Urvalsmetoder och Estimation 3

Sampling and Estimation 3

2011-01-31

Exam

- The result of the test will be presented 10 o´clock, Monday, March 28.

Summary of regression-type estimates

- Study variable Y, known X
- Find a predictor Y^*_i av Y € S, X € U (e.g by regression)
- Preliminary estimate $\Sigma_U Y^*_i$
- The error is $\Sigma_U (Y_i - Y^*_i) = \Sigma_U E^*_i$
- Estimate the error by $(N/n) \Sigma_S E^*_i$
- Final corrected estimate $\Sigma_U Y^*_i + (N/n) \Sigma_S E^*_i$
- With (approximate) variance $\frac{(N-n)}{n}\frac{N^2}{n}s_E^2$

- This holds for all the presented methods. But it is possible to correct the estimates of positive variables multiplicative too (ratio estimators)

- Preliminary estimate $\Sigma_U Y^*_i$

- The relative error is $\Sigma_U Y_i / \Sigma_U Y^*_i$

- Estimate the error by $R^* = \Sigma_S Y_i / \Sigma_S Y^*_i$

- Final corrected estimate $R^* \Sigma_U Y^*_i$

- With (approximate) variance $\frac{(N-n)}{n}\frac{N^2}{n}s_E^2$

- Where s^2_E is the estimated variance of $E_i = Y_i - R^* Y^*_i$, i€S

4. Stratified sampling

- The auxiliary variable X is categorical or a classification of a ordinal variable.

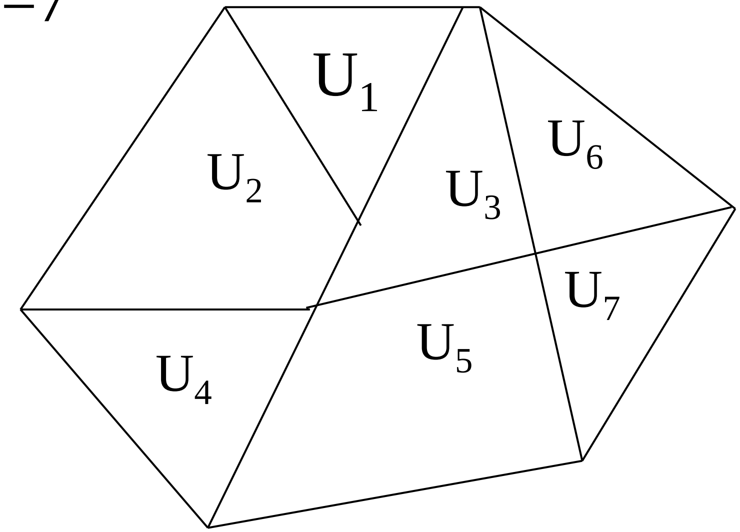
- Formally, the frame is a partition of U into a finite number, H, of groups/classes called strata, U_h. (e.g branch, marital status, region, age group)

- The partition/classification is known in advance

4.1 Simple stratified sampling

- First assume that the classification into strata is the only information (or the only information that will be used). Since there is no other information - SRS is the only possibility within stratum

U
H=7



- Take an SRS-sample of size n_h in each stratum.
- View each stratum as an SRS-problem of its own and estimate its total and variance.
- Estimate the population total by adding the stratum estimates and variances together!

Notation and formulas

- N_h units in stratum h, h = 1,…, H
- Take a sample, S_h from each stratum, n_h units from stratum h, h = 1,…, H (inclusion probability $\pi_i = n_h / N_h$)
- Estimate each stratum total by $t_{hy} = (N_h/n_h)\Sigma_{i€Sh}\ y_i$
- Estimate the population total by the sum: $t_{yst} = \Sigma_{h=1,\ \ldots,\ H}\ t_{hy}$

- Since sampling in the strata are independent, the variance of the sum is the sum of the variances
  $\mathrm{Var}(t_{yst}) = \Sigma_{h=1\ldots H}\ \mathrm{Var}(t_{hy}) = \Sigma_h\ (N_h\ (N_h-n_h)/n_h)\ \sigma_h^2$
  where σ_h^2 is the variance in stratum h

- An estimate of the variance is $\mathrm{Var}^*(t_{yst}) = \mathrm{Var}^*(\Sigma_h\ t_{hy}) = \Sigma_h\ \mathrm{Var}^*(t_{hy}) = \Sigma_h\ (N_h\ (N_h-n_h)/n_h)\ s_h^2$
  where s_h^2 is an estimate of the variance in that stratum.

