Urvalsmetoder och Estimation 8

Sampling and Estimation 8

2011-03-07

Practice at Statistics Sweden, summer 2011

- For persons actively studying statistics at higher levels. One requirement is speaking Swedish fluently

- For information, contact one of those persons at Statistics Sweden:
- Martin Axelson: 019-17 61 18
- Dan Hedlin: 08-50 69 43 34

9. Miscellaneous topics
9.1 Gauss- or Taylor-approximation

9.1.1 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

Several variables

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

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

$$\sum \sum_{i \neq j} \frac{\partial g(m_1, m_2, \ldots, m_p)}{\partial m_i} \frac{\partial g(m_1, m_2, \ldots, m_p)}{\partial m_j} Cov(X_i, X_j)$$

- Example: Ratio-estimation

9.1.2 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 ratio is $\ln(155*7940/(1318*587)) =$

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

(a very common measure of effect/relation. Odds is for example often used in gambling. 0 means no effect/independence)

- 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

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

- After some calculations it is easy to see that the variance can be approximated by

  $\Sigma_{ij}\ \mathrm{Var}(p^*_{ij})/p^{*2}_{ij} + \Sigma_{ij,kl}\ \mathrm{Cov}(p^*_{ij}, p^*_{kl})/(p^*_{ij}p^*_{kl}) =$

  $\Sigma_{ij}\ p^*_{ij}(1-p^*_{ij})/p^{*2}_{ij} + \Sigma_{ij,kl}\ -p^*_{ij}\, p^*_{kl}/(p^*_{ij}p^*_{kl}) = \ldots$

  $= \Sigma_{ij}\ 1/x_{ij} =$

  $1/155 + 1/587 + 1/1318 + 1/7940 \sim$

  $0.00904 \sim (0.095)^2$ (without correction for a finite population)
- An approximate 95% interval for the logoddsratio will thus approximately be

  $0.46 +/- 2*0.095 = (0{,}27, 0.65)$
- The corresponding test of independence is asymptotically equivalent with the usual chi2-test (but better since it can be made one-sided, and converges faster to the asymptotic distribution)

9.2 Resampling

- Heard about Jackknife, Bootstrap … ?

• 9.2.1 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 $ng(X_1, X_2, \ldots, X_p)- (n-1)g(X_1, X_{i-i}, X_{i+1} \ldots, X_p)$; i=1, … ,n

- (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).
- Can be used for also for finite population SRS-sampling. More difficult for sampling with varying inclusion probabilities.

9.2.2 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 (e.g. B=50 times).
- We can compute the empirical variance from these resamples (and also bias and full distribution and 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).

