

# Urvalsmetoder och Estimation 10

## Sampling and Estimation 10

2008-12-18/19

### Several variables

- $Y = g(X_1, X_2, \dots, X_p)$
- $E(Y) \sim g(m_1, m_2, \dots, m_p)$

$$Var(Y) \approx \sum_i \left( \frac{\partial g(m_1, m_2, \dots, m_p)}{\partial m_i} \right)^2 Var(X_i) + \sum_{i \neq j} \frac{\partial g(m_1, m_2, \dots, m_p)}{\partial m_i} \frac{\partial g(m_1, m_2, \dots, m_p)}{\partial m_j} Cov(X_i, X_j)$$

- Example: Ratio-estimation

## 10. Miscellaneous topics

### 10.1 Gauss- or Taylor-approximation

#### 10.1.1 Theory

- $X$  is a random variable
- $m$  is a central point. Usually, as here, the mean
- $Y = g(X)$  approximeras med
- $g(m) + (X-m) g'(m)$  eller
- $g(m) + (X-m) g'(m) + \frac{1}{2}(X-m)^2 g''(m)$
- ger  $E(Y) \sim g(m) + \frac{1}{2} Var(X) g''(m) \sim g(m)$
- $Var(Y) \sim (g'(m))^2 Var(X)$
- If  $Var(X) \sim C/n$ , these are good approximations for large  $n$

#### 10.1.2 Example logodds ratio

- 15 years ago 10 000 (out of 85 000) 65-year-olds were sampled and questioned 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 ratio is  $\ln(155 \cdot 7940 / 1318 \cdot 587) =$

$$\ln(X_{11}) + \ln(X_{22}) - \ln(X_{12}) - \ln(X_{21}) = 0.46$$

(a very common measure of effect/relation. Odds is for example often used in gambling)

- 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

$$1/p_{11}, 1/p_{22}, -1/p_{12} \text{ and } -1/p_{21}$$

## 10.2 Resampling

- After some calculations it is easy to see that the variance can be approximated by
- $\sum_{ij} \text{Var}(\hat{p}_{ij})/\hat{p}_{ij}^2 + \sum_{ij,kl} \text{Cov}(\hat{p}_{ij}, \hat{p}_{kl})/(\hat{p}_{ij}\hat{p}_{kl}) = \dots$   
 $\sum_{ij} 1/x_{ij} =$   
 $1/155 + 1/587 + 1/1318 + 1/7940 \sim$   
 $0.00904 \sim (0.095)^2$  (without correction for a finite population)
- A 95% interval for the logoddsratio will thus approximately be  
 $0.46 \pm 0.19 = (0.27, 0.65)$
- This test of independence is asymptotically equivalent with the chi2-test (but better in a way since it can be made one-sided)

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

### 10.2.2 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 jackknife, 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).

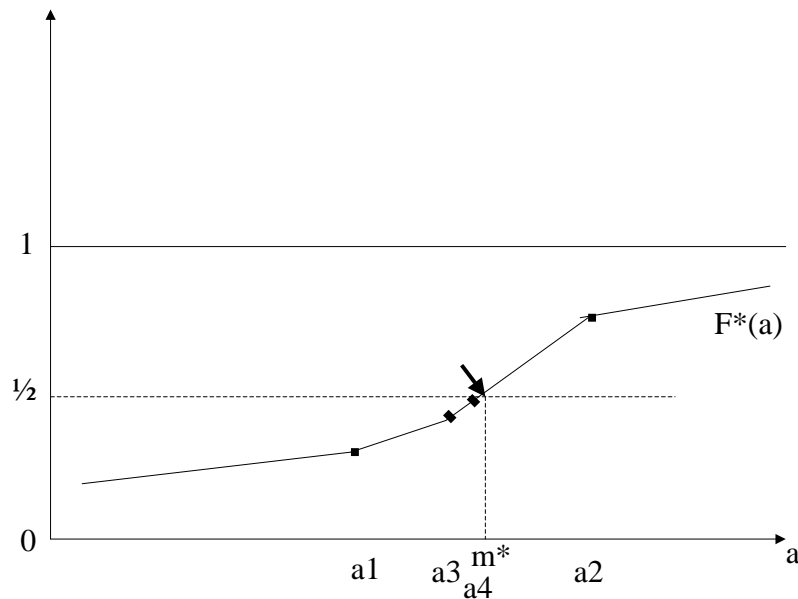
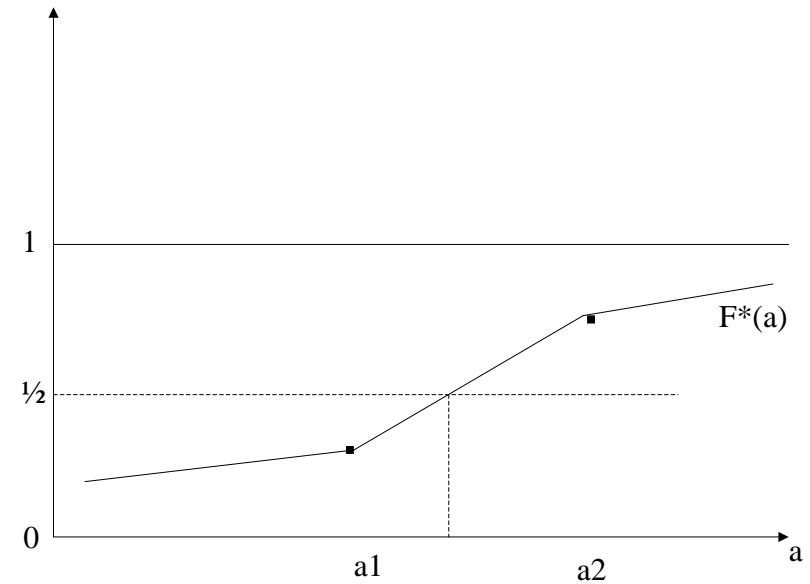
### 10.2.3 Balanced half-sampling Balanced Repeated Replications

