Section 7.
Nonparametric methods of uncertainty estimation
S-method, jackknife, bootstrap, cross-validation

Techniques broadly applicable, complex statistics
Generally justified by asymptotics
There exist singular and inappropriate cases
NECESSARY

S-method AKA method of linearization, propagation
of error, Taylor series method

Gauss (1815)

Basically one approximates functions by Taylor
expansions (usually linear) of basic random
variables.

Rao, Section 6a. 2

Sequence of k-dimensional statistics

\[ T_n = (T_{1n}, ..., T_{kn}) \quad n = 1, 2, ... \]

eg. \( (X, Y) \)
Derived statistic

\[ g(\bar{T}_n) = g(T_1, \ldots, T_n) \]

eg. \[ g(\bar{X}, \bar{Y}) = \frac{\bar{Y}}{\bar{X}} \]

Suppose \( \sqrt{n}(\bar{T}_n - \theta) \xrightarrow{d} N_{k}(0, \Sigma_{\theta}) \) / \( \rightarrow \)

Write

\[ g(\bar{T}_n) = g(\theta) + \frac{\partial g(\theta)}{\partial \theta} (\bar{T}_n - \theta) + \ldots \]

The entity of principal interest is typically \( g(\theta) \)

**Theorem.** If \( g() \) has a continuous first derivative

\[ \sqrt{n} \{ g(\bar{T}_n) - g(\theta) \} \xrightarrow{d} N_{k}(0, \frac{\partial g}{\partial \theta} \Sigma \frac{\partial g}{\partial \theta} ) \]
Convergence in distribution (law)

Sequence of r.v.'s \( \{X_n\} \)

\[ F_n(x) = \text{Prob}(X_n \leq x) \]

\[ X_n \overset{d}{\rightarrow} X \quad (X_n \overset{p}{\rightarrow} X) \quad \text{(weakly)} \]

\[ Y \to F_n(x) \to F(x) \text{ at all continuity points of } F \]

Equivalent to

\[ \int g dF_n \to \int g dF \quad \text{for all bounded continuous } g \]

Doesn't always hold if \( g \) is unbounded.

E.g., \( X_n = \mu + \sigma Z_i + C \) \( C \): Cauchy

\[ X_n \overset{d}{\rightarrow} X \sim N(\mu, \sigma^2) \]

\[ EX_n = \infty, \quad EX = \mu \]

\[ \text{Var } h(X_n) \to \text{Var } h(X) \quad \text{generally} \]
Result suggests estimating the covariance matrix of $g(T^*)$ by

$$
\left( \frac{\partial g}{\partial \theta} \right)_{\theta = T^*} \cdot \frac{1}{2} \left( \begin{array}{c} \frac{\partial^2}{\partial \theta^2} \\ \frac{\partial^2}{\partial \theta \partial \theta^*} \end{array} \right)
$$

but...

Procedure "works" provided $T^*$ is a neighborhood of $\theta$

Gives an approximating distribution. Could use for confidence intervals.
Example where not too useful (Cramer’s rule)

\begin{align*}
\Gamma &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \\
&= \frac{t_1 - t_3 + t_2}{\sqrt{(t_2 - t_3)(t_2 - t_1)}} = g(t_1, t_2) \\
+1 &= x/\bar{m} + b = x^2 + b = y^2 \\
+2 &= y/\bar{m} + a = xy
\end{align*}

Answer: An estimate of \( \rho \) is

\begin{align*}
\frac{r^2}{n} &\left[ \frac{\hat{\rho}_{00}}{\hat{\mu}_{00}^2} + \frac{\hat{\rho}_{04} + 2 \hat{\rho}_{22} + 4 \hat{\rho}_{22}}{\hat{\mu}_{10}^2 \hat{\mu}_{02}^2} + \frac{4 \hat{\rho}_{22} - 4 \hat{\rho}_{31}}{\hat{\mu}_{11} \hat{\mu}_{20}} \\
&- \frac{4 \hat{\rho}_{13}}{\hat{\mu}_{11}^2 \hat{\mu}_{02}} \right] \\
\end{align*}

But in normal case \( (1 - r^2)^2 \),

\[ \hat{\mu}_{gh} = \frac{1}{n} \sum (x_i - \bar{x})^g (y_i - \bar{y})^h \]
Remarks

1. Is $n$ large enough?

2. Is $g$ sufficiently linear? (differentiable?)

3. Is the algebra correct?

4. Is it approximately normal? (CI's etc.)

5. Note that this is not giving expected values.

6. The derivatives appearing may be approximated by finite differences.

\[
\frac{2g}{\theta_j} = \frac{g(\theta_1, \ldots, \theta_{j+1}, \theta_j + \delta, \theta_{j+2}, \ldots, \theta_p)}{g(\theta_1, \ldots, \theta_j - \delta, \ldots, \theta_p)} - \frac{28}{28}
\]

