Statistical Tests

A. Test Size = .10  
Confidence Level (.90)

B. Test Size = .05  
Confidence Level (.95)
Basic components of GMM

**Procrustes**
This aligns shapes and minimizes differences between them to ensure that only real shape differences are measured.

**PCA (primary use)**
This creates a shape space in which shape similarities and differences are easily seen. The first principal component (PC) accounts for most of the variation in shape, the second PC for the second most variation, etc. Variation on one PC is statistically uncorrelated with variation on another (they are independent components of shape variation).

**PCA (secondary use)**
PCA is also useful because it gives:

1. **scores** (the coordinates of each object in the shape space) that can be used as shape variables in other statistical analysis

2. **eigenvalues** that are the variances of the objects on each PC. The eigenvalues sum to the total variance in the data set

3. **eigenvectors** that are the rotation vectors from the PC space back to the landmark shape. Eigenvectors * scores + consensus give you a shape model for a particular point in the PC shape space.
Statistical analysis: partitioning variance

In GMM, variance / covariance = variation in shape

Purpose of statistical analysis is to ascertain to what extent part of that variance is associated with a factor of interest, aka partitioning variance.

1. **P-value:** indicates whether the association is greater than expected by random chance

2. **Regression parameters (slopes, intercepts):** indicate the axis in shape space associated with the factor, useful for modeling the aspect of shape associated with the factor

3. **Correlation coefficient ($R$):** indicates the strength of the association between the variance and the factor

4. **Coefficients of determination ($R^2$):** indicates the proportion of the variance that is associated with the factor
Statistical analysis

**Regression analysis:** for use with factors that are continuous variables

**Analysis of Variance (ANOVA):** for use with factors that are categorical (MANOVA if the test is multivariate)

Regression and MANOVA are the most common methods because they test for the association between shape and some other variable(s).
• *R* has special syntax for statistical models, and a lot of built-in capabilities

• Models represent (hypothesized) relationships among variables, usually one response (\(y\)) and one or more predictor (\(x\)) variables

**Linear regression:** \( y = \beta_0 + \beta_1 x + \epsilon \)

residual error
Linear Models

- Response variable is a linear function of predictor variable(s)
- Very common, includes regression & ANOVA

Model Notation: \( y \sim x_1 + x_2 - 1 \)

- \( y \sim \) separator
- Model inclusion
- Model exclusion

Continuous \( x \)
( numeric )

Categorical \( x \)
( factors )

intercept
Regression analysis

Regression measures (and assesses) the relationship between geometric shape and another continuous predictor variable.

Continuous variables are ones that can take nearly any value (e.g., temperature, latitude, body mass, age, etc.).

Results of regression can be used to predict geometric shape for new values of the predictor variable.
Regression example

Predictor variable (x) (also known as the independent variable)

Response variable (y) (aka, the dependent variable)
[geometric shape as measured by scores]

PC 1 Scores

Size
Linear Regression Basics

Linear regression finds the regression line that predicts variable Y from variable X.

\[ Y = a \, X + b + E, \]

where \( a \) is the slope of the line, \( b \) is the intercept on the Y axis, and \( E \) is the residual error around the regression line.

\[\begin{align*}
a &= 2.0 \\
b &= 0.5
\end{align*}\]
The Correlation Coefficient (R) is a measure of E

\[ Y = a \cdot X + b + E \]

\( R \) measures the tightness of fit to the regression line, roughly 1-E when E is measured as the standard deviation of the points from the line in the X axis and the data have been standardized so that E is never greater than 1.

\[
\begin{align*}
    a &= 2.0 \\
    b &= 0.5 \\
    R &= 0.92
\end{align*}
\]

\( R \) ranges from 1.0 (perfect correlation) to 0.0 (no correlation).
The Coefficient of Determination \((R^2)\) is related to \(E\)

\[ Y = a \, X + b + E \]

\(R^2\) also measures the tightness of fit to the regression line, but as the variance rather than the standard deviation of the points around the line. \(R^2\) can be interpreted as the proportion of the variance in \(Y\) that is explained by \(X\).

\[ a = 2.0 \]
\[ b = 0.5 \]
\[ R^2 = 0.85 \]

\(R^2\) also ranges from 1.0 (100% explained) to 0.0 (0% explained).
Correlation and Coefficient of Determination measure scatter of data around a linear relationship.

