Survey Data Analysis
week 6
“R practical – Combinations of stratification and clustering”

Today

- Discuss take home exercise
  - Your adopted survey
  - Questions: what do you encounter?
- Short lecture
  - Survey design:
    {population, question, frames} -> modes
  - How to stratify?
  - How to cluster?
- Set of class exercises

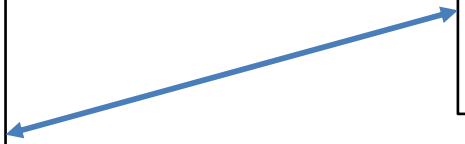
Which mode do I want to use?

What is my population?

- What modes are acceptable?

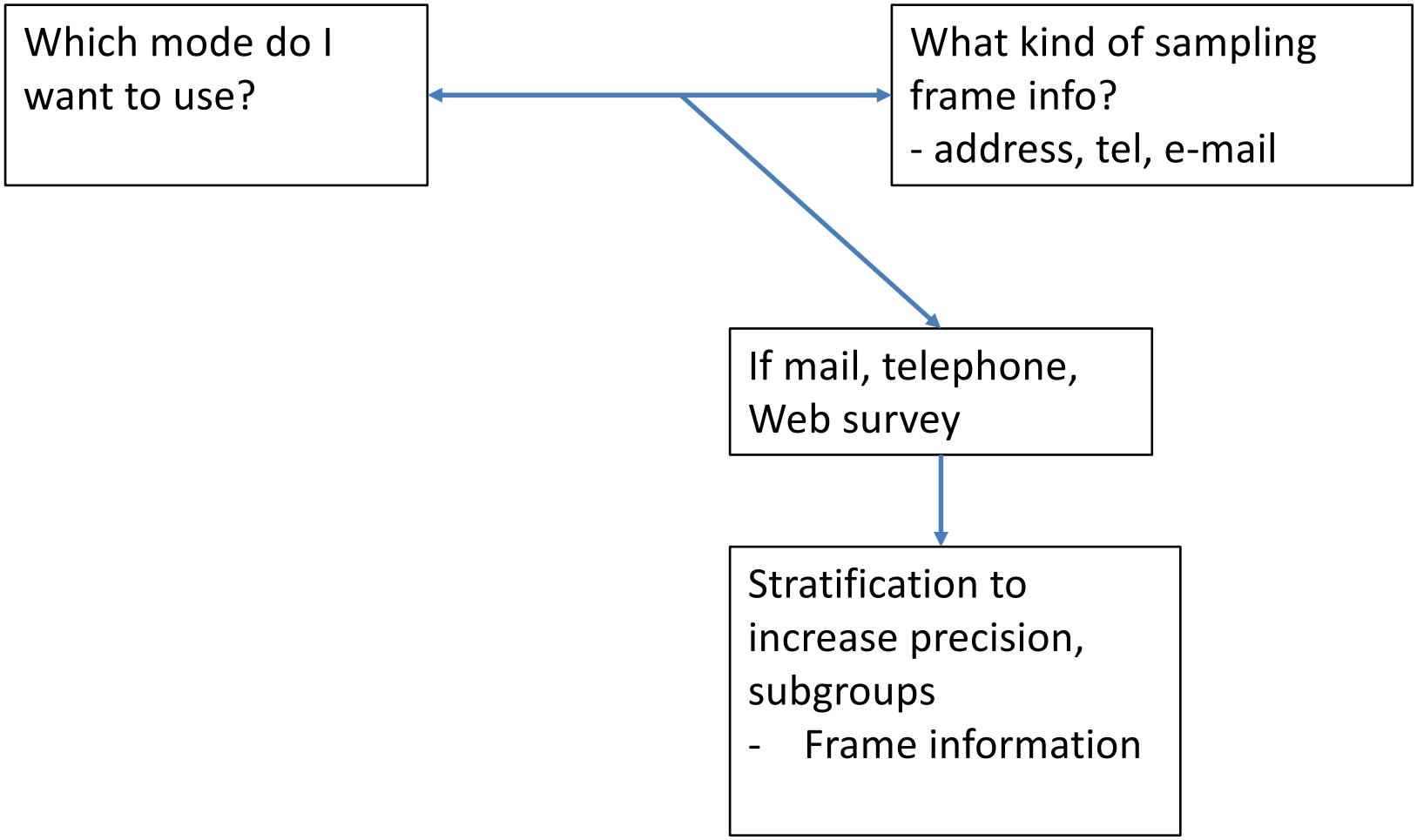
What is my question?