Different objectives

- Estimate the population total (mean)
- Estimate the stratum totals (means)
- Estimate the stratum differences (e.g. average male income – average female income)
  T_{h1y}/N_{h1} - T_{h2y}/N_{h1} by t_{h1y}/N_{h1} - t_{h2y}/N_{h2},
  with variance: $Var(t_{h1y})/N_{h1}^2 + Var(t_{h2y})/N_{h2}$
- One may even want to test that all stratum means are equal by an F-test, (e.g political opinion does not depend on age) but now we leave the usual sampling framework (not finite population parameters any more?)

Is stratification any good?

- A stratification is good if the means and/or variances vary much between strata but not within

- Example 1. Two equally large strata with means m_1 = 1 and m_2 = 2 and common s.d. σ = 1. Total sample size n gives variances:
  $\mathrm{Var}(m_{yst}^*) = 1/n$ and $\mathrm{Var}(m_{ysrs}^*) = 1.25/n$
  (Correction for finite population neglected)

- Example 2. Two equally large strata with s.dev. σ_1 = 1 and σ_2 = 2 but common mean 1. Total sample size n gives variances:
  $\mathrm{Var}(m_{yst}^*) = 2.25/n$ and $\mathrm{Var}(m_{ysrs}^*) = 2.5/n$
  (With (optimal) sample sizes n/3 and 2n/3, see below).

- Thus differences in means are usually more important.
- In practice means and variances often vary together

4.2 Allocation

4.2.1 Proportional allocation

- All sample sizes are proportional to stratum size. It is very often used.
- Proportional allocation is an example of self-weighted sampling i.e. all units have the same probability to be included π_k = n/N.
- Self weighted sample has the advantage that people who do not know sampling may analyse the data as SRS and ”all” estimates will be unbiased but the variance estimates will often (as in this case) be overestimates.
- In many cases one even says that surveys with proportional allocation is SRS (e.g. the Swedish Labour Fource Survey, LFS)

Is proportional allocation any good?

- Sampling fraction is denoted by f
- Var(t_{yst}) = Σ_h N_h σ_h^2(1-f)/f
- Var(t_{ysrs}) = N σ^2(1-f)/f

$$\sigma^2 = \frac{1}{N-1}\sum\nolimits_U (y_i - \bar{y})^2 = \frac{1}{N-1}\sum\nolimits_h \sum\nolimits_{U_h} (y_i - \bar{y})^2 =$$

$$\frac{1}{N-1}\sum\nolimits_h (\sum\nolimits_{U_h} (y_i - \bar{y}_h)^2 + N_h(\bar{y}_h - \bar{y})^2) =$$

$$\sum\nolimits_h \frac{N_h - 1}{N-1}\sigma_h^2 + \sum\nolimits_h \frac{N_h}{N-1}(\bar{y}_h - \bar{y})^2$$

Thus

$$N\sigma^2 - \sigma^2 = \sum\nolimits_h N_h \sigma_h^2 - \sum\nolimits_h \sigma_h^2 + \sum\nolimits_h N_h(\bar{y}_h - \bar{y})^2$$

- It is easily seen that under mild restrictions the stratified variance is the smallest (but not if all stratum means are exactly equal) (Omit all ”small” terms i.e. without an N)

4.2.2 Optimal allocation

- Assume that
  - all variances are known
  - the marginal cost, c_h, for one more unit in the sample does not depend on how many that are already taken
  - we want to minimise Var(t_{yst}) for fixed total cost C.

    $\Sigma_h (N_h(N_h-n_h)/n_h) \sigma_h^2 = \Sigma_h (N_h^2/n_h) \sigma_h^2 - \Sigma_h N_h\sigma_h^2$

- Use the Lagrange multiplicator technique to minimise

  $$\Sigma_h (N_h^2/n_h) \sigma_h^2 - \Sigma_h N_h\sigma_h^2 - \lambda(C - \Sigma_h c_h n_h)$$

- Differentiate with respect to n_h and set the derivatives equal to zero: $-(N_h^2/n_h^2) \sigma_h^2 + \lambda c_h = 0; \ h = 1, \dots H$

- which gives that the optimal stratum sample sizes, n_h, should be proportional to $N_h\sigma_h/c_h^{½}$