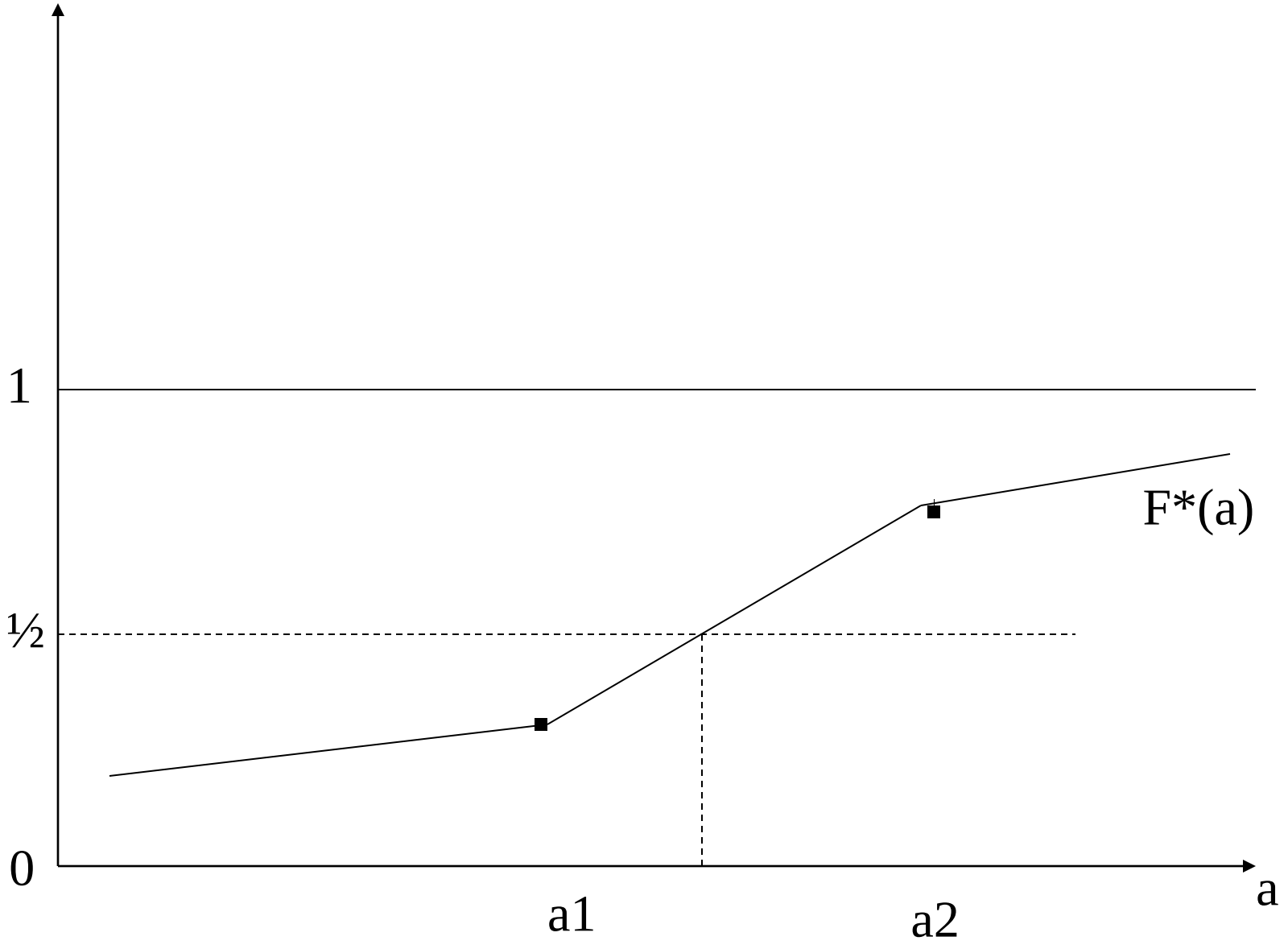
9.2.3 Balanced half-sampling

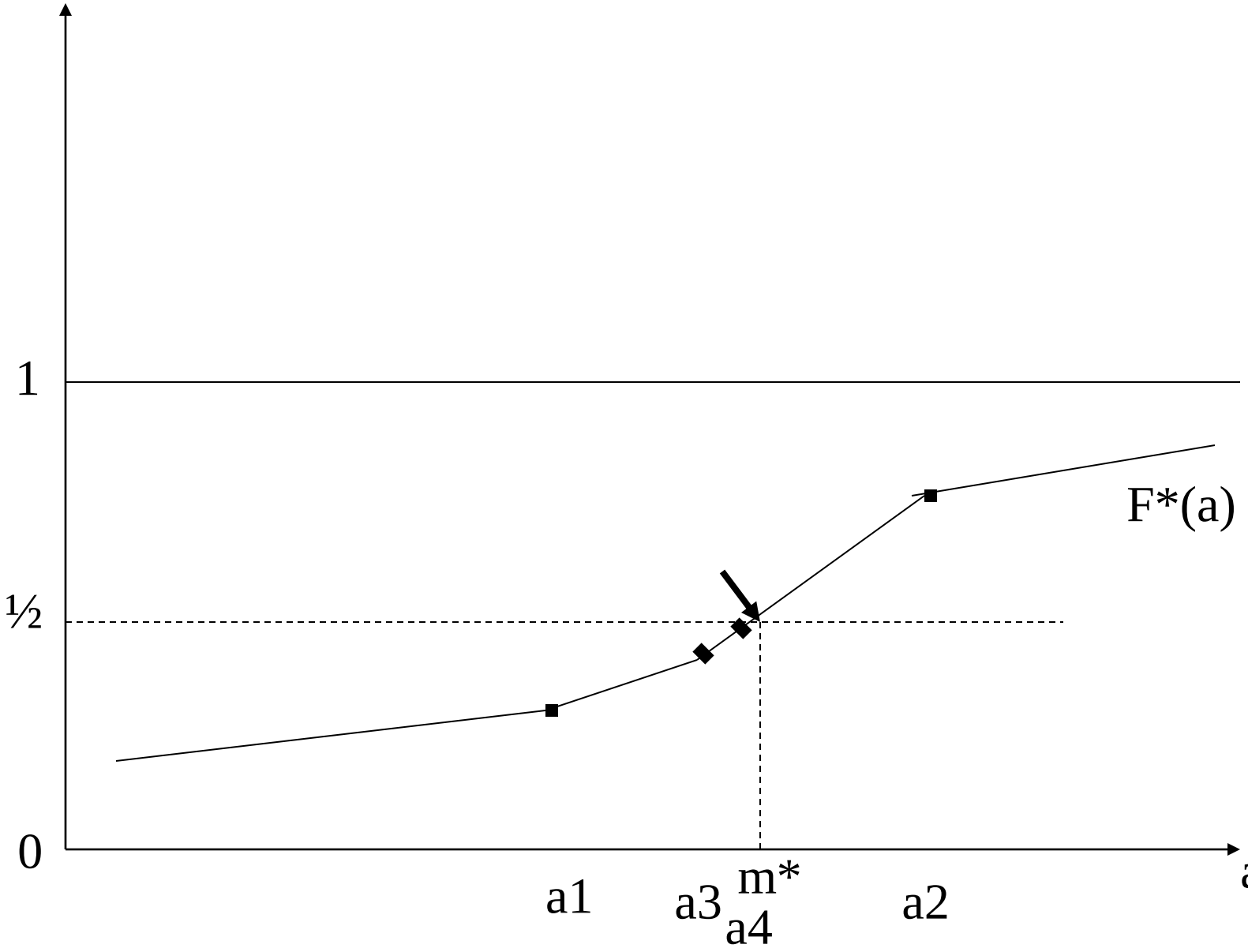
Balanced Repeated Replications

- 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 (i.e. assuming N=N/2). Then $(t_1 - t_2)^2$ is an estimate with 1 d.f of $Var(t_1 - t_2)$ and thus of $Var(t_1 + t_2) = Var(t)$ (Note that this holds regardless of the sampling fraction)
- Do this repeatedly with more random halves getting more d.f.
- This 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)

