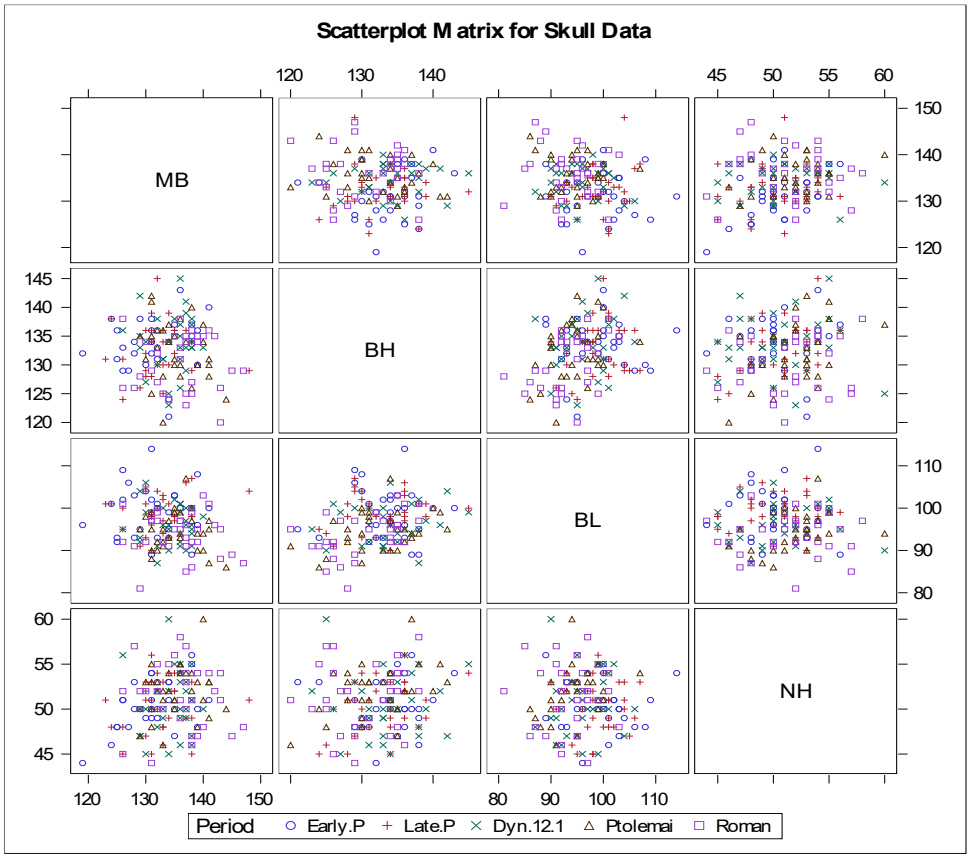
Background of Data Set

The skull data set gives four measurements on male Egyptian skulls: maximum breadth, basilbregmatic height, basialvelar length, and nasal height, all measured in millimeters. The data are taken from five different time periods.

Outliers

By time period for each variable, we consider boxplots for the univariate cases, create a scatterplot matrix for the bivariate cases, and calculate the standardized values and examine the generalized squared distances for the multivariate cases to identify outliers in the data set.



In the MB boxplot (not shown) by Period, Late.P has one outlier as it falls outside ±(1.5)IQR for the univariate case. The boxplots for BH has two outliers in Early.P and Ptolemaic while the boxplots for BL has one outlier during for Roman. The boxplot for NH shows no univariate outliers. The scatterplot matrix suggests there may be bivariate outliers for MB vs BH, MB vs BL, and BL vs NH as these graphs are not elliptical in shape from the origin. From SAS, the absolute standardized values $z_{ij} = \frac{x_{ij}-\overline{x_j}}{\sqrt{s_{jj}}}$ (i=1,…n and j=1,2,3,4) are all less than 3.5, indicating that there are no multivariate outliers in the data. For squared distances, we consider observations to be outliers if $d_j^2 > X^2_{p=4}(.99)=13.277$ (or if $d_j^2 > X^2_{p=4}(.995)=14.860$). In the data set, as the largest generalized squared distance (13.0225) is less than both of these values, this approach does not

detect any multivariate outliers either. As the multivariate analysis does not detect any outliers, we use the entire data set for the proceeding analysis.

