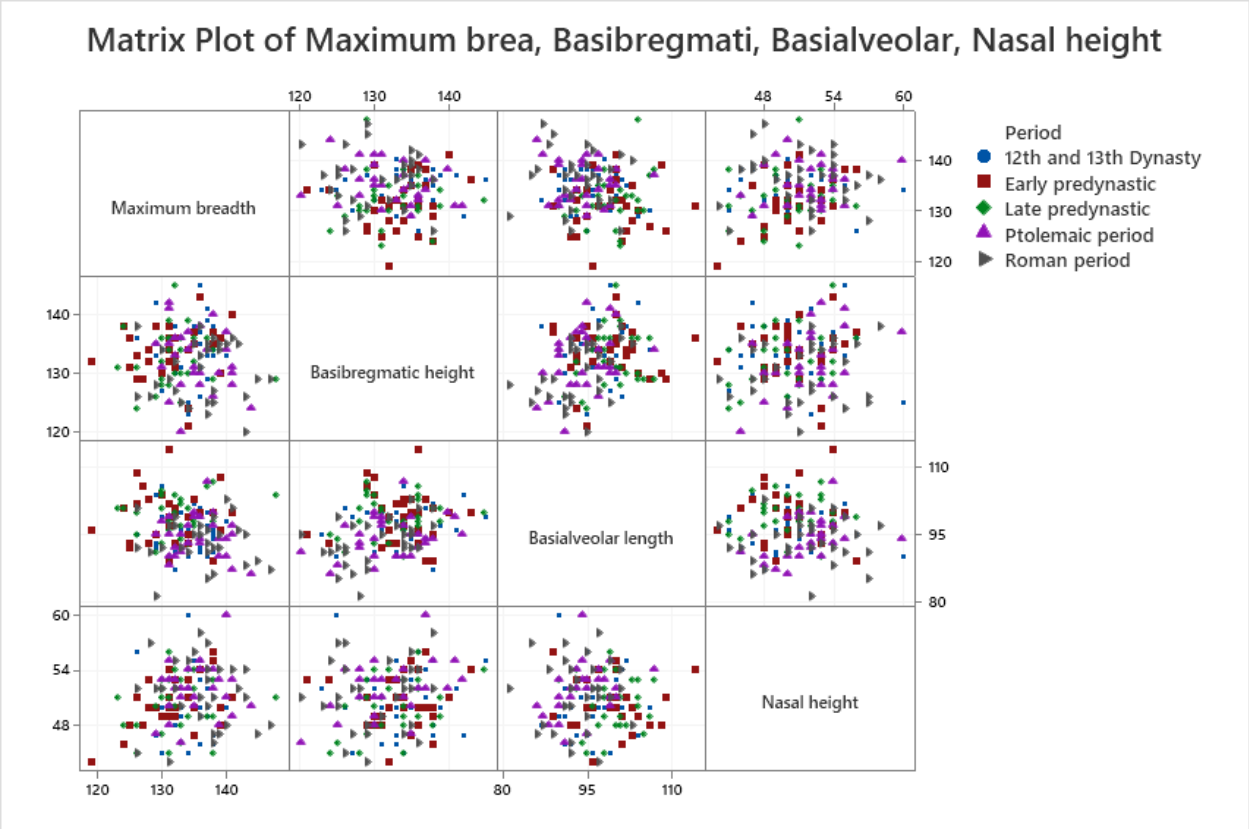
Additional Multivariate Exercise 1

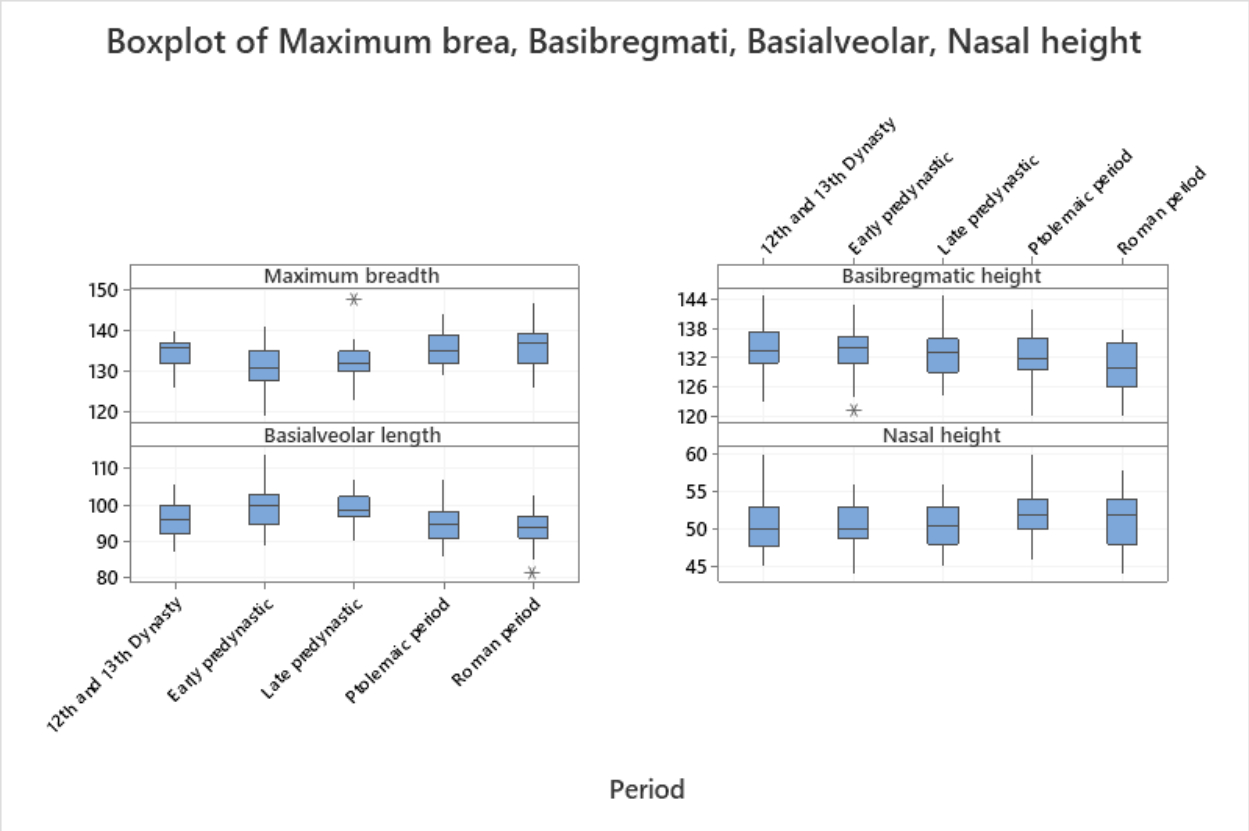
Data: The Egyptian skulls data were described in the Manly textbook. For each of the four variables, the data constitute a one-way ANOVA layout with 30 replications per group (period). When looking at one variable at a time, a natural notation is y_{ij} for the measurement (of a particular type) on the jth skull from period i; $i = 1, \ldots, 5$ and $j = 1, \ldots, 30$.

Initial graphical exploration: To depict both individual values and relationships between the 4 variables a scatterplot matrix plot seems the obvious choice. The periods may be indicated by different symbols, even at risk of cluttering up the plot too much. With a total of 150 points per scatterplot, it may be useful to reduce the size of the plotting symbols. Such a plot generated by Minitab is shown below.



The scatter plot matrix shows most of the pairwise plots as wide scatters without any strong patterns. No striking outliers are seen, and nor is it immediately obvious from the plot how the periods differ in their values.

The scatterplot does however not show the distribution for each of the variables well. Several options exist for this purpose, but because of the focus on comparison between the groups it may be attractive to display the distributions separately for the 5 periods. With only 30 observations per period, histograms are no longer that useful, and one may instead display the symbolic representation of the 5-number summary for each distribution in a boxplot. The boxplot below was generated by the menu in Minitab for "One Y" and "With Groups", when the multiple variables were displayed in separate panels of the same graph (under "Multiple Graphs").



The boxplots appear (perhaps surprisingly) regular, with only a few points indicated as suspected outliers and most boxes quite symmetrical about the median. The variability also appears to be rather similar across the periods. Generally speaking, normal distribution inference would seem quite acceptable, even if this will be assessed in more detail for each of the variables.