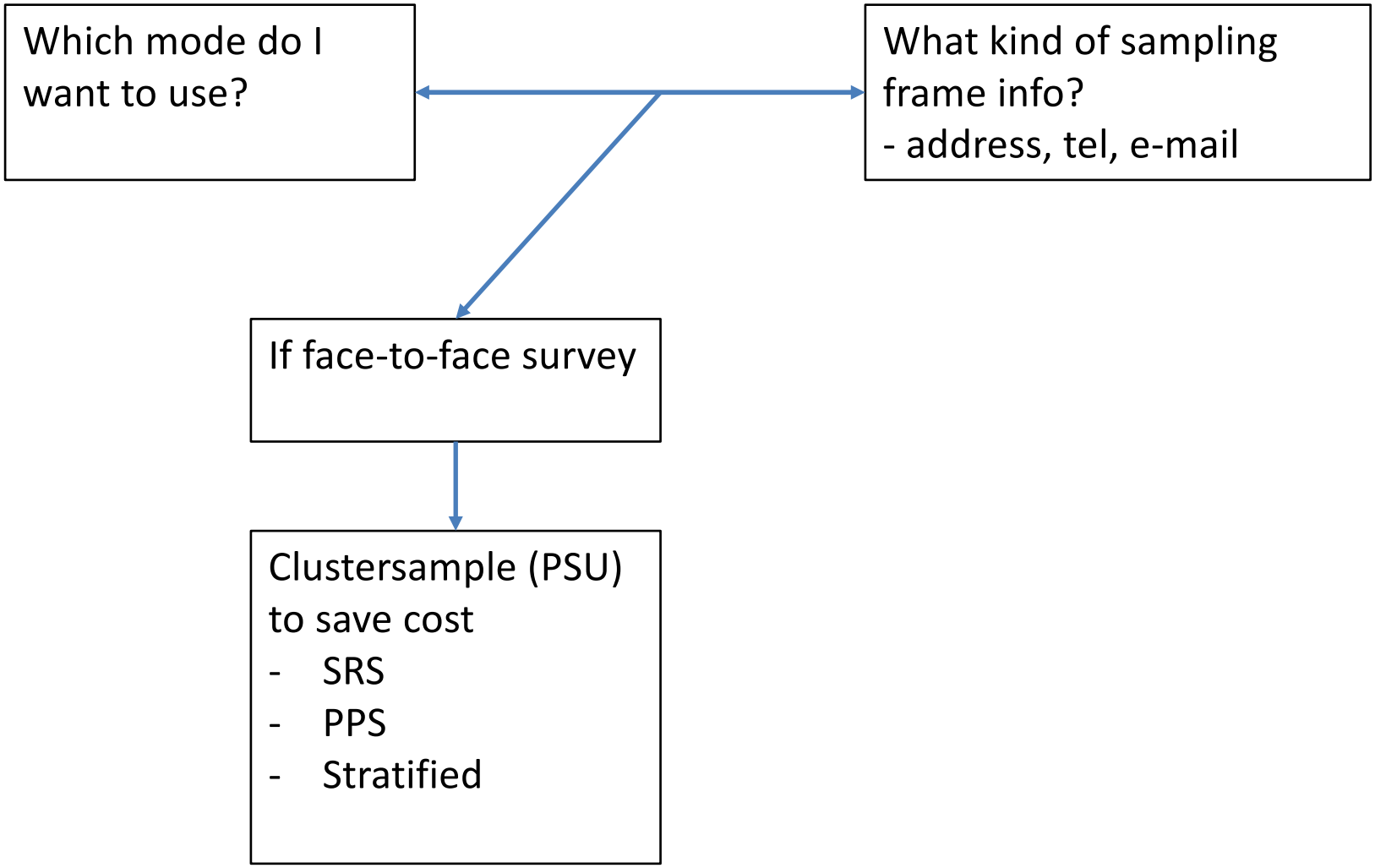
- Measurement