

Survey Analysis week 5

“Stratified and cluster sampling”

© Peter Lugtig

What have we done so far

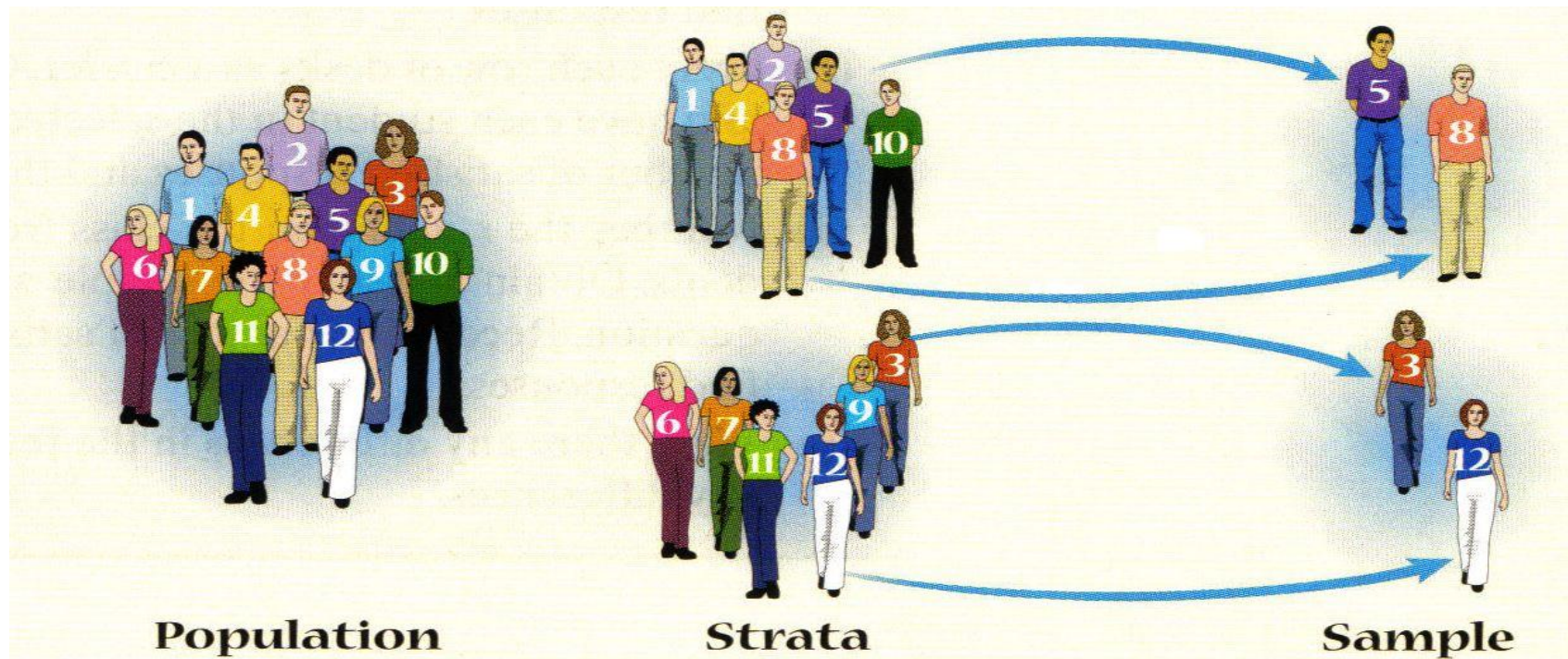
- Basics of survey error (TSE, bias, error)
- Mean, variance, standard error, proportions
 - SRS sampling – with and without replacement
 - $n-1$
 - F_{pc}
- R – survey and sampling packages
- Sample size calculations for mean
 - For simple random sample
 - Given CV, margin of error, Confidence interval
 - Given alpha, Beta, population variance

What will we do today

- Discuss THE week 3 in groups
 - Questions, issues?
- Lecture on stratification and clustering
 - Stratification and clustering: why?
 - Variance estimation, design effects
- 2 short class exercises
 - Set up svydesign objects

Stratified Random Sampling

- Population exists of pre-defined groups
- Use this information to optimize sampling design
- Idea – take a SRS from each group (*stratum*) and combine these for the final sample



Stratified sample design – why?

A population of companies in province of Utrecht

Sector/size (employees)	1-4	5-49	50+
Manufacturing	7000	1.800	335
Trade	14.000	3.000	250
Business	20.000	3.335	335
Government/education	1.600	282	79
Healthcare	1.100	625	224
Agriculture	682	118	3

Stratified sample design – why?

1. Increase precision in total (< s.e.)

2. Ensure precision in subgroups

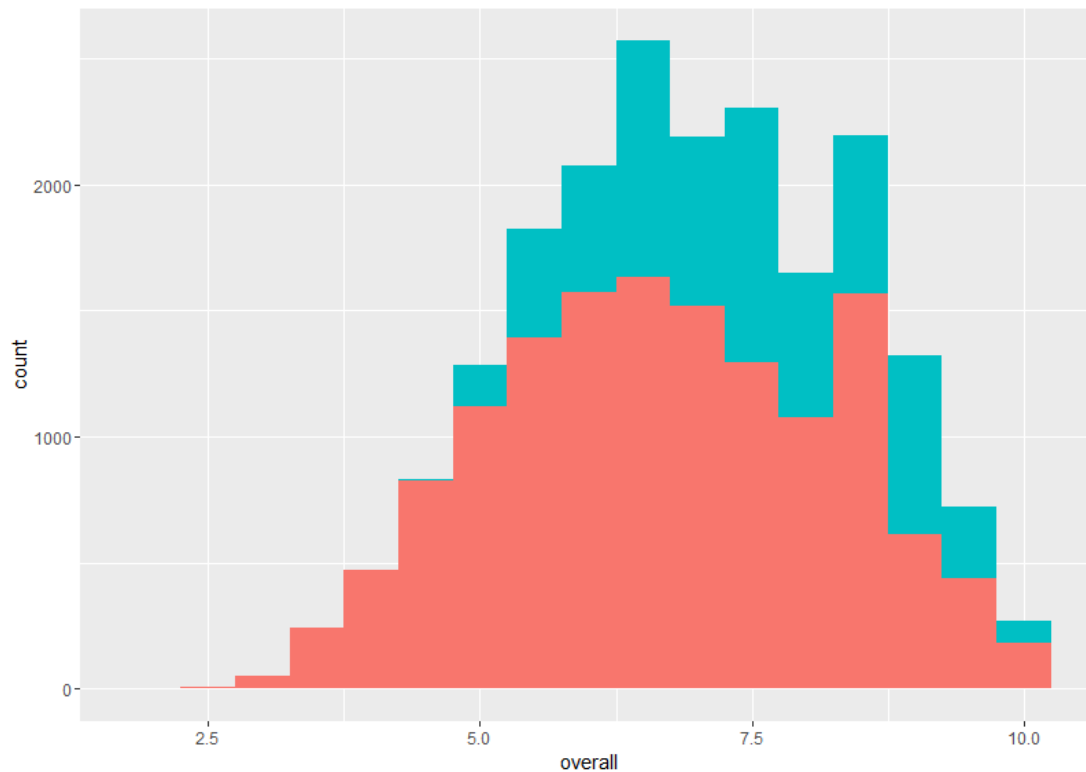
3. More practical when sampling frames are only available per subgroup

4. + to limit other surveys errors (later)

Sector/size (employees)	1-4	5-49	50+
Manufacturing	7000	1.800	335
Trade	14.000	3.000	250
Business	20.000	3.335	335
Government/education	1.600	282	79
Healthcare	1.100	625	224
Agriculture	682	118	3

An example: student grades at UU

- Total: 20000 students
 - 14000 BA, 6000 MA
- Interested in mean grade at UU



Why stratify on degree?