9.3 Derived quantities
9.3.1 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

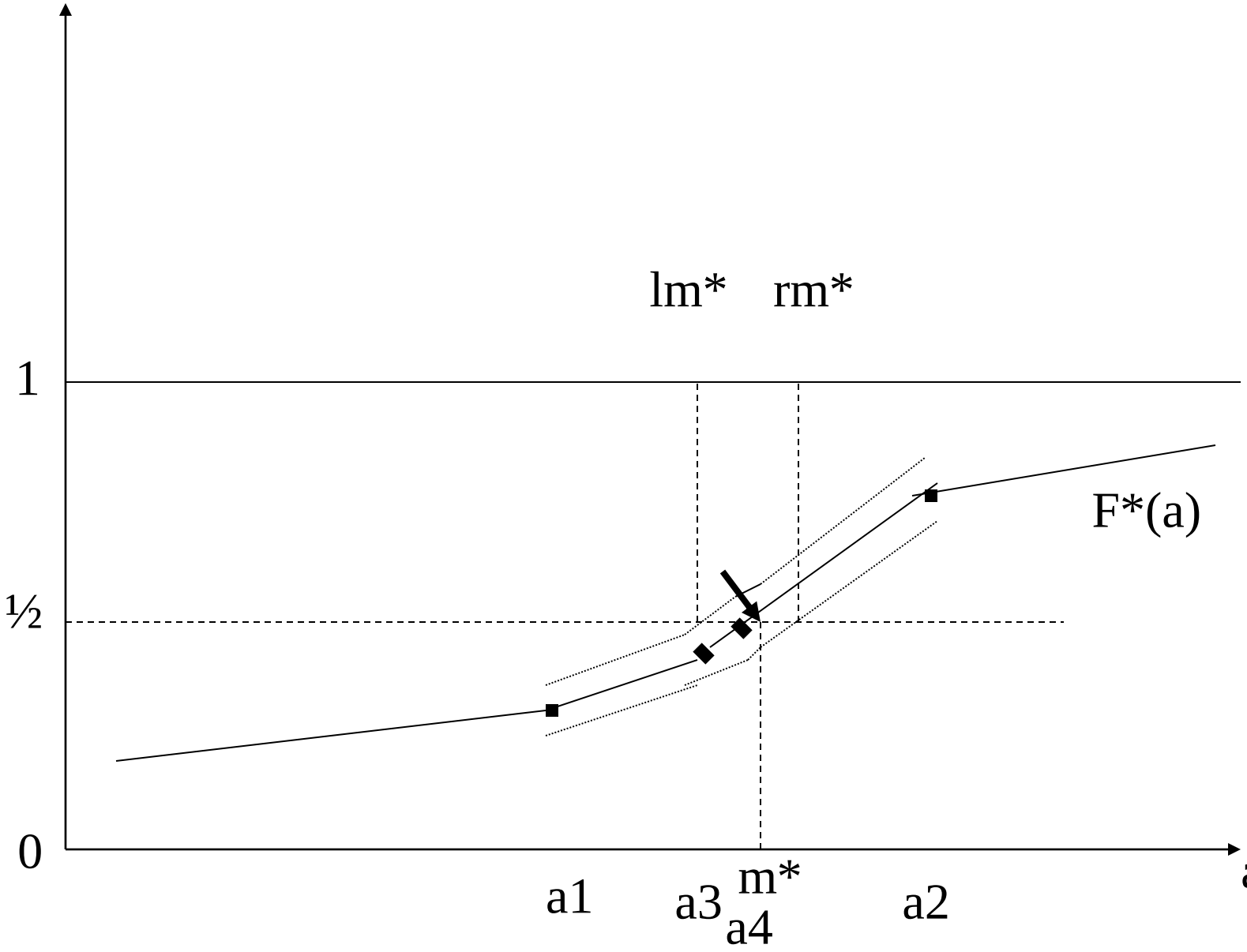

1
½
0
F*(a)
a1
a2
a


1
½
0
F*(a)
a1
a3
a4
m*
a2
a

Intervals

- Find confidence intervals for F(a)

