Urvalsmetoder och Estimation 10

Sampling and Estimation 10

2012-03-09

11.4 Hansen Hurwitz plan –

Subsampling in the non-response

- In a recent mail study on the number of dogs in Sweden, a random sample from the ordinary population was drawn and asked about their pets.
- In the first round a large non-response was observed after reminders (inclusion probability π_1. A subsample of the non-respondents were selected with inclusion probability π_2, and they were later contacted by phone).
- Estimate total by $\Sigma_{R1}\ Y_i/\pi_1 + \Sigma_{R2}\ Y_i/\pi_1\pi_2$
  (Assuming no non-response in the second phase)
- This gave a much lower estimate than the estimate without the second phase $\Sigma_{R1}\ Y_i/\pi_1\ /\ \Sigma_{R1}\ 1/\pi_1$
- Why? Do you think?

11.5 Editing

- Editing (Checking the answers) (Granskning) is an important topic in surveys in itself. For Statistics Sweden it accounts for 40 % of all data collection costs for business statistics.
- A good practice is to look at the sample. For each unit assess a probability of being incorrect and an estimate of the effect on the total estimate if incorrect.
  - Often only those with high probabilities and high potential effects are checked
- Another procedure is sampling:
  - Classify. Use this classification as an auxiliary varible for stratification.
  - Take a subsample in each stratum and call back to all in the sample.
  - Estimate the effect of calling back to the full sample

12. Large sample properties

12.1 Central limit theorem for independent variables.

- You have read about the central limit theorem (clt) and the law of large numbers. Do you know what the clt says?

- One variant is:
- Let X_1, X_2, X_3, … be a sequence of independent and identically distributed random variables with expected value m and variance σ^2. Then

$$P\left(\frac{\Sigma(X_i)-nm}{n^{1/2}\sigma} \le z\right) \to \Phi(z) \quad as \quad n \to \infty$$

where Φ is the standard normal distribution function.

- This is perhaps the most well known version and the result is what is called convergence in distribution

- There are bounds for the rate of convergence (named after Berry & Esseen). The difference between the right and left hand side is always less than

$$c * E(|X-m|^3)/(\sigma^3\sqrt{n})$$

- A general value on c is 2.05, but it may be much smaller for some distributions.
- When we construct confidence intervals we look at the points where the distribution function is 0,025 (α/2) or 0,95 (1-α/2) This expression tells us
  - what the maximal error in confidence level is
  - that it decreases as $n^{-1/2}$.
  - That generally speaking convergenc is slower for distributions with heavier tails

- Another variant with different distributions is:
- Let X_i, i = 1, 2, … be a sequence of independent random variables
  - with means m_i and variances σ_i^2 and
  - with a bounded third absolute moments (i.e. there is a number η such that $E(|X_i - m_i|^3) < \eta$ for all i) and
  - the total variance $\Sigma\sigma_i^2 \to \infty \ as \ n \to \infty$
- In that case

$$P\left(\frac{\Sigma X_i - \Sigma m_i}{(\Sigma\sigma_i^2)^{1/2}} \le z\right) \to \Phi(z) \quad as \quad n \to \infty$$

If you try to use these theorems on sampling, you will meet some problems. Which?

12.2 CLT and sampling

12.2.1 Differences; sampling and iid

- The random variables X_1, X_2, X_3, … are not independent
- It is impossible for n to tend to infinity since it can never be higher than N and when it is close to N the sum is certainly not Gaussian.

A special variant of the clt is thus needed.

Needed changes

- Drawing from a finite population implies a certain type of dependence.
- Both the population size N and the sample size n must tend to infinity.
- Since the consecutive addition of elements to the population changes it, we must allow for distributions varying with n.

12.2.2 A central limit theorem for sampling

- Let U_i ; i=1, 2, … be a sequence of populations with sizes N_i and bounded normalized absolute third moments

$$\frac{\Sigma_{U_i} | X_j - \overline{X} |^3}{N_i \sigma^3} < \eta \quad \textit{for some number} \ \eta$$

- Let N_i and n_i be such that both n_i and $(N_i – n_i)$ tend to infinity as i tends to infinity
- Take a srs sample S_i from each population with size n_i.
- Then

$$P(\frac{\Sigma_{S_i} X_j - n_i \overline{X}_{U_i}}{(\frac{n_i (N_i - n_i)}{N_i})^{1/2} \sigma_i} \leq z) \to \Phi(z) \ as \ i \to \infty$$

- Hajek was the first to give a practically useful formulation of a central limit theorem for sampling. There exist much sharper versions nowadays and also theorems dealing with the rate of convergence.