Mean BA: 6.69

Mean MA: 7.41

Var(BA): 1.49

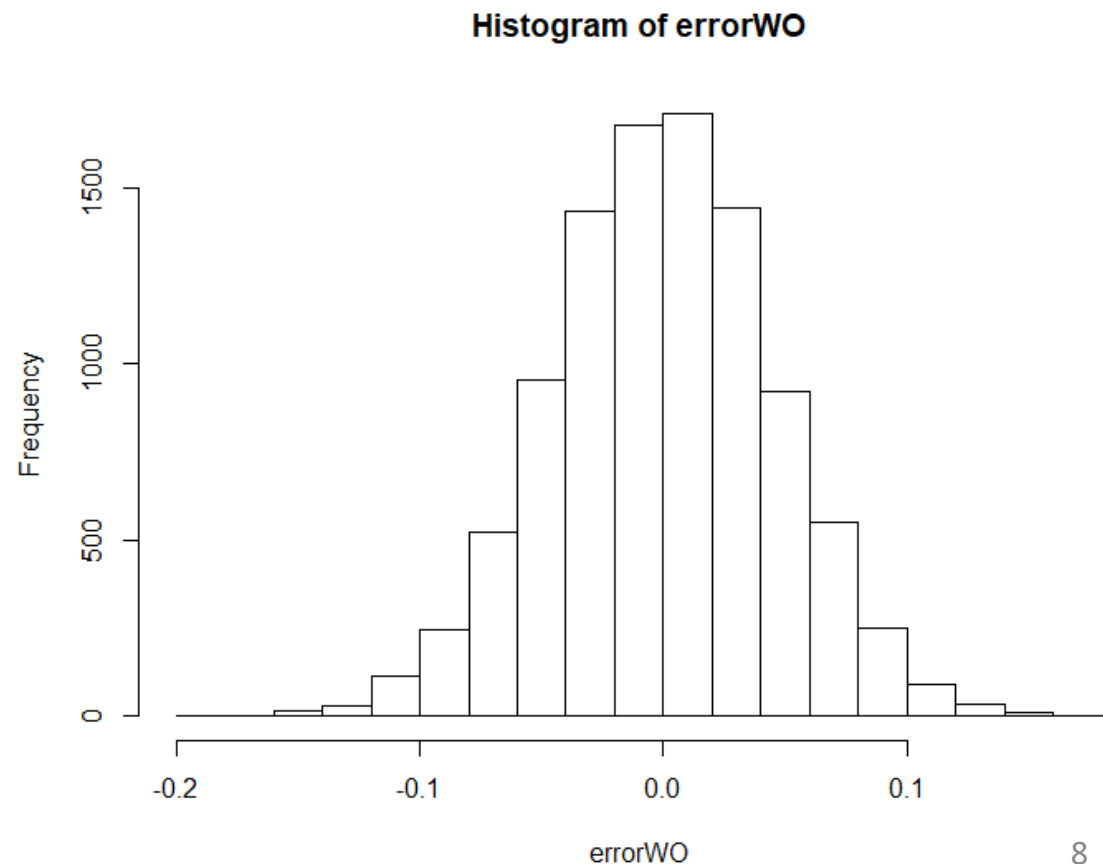
Var(MA): 2.36

Var(total): 2.20

Code of example on BB!

Simple Random sample

- Draw 10000 samples (replications)
 - $n=1000$



Simple Random sample

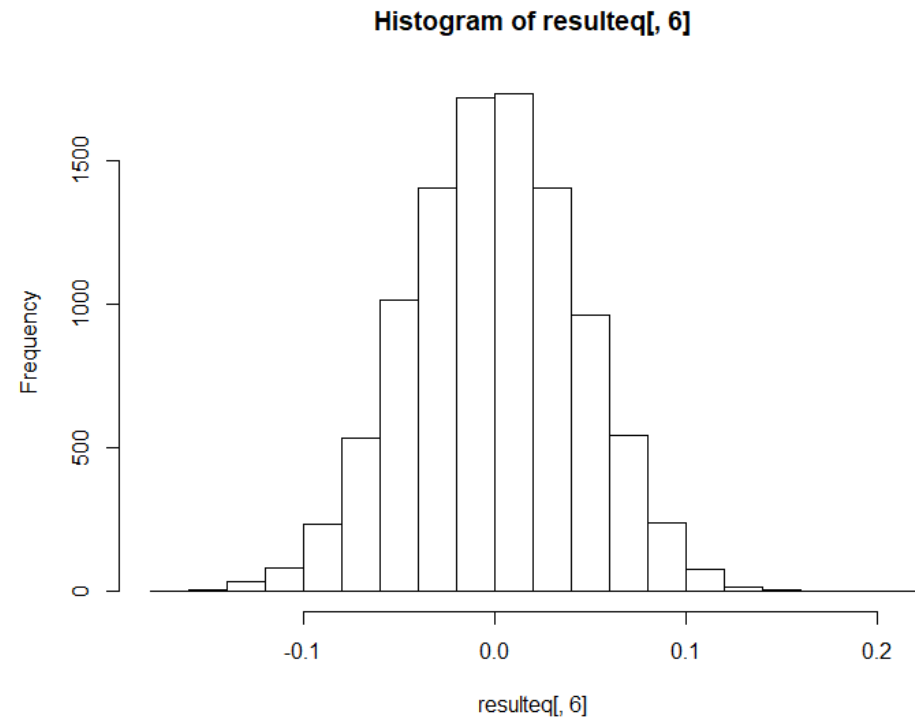
- Size=1000 out of 20000
 - Draw 10000 samples (replications)
 - **Roughly** 700 BA students, 300 Ma students

	Bias	Variance (s.e)
Simple random sample	-.026	.0021

- $MSE = -.026^2 + .0021 = .0027$
- How can we improve?

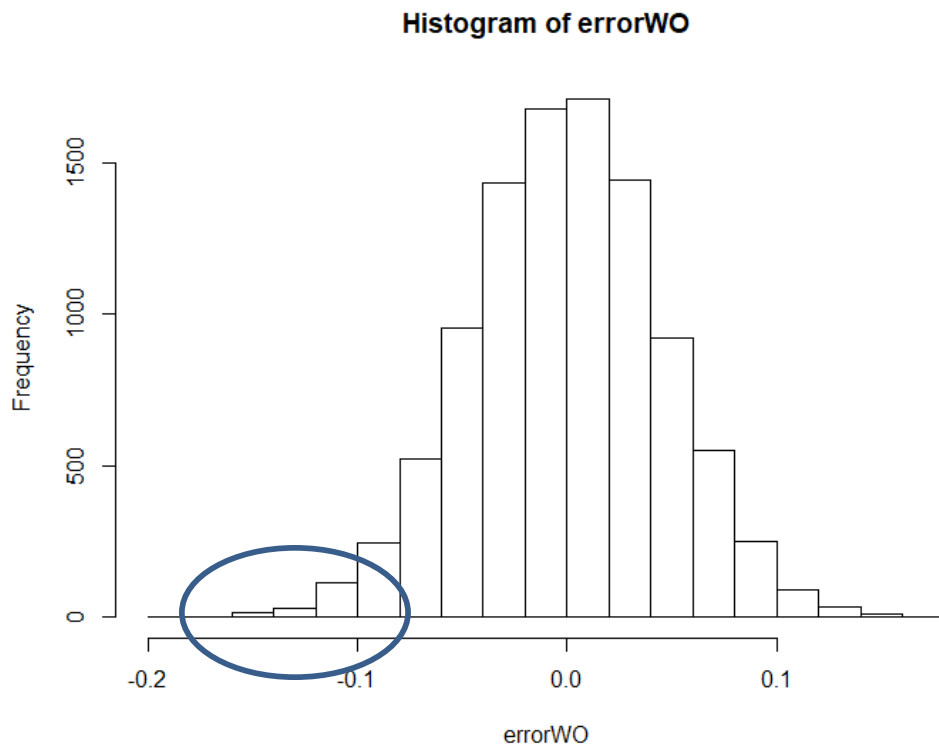
Stratification with equal probabilities

- Size=1000, 10000 replications
 - Exactly 700 BA students, 300 Ma students
 - Proportional to Size (PPS)
 - Error compared to population mean