- It is well-known that  $\text{Var}(X_1+X_2) = \text{Var}(X_1-X_2)$  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  $\text{Var}(t_1 - t_2)$  and also of  $\text{Var}(t_1 + t_2) = \text{Var}(t)$  (Note that this holds regardless of the sampling fraction)
- 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)

## 10.3 Derived quantities

### 10.3.1 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  $1/2$
- This is a 95% confidence interval for the median
- Similarly for any quantile



- If the sample is SRS the sample of pairs has the inclusion probabilities
- $\pi_{(i,j)} = \frac{n(n-1)}{N(N-1)}$  if  $i \neq j$   
 $\frac{n}{N}$  if  $i = j$
- $\pi_{(i,j)(k,l)} = \frac{n(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)}$   
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.

## 10.5 Quality

### 10.5.1 Introduction

- "Every industrial process should not only produce products of good quality, but also information on the process itself, which enables one to improve it even further" (George Box, famous statistician and quality specialist)
- Similar statements by other e.g. Deming (before becoming a quality guru he was one of the best known survey specialists). Other names are Ichikawa, Tagueuchi

### 10.5.2 Embedded experiments

- When you do a periodic survey you should experiment in the survey
- Small experiments not hazarding the statistical results but improving the knowledge.
- For example comparing different question formulations, introductory letters, forms of presentation of the survey etc
- Also document what happens during the survey so that you can estimate costs, when people are home etc.

### Example of embedded experiment

- Order effects in CATI-interviews. A stratified study.
- Q. Which are the three most important political questions for you in this election:
 

1. Immigration	5. Schools and education	9. Others, which
2. The economy	6. The environment	.....
3. Health	7. Housing	.....
4. Care of elderly	8. Gender equality	.....

Compare this with  
opposite order  
1-8 replaced by Wages, Law and order, Foreign policy and peace, Income inequality, Military defence, Taxation, Foreign aid, Child care  
And those in opposite order

- After having drawn the sample, divide it into four equal parts.
- Use standard methods for design of experiments: E.g. each stratum is divided equally, equally many males/females in each part. Randomise the interviews among the interviewers (if possible so that each interviewer gets equally many from each part).

- Analyse the pooled data from the survey as usual with percentages for the issues mentioned most often. (Use finite population correction)
- Analyse order effects and effect of being on the list read to the respondent. (Variance analysis may be a good method but often even simpler methods are sufficient)
- Remember this is not a finite population survey but an experiment and the population should be regarded as infinite (You are primarily not interested in what happened in this population, but what will happen in the future).

### 10.5.3 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  $\pi_1$ . A subsample of the non-response was selected probability  $\pi_2$  and called).
- Estimate total by  $\sum_{R1} Y_i/\pi_1 + \sum_{R2} Y_i/\pi_1\pi_2$   
(Assuming no non-response in the second phase)
- This gave a much lower estimate than the estimate without the second phase  $\sum_{R1} Y_i/\pi_1 / \sum_{R1} 1/\pi_1$
- Why? Do you think?

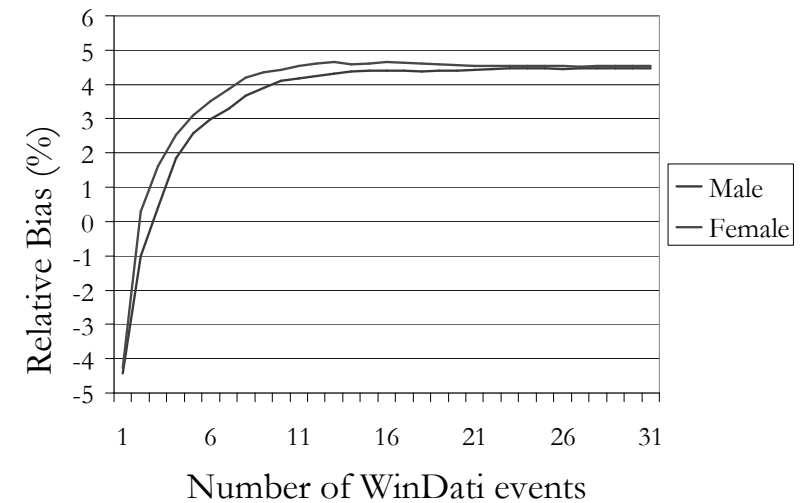
### 10.5.4 Example 2 - Editing

- Editing (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 variable 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