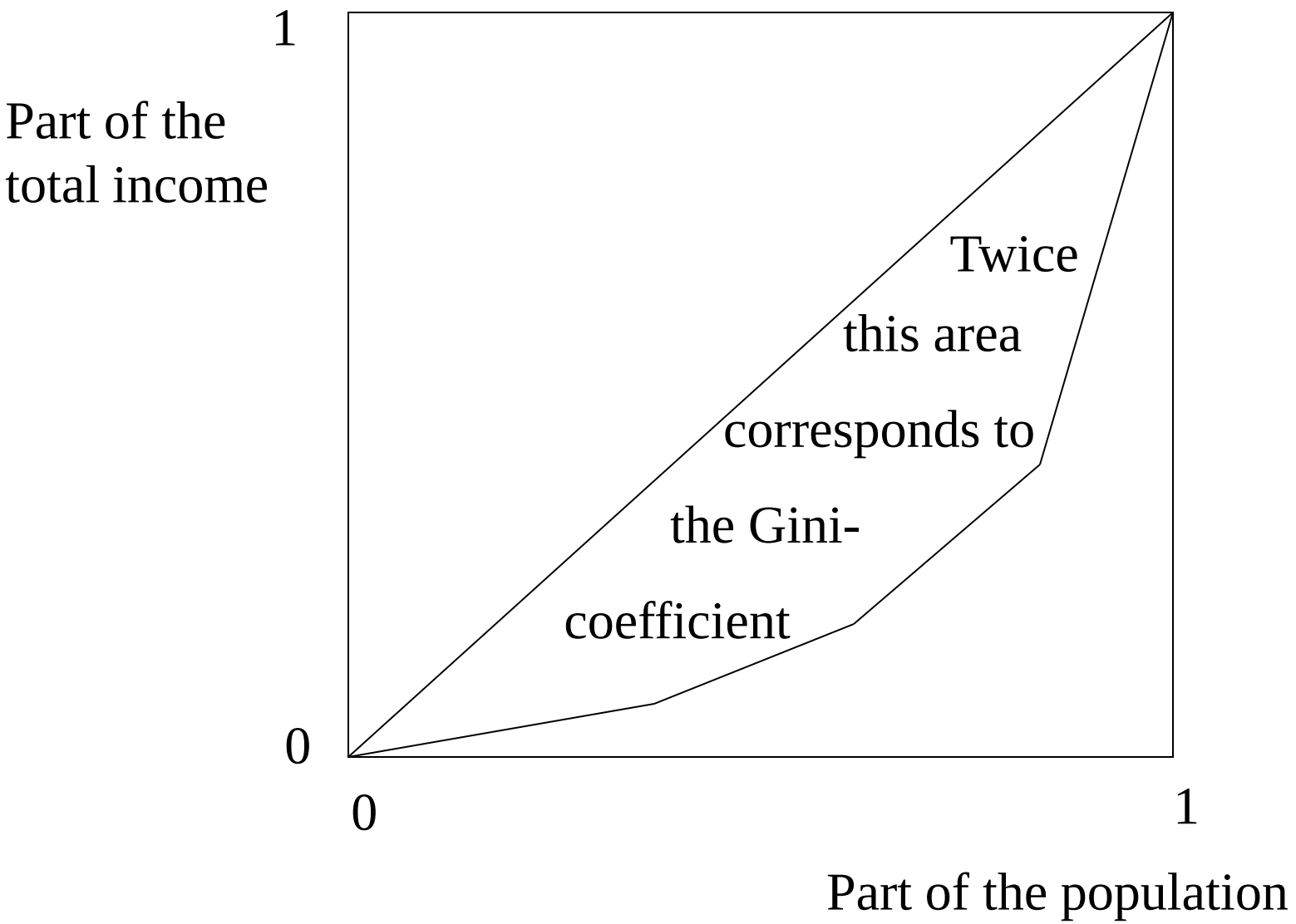
  for some values of a including m*.
- Connect them and find their intersections with the line ½
- This is a 95% confidence interval for the median
- Similarly for any quantile


lm*
rm*
1
½
0
F*(a)
a
a1
a3
m*
a4
a2

9.3.2 The Gini coefficient

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

Picture of the income distribution



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.
  $= n(n-1)(n-2)/n(N-1)(N-2)$
  if two indices in are equal

- 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.

9.4 Quality
9.4.1 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, Tageuchi

9.4.2 TQM – TSE

Total Quality Management – Total Survey Errors

- How to weight between different aspects. How to give the reader the best information given a limited budget.
  - How much of the budget and time should be on questionnaire design, length of interview, mode, sample size, reminders, interviewer education, choice of frame (e.g. RDD versus RTB), non-response compensation, presentation etc.
  - For example weight relevance against response rate. Is it better with to ask for the monthly salary from the main job or to ask for total yearly income from all sources. You may guess that the monthly salary has a 50 % smaller variance (per year) but that it is an underestimate with between 5 and 15 % (bias) and that the item non response rate will increase from 0 % for monthly salary to 5 % for total income.
  - Use elements of decision theory and subjective distributions. E.g.
    - Variance may be $0{,}5/n + (0.1/4)^2$ (many assumptions e.g.standard deviation = average level)
    - Variance may be $1/(0.95\ n)$ (many assumptions e.g. non-response is MCAR)
    - Chose the second method if $n > 783$

9.4.3 Embedded experiments

- When you do a periodic survey you should experiment in the survey
- Small experiments not hazardous to 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 interwiews among the interviewerws (if possible so that each interviewer gets equally many from each part).
- Easy to do with CATI and also with web surveys

- 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 the experiment is not a finite population survey but an experiment and the population should be regarded as infinite (You are not primarily interested in what happens in this population, but what will happen in similar studies in the future).

9.4.4 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?