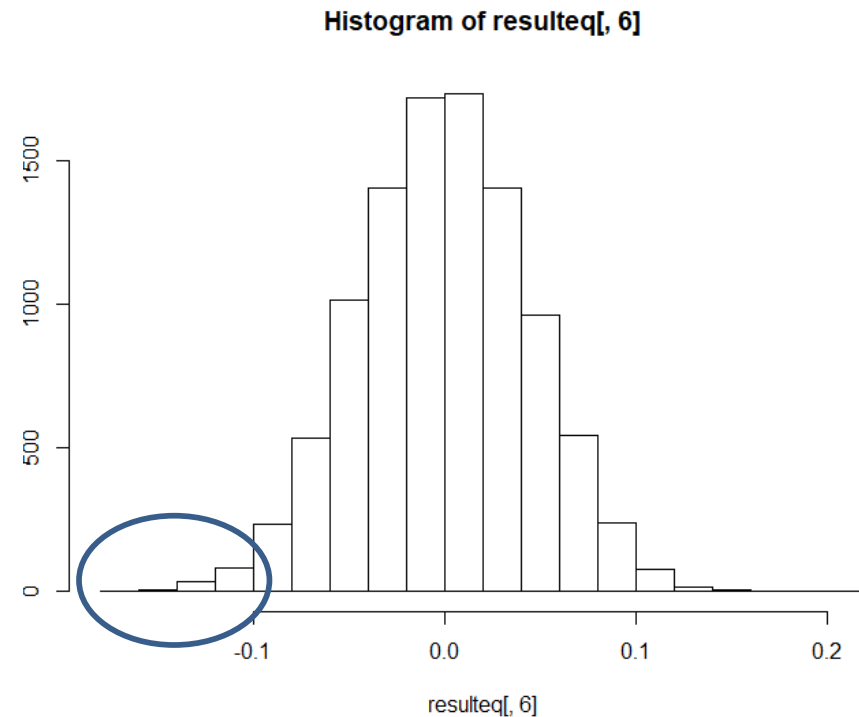


Comparison of standard errors

SRS



Stratification pps



Slightly fewer extreme samples: stratification reduces s.e.

Stratification with equal probabilities

- Size=10000
 - Exactly 700 BA students, 300 Ma students
 - Proportional to Size (PPS)

	Bias	Var(mean)	Var(mean) Bachelor	Var(mean) Master
Simple random sample	-,026	,0021		
Stratification equal probabilities	-,031	,001966	,0031	,0047


Formulas

- 1. Mean with SRS in every stratum (pps)

$$\bar{y}_h = \frac{1}{n_h} \sum y_{hj}$$

- 2. Combining means (pp2):

$$\bar{y}_{\text{str}} = \sum \frac{N_h}{N} \bar{y}_h$$



Size of stratum in
population

Formulas

- 1. Mean with SRS in every stratum (pps)

$$\bar{y}_h = \frac{1}{n_h} \sum y_{hj}$$

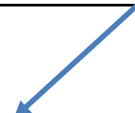
- 2. Combining means (pps):

$$\bar{y}_{\text{str}} = \sum \frac{N_h}{N} \bar{y}_h$$

- 3. Variance (pps):

$$\hat{V}(\bar{y}_{\text{str}}) = \sum \left(1 - \frac{n_h}{N_h}\right) \bar{y}_h \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}$$


Variance in
stratum



Fpc per stratum



(Size of stratum
in population)²



Design effect

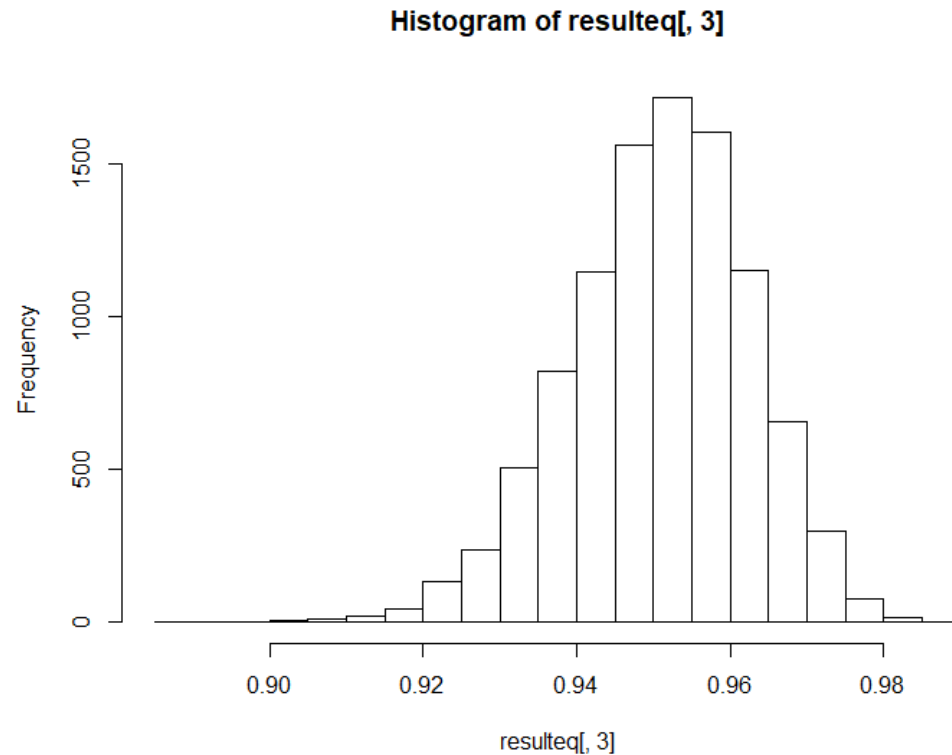
- If we have a measure of the bias in standard (SRS) estimates of the sampling variance this can be used to adjust the variances or standard errors from SRS
- $D_{\text{eff}} = \frac{\text{Variance under specific design}}{\text{Variance under SRS}}$
- D_{eff} = inflation factor for variance
 - Can be used for estimating effective sample size
 - $n_{\text{eff}} = N/D_{\text{eff}}$
- $D_{\text{eft}} = \sqrt{D_{\text{eff}}} =$ inflation for standard error

Design effect (3)

- $D_{\text{eff}} = .00197/.0021 = .9523$
 - $D_{\text{eff}} = .952^2$. We need **.906** as many respondents for same precision
 - $N_{\text{eff}} = 1000/.906 = 1103$
- Stratified sample: $D_{\text{eff}} < 1$
- Cluster sample: $D_{\text{eff}} > 1$
- R produces Design effect as well
 - For one sample against same sample under SRS
 - (Pooled) d_{eff} in R: **.951**
 - Rounding error in manual computation above

