XI. Multivariate Analysis:

A. Problem:

1. \( X_{ij} \): Multiple \( J \) responses for each subject \( i \).
2. Explain whole distribution of \( X_{ij} \) in terms of covariates.
   a. covariates often indicate group membership

B. Simple case: one sample.

1. Model: \( X_i \sim F_X(x) \)
2. Normal case, known variance \( \Sigma \):
   a. Let \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \)
3. Null hypothesis: marginal medians take on prespecified values
   a. WLOG, value is zero.
   b. That is, \( F_X(\infty, \ldots, \infty, 0, \infty, \ldots, \infty) = \frac{1}{2} \) for each possible slot for zero.
   c. Less sloppily, let
      i. \( z_j \) be the vector with \( j \) components,
      ii. with all components \( \infty \) except for that in component \( j \)
      iii. \( H_0 : F_X(z_j) = \frac{1}{2} \)
4. Normal case:

a. Let $\bar{X} = \sum_{i=1}^{n} X_i / n$

b. $\Sigma$ known: $\bar{X}^\top \Sigma^{-1} \bar{X} \sim \chi^2_J$, if $\Sigma$ nonsingular.

c. $\Sigma$ unknown: estimate using usual sum of squares:

\[ \bar{X}^\top \Sigma^{-1} \bar{X} \sim F_{J,?}, \text{ if } \Sigma \text{ nonsingular}. \]

d. $\Sigma$ unknown, but sum of squares estimator not appropriate:

\[ \bar{X}^\top \Sigma^{-1} \bar{X} \sim \chi^2_J, \text{ approximately.} \]

e. Techniques require multivariate normality, which is stronger than marginal normality. [Mark A R]


a. Sign test, or signed rank test assuming symmetry (often in the context of paired data).

i. Solution depends on approximate normality of $T$

ii. We’ve claimed before that the components of $T$ are separately approximately normal, by reference to a CLT

iii. Similar arguments work for the vector as a whole.

b. Difficulties:

i. Separate tests are generally dependent, and dependence
structure depends on distribution of raw observations.

- We will have to estimate this.

ii. Null hypothesis dependent on the coordinate system for variables, but analysis does not.

- Ex. If \((X_i, Y_i) \sim \mathcal{N}(\mu, \Sigma)\) with \(\Sigma\) known, and \(H_0 : \mu = 0\), then the canonical test is \((X_i, Y_i)\Sigma^{-1}(X_i, Y_i)^\top\), and it is unchanged if we base test on \((U_i, V_i)\) for \(U_i = X_i + Y_i\) and \(V_i = X_i - Y_i\).

c. Solution for sign test:

i. Let \(T_j = \sum_i \hat{s}(X_{ij})\) for \(\hat{s}(u) = \begin{cases} 1 & \text{if } u > 0 \\ -1 & \text{if } u < 0 \end{cases}\).

ii. Then under \(H_0\), \(T_j / \sqrt{n} \approx \mathcal{N}(0, 1)\).

iii. Estimate \(\text{Cov}[\hat{s}(X_{ij}), \hat{s}(X_{ij'}))] = E[\hat{s}(X_{ij})\hat{s}(X_{ij'}))]\) by \(\hat{\sigma}_{jj'} = \sum_i \hat{s}(X_{ij})s(X_{ij'})/n\).

iv. So test using \(T^\top \begin{pmatrix} 1 & \hat{\sigma}_{12} & \cdots \\ \hat{\sigma}_{21} & 1 & \cdots \\ \vdots & \ddots & \ddots \end{pmatrix}^{-1} T/n \sim \chi^2_J\) under \(H_0\).

d. Solution for Wilcoxon signed rank test is similar. [Mark B R]

[Mark B sas]

6. Permutation Solution:

a. Select an existing test statistic
C. Confidence Regions for parameter vector $\mu$

1. Introduce shift parameter to move data to conventional null hypothesis; for ex.,

   a. One-sample: $X \mapsto X - 1_n \otimes \mu$

      i. $1_n$ is vector of ones of length $n$

      ii. $\otimes$ is outer product, so $1 \otimes \mu_n$ is the matrix with entry $\mu_j$ in column $j$ for all rows.

   b. Two-sample: $(X, Y) \mapsto (X, Y - 1_n \otimes \mu)$

2. Calculate test statistic $T(\mu)$ using shifted data

3. Determine critical value for test statistic, $t_{1-\alpha}$

4. Report as confidence region $\{\mu | T(\mu) \leq t_{1-\alpha}\}$.

5. A case for which the test inversion approach is problematic.

   a. Setup: $\bar{X}, \bar{Y}$ independent normals, mean $\mu$ and $\nu$ and
variance $\sigma^2/n$

b. CI for $\rho = \mu/\nu$.