9.4.5 Editing

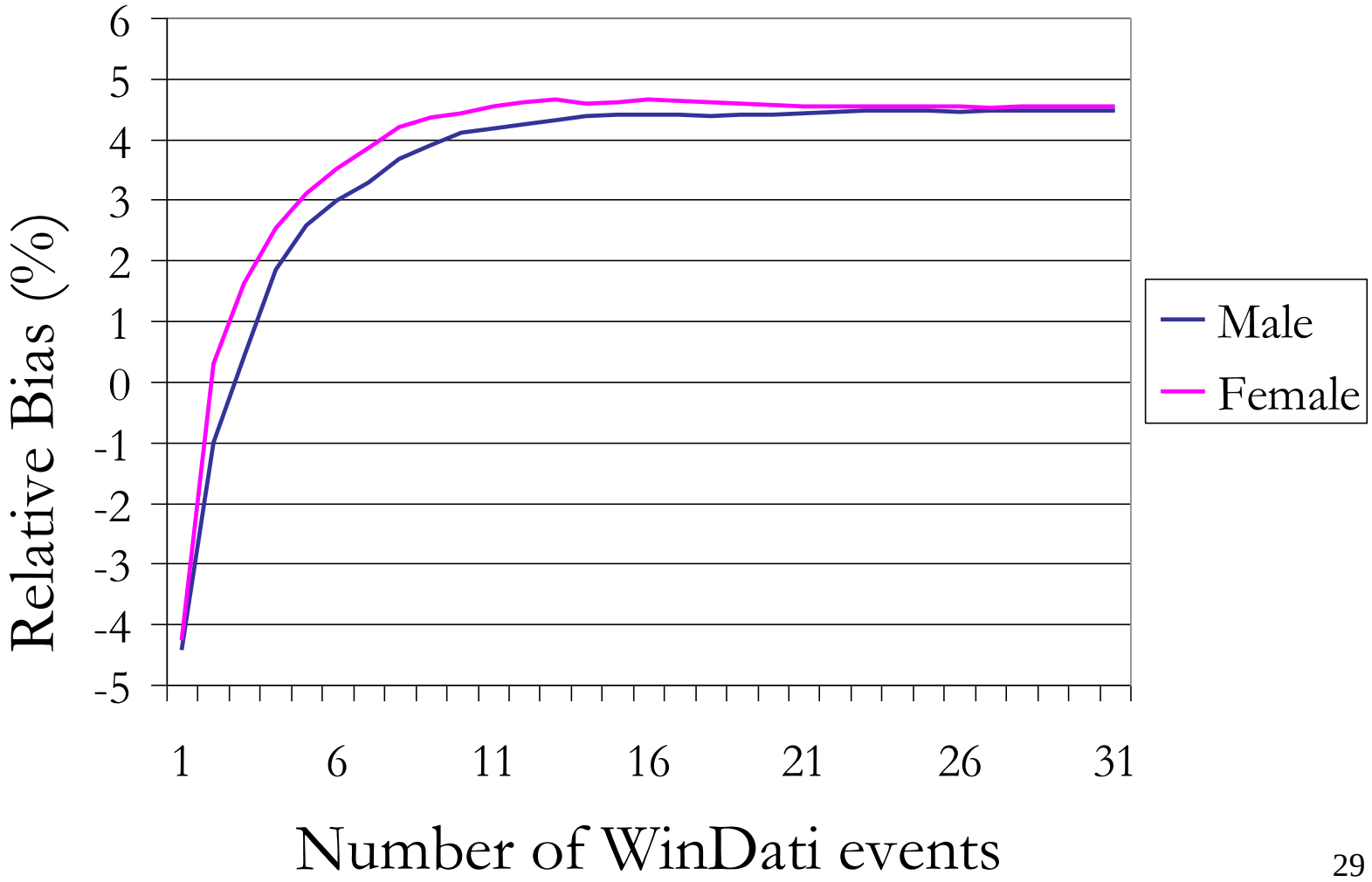
- Editing (Checking the answers) (Granskning) is an important topic in surveys in itself. For Statistics Sweden it accounts for 40 % of all data collection costs for business statistics.
- A good practice is to look at the sample. For each unit assess a probability of being incorrect and an estimate 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

