Large Sample Inference

- The $T^2$ and Bonferroni procedures are valid for all sample sizes $n$ and dimensions $p$, provided the data are sampled from a multivariate normal population.

- If $n$ is large enough and $p$ not too large, the $t$ and $F$ distributions used for testing hypotheses and constructing confidence intervals are approximately normal and $\chi^2$, respectively.

- The resulting large sample approximate methods are also valid for data sampled from non-normal populations, because of the Central Limit Theorem.
Missing Observations: Two Strategies

• Exclusion strategy:
  
  – exclude items with missing data, i.e. analyze only completers;
  
  – used in proc glm, in multivariate mode;
  
  – no modification of full-data methods such as ML.
• Inclusion strategy:
  
  – use all observed data, including partial observations;
  
  – used in \texttt{proc mixed};
  
  – likelihood function no longer easy to maximize—can use full-data likelihood in \textit{EM} algorithm, or brute force.
Missing Observations: Two Mechanisms

• Missing Completely At Random (MCAR):
  – missingness is independent of data;
  – both strategies give valid inferences;
  – analyzing only completers uses less information than inclusion strategy.
• Missing At Random (MAR):
  – given the *observed* data, missingness does not depend on the *unobserved* data;
  – exclusion strategy (using only completers) may give biased estimators;
  – inclusion strategy still valid.

• Inclusion strategy is more efficient and more widely valid, but more sensitive to departures from multivariate normality.
Correlated Observations

- We have assumed that the data $X_1, X_2, \ldots, X_n$ are a random sample from $N_p(\mu, \Sigma)$.

- Sometimes we would like to analyze other multivariate data, e.g. from a multiple time series.

- Correlation of $X_j$ with $X_k$ changes sampling distributions.
• Univariate problem: estimating \( \mu = E(X_t) \) from correlated observations.

• We can still use \( \bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n \), though it may be inefficient (usually only slightly).

• Assume second-order stationary:
  \[
  \text{corr}(X_t, X_{t+h}) = \rho(h)
  \]
does not depend on \( t \).

• Then
  \[
  \text{Var}(\bar{X}_n) \approx \frac{\sigma^2_X}{n} \left[ 1 + 2 \times \sum_{h=1}^{\infty} \rho(h) \right].
  \]
• In particular, if $\rho(h) = \phi|h|$, 

$$\text{Var}(\bar{X}_n) \approx \frac{\sigma^2_X}{n} \times \frac{1 + \phi}{1 - \phi}. $$

• Crude adjustment ("effective number of observations"): 

$$n' = n \times \frac{1 - \phi}{1 + \phi},$$

where $\phi = \rho(1) = \text{corr}(X_t, X_{t+1})$. Then 

$$\text{Var}(\bar{X}_n) \approx \frac{\sigma^2_X}{n'}. $$

• A similar approximation is sometimes used for correlated multivariate data.