Univariate analyses: The most natural estimates are means and standard deviations per period for each of the four variables. For statistical inference, the first choice would be analysis by one-way ANOVA models: $y_{ij} = \mu_i + \varepsilon_{ij}$, with the overall F-test for equality between periods, pairwise comparisons as needed, and the usual checks of model assumptions. Because these methods are not related to multivariate statistics, they will be reported briefly for this solution. The table below gives P-values for the overall test for periods, the specific period means (and their SE) with grouping indicators corresponding to unadjusted (Fisher) pairwise comparisons. All residual plots looked nice, and the standard deviations varied far less than by a factor of 2 between the periods. Thus the assumptions behind the one-way ANOVAs were all met.

	ANOVA F	Means for period					SE for
Measurement	P-value	1	2	3	4	5	mean
Maximum breadth	< 0.0005	134.5^{ab}	131.4^{c}	132.4^{bc}	135.5^{a}	136.2^{a}	0.84
Basibregmatic height	0.049	133.8^{a}	133.6^{a}	132.7^{ab}	132.3^{ab}	130.3^{b}	0.89
Basialveolar length	< 0.0005	96.0^{bc}	99.2^{a}	99.1^{a}	94.5^{bc}	93.5^{c}	0.90
Nasal height	0.20	50.6	50.5	50.2	52.0	51.4	0.58

Two of the variables (Maximum breadth and Basialvenlar length) had strongly significant differences between periods, but neither of them showed a monotonic trend of means across periods. In fact, for both of them the periods 2 and 3 appeared to be separated from the other periods. Only Basibregmatic

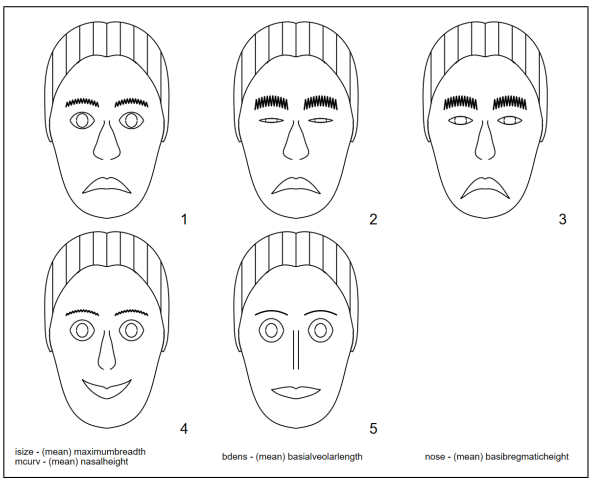
height showed a monotonic trend over time (decreasing). The means can be shown graphically against periods by Interval plots (Minitab) or Margins plots (Stata); these plots are not included here.

Relationships between variables: With all variables normally distributed, the Pearson correlation coefficient r is the natural statistic to quantify association between variables. The statistic can be computed for each period or overall. If an association is reasonably constant across periods, it can be expected to appear stronger across periods for variables with period differences. This is because the period differences will generate a larger spread of the points. For these data this however does not appear to be the case, as seen in the table of overall pairwise correlations below (with ranges for r across the five periods indicated in parentheses).