9.4.6 A study of non-response

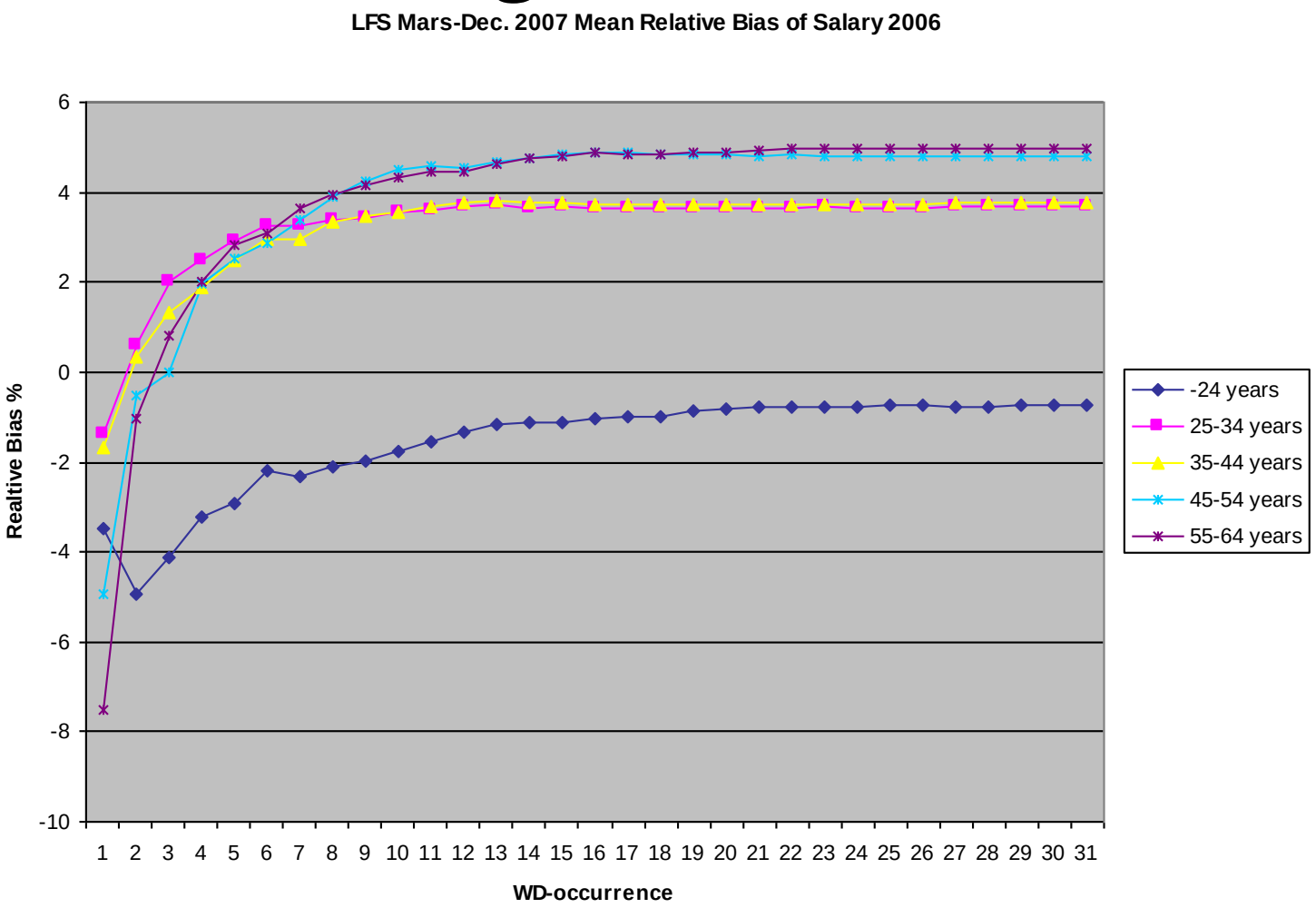
- Some years ago I was involved in an experiment where we tried to measure the effect of different call algorithms (which persons in a sample should be contacted first and at what times of the day and how many times)
- A very over-generalised description is that the population consists of three groups
  - Home-sitters. Stugsittare (people at home and easy to contact. Home with children, unemployed ordinary people, not out at nights or free time activities)
  - Ordinary, mainly occupied people (People difficult to reach but they will be reached eventually after ten or thirty days or so. Often away on travels, conferences, or out during nights on sports or political activities a.s.o)
  - Homeless, backpackers, some youngsters etc. Will never be reached
- The middle group is the group with highest income and best living conditions.