  (if the boundary conditions are met i.e. if $n_h \leq N_h$)

- ”Proportional to $N_h\sigma_h/c_h^{½}$” gives that n_h can be written $k * N_h\sigma_h/c_h^{½}$ for some k

- Total cost will be $C = \Sigma_h\, kN_h\sigma_h c_h^{½}$

- Thus $k = C / \Sigma_h\, N_h\sigma_h c_h^{½}$

- Thus $n_h = C * (N_h\sigma_h/c_h^{½}) / (\Sigma_g\, N_g\sigma_g c_g^{½})$

- Total variance will be $\Sigma_h\, (N_h^2/n_h)\, \sigma_h^2 - \Sigma_h\, N_h\sigma_h^2 = (\Sigma_h\, N_h\sigma_h/c_h^{½})(\Sigma_g\, N_g\sigma_g c_g^{½}) / C - \Sigma_h\, N_h\sigma_h^2$.

- This is called Neyman allocation from Neyman (1934)

Constant costs across strata

- Often the cost per sampled unit is the same in all strata, i.e. c is assumed constant.

- Then the sample sizes should be proportional to $N_h\sigma_h$

- and the total variance will be

  $(\Sigma_h N_h\sigma_h)^2/ (C/c) - \Sigma_h N_h\sigma_h^2 = (\Sigma_U \sigma_i)^2/ (C/c) - \Sigma_U \sigma_i^2$

- When also all variances are constant, proportional allocation $n_h = C * N_h /\Sigma_g N_g$ is optimal

Minimum cost given the precision

- Suppose that the precision of the final estimate is the dimensioning side condition (the textbook criteria for dimensioning surveys) and that the goal of stratification is to minimise costs.

- Also here n_h should be proportional to $N_h\sigma_h/c_h^{½}$ (This is proved in the same way)

- Thus since $V = \Sigma_h (N_h^2/n_h) \sigma_h^2 - \Sigma_h N_h\sigma_h^2$ optimality gives that $n_h = ((N_h\sigma_h)/c_h^{½})*(\Sigma_g(N_g\sigma_g c_g^{½})/(V+\Sigma_g N_g\sigma_g^2))$.

Considerations

- It is seldom only the total estimate of one variable we are interested in. We are often interested in stratum estimates or estimates for other groups (often called domains)
- Multi-purpose surveys. For example we may survey farms and be interested both in economic conditions and environmental impact. Or a survey to individuals may ask about health, child care and income
- Some people say that one should minimise

$$\Sigma_h (N_h^2/n_h) \Sigma_j \omega_j\sigma_{hj}^2 - \Sigma_h N_h \Sigma_j \omega_j\sigma_{hj}^2 - \lambda(C - \Sigma_h c_h n_h)$$

where ω_j and σ_{hj} are the weights (importance) and variances attached to variable j.

More considerations

- The optimal size formulas do not give integer sample sizes. One must round the figures and the rounding is not always to the closest integer.

- If variances/costs are unknown, one must insert guessed/estimated values. Just pick any reasonable value. The maximum is usually flat

- Resources are often much better used on other statistical issues then minimising variance with the last few percents. (Nonresponse, questionnaire construction, editing, …)

Domains

- Domains are groups who are not known in advance in the frame, but where you want to report statistics for that partition
  - Poor people – below the poverty line, female chairman of the board, homosexuals, a.s.o.
- Construct new varables I_i = 1 if i € domain and $Y'_i = I_i * Y_i$
- Estimate the total as usual or the average by a ratio estimator $\Sigma\ (Y'_i / \pi_i) / \Sigma\ (I_i / \pi_i)$
- Variance as ratio estimators

4.3 Number of strata

- If the auxiliary data are on a continuous scale the data are often grouped and intervals used as strata.
  - E.g. age groups, farm size (hectares arable land), number of employees, income groups, ...

- How large should the strata be?

(several answers I will give five different)

Answer 1.

- It is the first divisions which are most important. You often gain more than half of what you can gain with only 2 strata. (e.g. if X and Y come from a multivariate normal distribution you gain $2/\pi$ of what is possible)

- It is usually enough with between five and ten intervals. In most cases you gain more than 90 % of what a stratification can give with 10 strata.

- But with very informative auxiliary data (often with skew distributions, economic data). It can be motivated to divide into even more strata. (Even if you already have gained 90 %, it may be good to gain 50 % of the remaining 10 %).