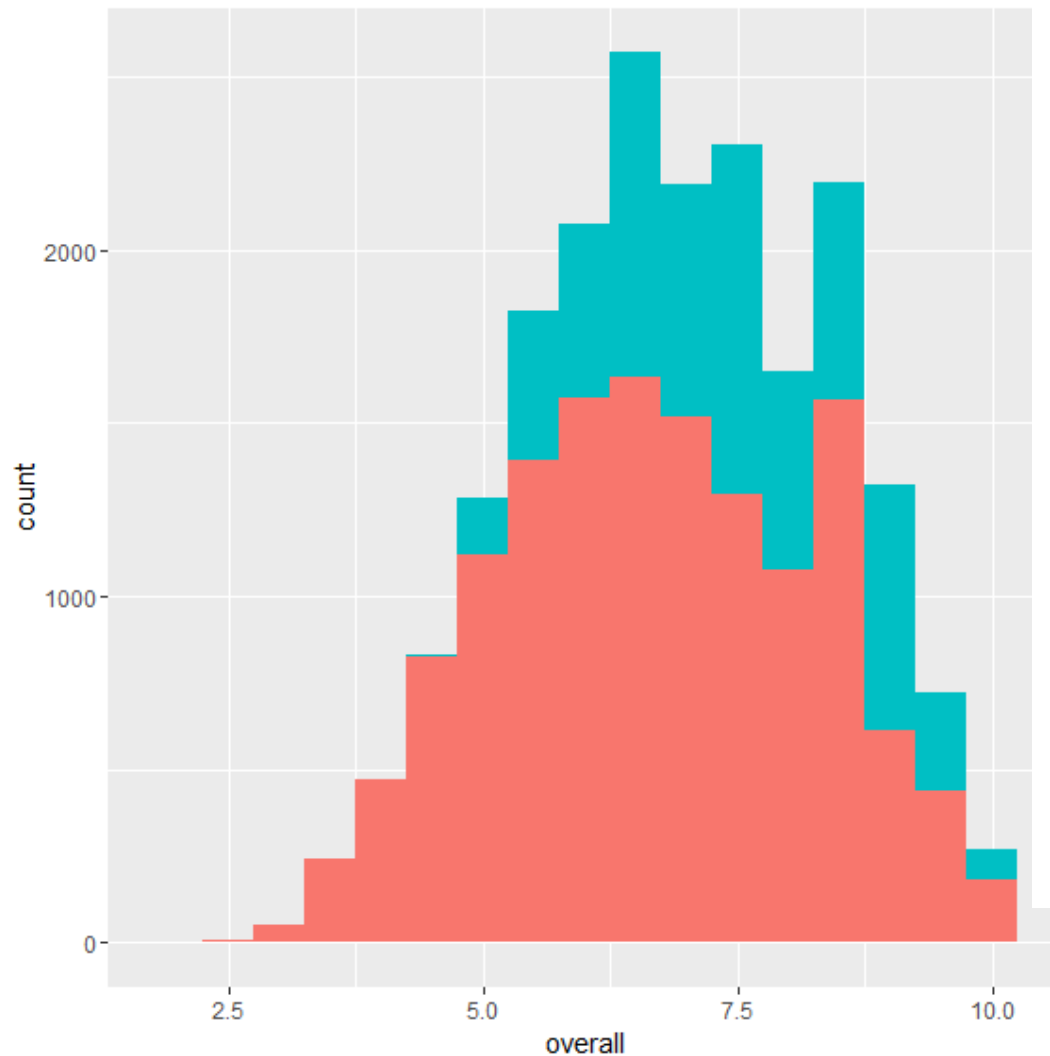
Design effect (3)

- R produces Design effect as well



– Note this is skewed (why?)

Unequal probabilities



- Why?
 - Increase precision for small group in population (Ma students)
 - Optimize the precision for the total mean grade?

Unequal probabilities

- Population: 14000 BA, 6000 MA
 - We can oversample MA students
 - Because there is more nonresponse, or we want increased precision in this group
 - Imagine we sample 500 students from each group
 - $\pi(\text{BA}) = 500/14000 = .035714$
 - $\pi(\text{MA}) = 500/6000 = .083333$

Stratification with unequal probabilities

- Size=1000
 - 500 BA students, 500 Ma students

	Bias	Var(mean)	Var(mean) Bachelor	Var(mean) Master
Simple random sample	-,026	,0021		
Stratification PPS	-,031	,001966	,0031	,0047
500 in each	,023	,0025	,0047	,0027

- #\$\$%&: wasn't stratified sampling supposed to decrease the variance?

Optimal allocation

- How large should samples within strata be so that s.e. for population is minimized?
- Neyman allocation:
 - N_h = sample size stratum
 - S_h = standard deviation in stratum
- Needed?
 - N_l = Population size
 - S_l = Standard deviation in population (!)
- Often costs are included (more complicated)

$$n_h, \text{Neyman} = \frac{N_h S_h}{\sum N_l S_l}$$

Optimal allocation: example

- Neyman allocation:
– N_h = sample size stratum
– S_h = standard deviation in stratum
$$n_h, \text{Neyman} = \frac{N_h S_h}{\sum N_l S_l}$$
- Population variance: 2,20
- Stratum variances : 2,36 and 1,49
– See slide 7
- Stratum standard deviations = $\sqrt{2,36}$ and $\sqrt{1,49}$

Optimal allocation: example

- Neyman allocation: $n_h, \text{Neyman} = \frac{N_h S_h}{\sum N_l S_l} n$
 - N_h = sample size stratum
 - S_h = standard deviation in stratum

$$N_{MA} = \frac{(6000 * \sqrt{1.49})}{(6000 * \sqrt{1.49}) + (20000 * \sqrt{2.20})}$$

$$= \frac{7324}{28831} = 0.254 * 1000$$

- $N_{BA} = \frac{(6000 * \sqrt{1.49})}{(6000 * \sqrt{1.49}) + (20000 * \sqrt{2.20})}$

$$= \frac{21507}{28831} = 0.746 * 1000$$

Optimal allocation to strata

- Size=1000
 - 746 BA students, 254 Ma students

	Bias	Var(mean)	Var(mean) Bachelor	Var(mean) Master
Simple random sample	-,026	,0021		
Stratification equal probabilities	-,031	,001966	,0031	,0047
500 in each	,023	,0025	,0047	,0027
Neyman	,04	,001962	,0029	,0055

Optimal allocation to strata

- Unequal selection probabilities
 - 746 BA students, 254 MA students
 - $D_{\text{eff}} = .01962 / .0021 = .934$
 - $n_{\text{eff}} = 1000 / .93^2 = 1145$
- In words: we optimize stratification when:
 - The stratum accounts for a large part of the population
 - BA = 0.7, MA = 0.3
 - The variance within the stratum is large; we sample more heavily to compensate
 - variance BA: 2.36, Variance MA: 1.49
 - -> leads to **746 BA students**

2 ways to deal with unequal selection

- 1. We can combine means from n_h strata
 - If we know the population sizes (see earlier slides)
- 2. We can use sampling weights
 - With SRS, sampling weights are equal for all i
 - In Neyman allocation example:

- $\pi_{ba} = \frac{746}{14000} = .053$. Sampling weight $w_{ba} = \frac{1}{.053} = 18.76$

- $\pi_{MA} = \frac{254}{6000} = .042$. Sampling weight $w_{ma} = \frac{1}{.042} = 23.61$

