Urvalsmetoder och Estimation 10

Sampling and Estimation 10

2007-12-14

Skrivning

- Ändrad till 15 januari kl 9-14 i B 419

Extra tid

- För redovisning av inlämningsuppgift artikelläsning
- 11 januari 2008 kl 10-12 i B 419

Gauss- or Taylor-approximation

- X is a random variable
- m is a central point. Usually, as here, the mean
- Y = g(X) approximeras med
- g(m) + (Y-m) g'(m) eller
- g(m) + (Y-m) g'(m) + ½(Y-m)^2 g"(m)
- ger E(Y) ~ g(m) + ½ Var(Y) g"(m) ~ g(m)
- Var(Y) ~ (g'(m))^2 Var(Y)
- If Var(Y) ~ C/n, these are good approximations for large n

Several variables

- Y = g(X_1, X_2, …, X_p)
- E(Y) ~ g(m_1, m_2, …, m_p)

$$Var(Y) \approx \sum_i \frac{\partial g^2(m_1, m_2, ..., m_p)}{\partial m_i^2} Var(X_i) +$$

$$\sum \sum_{i \neq j} \frac{\partial g(m_1, m_2, ..., m_p)}{\partial m_i^2} Var(X_i)$$

- Example: Ratio-estimation

Example logodds ratio

- 15 years ago 10 000 (out of 85 000) 65-year-olds were sampled and questionned on their drinking behaviour. We now combine this data with the death register and get the table
	Dead	Alive	Total
Drinkers (> 25 cl/week)	155	587	742
Non-drinkers	1318	7940	9258
- logodds is ln(155*7940/1318*587) =
  $\ln(X_{11}) + \ln(X_{22}) - \ln(X_{12}) - \ln(X_{21}) = 0.46$
- Find its variance!
- The logodds can be written
  $\ln(p_{11}) + \ln(p_{22}) - \ln(p_{12}) - \ln(p_{21})$
- It is simple to find the partial derivatives
  $1/p_{11}$, $1/p_{22}$, $-1/p_{12}$ and $-1/p_{21}$

Example logodds ratio

- After some calculations it is easy to see that the variance can be approximated by
- $\Sigma_{ij} \mathrm{Var}(\hat{p}_{ij})/\hat{p}_{ij}^2 + \Sigma_{ij,kl} \mathrm{Cov}(\hat{p}_{ij},\hat{p}_{kl})/(\hat{p}_{ij}\hat{p}_{kl}) = \ldots$
  $\Sigma_{ij} 1/x_{ij} =$
  $1/155 + 1/587 + 1/1318 + 1/7940 \sim$
  $0.00904 \sim (0.095)^2$ (utan korrektion för ändlig population)
- Ett 95% intervall för logoddsratio blir alltså ungefär
- 0.46 +/- 0.19 = (0,27, 0.65)
- Detta test av oberoende är asymptotiskt ekvivalent med chi2-testet

Resampling

- Heard about Jackknife, Bootstrap … ?
- Jackknife.
  - Idea: We have an estimate $g(X_1, X_2, \ldots, X_p)$. Estimate its precision by seeing what happens when one observation is removed at a time e.g. $g(X_2, X_3, \ldots, X_p)$
  - i.e. base the variance estimate on $ng(X_1, X_2, \ldots, X_p)-(n-1)g(X_1, X_{i-i}, X_{i+1} \ldots, X_p)$
- (Check, what happens for X-bar!)
- Good method if g is a nice function (twice continuously differentiable with bounded second derivative and the sample is SRS).

Bootstrap

- Each observation can be thought of as representing $1/\pi_i$ elements
- A reasonable model for the whole population can thus be N elements where y_i is repeated $1/\pi_i$ times.

- This population is known and we can draw independent samples from it repeatedly with the same design as originally.
- We can compute the empirical variance from these resamples (and also bias and full distribution and also make confidence intervals).