MANOVA: comparing the mean vectors of skulls from the five different periods

Source of variation	Matrix of sum of squares and cross products	Degrees of freedom
Treatment	B	g-1=5-1=24
Residual (error)	W	$\sum_{l=1}^{g} n_l - g$ =150-5=145
Total (corrected for the mean)	B+W	$\sum_{l=1}^{g} n_l - 1$ =150-1=149

$$\boldsymbol{B} = \sum_{\boldsymbol{l=1}}^{\boldsymbol{g}} \boldsymbol{n_l}(\bar{\boldsymbol{x}}_{\boldsymbol{l}} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{\boldsymbol{l}} - \bar{\boldsymbol{x}})'$$

B				
	MB	**BH**	**BL**	**NH**
MB	502.82666667	-228.1466667	-626.6266667	135.43333333
BH	-228.1466667	229.90666667	292.28	-66.06666667
BL	-626.6266667	292.28	803.29333333	-180.7333333
NH	135.43333333	-66.06666667	-180.7333333	61.2

$$\boldsymbol{W} = \sum_{\boldsymbol{l=1}}^{\boldsymbol{g}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n_l}} (\boldsymbol{x_{lj}} - \bar{\boldsymbol{x}}_{\boldsymbol{l}})(\boldsymbol{x_{lj}} - \bar{\boldsymbol{x}}_{\boldsymbol{l}})'$$

W				
	MB	**BH**	**BL**	**NH**
MB	3061.0666667	5.3333333333	11.466666667	291.3
BH	5.3333333333	3405.2666667	754	412.53333333
BL	11.466666667	754	3505.9666667	164.33333333
NH	291.3	412.53333333	164.33333333	1472.1333333

MANOVA: the analysis

We want to compare the mean vectors of MB, BH, BL, and NH measurements from the five different time periods or populations. We use one-way MANOVA as we only want to analyze one response variable (time period) with g=4 levels (groups) and p=4 variables.

Model:

$$X_{lj} = \boldsymbol{\mu}_l + \boldsymbol{\epsilon}_{lj} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \boldsymbol{\epsilon}_{lj}$$

where $\boldsymbol{\epsilon}_{ij} \sim \text{iid} N_p(0, \Sigma)$

and $\sum_{l=1}^{g} \boldsymbol{\tau}_l = 0$ for l=1,,g and j=1,…n

Hypothesis Test:

H_o: $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\boldsymbol{\mu}_3=\boldsymbol{\mu}_4$ vs

H_a: $\mu_{ij} \neq \mu_{jk}$ for at least one i ≠j and at least one variable k (that at least one mean is different)

Assumptions:

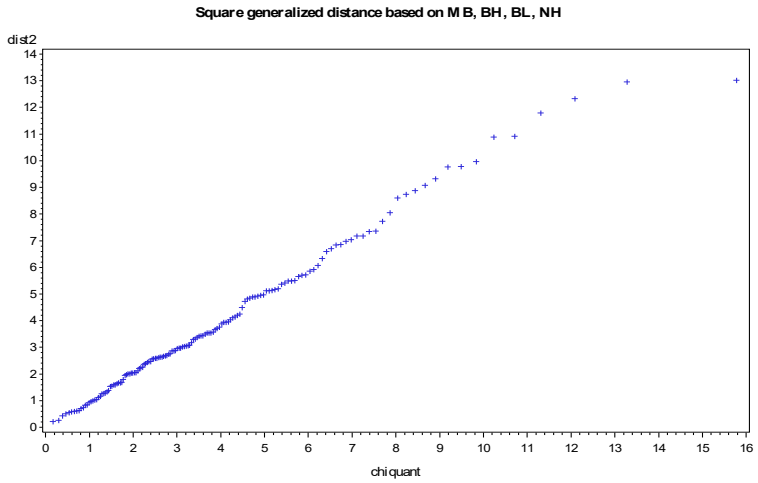
1) $\boldsymbol{X}_{l1}, \boldsymbol{X}_{l2}, \ldots \boldsymbol{X}_{ln\_l}$, is a random sample of size n_l from a population with mean $\boldsymbol{\mu}_l$, l=1,…,g. The random samples from different population are independent.

2) All populations have a common variance-covariance matrix Σ.
3) Each population is multivariate normal

The random samples from different time periods are independent as a skull from Early.P does not depend on the other time periods; skulls from different time periods to do depend on skulls from the other time periods.

Chi-Square	DF	Pr > ChiSq
65.426957	84	0.9335

Using Bartlett's test in SAS above, we conclude at 10% significance level that all populations have a common variance-covariance matrix.