- **Slope = 2.0**
  - $R = 1.0$
  - $R^2 = 1.0$

- **Slope = 2.0**
  - $R = 0.95$
  - $R^2 = 0.90$

- **Slope = 2.0**
  - $R = 0.77$
  - $R^2 = 0.60$
Regression analysis thus reveals the following

\[ Y = a \, X + b + E \]

- The slope of the relationship between \( x \) and \( y \) (coefficient \( a \), often symbolized as \( \beta \))
- The statistical significance and standard error of the slope
- The intercept of the regression line (coefficient \( b \), often symbolized as \( c \))
- The statistical significance and standard error of the intercept
- The amount of variance in \( Y \) explained by \( X \) (Sum of Squares of the Model)
- The amount of variance in \( Y \) not explained by \( X \) (SS Error, also known as SS Residual)
- The percentage of variance explained by the regression line (\( R^2 \) can be interpreted as the percentage of the variance in \( Y \) that is explained by \( X \))
Note that regression equations can be varied to suit new problems

\[ Y = a \, X + b + E \]
\[ Y = a \, X^2 + b + E \]
\[ Y = a_1 \, X_1 + a_2 \, X_2 + b + E \]
\[ Y_2 + Y_2 = a \, X + b + E \]
\[ Y = a \, X \times Z + b + E \]
Regression example for GMM

\[ y = 2 \cdot x + 0.5 \]

PC 1 score = 2 body mass + 0.5
The function `lm()`

- Evaluates linear models for best fitting coefficients
- Returns a list with lots of information
- Example: linear regression

```r
plot(mg.temp, valve.length)
w <- lm(valve.length ~ mg.temp, data=cope)

w  # gives only coefficients & formula
abline(w)  # adds regression line to plot
summary(w) # gives much more info: stats, P, etc.
```
> summary(w)      ## summary of linear regression

Call:                  
  lm(formula = valve.length ~ mg.temp)

Residuals:             
  Min       1Q   Median       3Q      Max
-115.369  -44.431    8.367   45.486  104.234

Coefficients:          
                         Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)             764.639     10.125  75.522  < 2e-16 ***
mg.temp                -19.530      2.466  -7.918 2.77e-12 ***
---                      
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.77 on 104 degrees of freedom 
Multiple R-Squared: 0.3761, Adjusted R-squared: 0.3701
F-statistic: 62.7 on 1 and 104 DF,  p-value: 2.774e-12

Some information extracting functions
          resid(w) coef(w) fitted(w) confint()
Analysis of Variance (ANOVA)

Also, Multivariate Analysis of Variance (MANOVA)

ANOVA assesses the relationship between geometric shape and a categorical predictor variable.

Categorical variables are ones that define groups and are not ordered (e.g., male/female, herbivore/carnivore, island/continent)

ANOVA tests for differences in means between the groups. The mean of each group can easily be modelled and illustrated as a deformation of one to the other.
ANOVA uses F-test for overall difference among groups, where F is based on proportion of variance within groups to that between groups.

Pairwise comparisons give $p$ values for differences between specific pairs of groups if there are more than two groups.

Overall results relate to differences between any group.

$F = 10.29$

$p < 0.01$
ANOVA gives the following useful results

The means of the groups

1. The among-group variance (SS Model)
2. The within-group variance (SS Residual)
3. The total variance (SS Total)
4. The statistical significance of the difference between the two
5. The percentage of variance explained by the group difference ($R^2$).
Make fake groupings

\[ gg \leftarrow \text{rep}(c(\text{"a"}, \text{"b"}), \text{length.out}=\text{nrow(cope)}) \]
\[ \text{ggf} \leftarrow \text{as.factor(gg)} \quad \# \text{convert from character to factor} \]
\[ \text{w.a} \leftarrow \text{lm(\text{valve.length} \sim \text{ggf}, \text{data=cope})} \]

A more meaningful example

\[ \text{w.sp} \leftarrow \text{lm(\text{valve.length} \sim \text{species}, \text{data=cope})} \]
\[ \text{summary(w.sp)} \quad \# \text{coefficients and p-values} \]
\[ \text{anova(w.sp)} \quad \# \text{standard ANOVA table (SSQ, MSQ, F-test)} \]
Advanced Topics

1. More complicated models
   • Variables can be nested
   • Mix of continuous & categorical predictors
   • Interaction terms
   • There are more rules for model notation

2. Model Selection
   • F-tests (nested models)
   • Akaike Information Criterion: $\text{AIC()}$
### Generalizing the Linear Model

- Linear models assume: residuals are *independent*, *normal*, and with *equal variance*

\[ y = \beta_0 + \beta_1 x + \epsilon \]

<table>
<thead>
<tr>
<th>Assumption to be relaxed</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal variance</td>
<td>Weighted least squares <code>lm(weights=)</code></td>
</tr>
<tr>
<td>Normally distributed</td>
<td>Generalized linear models <code>glm()</code></td>
</tr>
<tr>
<td>Independent</td>
<td>Generalized least squares <code>gls()</code></td>
</tr>
</tbody>
</table>
On statistical tests

*A difference to be a difference must make a difference.* - Gertrude Stein

Always consider two aspects of a statistical test:

1. Does the P-value show the association to be significantly stronger than random?

2. Does the $R^2$ value show that a substantial part of the variance is associated with the factor?
Example Questions:

How do we determine whether mandible shape is related to skull length?

How do we determine how much of mandible shape is related to skull length?

How do we determine whether mandible shape is related to sex?

How do we determine how much of mandible shape is related to sex?

How do we determine if mandible shape is related to depositional environment?