Answer 2.

- Use simple standard boundaries which also will be used for presentation.

- It is seldom motivated to divide further, but may be motivated in the largest classes for skew distributions.

- There is also an argument for using stratum boundaries that have been used before. It simplifies comparisons

- For presentation purposes a rule of thumbs says: Use roughly (proportional to) $n^{2/5}$ (or $n^{1/3}$) strata. More will give a diagram (histogram, frequency polygon) that fluctuates too much.

Answer 3.

- Take as many strata as you can. The variance can never increase by dividing into more strata (apart from rounding problems and, as we saw, in pathological cases).

- Usually the <u>mathematically</u> best method is to divide the auxiliary variable into intervals so that exactly one unit should be taken in each stratum (still apart from rounding problems).

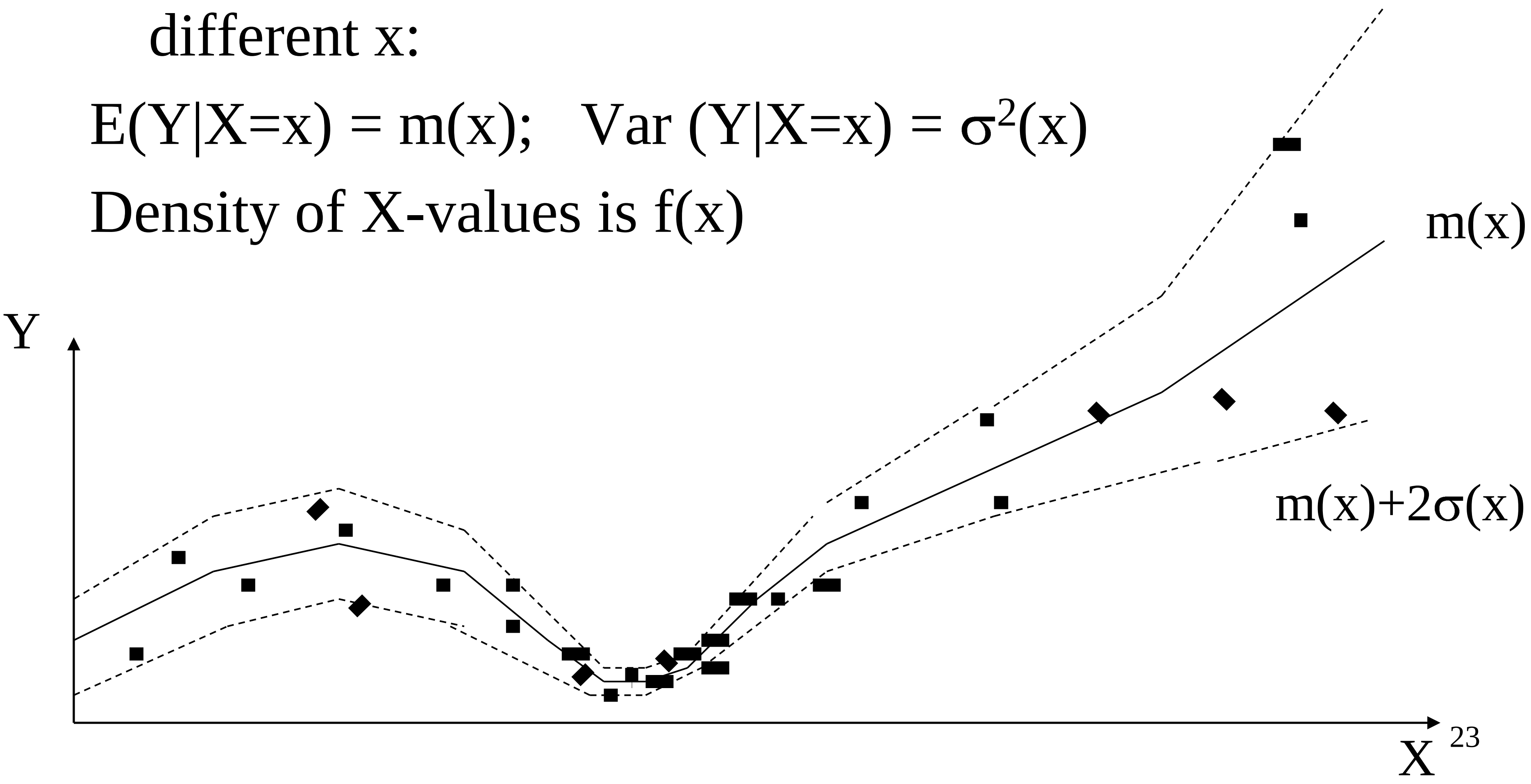
- A relevant aspect though: With only one unit in every stratum sample, the variance cannot be unbiasedly estimated (design-unbiased, but there are good methods to obtain an upper bound of that variance or to estimate the model-variance)