As the X^2 plot above is nearly straight line and passes through the origin, we conclude that the populations are multivariate normal. We do note the presence of one outlier that falls to the right on the graph.

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Period Effect
H = Type III SSCP Matrix for Period
E = Error SSCP Matrix

S=4 M=-0.5 N=70

Statistic	Value	F Value	Num DF	Den DF
Wilks' Lambda	0.66358580	3.90	16	434.45
Pillai's Trace	0.35330557	3.51	16	580
Hotelling-Lawley Trace	0.48181908	4.25	16	278.06
Roy's Greatest Root	0.42509538	15.41	4	145

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Period Effect
H = Type III SSCP Matrix for Period
E = Error SSCP Matrix

S=4 M=-0.5 N=70

Statistic	Pr > F
Wilks' Lambda	<.0001
Pillai's Trace	<.0001

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Period Effect H = Type III SSCP Matrix for Period E = Error SSCP Matrix S=4 M=-0.5 N=70	
Statistic	Pr > F
Hotelling-Lawley Trace	<.0001
Roy's Greatest Root	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.	

The Wilks' lambda test statistic is 0.6636 with $p < 0.0001$. For large n, we reject H_0 at significance level α if $-\left(n-1-\frac{p+g}{2}\right) ln(\Lambda^*) > \chi^2_{p(g-1)}(\alpha)$. For p=4, g=4, and n=150 at 1% significance level we reject H_0 as $59.4641 > 26.22$ and conclude that at least one mean is different. We would then need to use multiple comparisons to determine which pairs of groups have different mean vectors and which components of these mean vectors differ. This would be done using the Bonferroni confidence intervals for all $p\binom{g}{2}$ possible pairwise comparisons.

More Analysis: MB and NH mean vector comparison from early and late predynastic periods

To compare the mean vector of MB and NH measurements between the early and late predynastic periods, we construct a 95% confidence region for the difference of the two mean vectors.

Test: H_o: $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ vs H_a: $\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$
This is equivalent to H_o: $\boldsymbol{\mu}_1-\boldsymbol{\mu}_2=\boldsymbol{\delta}_0$ vs H_a: $\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\neq\boldsymbol{\delta}_0$, where $\boldsymbol{\delta}_0=0$ in this case.

Assumptions (for n_1 and n_2 large):

1) The sample $\mathbf{X}_{11}, \mathbf{X}_{12},\ldots,\mathbf{X}_{1n\_1}$, is a random sample of size n_1 from a *p*-variate population with mean vector $\boldsymbol{\mu}_1$ and covariance matrix Σ_1.
2) The sample $\mathbf{X}_{21}, \mathbf{X}_{22},\ldots,\mathbf{X}_{2n\_2}$, is a random sample of size n_2 from a *p*-variate population with mean vector $\boldsymbol{\mu}_2$ and covariance matrix Σ_2.
3) Also, $\mathbf{X}_{11}, \mathbf{X}_{12},\ldots,\mathbf{X}_{1n\_1}$ are independent of $\mathbf{X}_{21}, \mathbf{X}_{22},\ldots,\mathbf{X}_{2n\_2}$.

The two samples are random, both with sample size 30 from a 4-variate population with corresponding mean vectors and covariance matrices. The samples are independent as a skull from the early predynastic period does not depend on skulls from the late predynastic period. We use Hotelling's T^2 test:

$$T^2 = [\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)]' \left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right) S_p\right]^{-1} [\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)]$$