Sketch of a proof for model-based sampling (using the ordinary clt)

- Let Y_1, Y_2, Y_3, … and X_1, X_2, X_3, … be two independent iid sequences both with the same distribution F, which has three moments.
- Let the population U consist of {Y_1, Y_2, …. Y_n, X_1, X_2, … X_{N-n}) and the sample of the first n.
- Then the mean in the sample, $\vec{y}$, and of the remaining units, $\vec{x}$, are independent and both tend to normal distributions.
- A linear combination of two independent normal variables is also normal.
- Thus

$$\frac{1}{n}\Sigma_S Y_i - \frac{1}{N}(\Sigma_S Y_i + \Sigma_{U-S} X_i) = \frac{N-n}{N}\Sigma_S Y_i - \frac{1}{N}\Sigma_{U-S} X_i$$

must also have an asymptotic normal distribution

- The proof for designbased is much more complicated

12.2.3 An example: The Wilcoxon rank sum

- We have two independent samples x_1, …, x_n and y_1, …, y_m of iid random variables and we want to test whether they come from the same population
- The combined sample of N = n+m units is ordered and the ranks of all x-units is added giving a rank sum.
- If the distributions were the same, all orderings are equally likely. In other words the x-ranks is a sum of a simple random sample of size n from U={1, 2, …, N}
- The mean, variance and absolute third moment of the values in U are (N+1)/2, (N+1)(N-1)/12 and $(N-1)^2(N+1)^2/(32N)$. The conditions of the theorem applies.
- The limiting distribution of the Wilcoxon rank sum statistic is thus normal with mean m(N+1)/2 and ((N-m)/(N-1))*m(N-1)(N+1)/12= m(N-m)(N+1)/12

12.3 What about the variance estimator?

- 10.3.1 The standard iid case.

- In the iid case, what do we know about the variance estimator?

- It is consistent i.e. $s^2 \to \sigma^2 \; as \; n \to \infty$. When and why?

- It is consistent i.e. $s^2 \to \sigma^2 \; as \; n \to \infty$. When and why?
- The law of large numbers, since

$$s^2 = \frac{1}{n-1}\Sigma(X_i - \overline{X})^2 = \frac{n}{n-1}(\Sigma X_i^2 / n - \overline{X}^2)$$

- The first term here is an arithmetic mean of a sum of random variables (X^2) and the LLN says that it converges to E(X^2) = $\sigma^2 + m^2$, in probability, if Var(X^2) exists.
- Since the mean of X also converges (to *m* in probability), the whole expression converges to σ^2.

What about intervals (for iid)?

- Since $\frac{\overline{X} - m}{\sigma / n^{½}} \to N(0,1) \; and \;\; s \to \sigma$,
- also $\frac{\overline{X} - m}{s / n^{½}} \to N(0,1)$
- Remember: It is only for originally normal variables that one can use t-distributions.
  - Otherwise one has to use the central limit theorem to obtain normality and when *n* is large s^2 has converged.
  - But on the other hand it is reasonable to use wider intervals when the variance is estimated. Thus use t-values or something larger.

12.3.2 Variance estimation and intervals for sampling

How much of the results for the iid case holds also for sampling from finite populations

?

Everything holds!

- Under the same conditions as above (a sequence of populations with uniformly bounded relative absolute third moments)
- Then s^2_S / s^2_U -> 1 w. p. 1
- And

$$P\left(\frac{\overline{X}_S - \overline{X}_U}{\frac{N-n}{nN} s_S} \leq z\right) \rightarrow \Phi(z) \text{ } as \text{ } i \rightarrow \infty$$

- But the convergence is sometimes slow and t-intervals are seldom used.
- Many books use only a number 2 and do not call the intervals confidence intervals, since the coverage probabilities can differ much from exactly 0,95.
- The usual rules of thumbs for when a normal approximation is allowed from standard statistical text-books seldom applies, since variables are often very skew (in particular economic variables incomes, turnover, number of employees etc) (But for binary variables (proportions) they can be used).

The following is a Berry-Esseen type version holding for every finite population and SRSwor. It treats the case when the distribution is normalised with the estimated variance. In the real life you almost never knows the true variance.

$$\sup \left| P\left(\left\{\sqrt{\frac{n}{(1-n/N)}}\, \frac{\overline{y}_S - \overline{y}_U}{s} \leq x\right\} - \Phi(x)\right) \right| \leq$$

$$\leq c \, \frac{\sum |y_i - \overline{y}_U|^3 / N}{\sqrt{\frac{(N-n)n}{N}}\, \sigma^3} \qquad (Bloznelis, 1999)$$

The same theorem with the true variance σ instead of the estimated standard deviation s was shown already 1975 by Höglund.

Not covered

- Area sampling
- Small area estimation
- Analyses of survey data
- Adaptive sampling
- …

- *Thanks for your patience*

- *Good luck!*