- $\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$

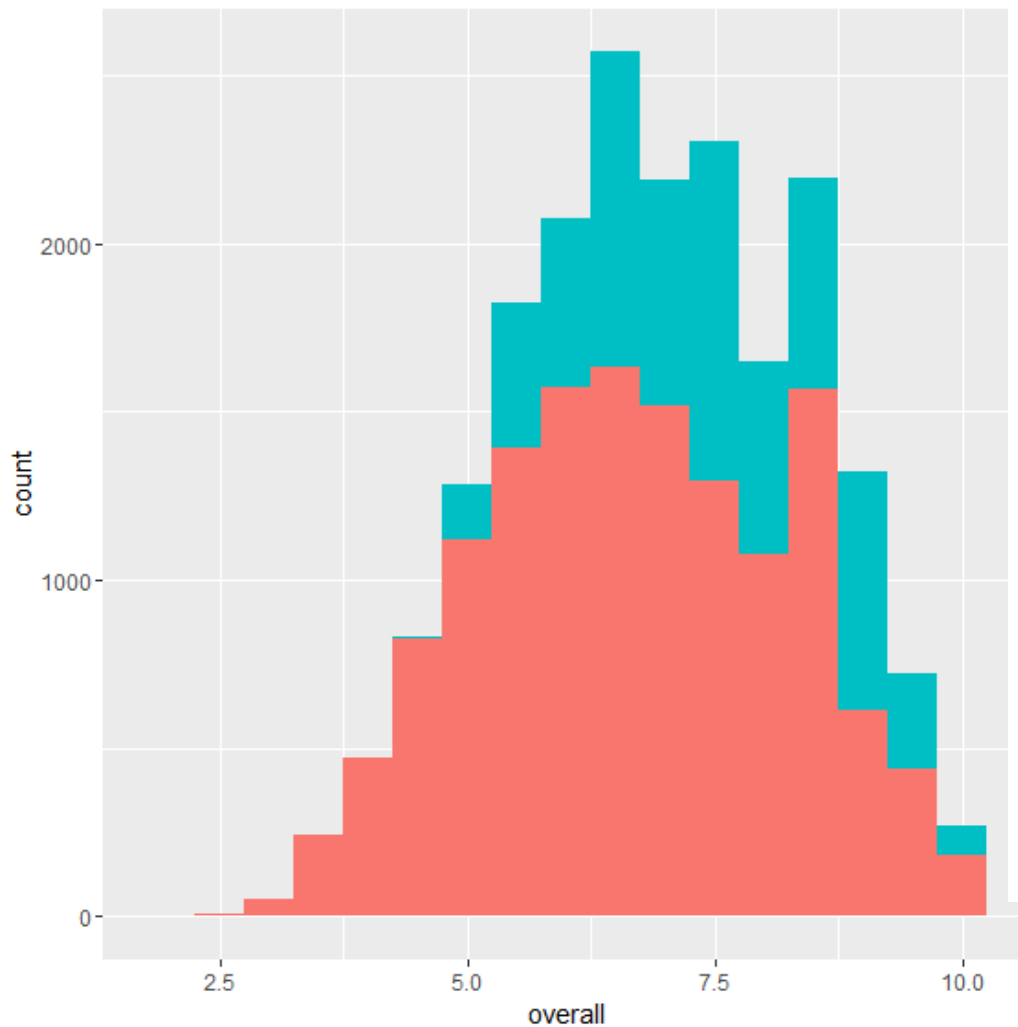
Bringing Nonresponse in

(more in week 9-11)

- We often oversample from specific groups
 - Nonresponse
 - Imagine:
 - Response rate of 50% among BA students
 - And 20% among Ma students
 - Achieved sample = 399 BA , 40 MA students
 - We can correct by using Nonresponse weights, but this is inefficient

Should we actually stratify on Ba/MA?

- What variable to stratify on?
 - BA/MA, Faculties, Programme
 - Social science, humanities, etc.
 - Gender, Age, living in Utrecht
 - Member of student union
 - Etc.
- We can stratify on multiple variables
- Survey mode:
 - E-mail: stratification easy
 - Face-to-face: costs!
 - Cluster sampling?

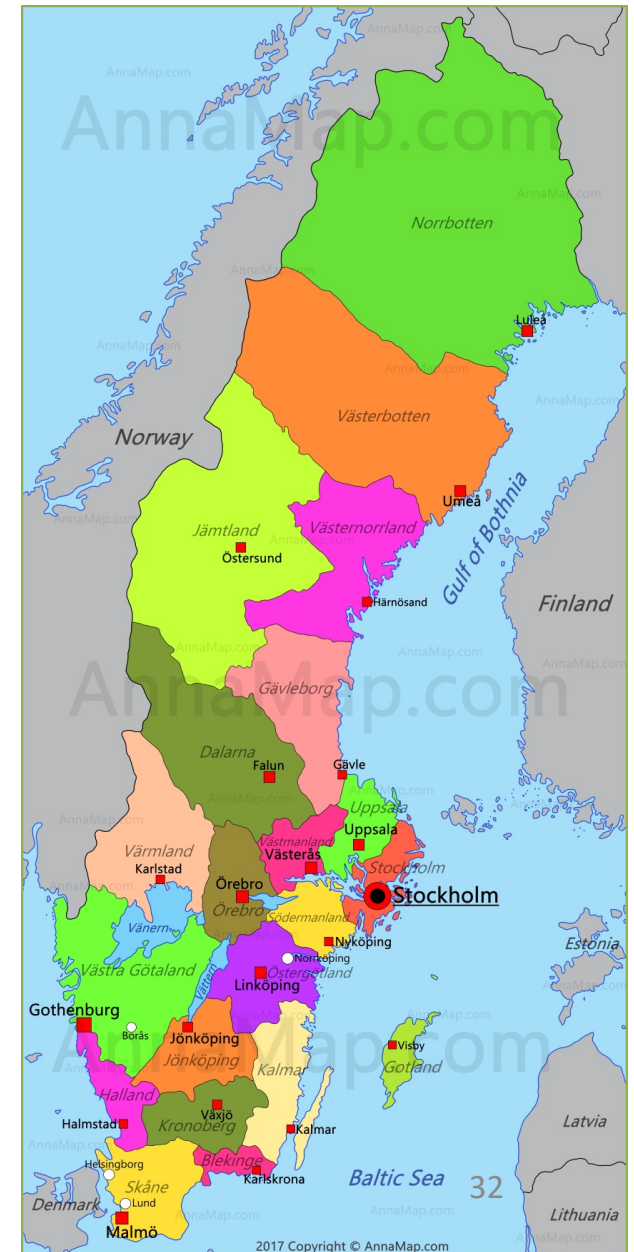


Class exercise 1

- Draw and analyse a stratified sample from a simplified dataset
 - Understand the basic R code for specifying **stratified** survey design object
 - Which variable to stratify on?
- See Class exercise document (Blackboard)

Adding in clusters

- What if we don't use e-mail?
 - Cost per case: € 0,50
- But face-to-face
 - Costs per programme: € 50,-
 - Costs per case: €10,-
- Households, businesses
- Hospitals, schools
- Towns!