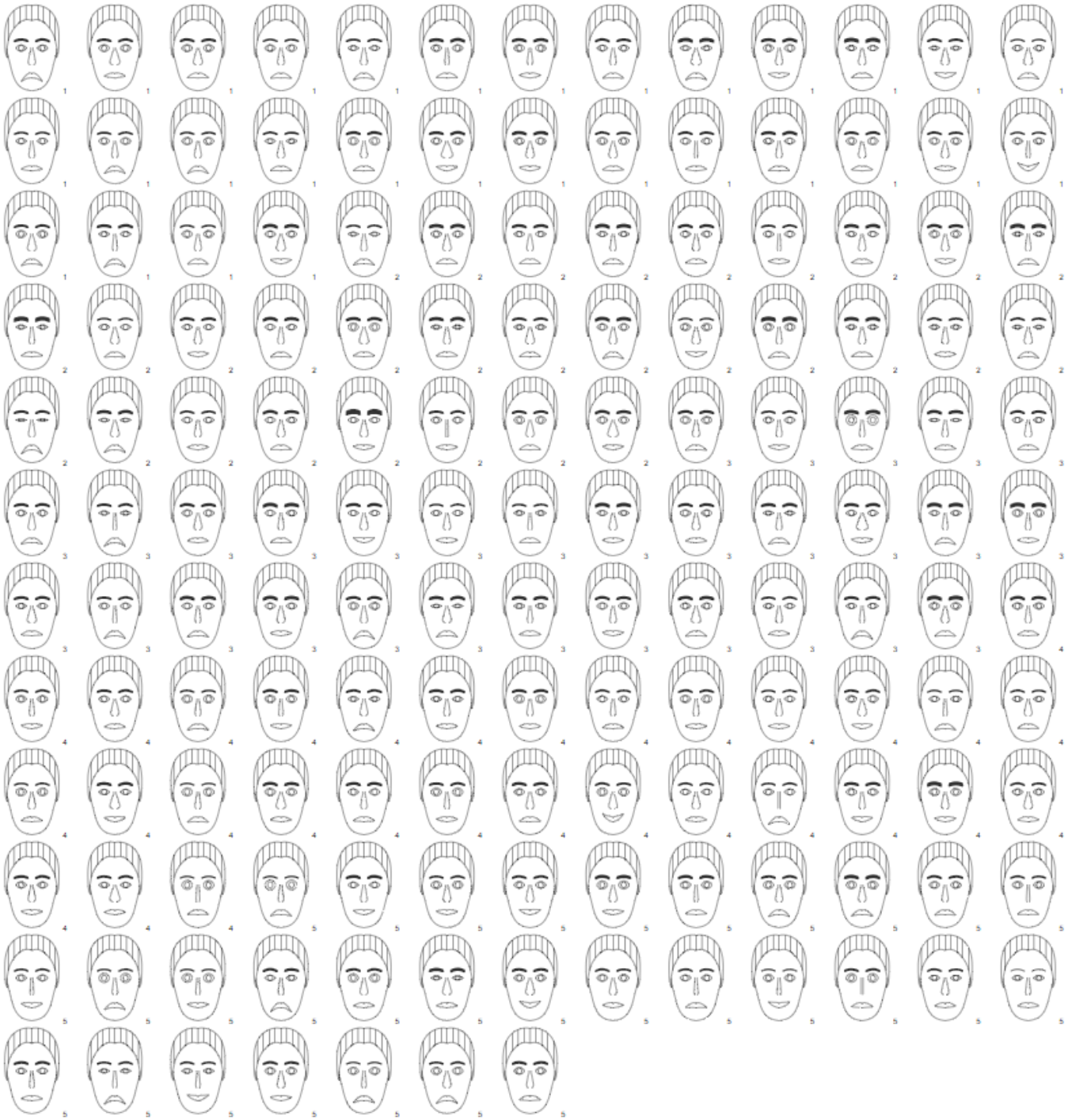
	Maximum breadth	Basibr_height	Basial_length	Nasal_h
Maximum breadth	1.00			
Basibregmatic height	-0.06 (-.28,.18)	1.00		
Basialveolar length	-0.16 (-.12,.25)	0.26 (-.03,.47)	1.00	
Nasal height	0.18 (-.10,.51)	0.15 (-.01,.42)	-0.01 (-.12,0.41)	1.00

Overall, the correlations are quite weak, even if some periods have correlations nearing 0.50.

Symbolic representation of periods and samples: Chernoff plots were generated in Stata with the following features for the four variables (in their above order): eye size, nose line, brow density and mouth curvature. The choice of features will naturally affect the appearance of the faces.



The differences between periods 2-3 and the rest are seen in the eye size and brow density features. The value for period 5 on Basibregmatic height is seen in the nose line.



isize - Maximum breadth
mcurv - Nasal height

bdens - Basialveolar length

nose - Basibregmatic height

It is perhaps a bit overwhelming to try to extract patterns among such a large number of faces. The first impression is probably that they look surprisingly similar. Specifically, the distinguishing features in the Chernoff faces based on the means are not very consistent within periods, reflecting the strong overlap between the distributions in the boxplots.