c. Invert $\sqrt{n}(\bar{X} - \rho\bar{Y})/(\sigma\sqrt{1 + \rho^2})$

d. CI is $\{\rho : n(\bar{X} - \rho\bar{Y})^2/\sigma^2(1 + \rho^2) \leq z^2\}$.

i. Simplest case $\sigma = 1, n = 1$.

ii. $\rho \in (\bar{X}/\bar{Y} \mp z/\bar{Y}\sqrt{(\bar{X}/\bar{Y})^2 + 1 - (z/\bar{Y})^2})/(1 - (z/\bar{Y})^2)$, only if $\bar{Y}^2 > z^2$

iii. If $\bar{X}^2 + \bar{Y}^2 < z^2$ no solution exists, and CI is entire line, since inequality never holds with equality.

iv. If $\bar{X}^2 + \bar{Y}^2 \geq z^2$ but $\bar{Y}^2 \leq z^2$, then CI is a union of two rays.

e. Similar approach holds in cases when $\bar{X}$ and $\bar{Y}$ are bivariate normal with nonzero correlation, and when $\sigma$ is unknown.

D. Next Easiest case: covariate represents membership in one of two groups

1. In this case, represent group one as $X_{ij}$ and group 2 as $Y_{ij}$.

   a. $H_0$: mean vectors are the same,

   b. Sample sizes $m$ and $n$. 
2. Traditional normal-theory approach:

i. Let $C_{X,u,v}$ be the sample covariance for the $X$'s between responses $u$ and $v$: 
$$
\sum_{i=1}^{m} (X_{iu} - \bar{X}_u)(X_{iv} - \bar{X}_v)/(m - 1).
$$

ii. Let $C_{Y,u,v}$ be the sample covariance for the $Y$'s between responses $u$ and $v$: 
$$
\sum_{i=1}^{n} (Y_{iu} - \bar{Y}_u)(Y_{iv} - \bar{Y}_v)/(n - 1).
$$

iii. Let $C_{u,v}$ be the pooled sample covariance for all observations: 
$$
C_{u,v} = ((m - 1)C_{X,u,v} + (n - 1)C_{Y,u,v})/(m + n - 2).
$$

a. Hotelling's $T^2 = \frac{mn}{m+n}(\bar{X} - \bar{Y})^\top C^{-1}(\bar{X} - \bar{Y})$ measures difference between sample mean vectors,

i. In a way that accounts for sample variance,

ii. and combines the response variables.

b. 
$$
\frac{m+n-J-1}{(m+n-2)J}T^2 \sim F_{J,m+n-J-1} \text{ if }
$$

i. $(X_{i1}, \ldots, X_{iJ})$ and $(Y_{i1}, \ldots, Y_{iJ})$ multivariate normal

ii. variance matrices are the same. \[\text{Mark E R}\] [Mark E sas]

3. permutation test

a. Under $H_0$, vectors are all independent and identically distributed.
b. Can calculate p-value by counting the permutations between groups (keeping vector together) that gives as large or larger $T^2$.

c. Other test statistics combining component-wise results:

i. Max $t$-statistic:
   - Do univariate $t$-statistics for each response.
   - Report maximum.

ii. Max absolute value of $t$-statistic:
   - Like above, but take $|\cdot|$ before optimizing.

iii. Max of Wilcoxon statistics or absolute value of Wilcoxon statistics

iv. Rank version: sub ranks for data values, and proceed as before.
   - Makes statistic less sensitive to extreme values.
   - Doesn’t appear to make it fit distributional assumptions.
   - Can also use rank scores. [Mark F R] [Mark F sas]: 6.2

4. Normal-theory Rank based approach

   a. Let $W_j$ be Mann-Whitney-Wilcoxon statistic using manifest variable $j$, for $j \in \{1, \ldots, J\}$. 
Let $\mathbf{W} = (W_1, \ldots, W_J)$, $\mathbf{\Psi} = \text{Var}[\mathbf{W}] = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1J} \\ \vdots & \ddots & \vdots \\ \sigma_{J1} & \cdots & \sigma_{JJ} \end{pmatrix}$.

- $\sigma_{jj}$ are all known (to equal $mn(n + m + 1)/12$, but that’s not important here).

i. Remaining entries of $\mathbf{\Psi}$ must be estimated.

- For $i = 1, \ldots, m + n$, let $F_{ij}$ be the number of observations in group 2 that beat observation $i$ on variable $j$ if $i$ is in group 1, and the number of observations in group 1 that $i$ beats on variable $j$, if $i$ is in group 2.

- $4/(n + m)$ times Covariance matrix for $F$ estimates the variance matrix of $\mathbf{W}$.

ii. Remember that F approximation requires multivariate normality for response variables.

iii. Rank scores don’t fix nonnormality in joint distribution structure.

c. Normal-theory Component-wise maxima require

i. Known variance case: Multivariate normal CDF with arbitrary variance-covariance matrix: doable.
ii. Unknown variance case: Multivariate normal CDF with arbitrary variance-covariance matrix: much harder.