4. Under some more restrictive conditions, can also use this to estimate variance.

a. Assume that $T_n$ acts sort of like mean

b. If $T_n$ really were a mean, then $\tilde{T}_{n,i} = nT_n - (n - 1)T_{n-1,i}$ would be the observations that make it up.

c. Then $\text{Var}[\tilde{T}_{n,i}] = n\text{Var}[T_n]$, and $\tilde{T}_{n,i}$ iid.

d. Estimate $\text{Var}[T_n]$ by $\frac{n-1}{n} \sum_{i=1}^{n} (\tilde{T}_{n,i} - \frac{1}{n} \sum_{j=1}^{n} \tilde{T}_{n,j})^2$.

5. Example where jackknife fails

a. Expansion of expectation fails for rank-based estimators

b. Ex., median for continuous data

c. $n$ even:

i. $T_n = \left( X_{(n/2)} + X_{(n/2+1)} \right) / 2$.

ii. $T_{n-1,i}^* = \begin{cases} X_{(n/2+1)} & \text{if } i \leq n/2 \\ X_{(n/2)} & \text{if } i \geq n/2 + 1 \end{cases}$, with removed observation from ordered observations.

iii. $\bar{T}_{n-1}^* = (X_{(n/2)} + X_{(n/2+1)}) / 2 = T_n$.

iv. Bias estimate always 0.

d. $n$ odd:

i. $T_n = X_{((n+1)/2)}$. 
ii. \( T_{n-1,i}^* = \begin{cases} \frac{(X((n-1)/2) + X((n+1)/2))}{2} & \text{if } i > (n + 1)/2 \\ \frac{(X((n+3)/2) + X((n+1)/2))}{2} & \text{if } i < (n + 1)/2, \\ \frac{(X((n+3)/2) + X((n-1)/2))}{2} & \text{if } i = (n + 1)/2 \end{cases} \)

with removed observation from ordered observations.

iii. Mean of results from the smaller sample is \( \bar{T}_{n-1}^* \)

\[
= \frac{n+1}{4n} X((n-1)/2) + \frac{n+1}{4n} X((n+3)/2) + \frac{n-1}{2n} X((n+1)/2)
\]

iv. \( \bar{T}_{n-1}^* - T_n \)

\[
= \left( \frac{n+1}{4n} X((n-1)/2) + \frac{n+1}{4n} X((n+3)/2) + \frac{n-1}{2n} X((n+1)/2) \right) - X((n+1)/2)
\]

\[
= (n + 1) \left( \frac{1}{4}X((n-1)/2) + \frac{1}{4}X((n+3)/2) - \frac{1}{2}X((n+1)/2) \right) / n
\]

v. Bias estimate is \( B = (n - 1)\bar{T}_{n-1}^* - T_n = (n + 1) \left( \frac{1}{4}X((n-1)/2) + \frac{1}{4}X((n+3)/2) - \frac{1}{2}X((n+1)/2) \right) / n \)

6. Works better for smooth functions of data like \( \alpha \) trimmed mean

a. that is, mean of observation after smallest proportion \( \alpha \) removed, and largest proportion \( \alpha \) removed.

: 10.1-10.2

XIII. Density and Regression Function Estimates

A. Estimate density

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

2. Most elementary approach: histogram
Lecture 11

a. Represent distribution as a bar chart

b. Bars butt up against neighbors

c. Heights set so that area equals sample proportion in that region

d. So area over a region may be used to approximate proportion in the region

e. Generally choose number of bars so that divisions are “round numbers”

f. Method of Scott (1992): bar width \( \approx 3.5S n^{−1/3} \)

g. Advantages

i. Easy

ii. If region of interest begins and ends exactly at division between bar, area in region is exactly proportion in region.

h. Disadvantages

i. Results may depend on choice of where first bar starts, as well as on the bar width.

ii. Choppy shape generally does not reflect our expectations about true density.

3. More sophisticated approach: Kernel density estimate

a. Pick a density \( w \) that you want to use for to build estimate
i. That is, a non-negative function integrating to 1

ii. Called the kernel

iii. Choices:
   • (Standard) Normal density
   • Quadratic \( w(x) = \frac{3}{4}(1 - x^2) \) (R calls it Epanechnikov, with a different scale)
   • Triangle \( w(x) = 1 - |x| \).
   • Generally, any other density, generally symmetric.