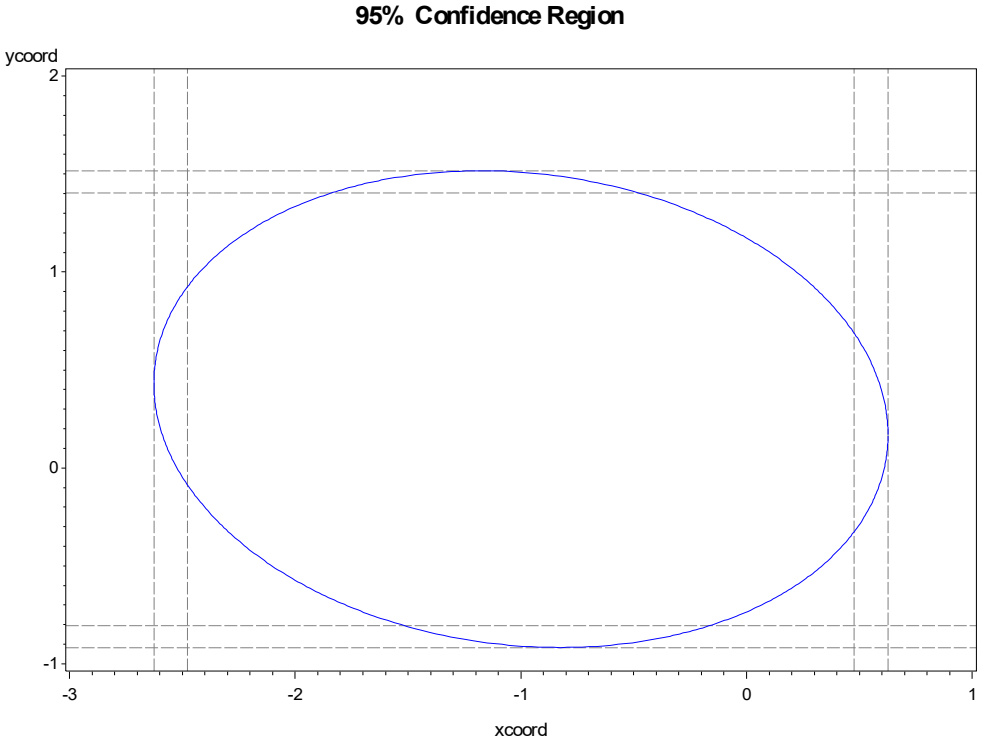
$$\sim \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1}$$

As the populations have equal variance-covariance matrices, the 100(1- α)% confidence region (ellipsoid) for $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is given by all $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ satisfying:

$$[\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p\right]^{-1} [\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)] \leq \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p,n_1+n_2-p-1}$$

We reject the null hypothesis if $T^2 > \frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} F_{p,n_1+n_2-p-1}(\alpha) = c^2$.
Using SAS, T^2=1.5155498 and c^2=2.53543. As $T^2 < c^2$ we fail to reject H_0 and conclude that the two mean vectors are equal.

95% Confidence Ellipsoid



As the ellipse does contain the zero vector, we can conclude that the two mean vectors are the same, supporting our conclusion reached in part D. The lengths and directions for the axes of the ellipsoid are given by the eigenvalue and eigenvector pairs of S_p. The longest axis is λ_1=25.88949 units along e'_1=[-0.9686,0.2486] and the shortest axis is λ_2=7.0203 units along e'_2=[-0.2486, -0.9686]. As $\boldsymbol{\mu}_1$-$\boldsymbol{\mu}_2$=**0** is included in the ellipse, we can also verify that the two mean vectors are equal.

The simultaneous confidence intervals are given by $\boldsymbol{a}'(\overline{\boldsymbol{x}_1} - \overline{\boldsymbol{x}_2} \pm c\sqrt{\boldsymbol{a}'\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{S}_p\boldsymbol{a}}$ and will cover $\boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for all $\boldsymbol{a}'$. The Bonferroni confidence intervals for the p population difference is given by

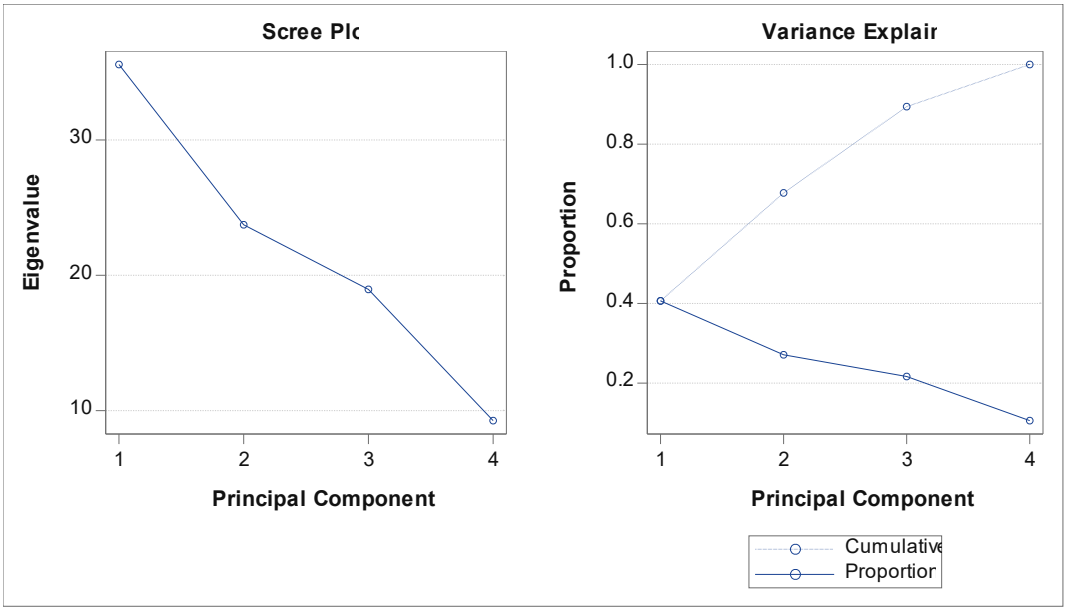
$$(\overline{x_{1i}} - \overline{x_{2i}}) \pm t_{n_1+n_2-2}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{i,i,p}}$$