Answer 4. Model assisted stratification

Assume a model, where the y-values lie around a curve with different spread for different x:

$E(Y|X=x) = m(x)$; $\mathrm{Var}\,(Y|X=x) = \sigma^2(x)$

Density of X-values is f(x)



- We saw above that the sampling frequency in a stratum should be proportional to $N_h\sigma_h$ (if marginal unit costs were equal and optimal estimate was used) and that the total variance is $(\Sigma_U\ \sigma_i)^2/\ (C/c)\ \ -\ \ \Sigma_U\ \sigma_i^2$
- The variance in a small stratum of length L(x) around x is approximately $\sigma^2(x) + L^2\,(x)\ m'^2(x)/12$.
- We should thus try to minimise

$$\Sigma_U\ (\sigma^2(x_i) + L^2\,(x)\ m'^2(x_i)/12)^{1/2} =$$

$$\Sigma_U\,\sigma(x_i)(1 + L^2\,(x)\ m'^2(x_i)/(24\ \sigma^2(x_i))) =$$

$$\Sigma_U\,\sigma(x_i) + \Sigma_U\ L^2\,(x)\ m'^2(x_i)/(24\ \sigma^2\ (x_i)))$$

- This can be minimised using Lagrange multipliers (not easy) giving that L(x) should be proportional to $\sigma(x_i)/(m'^2(x_i)f(x_i))$

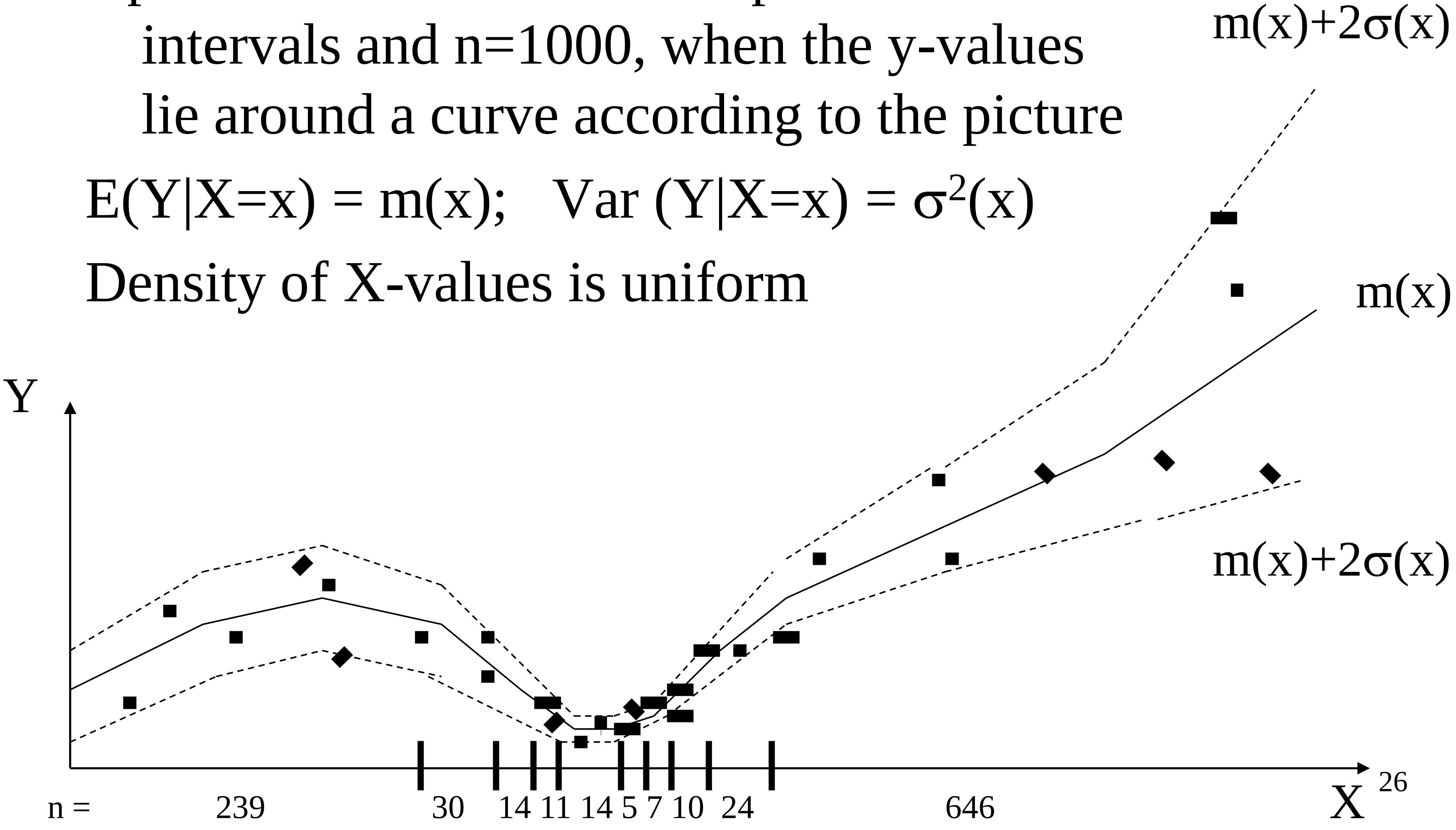
- The variance was minimised by an L(x) proportional to $\sigma(x_i)/(m'^2(x_i)f(x_i))$
- i.e. shorter intervals when the mean changes fast relative to the standard deviation.
- Many approximations, assuming that $\sigma(x)$ and $m'^2(x)$ can be guessed and are positive, and that between strata. But sensible result if optimal allocation is used.