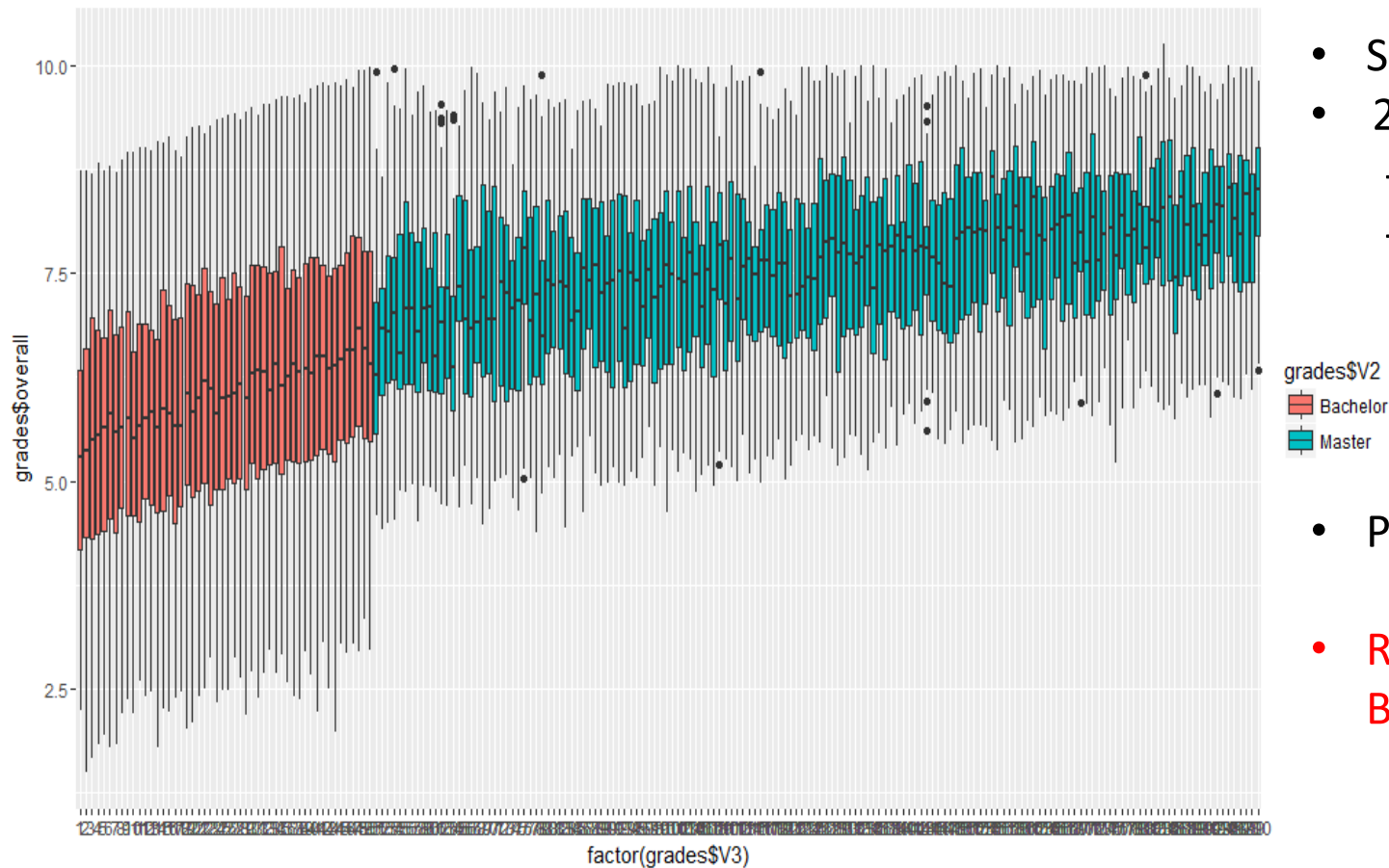


Terminology

- Primary sampling units (psu)
- Clusters
- Secondary sampling units (ssu)
- Intra-Cluster(Class) Correlation Coefficient (ICC)
One-stage sample: whole cluster interviewed
- Two-stage sample: further sampling within clusters
- Ratio estimation (week 8)

Example – 150 programmes (Ba/MA)

simulated data



- Student grades (y)
- 200 programmes (x)
 - 50 BA, n=280 each
 - 150 MA, n=40 each

- Population mean: 6.52

- R-code is available on Blackboard

Cluster sampling – why and how

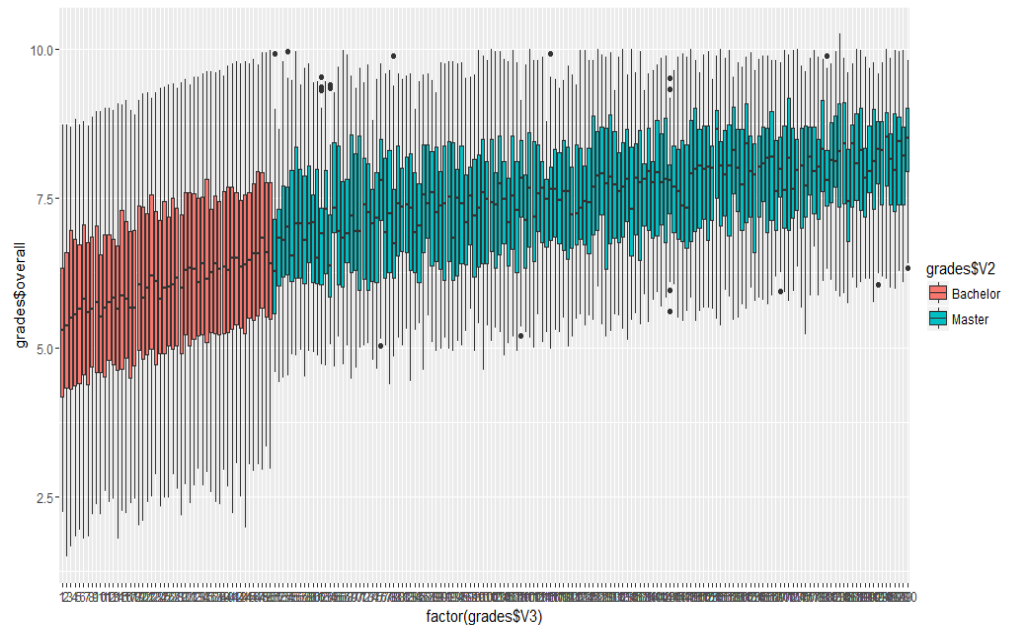
- How?
 - One-stage: ask everyone in cluster
 - Educational surveys: ask every pupil at school (or class)
 - Two stage: do a sample within every cluster
 - Multi-stage: e.g.
 - 1. stratify on income in neighborhood
 - 2. neighbourhoods (PSU)
 - 3. households (SSU)
 - 4. individual in hh

Why **not** do a cluster sample

- It is inefficient from a statistical perspective
 - ICC: measure of relative variance between clusters