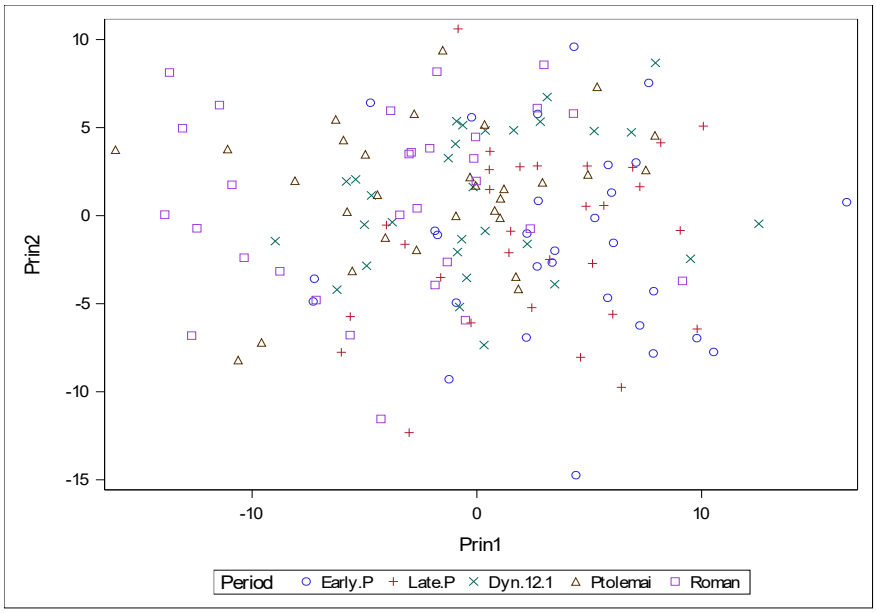
The 95% simultaneous T^2 intervals are shadows or projects of the confidence ellipse on the axes of the component means. The T^2 intervals are slightly wider (more conservative) than the Bonferroni intervals, however, the T^2 intervals can be used to create confidence intervals for other linear combinations of the components μ_i. Therefore, the Bonferroni intervals are contained within the simultaneous confidence intervals.

Principal Component Analysis

As all four measurements are in millimeters (are on the same scale) we do not need to standardize the data. We therefore use the variance-covariance matrix to conduct our principal component analysis.

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	35.5758984	11.8420591	0.4065	0.4065
2	23.7338393	4.7763104	0.2712	0.6776
3	18.9575289	9.6968761	0.2166	0.8942
4	9.2606528		0.1058	1.0000





From the first table, we see that the cumulative percentage of the total sample variance explained by the first three components is 89.42% The scree plot also suggests that three components are appropriate to effectively summarize the sample variability. In the plot of Prin2 vs Prin1, we are unable to clearly distinguish between the groups representing the five different time periods; no groups clearly stand out from the others in the plot. Three outliers also appear to be present. One point is from Early.P towards the bottom of the graph and another from Early.P to the far right. Also, one point from Ptolemai is to the far left.

Classification Rule

We use all four measurements to develop a classification rule for the five time periods assuming equal prior probability and equal costs of misclassification. We use Bartlett's test to determine if linear or quadratic discriminant is appropriate.

Chi-Square	DF	Pr > ChiSq
45.667228	40	0.2483

From Bartlett's test, as p > 0.05, we conclude that the variance-covariances are equal and so a linear discriminant analysis is appropriate. We also noted that the variance-covariance were equal in the MANOVA section.