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

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

C. \( \Delta \) is called band width.

i. \( \Delta \) should depend on spread of data, and \( n \).
   • SD, IQR, range, etc.
   • If bandwidth is too high, density estimate will be too smooth, and hide features of data.
   • If bandwidth is too low, density estimate will provide too much clutter to make understanding the distribution possible.
e. \( \Delta \) chosen to minimize MSE
\[
\text{MSE} = \text{Var} \left[ \hat{f}(x) \right] + (E[\hat{f}(x)] - f(x))^2
\]
f. Box kernel:
\[
\text{Var} \left[ \hat{f}(x) \right] = p(1 - p)/(n\Delta^2) \text{ for } p = F(x + \Delta/2) - F(x - \Delta/2), \text{ and bias is } p/\Delta - f(x) = f(x^*) - f(x) \text{ for some } x^* \in [x - \Delta/2, x + \Delta/2]
\]
i. If \( \Delta \not\to 0 \) then bias \( \not\to 0 \).

ii. If \( \Delta = O(1/n) \) then variance \( \not\to 0 \).

g. More generally,

i. For explanation of power in \( n \), see Silverman (1986) section 3.3.

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

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

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

h. Higgins suggests \( 1.06Sn^{-1/5} \), or replace \( S \) by IQR/1.34.

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

B. Estimate regression function

1. Setup: Model \( Y_j \) as a function of \( X_j \):
\[
Y_j = g(X_j) + \epsilon_j.
\]
2. Most restrictive: \( Y_j = \beta_0 + \beta_1 X_j \)

3. Least restrictive: \( g(x) = \text{mean of } Y_j \text{ with } X_j = x \)
   a. Mostly this will be a single \( Y_j \)

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

5. Kernel smoothing:
   a. Get an expression that is explicit rather than implicit:
      \[
      \hat{g}(x) = \frac{\sum_{j=1}^{n} Y_j w((x - X_j)/\Delta)}{\sum_{j=1}^{n} w((x - X_j)/\Delta)}.
      \]
   b. Weight function can be
      i. the same as above
      ii. Often a normal density.
      iii. Often uniform density centered at 0. [Mark B R] [Mark B sas]

6. LOESS
   a. \( f(x) \) fitted value at \( x \) for low-degree (viz., linear or quadratic) regression of points with \( X_j \) near \( x \).
   b. Specify the number of points \( k \)
   c. Upweight points near \( x \) and downweight them away from \( x \)
   d. Weighting function scaled to make point in neighborhood farthest from \( x \) have weight going down to zero.
i. This keeps the curve smooth as $x$ moves.

e. Common weight function is $w(x) = (1 - |x|^3)^3$.

f. So $f(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ for

i. $\hat{\beta} = \text{argmin} \left( \sum_{j \in N(x)} (Y_j - \beta_0 - \beta_1 X_j - \beta_2 X_j^2)^2 w((x - X_j)/\Delta)) \right)$ for

ii. $N(x) =$ indices of $k$ closest points to $x$, and

iii. $\Delta = \max \{ |X_j - x| \mid j \in N(x) \}$.

g. Procedure formerly Lowess, Locally Weighted Sum of Squares.

h. Result can not be expressed as a simple formula. \[\text{[Mark C R]}\]

\[\text{[Mark C sas]}\]

7. Related technique: Isotonic regression

a. Fit nonparametric relationship between variables, assuming that it is non-decreasing.

b. Choose $\hat{Y}_j$ to minimize $\sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2$ subject to $\hat{Y}_j \geq \hat{Y}_i$ whenever $X_j \geq X_i$. \[\text{[Mark D R]}\]

8. Spline:

a. A way to draw a smooth curve between two points $x_0$ and $x_N$:

i. Pick $N - 1$ intermediate points $x_1 < x_2 < \cdot \cdot \cdot < x_{N-2} < x_{N-1}$ (called knots).
ii. Define a polynomial of degree $M$ between $x_{j-1}$ and $x_j$

iii. Constrain so that the derivatives of order up to $M - 1$ match up at knots.

b. Use to fit pairs of points $(X_1, Y_1), \ldots, (X_n, Y_n)$.

i. Taken to an extreme, if all $X_j$ are unique, then we can fit all $n$ points with a polynomial of degree $n - 1$.

ii. Denote by $\hat{\mu}(x)$

iii. Choose to minimize $\sum_{j=1}^{n} (Y_j - \hat{\mu}(X_j))^2 + \lambda \int_{X_1}^{X_n} \hat{\mu}''(x) \, dx$