Answer 4. Model assisted stratification

Optimal boundaries and sample sizes for 10 intervals and n=1000, when the y-values lie around a curve according to the picture

$E(Y|X=x) = m(x)$; $\text{Var}\ (Y|X=x) = \sigma^2(x)$

Density of X-values is uniform



Answer 5. Hodges-Dalenius’ cum square root of frequency rule

- A rule of thumbs. Useful, but with only weak theoretical support. Should be mentioned at this department! Example:
- Suppose we have decided on H (e.g. 3) strata?
- Classify into small strata of same lengths and give frequencies (end-points described by first digits)
- 1-2 3-4 5-6 7-8 9-10 11-12 13-14 15-16 17-18 19-20 21-22
- 50 142 213 328 397 212 105 47 12 5 1
- Take roots and cumulate
- 7.1 11.9 14.7 18.2 19.9 14.7 10.2 6.7 3.5 2.2 1
- 7.1 19.0 33.7 51.9 71.8 86.5 96.5 103.2 106.7 108.9 109.9
- Divide into 3 equal intervals 36.6, 73.3
- Strata 1-6, 7-10 and 11-22 (more exact 1 – 7.16 7.17 – 10.15, 10.16 -. But often practical with even end-points)

Many auxiliary variables

- There are often many auxiliary variables that can be used for stratification
- It is seldom feasible to stratify (cross-stratify) after all variables simultaneously
- In practice it is usually enough to stratify after the three most important and to use no more than 25 strata (if the purpose is to estimate the total)
- Another approach is to construct two helpfunctions trying to predict mean and variance and then to stratify after them
- Problems with many different study-variables, though. Their means and variances may be different, calling for different stratifications (taking this into account more than 25 strata may be called for)

4.4 Poststratification

4.4.1 Description

- Often the stratification is made after that the sample is taken.
- One reason is that one may not know the X-values in the frame (e.g. a list of surnames or telephone numbers) but you may know the total in the population (e.g. the number of females). For those in the sample you can ask about their value on the auxiliary variable.
- Another reason is that the stratification variables you want to use will be irrelevant for other study-variables
- And finally, the third reason may be that you just did not stratify, even though it would have been motivated.

- The estimates are the same as for ordinary stratification $t_{yst} = \Sigma_{h=1, \ldots, H}\, t_{hy} = \Sigma_{h=1, \ldots, H}\, (N_h/n_h)\Sigma_{i€Sh}\, y_i$ where n_h is the random number of sample units from stratum h.