We use the Estimated Minimum TPM Rule for Equal-Covariance Normal Populations as each population (time period) is normal and the variance-covariance matrices of the time periods are equal. We allocate **x** to π_k if the linear discriminant score $\widehat{d_k}(x)$=the largest of $\widehat{d_1}(x), \ldots, \widehat{d_5}(x)$ with $\widehat{d_i}(x) = \bar{\boldsymbol{x}}_i' \boldsymbol{S}_p^{-1} \boldsymbol{x} - \frac{1}{2} \bar{\boldsymbol{x}}_i' \boldsymbol{S}_p^{-1} \bar{\boldsymbol{x}}_i + ln\, p_i$ for i=1,2,3,4,5.

Number of Observations and Percent Classified into Period						
From Period	Dyn.12.1	Early.P	Late.P	Ptolemai	Roman	Total
Dyn.12.1	15 50.00	4 13.33	4 13.33	2 6.67	5 16.67	30 100.00
Early.P	4 13.33	12 40.00	8 26.67	4 13.33	2 6.67	30 100.00
Late.P	5 16.67	10 33.33	8 26.67	4 13.33	3 10.00	30 100.00
Ptolemai	7 23.33	3 10.00	3 10.00	5 16.67	12 40.00	30 100.00
Roman	4 13.33	2 6.67	4 13.33	9 30.00	11 36.67	30 100.00
Total	35 23.33	31 20.67	27 18.00	24 16.00	33 22.00	150 100.00
Priors	0.2	0.2	0.2	0.2	0.2	

Error Count Estimates for Period						
	Dyn.12.1	Early.P	Late.P	Ptolemai	Roman	Total
Rate	0.5000	0.6000	0.7333	0.8333	0.6333	0.6600
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

Apparent error rate is the fraction of the observations in the training sample that are misclassified by the sample classification function and is given by

$$APER = \frac{n_{1M} + n_{2M} + n_{3M} + n_{4M} + n_{5M}}{n_1 + n_2 + n_3 + n_4 + n_5} = 0.6600$$

This states that 66% of observations in the training sample were misclassified.

Number of Observations and Percent Classified into Period						
From Period	Dyn.12.1	Early.P	Late.P	Ptolemai	Roman	Total
Dyn.12.1	12 40.00	6 20.00	4 13.33	2 6.67	6 20.00	30 100.00
Early.P	5 16.67	9 30.00	10 33.33	4 13.33	2 6.67	30 100.00
Late.P	5 16.67	11 36.67	7 23.33	4 13.33	3 10.00	30 100.00
Ptolemai	7 23.33	3 10.00	3 10.00	5 16.67	12 40.00	30 100.00
Roman	4 13.33	2 6.67	4 13.33	10 33.33	10 33.33	30 100.00
Total	33 22.00	31 20.67	28 18.67	25 16.67	33 22.00	150 100.00
Priors	0.2	0.2	0.2	0.2	0.2	

Error Count Estimates for Period						
	Dyn.12.1	Early.P	Late.P	Ptolemai	Roman	Total
Rate	0.6000	0.7000	0.7667	0.8333	0.6667	0.7133
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

The estimated expected actual error rate is given by $\hat{E}(AER) = \frac{\sum_{i=1}^{g} n_i^{(H)} M}{\sum_{i=1}^{g} n_i}$. Therefore $\hat{\mathrm{E}}(\mathrm{AER}) = 0.7133$, an unbiased estimate of the expected actual error rate.

Posterior Probability of Membership in Period							
Obs	Classified into Period		Dyn.12.1	Early.P	Late.P	Ptolemai	Roman
1	Ptolemai		0.2516	0.0967	0.1005	0.3193	0.2319

If given the new observed vector of the four measurements of a male Egyptian skull being $(136,135,94,53)^{T}$, we assign the new observed vector to π_4 as the distance from the observation to the group mean $\overline{x_4}$ is smallest and conclude that this observation is from the Ptolemai period.

Conclusion

To analyze the skull data set, we first determined that there were no multivariate outliers. Using one-way MANOVA, we concluded that at least one mean vector of the four measurements from the five different time periods is different. To determine if one or all vectors differ, we would do multiple comparisons to determine which pairs of groups have different mean vectors and which components of these mean vectors differ. Using principal component analysis, we determined that three components are appropriate to effectively summarize the sample variability and, based on the first and second principal components, it would be difficult to distinguish between the five time periods. For classification analysis, we determined that linear discriminant analysis is appropriate using Estimated Minimum TPM Rule for Equal-Covariance Normal Populations. Further analysis may include interpretation of the principal components and exploration into the high apparent error rate and estimated expected actual error rate.