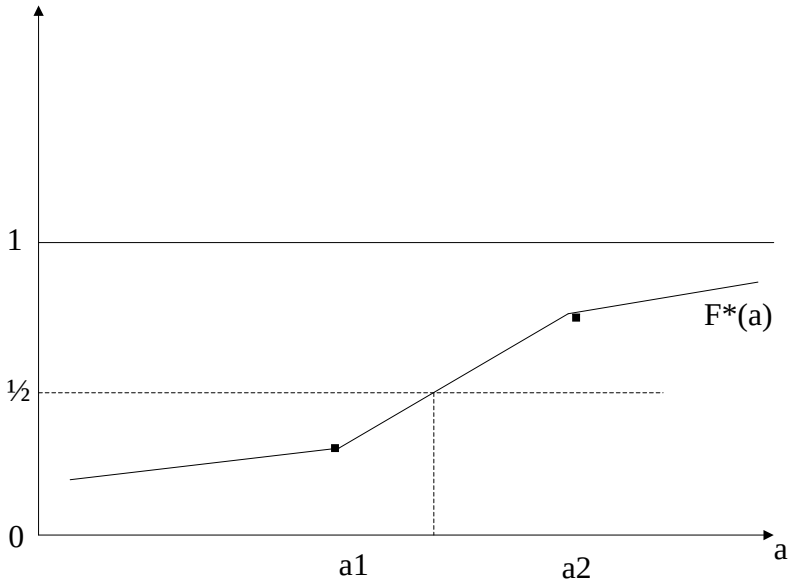
- The bootstrap can be used more often than the jackkife but is not so good when the conditions for the jackknife hold. Be careful with small strata or when second order inclusion probabilities play an important role. (E.g. Does not work with systematic sampling).

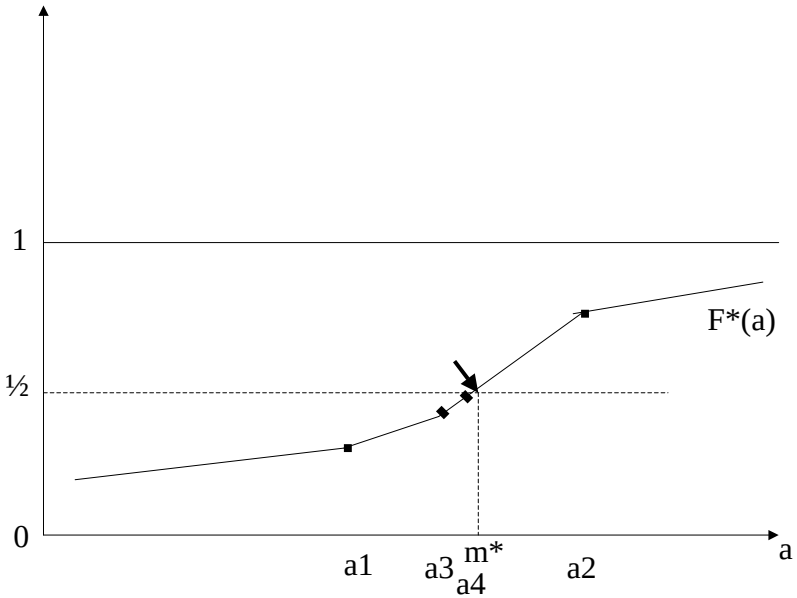
Balanced half-sampling
Balanced Repeated Replications

- It is well-known that $Var(X_1+X_2) = Var(X_1-X_2)$ for independent variables. We will use this!
- Divide the sample in two random parts so that each stratum is divided equally. Estimate half the total from both parts t_1 and t_2. Then $(t_1 - t_2)^2$ is an estimate with 1 d.f of $Var(t_1 - t_2)$ and also of $Var(t_1 + t_2) = Var(t)$.
- Do this repeatedly with more random halves getting more d.f.
- Works well for many methods. But not for cluster sampling since the between cluster variance is not estimated. (One may modify the procedure to cover this)

Derived quantities
Quantile estimation

- We illustrate by median
- The proportion F(a) of units less than a specified value, a, say, can be estimated by looking at the indicator $Y_{ai} = I(Y_i \leq a)$ and using ordinary formulas getting F*(a)
- Find m* such that F*(m*) =1/2 by trial and error and interpolation



