=a mean=mean;  run;
proc gplot;
  axis1 length=4 in;
  axis2 length=5 in;
  plot mean*condition=gender / vaxis=axis1 haxis=axis2;
  symbol1 v=J f=special h=2 i=join color=black;
  symbol2 v=K f=special h=2 i=join color=black;
  symbol3 v=L f=special h=2 i=join color=black;
  symbol4 v=M f=special h=2 i=join color=black;
  run;

proc iml;
start hotel;
mu0={0, 0};
one=j(nrow(x),1,1);
ident=i(nrow(x));
ybar=x`*one/nrow(x);
s=x`*(ident-one*one`/nrow(x))*x/(nrow(x)-1.0);
print mu0 ybar;
print s;
t2=nrow(x)*(ybar-mu0)`*inv(s)*(ybar-mu0);
f=(nrow(x)-ncol(x))*t2/ncol(x)/(nrow(x)-1);
df1=ncol(x);
df2=nrow(x)-ncol(x);
p=1-probf(f,df1,df2);
print t2 f df1 df2 p;
finish;
use driver;
read all var{y1 y2} into x;
run hotel;
quit;
```

Problem 3(*adapted from J&W Exercise 11.28*)

a) The plot below illustrates somewhat elliptical scatters for each group individually, suggesting bivariate normality for each group individually, but the orientations of the ellipses suggest different covariance matrices.



b) Using $\log Y_1$ only, the APER is $49/150 = .3267$ (49 observations were misclassified into the wrong population), and the discriminant scores are

$$\hat{d}_1(y) = -8.92 + 40.90y$$
$$\hat{d}_2(y) = -32.40 + 81.84y$$
$$\hat{d}_3(y) = -35.03 + 85.20y$$

Using $\log Y_2$ only, the APER is $34/150 = .2267$, and the discriminant scores are

$$\hat{d}_1(y) = -29.83 + 30.93y$$
$$\hat{d}_2(y) = -12.54 + 19.52y$$
$$\hat{d}_3(y) = -9.64 + 16.87y$$

Using $\log Y_1$ and $\log Y_2$, the APER is $26/150 = .1733$, and the discriminant scores are

$$\hat{d}_1(\mathbf{y}) = -33.06 + 26.81y_1 + 28.90y_2$$
$$\hat{d}_2(\mathbf{y}) = -37.92 + 75.10y_1 + 13.82y_2$$
$$\hat{d}_3(\mathbf{y}) = -38.40 + 79.94y_1 + 10.80y_2$$

c) The estimated expected AERs when using only $\log Y_1$ or $\log Y_2$ are identical to the APERs. However, when using both variables, the holdout (cross-validation) method misclassifies one additional observation so that its estimated expected AER is $27/150 = .18$.

The error rates for these shape variables are much higher than those in the previous example. This is due to the overlap between populations 2 and 3 seen in the plot above. So shape alone is not as an effective discriminator for these species.

d) The outlier (seen in the plot above) from population 1 is misclassified with the holdout method because it is not used to construct the discriminant rule. When using re-substitution, its influence on the sample 1 mean is apparently enough to get it (correctly) classified into population 1.

SAS code:

```
data iris;
  infile 'v:\505\datasets\iris.dat';
  input slength swidth plength pwidth type $;
  y1=log(slength/swidth);
  y2=log(plength/pwidth);
  run;
proc gplot data=iris;
  plot y2*y1=type;
  symbol1 v=b f=special h=2 i=join color=red interpol=none;
  symbol2 v=a f=special h=2 i=join color=green interpol=none;
  symbol3 v=c f=special h=2 i=join color=blue interpol=none;
  run; quit;
proc discrim data=iris pool=yes crossvalidate;
  class type;
  var y1;
  priors '1'=1 '2'=1 '3'=1;
  run;
proc discrim data=iris pool=yes crossvalidate;
  class type;
  var y2;
  priors '1'=1 '2'=1 '3'=1;
  run;
proc discrim data=iris pool=yes crossvalidate;
  class type;
  var y1 y2;
  priors '1'=1 '2'=1 '3'=1;
  run;
```