Large Samples

• Suppose that $X_1, X_2, \ldots, X_n$ is random sample from some multivariate population with finite variances.

• Laws of Large Numbers; as $n$ becomes large:
  
  – $\bar{X}$ converges in probability to $\mu$;
  
  – $S$ converges in probability to $\Sigma$;
• Central Limit Theorem; as \( n \) becomes large:
  
  \[- \sqrt{n} \left( \bar{X} - \mu \right) \text{ is approximately } N_p(0, \Sigma), \]

  \[- \text{whence } n \left( \bar{X} - \mu \right)' \Sigma^{-1} (\bar{X} - \mu) \text{ is approximately } \chi^2_p; \]

• and with the Law of Large Numbers for \( S \), this implies that:

  \[- \left( \bar{X} - \mu \right)' \left( \frac{1}{n} S \right)^{-1} (\bar{X} - \mu) = n \left( \bar{X} - \mu \right)' S^{-1} (\bar{X} - \mu) \text{ is also approximately } \chi^2_p. \]
Assessing Normality

- Marginal univariate distributions:
  
  - quantile-quantile plots;
  
  - goodness-of-fit test based on correlation coefficient in qq-plot;
  
  - Shapiro-Wilk test is more conventional, very similar.
• Marginal bivariate distributions: scatter plots.

• Marginal distributions with $2 \leq p' < p$, and full joint distribution ($p' = p$): $\chi^2$ plots:

  – squared generalized distances

  \[ d_j^2 = (x_j - \bar{x})' S^{-1} (x_j - \bar{x}), \quad j = 1, 2, \ldots, n \]

  are approximately $\chi^2_{p'}$;

  – make a quantile-quantile plot against theoretical quantiles for $\chi^2_{p'}$.

• Note: if the data are multivariate $t_\nu$ instead of multivariate normal, the squared generalized distances are $p' F_{p', \nu}$ instead of $\chi^2_{p'}$; make a modified q-q plot to detect this.
Outliers

- Outlier: an unusual observation that does not follow the pattern of the bulk of the data.

- May be erroneous, but may be important; in either case, they may distort the statistical properties of the bulk of the data.

- Detect using the same scatter plots and qq-plots.

- Note: a multivariate outlier may not look extreme in 1- or 2-dimensional marginal plots—need $\chi^2$ plots.
Transformation

• Many types of data have better statistical properties when expressed on a scale different from the one on which they are measured.

• For instance:
  
  – more symmetric distribution (closer to normal);
  
  – less variation in dispersion (closer to homoscedastic);
  
  – more additive effects of experimental factors (smaller interactions).
Most common transformations:

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts, $y$</td>
<td>$\sqrt{y}$</td>
</tr>
<tr>
<td>Proportions, $\hat{p}$</td>
<td>$\text{logit}(\hat{p}) = \frac{1}{2} \log \left( \frac{\hat{p}}{1-\hat{p}} \right)$</td>
</tr>
<tr>
<td>Correlations, $r$</td>
<td>$\text{Fisher's } z(r) = \frac{1}{2} \log \left( \frac{1+r}{1-r} \right)$</td>
</tr>
</tbody>
</table>

Square roots are a special case of Box-Cox transformation:

$$x(\lambda) = \begin{cases}  \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln x & \lambda = 0. \end{cases}$$

The most common transformation, to logarithms, is another special case.
• Box-Cox power $\lambda$ can be estimated by likelihood methods, if we assume that the transformed data are normally distributed:

– for each variable separately, with graph;

– for pairs of variables, with contour plot;

– for $2 < p' \leq p$ variables, by optimization.