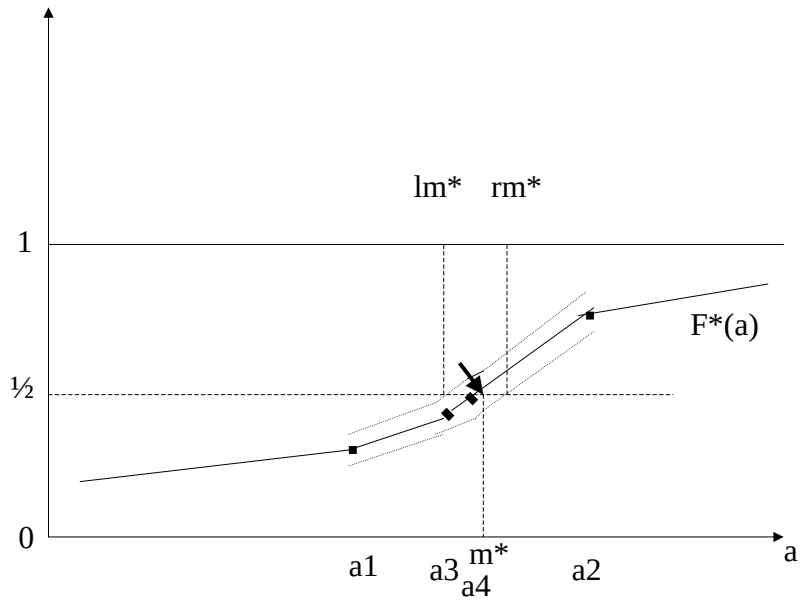


Intervals

- Find confidence intervals for F(a)

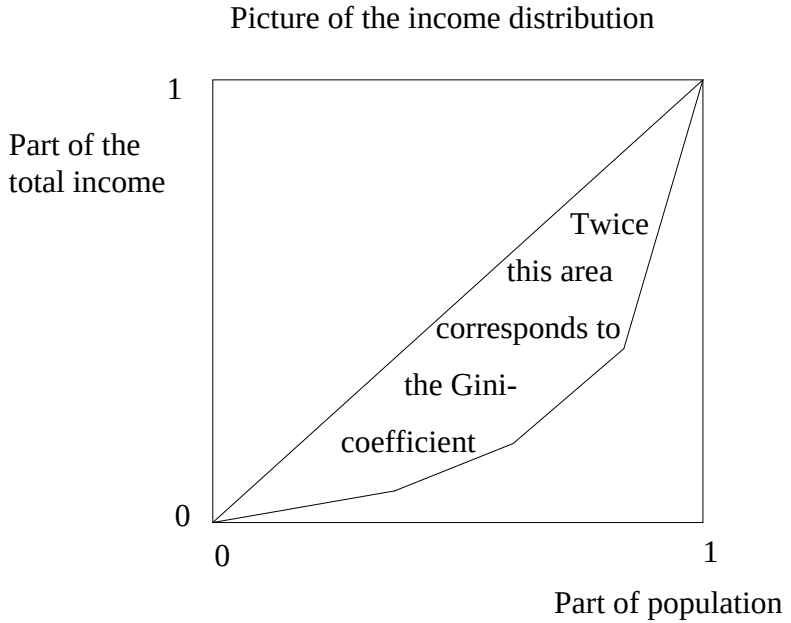
For some values of a including m*.

- Connect them and find their intersections with the line ½
- This is a 95% confidence interval for the median
- Similarly for any quantile



Derived quantities
The Gini coefficient

- The Gini coefficient is the best known measure of inequality in welfare distributions (e.g. In incomes, fortunes …)
- Order the persons after increasing incomes.
- Plot the cumulative income (percent of total income) against the percentage of people



Gini coefficient

- It is possible to show that the area under the curve is

$$\sum_{all\ \ pairs}\ \sum_{(i,j)} \min(\ y_i, y_j\ ) / n \sum_i y_i =$$

$$\sum_{all\ \ pairs}\ \sum_{(i,j)} \min(\ y_i, y_j\ ) / \sum_{all\ \ pairs}\ \sum_{(i,j)} y_i$$

- Check that the Gini coefficient is 1minus this expression
- Now we will consider the problem of sampling pairs and use what we already know about sampling

- If the sample is SRS the sample of pairs has the inclusion probabilities

- $\pi_{(i,j)}$ = n(n-1)/N(N-1) if i#j
  n/N if i=j
- $\pi_{(i,j)(k,l)}$ = n(n-1)(n-2)(n-3)/n(N-1)(N-2)(N-3)
  if all indices are unequal.
  if some are equal the number of factors in numerator and denominator decreases correspondingly

- It is now easily seen that we can estimate the Gini coefficient by a HT-ratio estimator and its variance correspondingly with SYG variance estimators as building blocks.