- For each respondent we know how much efforts were made to reach him and if he eventually responded.
- For all we knew from registers their assessed income last year.
- We can thus estimate the bias if we asked for last years income and put in a certain amount of effort.

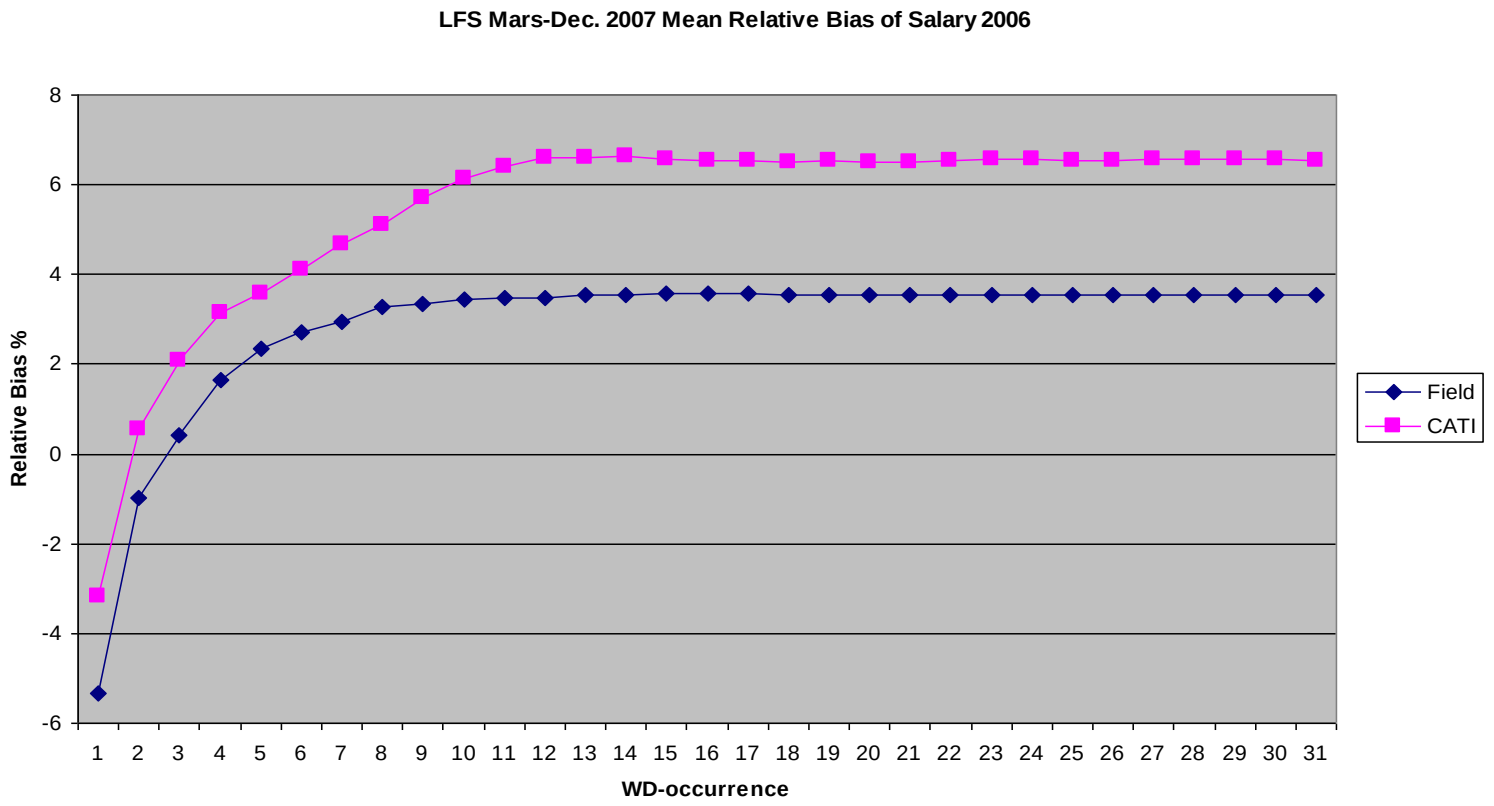
Relative Bias, Annual Salary



Mean relative bias of salary after age, 2006



Rel Bias - after type of interviewer



Response Rates, April 2007