7. While often forgotten “learning” the bias of an estimate can be important.
Example where useful.

Ratio estimator as in survey sampling:
\[ \hat{R} = \frac{\bar{y}}{\bar{x}} \quad \text{(sampling without replacement)} \]
\[ \hat{R} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} = \left( \frac{\bar{y} - R\bar{x}}{\bar{x}} \right) \left( \bar{x} + (\bar{x} - \bar{x}) \right) \]

Writing cap letters for population values:
\[ \hat{R} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} \left[ 1 - \frac{\bar{x} - \bar{x}}{\bar{x}} + \left( \frac{\bar{x} - \bar{x}}{\bar{x}} \right)^2 + \ldots \right] \]
\[ \approx \frac{\bar{y} - R\bar{x}}{\bar{x}} \quad \text{linear approx.} \]
\[ \approx \frac{\bar{y} - R\bar{x}}{\bar{x}} - \frac{\bar{y} - R\bar{x}}{\bar{x}} \left( \frac{\bar{x} - \bar{x}}{\bar{x}} \right) \quad \text{quadratic} \]

\[ \text{var} \hat{R} \approx \text{var} \left( \frac{\bar{y} - R\bar{x}}{\bar{x}} \right) = \frac{1 - \frac{f}{\bar{x}}}{\frac{\bar{x}^2}{N}} \sum_{i=1}^{N} \left( y_i - Rx_i \right)^2 / (N-1) \]
\[ f = \frac{N}{\bar{x}} \]

bias \[ E(\hat{R} - R) \approx E \left( \frac{\bar{y} - R\bar{x}}{\bar{x}} \right) - E \left( \frac{\bar{y} - R\bar{x}}{\bar{x}} \right)^2 \]
\[ \approx - \frac{1 - \frac{f}{\bar{x}}}{\frac{\bar{x}^2}{N}} \left( \rho S_y S_{\bar{x}} - R S_{\bar{x}}^2 \right) \]
Can use more terms for a (possibly) better approximation:

\[ g(T_x) = g(\theta) + g'(\theta) (T_x - \theta) + \frac{1}{2} g''(\theta) (T_x - \theta)^2 + \ldots \]

and \( g(T_x) \) ?

\[ E T_x = \theta + \frac{b \theta}{n} + \ldots \] \text{ biased}

\[ \text{var} T_x = \frac{\sigma^2}{n} + \ldots \]

\[ E (T_x - \theta)^2 = \left( \frac{b \theta}{n} \right)^2 + \frac{\sigma^2}{n} + \ldots \]

\[ \text{var} g(T_x) = g(\theta) + g'(\theta) \frac{b \theta}{n} + \frac{1}{2} g''(\theta) \frac{\sigma^2}{n} + \ldots \]
Variance stabilizing transformation

Correlation coefficient

$\sqrt{n} (r - \rho) \xrightarrow{d} N(0, (1-\rho')^2)$

Look for $g()$ such that variance of the large sample distribution is approximately constant

With $[g'(\rho)]^2 (1-\rho)^2 = c$

Take $g'(\rho) = \frac{c}{1-\rho^2}$

$g(\rho) = \int \frac{c}{1-\rho^2} \, d\rho$

$= \frac{c}{2} \int \left( \frac{1}{1+\rho} + \frac{1}{1-\rho} \right) \, d\rho$

$= \frac{c}{2} \left( \log(1+\rho) - \log(1-\rho) \right)$

$= c \tanh^{-1} \rho$

Closing: makes variable more gaussian relationships more additive
An oddity. \( r^2 \) when \( \rho = 0 \)

\[
\text{Var}(r^2) \sim \frac{[2\rho]^2}{m^2} \left(1 - \rho^2\right)^2
\]

= 0 when \( \rho = 0 \)

Need more terms in expansion

\[
\text{Var}(r^2) \sim \frac{2}{m^2} \quad \text{when} \quad \rho = 0
\]

Singular. \[ \frac{\partial q}{\partial \theta} \bigg|_{\theta = \theta_0} \] does occur in practical situations.

E.g. estimating MI

Distributions become \( \chi^2 \), not normal.
There are functional forms of these results, e.g., using Frechet or Gateaux derivatives.

Consider \( \frac{y}{x} \).