$$ICC = \frac{s_b^2}{(s_b^2 + s_w^2)}$$

- S_b^2 : Variance between clusters
- S_w^2 : Variance within clusters



One-stage cluster sampling

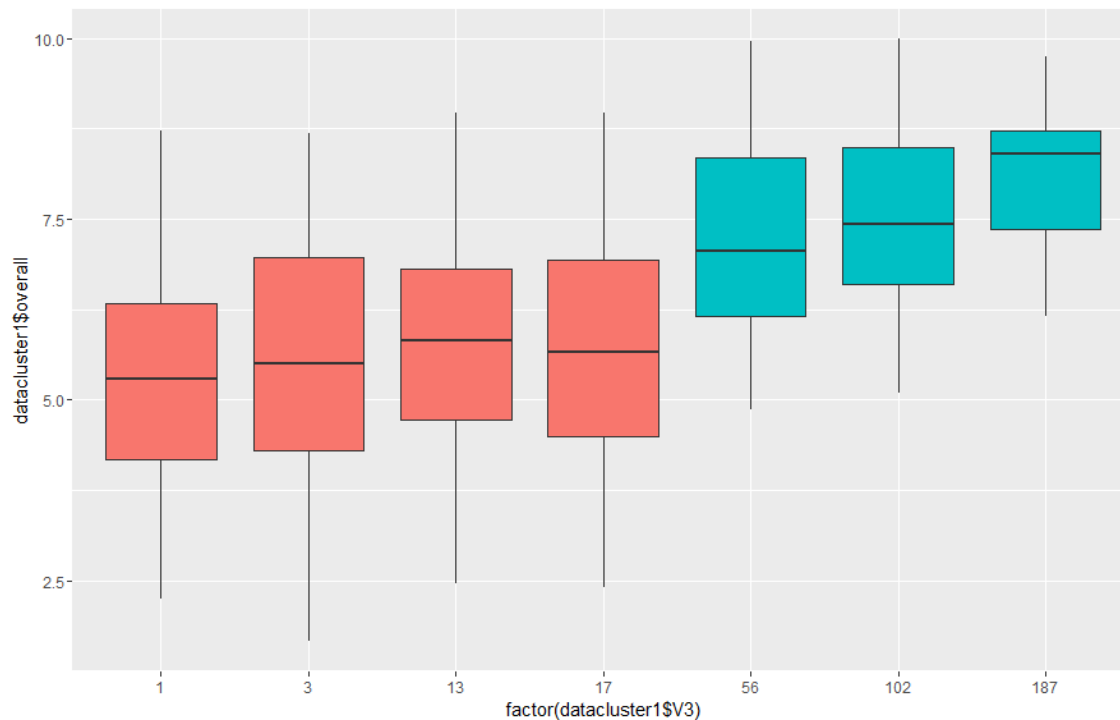
Clusters of equal size:

- Selection probabilities equal for all $i \rightarrow$ SRS of clusters
- Rarely the case

Clusters of unequal size:

1. draw clusters with unequal selection probabilities
 - Proportional to Size (PPS)
 - Self-adjusting sample
 2. Or, draw SRS of clusters
 - Use weights to correct for unequal probabilities
- Example: draw 7 clusters out of 150 with SRS
 - Average cluster size = 133, expected sample size = 933

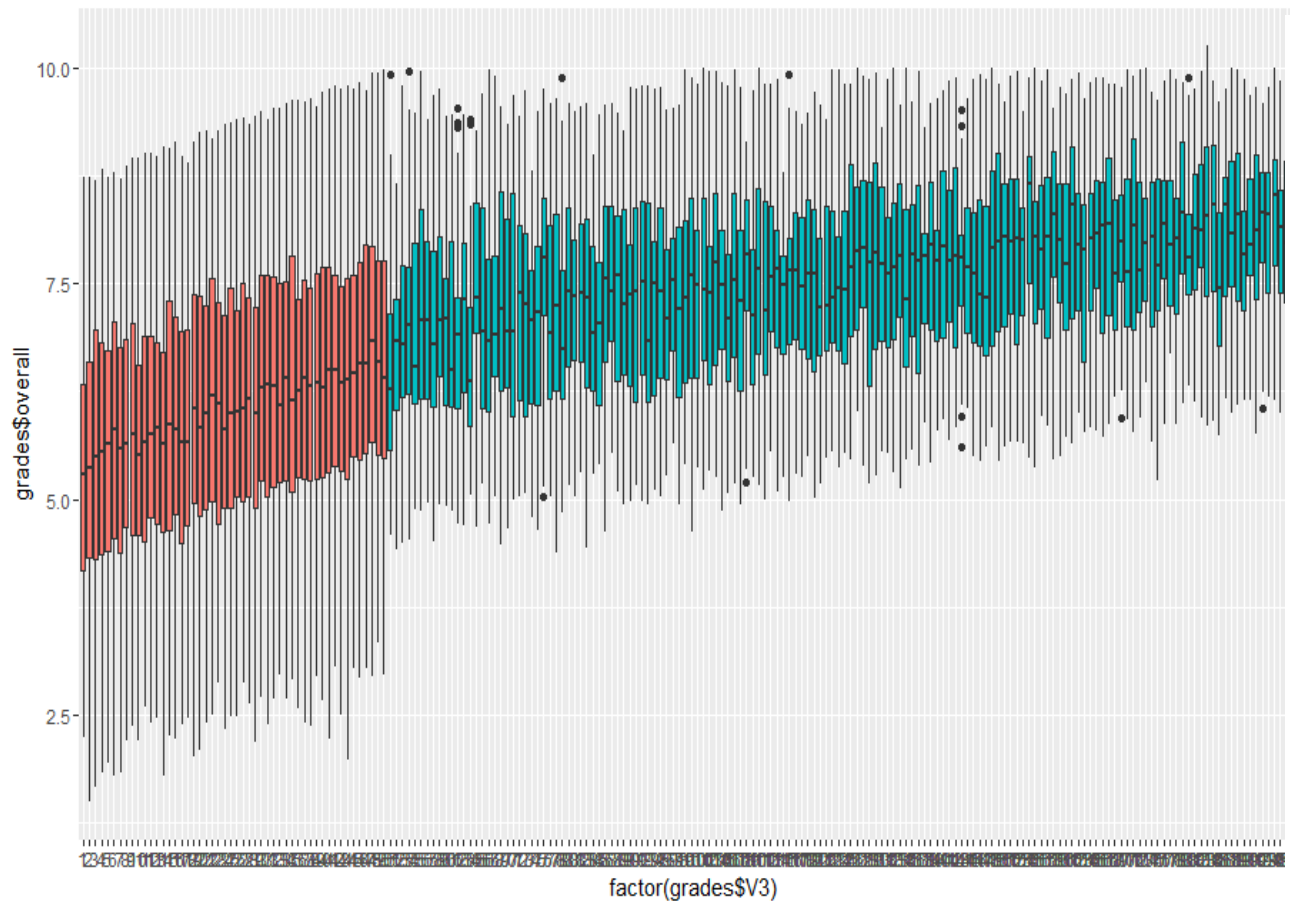
1 draw for One-stage cluster sample



Results from R (1 sample)

- $n=1240$
- mean: **5.79** (!) (population = 6.52)
- s.e. SRS .045624
- s.e. cluster .17397
- $D_{\text{eff}} = .17397 / .045624 = \mathbf{3.81}$
- Why are results so bad?

Problems in 1-stage cluster sampling



In our example:

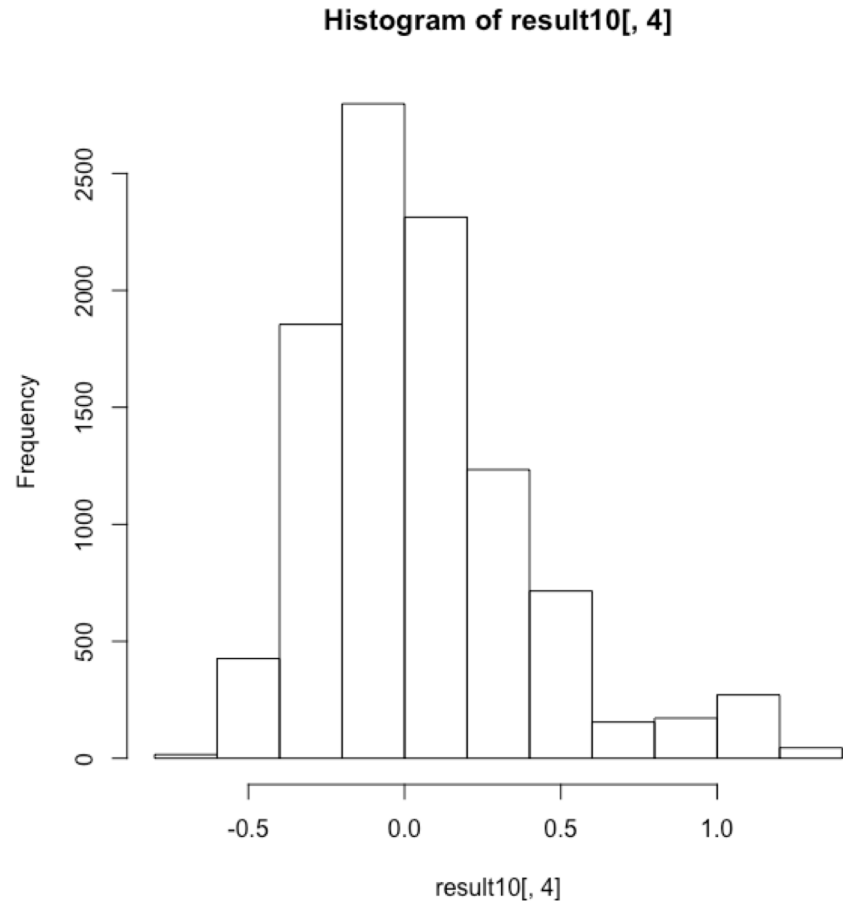
- Bad luck: Clusters selected:
 - 1,3,13,39,56,102,187
- Small no. of clusters sampled
- Large difference between cluster means (large ICC)
 - ICC population:
 - For Ba/MA: .18
 - For Programme: .21