## 10.5.5 A study of non-response

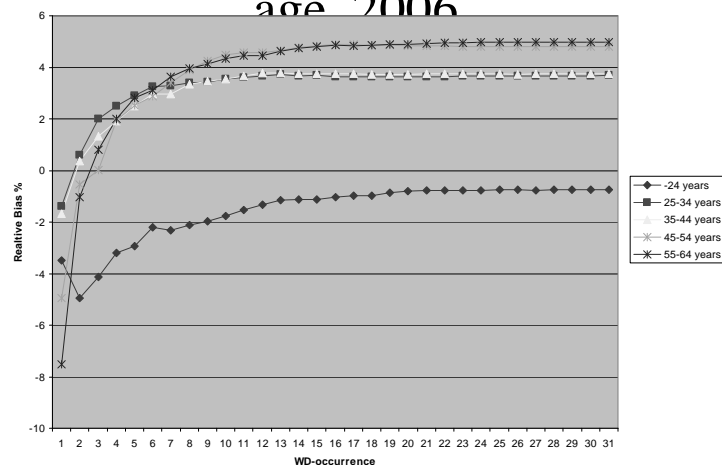
- I am currently involved in an experiment where we try to measure the effect of different call algorithms (which persons in a sample should be contacted first and at what times of the day and how many times)
- A very general and overgeneralised description is that the population consists of three groups
  - Home-sitters. Stugsittare (people at home and easy to contact. Home with children, unemployed ordinary people, not out at nights or freetime)
  - Ordinary busy people (People difficult to reach but they will be reached eventually after 10 or thirty days or so. Often away on travels, conferences, or out during nights on sports or political activities a.s.o)
  - Utslagna (Homeless, backpackers etc. Will never be reached)
- The middle group is the group with highest income and living conditions.

## Relative Bias, Annual Salary, Monthly Averages



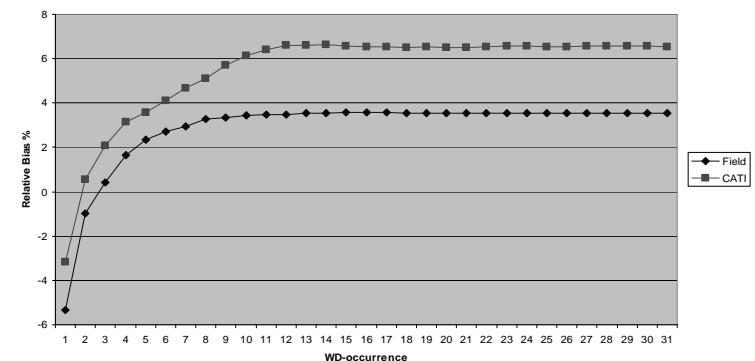
## Mean relative bias of salary after age 2006

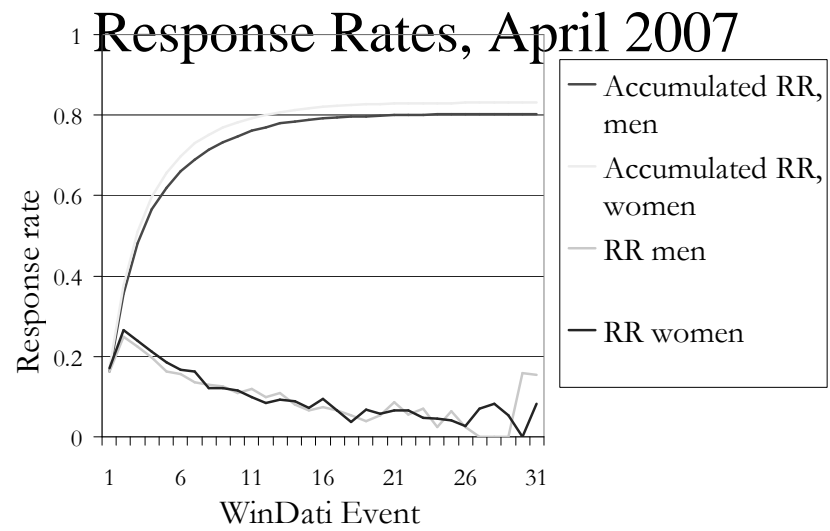
LFS Mars-Dec. 2007 Mean Relative Bias of Salary 2006



## Rel Bias - after type of interviewer

LFS Mars-Dec. 2007 Mean Relative Bias of Salary 2006





*Merry Christmas  
and  
A Happy New Year*

*God Jul  
och  
Gott Nytt år*