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 the 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. $\frac{m+n-J-1}{(m+n-2)J} T^2 \sim F_{J,m+n-J-1}$ if
i. \((X_{i1}, \ldots, X_{iJ})\) and \((Y_{i1}, \ldots, Y_{iJ})\) multivariate normal

ii. variance matrices are the same,

: 6.1

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.

: 6.2

4. Normal-theory Rank based approach

a. Let \( W_j \) be Mann-Whitney-Wilcoxon statistic using manifest variable \( j \), for \( j \in \{1, \ldots, J\} \).

b. Let \( W = (W_1, \ldots, W_J) \), \( \Psi = \text{Var}[W] = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1J} \\ \vdots & \ddots & \vdots \\ \sigma_{J1} & \cdots & \sigma_{JJ} \end{pmatrix} \)

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

i. Remaining entries of \( \Psi \) 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 \( W \)

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.

XII. Density Estimates

A. Setup $X_1, \ldots, X_n$ iid from density $f(x)$.

B. Most elementary approach: histogram

1. Represent distribution as a bar chart
2. Bars butt up against neighbors
3. Heights set so that area equals sample proportion in that region
4. So area over a region may be used to approximate proportion in the region
5. Generally choose number of bars so that divisions are “round numbers”
6. Method of Scott (1992): bar width $\approx 3.5Sn^{-1/3}$

7. Advantages
   a. Easy
   b. If region of interest begins and ends exactly at division between bar, area in region is exactly proportion in region.

8. Disadvantages
   a. Results may depend on choice of where first bar starts, as well as on the bar width.
   b. Choppy shape generally does not reflect our expectations about true density.

C. More sophisticated approach: Kernel density estimate

1. Pick a density $w$ that you want to use for to build estimate
   a. That is, a non-negative function integrating to 1
   b. Called the kernel
   c. Choices:
      i. (Standard) Normal density
      ii. Quadratic $w(x) = \frac{3}{4}(1 - x^2)$ (R calls it Epanechnikov, with a different scale)
      iii. Triangle $w(x) = 1 - |x|$.
iv. Generally, any other density, generally symmetric.

2. Build a location-scale family \( w(x; \mu, \Delta) = w((x - \mu)/\Delta)/\Delta \)

3. Report the average of kernels centered at data points:
\[
\hat{f}(x) = (\Delta n)^{-1} \sum_{i=1}^{n} w((X_i - x)/\Delta)
\]

4. \( \Delta \) is called band width.
   a. \( \Delta \) should depend on spread of data, and \( n \).
      i. SD, IQR, range, etc.
      ii. If bandwidth is too high, density estimate will be too smooth, and hide features of data.
      iii. If bandwidth is too low, density estimate will provide too much clutter to make understanding the distribution possible.

5. \( \Delta \) chosen to minimize MSE
\[
\text{MSE}[\hat{f}(x)] = \text{Var}[\hat{f}(x)] + (E[\hat{f}(x)] - f(x))^2
\]

6. Box kernel:
\[
\text{Var}[\hat{f}(x)] = p(1 - p)/(n\Delta^2) \quad \text{for} \quad p = F(x + \Delta/2) - F(x - \Delta/2), \quad \text{and bias is}
\]
\[
p/\Delta - f(x) = f(x^*) - f(x) \quad \text{for some} \ x^* \in [x - \Delta/2, x + \Delta/2]
\]
   a. If \( \Delta \not\to 0 \) then bias \( \not\to 0 \).
   b. If \( \Delta = O(1/n) \) then variance \( \not\to 0 \).

7. More generally,
For explanation of power in $n$, see Silverman (1986) section 3.3.

b. Variance of the estimator is $\approx (n\Delta)^{-1} f(x) \int_{-\infty}^{\infty} w(t)^2 \, dt$.

c. Bias of estimator is $1/2\Delta^2 f''(x) \int_{-\infty}^{\infty} t^2 w(t) \, dt$.

d. Balancing variance and squared bias means $(n\Delta)^{-1} \propto \Delta^4$, or $\Delta \propto n^{-1/5}$.

8. Higgins suggests $1.06S n^{-1/5}$, or replace $S$ by IQR/1.34.

a. For explanation of constant, see Sheather and Jones (1991).

XIII. Regression Function Estimates

A. Setup: Model $Y_j$ as a function of $X_j$: $Y_j = g(X_j) + \epsilon_j$.

1. Most restrictive: $Y_j = \beta_0 + \beta_1 X_j$

2. Least restrictive: $g(x) =$ mean of $Y_j$ with $X_j = x$

   a. Mostly this will be a single $Y_j$

3. Intermediate: $g(x)$ continuous and differentiable, with curves that turn quickly discouraged

B. Kernel smoothing:

1. Get an expression that is explicit rather than implicit:

   \[ \hat{g}(x) = \frac{\sum_{j=1}^{n} Y_j w((x - X_j)/\Delta)}{\sum_{j=1}^{n} w((x - X_j)/\Delta)}. \]

2. Weight function can be

   a. the same as above
Lecture 10

b. Often a normal density.

c. Often uniform density centered at 0.