In general:

- hard to control sample size
- Design effect could be large if variance differs across clusters

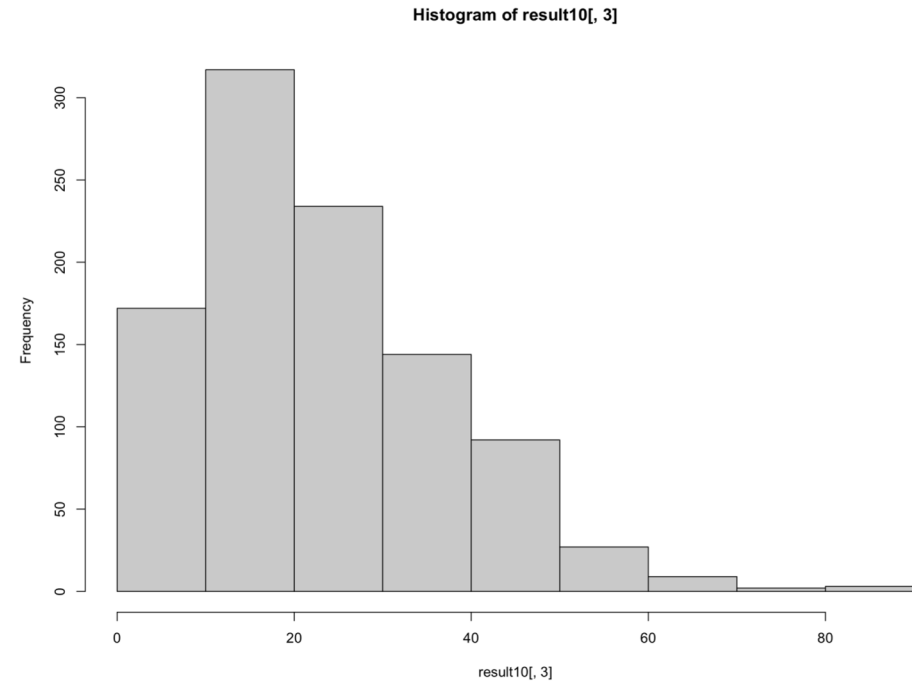
Were we unlucky? a simulation

- 10000 replications
- 10 clusters
 - $n=1000$ on average
 - but large variation
- Bias
 - ,0005 on average
 - Long tail
 - when we sample only MA or BA



a simulation for 1-stage cluster (2)

- 10000 replications
- 10 clusters
 - n=1000 on average
 - but large variation
- Design effect (D_{eff}):
 - 23 (!) on average
 - We would need sample of 23.000!



Better: A 2-stage cluster sample

- Double the clusters, sample half
 - 1. Sample 20 clusters (instead of 10) using SRS
 - 2. Sample $\frac{1}{2}$ of i within every cluster > **expected sample size= 1000**
 - Or do 50 clusters and sample $\frac{1}{5}$ of individuals, etc.
- Optimal allocation given ICC and costs
 - But need good estimates for:
 - Costs?
 - population total? , cluster sizes? (\sim fpc),
 - Often: use of “pseudoclusters”

Two-stage cluster simulation

- 10000 replications
- 20/50 clusters
 - Second stage sample $\frac{1}{2}$ or $\frac{1}{5}$
 - $n=1000$ on average (but less variation)
- Bias
 - 20 clusters: $-.0034$
 - 50 clusters: $-.0322$

Two-stage cluster simulation

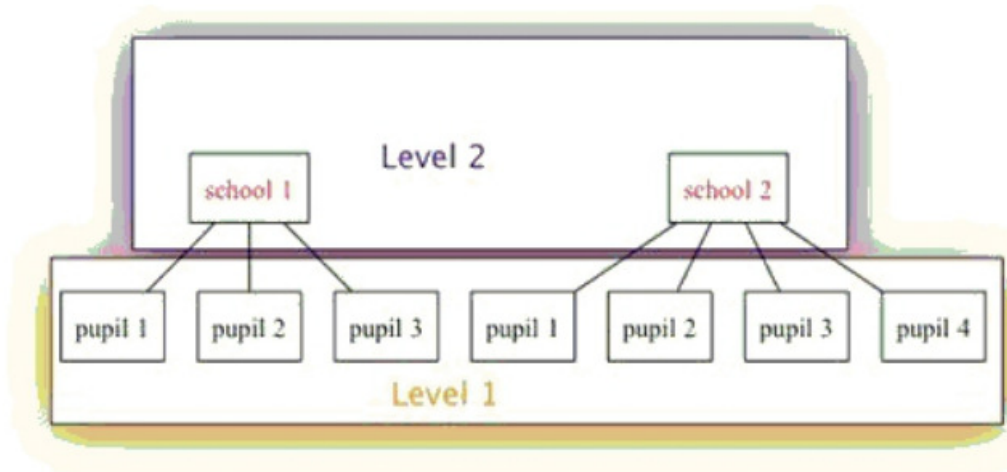
- 10000 replications
- 20/50 clusters
 - Second stage sample $\frac{1}{2}$ or $\frac{1}{5}$
 - $n=1000$ on average (but less variation)
- Design effect (deft)
 - 20 clusters: 16
 - 50 clusters: 6.8

Two-stage cluster estimation

- Uncertainty now at 2 levels
 - Sampling of clusters
 - Sampling of individuals in clusters
- Ok for Means, regression coefficients
 - Imagine we do SRS for selection of clusters and individuals
- Complex for variances
 - Rely on R for variance estimation

Analysis of cluster samples

1. Use the survey package
2. Do multilevel analysis (in semester 2)



Equation for Level 1

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

Equation for Level 2

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

3. Economists use Huber-White “robust” standard errors
 1. Pretend your dataset is SRS,
 2. estimate d_{eft} and
 3. multiply your s.e. with defl

Class exercise 2

- Analyse a cluster sample for the “boys” dataset
- See Class exercise document (Blackboard)
- If unfinished, finish at home before practical next week

Next week:

- Finish: Class exercises – different sampling designs
 - discussion about problems in class next week.
- Take Home Exercise:
 - Obtain the data: you often need to register at the data archive
- Next week: practical (in class time)
 - Mix SRS, stratified and cluster sampling
 - Design weights
 - Horvitz-Thompson estimator
 - Multistage sampling
 - ...