- One usually also uses the same formulas for variance and variance estimation (even though this is not formally correct. It will be a type of conditional variance)

- The poststratified estimates are not ”formally” design-unbiased since n_h may be 0 (but in practice they are)

- If the poststratification is strongly influenced by the sample the estimated variance may become too small due to overfitting. But the problems are very small if strata are defined in advance or the considered stratifications not too many (overfit: e.g. plot x and y on a scatter plot and put the boundaries where the gain seems to be largest)

- Poststratification may also be thought of as a regression estimator (GREG), where each stratum-indicator is an explaining auxiliary variable.

- If you ”post-stratify” after several auxiliary variables using GREG you can omit the interaction terms. Three variables with four groups each, result in 10 explaining variables (9 regressors + intercept) instead of 64 strata (4*4*4).

- It is often enough to skip the interactions but not always. If you are allowed to, the risk for spurious variance estimation due to overfitting will decrease considerably

4.4.2 Post or Prestratify

- Usually post-stratification works almost as good as prior stratification if the sample is taken with proportional allocation. (Only a minute increase in variance occurs especially for large samples)

- But if optimal sampling fractions should differ between strata, that cannot be obtained afterwards by poststratification.

4.4.3 Raking

- Suppose that you have two possible post-stratification grounds, which both are promising. But that they cannot be used at the same time since there will be too many empty cells or you do not have population figures for the joint distribution of the variables
- Then you can solve this by raking or equivalently by iterative proportional fitting (IPF)
  - Poststratify after variable 1, getting new weights.
  - Now the margin of variable 1 is not correct but not after 2, Reweight after variable 2 so that its margins become correct
  - Now the margin of variable 2 is not correct but not after 1, Reweight after variable 1 so that its margins become correct
  - Repeat until convergence - usually less than three iterations.

Observed number/proportions in cells

Known marginals 25, 50, 25 and 33.3, 33.3, 33.3, resp

12	11	7	30
7	20	4	31
9	12	18	39
28	43	29	100

First half iteration 12*25/28 = 10.7 a.s.o.

10.7	12.8	6.0	29.5
6.3	23.3	3.4	33.0
8.0	14.0	15.5	37.5
25	50	25	100

12.1	14.4	6.8	33.3
6.3	23.5	3.5	33.3
7.1	12.4	13.8	33.3
25.6	50.4	24.0	100

Second half iteration 10.7*33.3/29.5 = 12.1

After three iterations we get the adjusted weights (%) (In fact with two decimal places)

11.9	14.4	7.1	33.3
6.2	23.5	3.7	33.3
6.9	12.2	14.2	33.3
25	50	25	100

Raking is post-stratifying after only marginals

Observed mean values in each stratum

24	32	47
41	43	58
30	40	51

Unstratified mean (12*24+11*32+47*7+41*7+ …)/100=40,16

Raked mean (12*11.9+11*14.4+47*7.1+ …)/100 = 39,74

Mean stratified after x is 39.97

Mean stratified after y is 39,85

4.5 Quota sampling

- **Stratified sampling.** Decide how many you want in each group. Draw the sample with probability sampling.
- **Quota sampling.** Decide how many you want in each group. Draw the sample one at a time with a sequential procedure e.g. from a list, queue. When a group is filled continue but reject every unit belonging to an already filled group. Usually not a probability sample. (But it is, if the list is in random order)
- Quota sampling is an object of detest (”hatobjekt”) for every orthodox sample statistician

- One may group after several marginals. E.g. reject a retired female if the quota of females or the quota of retired persons is full. The full sample may thus contain too many retired females if this is compensated by fewer retired males and fewer non-retired females.

- Quota sampling is often used in market research. E.g. the Swedish television audience survey attempts to be a quota sample. The reason is that too many persons refuse to comply.

- Some of you may have experienced this in telephone interviews, where the interviewer after asking about background variables says that you do not belong to the focus group

- Quota samples are usually analysed as if they were stratified samples
- Advantages: Much better than just taking the n first from the list. The result is balanced for at least the stratification variables.
- If the list is randomly ordered you may get a true stratified sample in this way even when your frame does not contain the auxiliary variable.
- Disadvantages: If the list is biased in some other way, the bias will remain. Very often the list does not even contain every object in the population. (E.g. Web panels)

Quota sampling follows only the first part of the statistical design rule:

Take into account what you know

(Balance, model, compensate, ...)

-

Randomise what you do not know