Suppose c.d.f. of \((x,y)\) is \(F(x,y)\) and empirical c.d.f. is

\[
F_n(x,y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq x, y_i \leq y}
\]

Then

\[
\hat{\theta} = \frac{\sum_{i=1}^{n} y_i dF(x_i,y_i)}{\sum_{i=1}^{n} x_i dF(x_i,y_i)}
\]

and

\[
\hat{\theta} = \frac{\sum_{i=1}^{n} y_i dF_n(x_i,y_i)}{\sum_{i=1}^{n} x_i dF_n(x_i,y_i)}
\]

One is considering

\[
\theta = t(F) \quad \text{and} \quad \hat{\theta} = t(F_n) \quad (\ast)
\]

Might use a density estimate \(f_n\) instead of \(F_n\). Using (\ast) defines \(\hat{\theta}\), consistently, for all \(n\).
Replicated sub-samples

Incorporating sub-samples (Mahalanobis)

One computes $\tilde{Y}_i = t(\tilde{T}_i)$ for the i-th sub-sample.

$L = \bigcup_i L_i \quad L_i \cap L_{i'} = \emptyset \quad i \neq i'$

**Purposes**

1. To estimate sampling variances when sample design is complicated and exact estimators are unavailable or cumbersome.

2. To control field work.

3. To measure components of non-sampling variances (e.g., enumerators) particularly useful for the study of correlated errors.

$\bar{Y}$ overall sample mean

$\bar{Y}_i$ mean of i-th sub-sample

Estimate var $\bar{Y}$ by

$$\frac{1}{I} \sum_{i=1}^{I} (\bar{Y}_i - \bar{Y})^2 / (I-1)$$

Faster
<table>
<thead>
<tr>
<th>Term</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between sub samples</td>
<td>$\sum \sum (\bar{Y}_i - \bar{Y})^2$</td>
<td>I-1</td>
<td>$s^2_b$</td>
</tr>
<tr>
<td>Within sub samples</td>
<td>$\sum \sum (Y_{ij} - \bar{Y}_i)^2$</td>
<td>I(J-1)</td>
<td>$s^2_w$</td>
</tr>
<tr>
<td>Total</td>
<td>$\sum \sum (Y_{ij} - \bar{Y})^2$</td>
<td>IJ-1</td>
<td></td>
</tr>
</tbody>
</table>

Perhaps $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$; $\tau_i$ random

$\tau_i$ make values in $i$-th sub sample correlated

$corr\{Y_{ij}, Y_{ij}\} = \frac{s^2_\tau}{s^2_\tau + s^2_\epsilon}$

$E\delta_w^2 = s^2_\epsilon$

$E\delta_b^2 = s^2_\epsilon + J s^2_\tau$
Advantages of the jackknife.

"Like the Boy Scout's knife, it can be used to do many jobs..."

Just need a program to evaluate the estimates of interest.
The jackknife. \( n = 11 \), 2 grouped

\[ \hat{\theta} \] based on all the data

\[ \hat{\theta}_{pi} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{-i} \]

\[ \text{and } \left( I \hat{\theta} - (n-1) \hat{\theta}_{-i} \right) = \theta + \frac{c}{(n-1)} \]

\[ \theta - \text{var} \theta \] has reduced bias, in an asymptotic sense

\[ \hat{\theta} - \theta = -(n-1) \left[ \hat{\theta} - \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{-i} \right] \]

estimate, the bias

\[ s^2 = \frac{\sum_{i=1}^{n} (\hat{\theta}_{pi} - \bar{\theta})^2}{(n-1)} \]

Estimates of \( \text{var} \theta \), var \( \theta \) biased by

\[ \frac{\lambda \theta^2}{\lambda^2} \]

J.W. Tukey, M. Quenouille
Example: \[ \hat{\theta} = \frac{X_1 + \ldots + X_n}{n} = \bar{X} \]
\[ = \frac{\sum_{i=1}^{n} X_i}{n} \]
\[ = \frac{\sum_{i=1}^{n} \bar{X}_i}{n} \]
\[ = \frac{\bar{X}_1 + \ldots + \bar{X}_n}{n} \]

\[ \hat{\theta}_{i.i} = \frac{\sum_{i=1, i \neq i}^{n} \bar{X}_i}{(n-1) \bar{X}_i} - \bar{X}_i \]
\[ = \frac{\sum_{i=1}^{n} \bar{X}_i - \bar{X}_i}{(n-1) \bar{X}_i} \]

\[ \hat{\theta}_{p.i} = \bar{X}_i \]

\[ \bar{\theta} = \hat{\theta} = \bar{X} \text{ here} \]
\[ \hat{\theta} - \bar{\theta} = 0 \]
\[ \hat{\delta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)} \]
\[ \text{var } \bar{X} \text{ estimated by } \hat{\delta}^2/(n-1) \]
Other justifications are asymptotic, e.g.
for $\hat{\theta} = g(A, B, \ldots)$
(function of means)

Which asymptotics?
I fixed $J \to \infty$ easy
$I \to \infty$, $J$ fixed harder
   e.g. $J = 1$

The estimate is inconsistent for the sample median when $J = 1$. (Not a regular enough functional)

Might compute for: histogram, qq plot, ...

There are also weighted jackknives, e.g. for regression

Tukey suggested $I = 10$