error

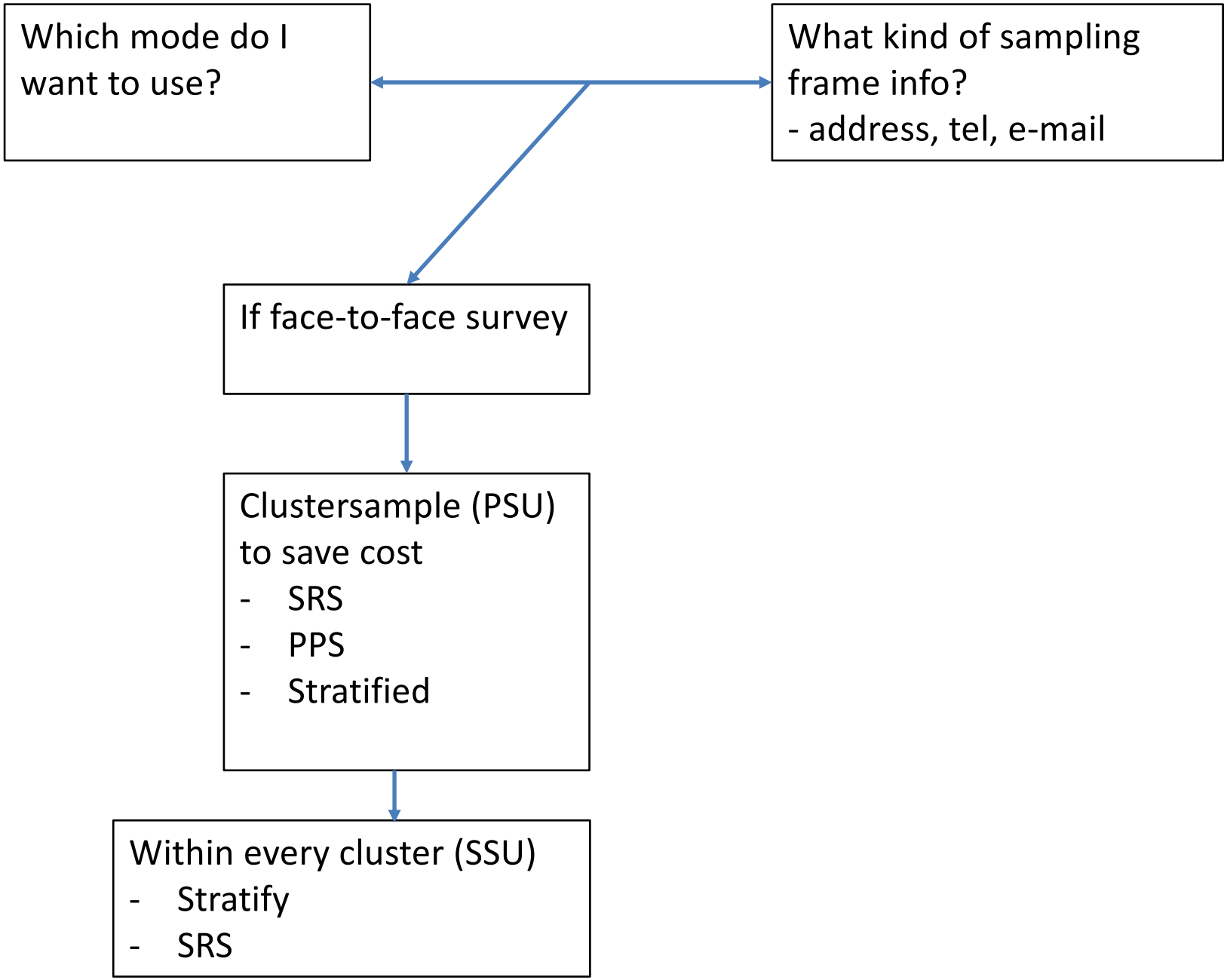


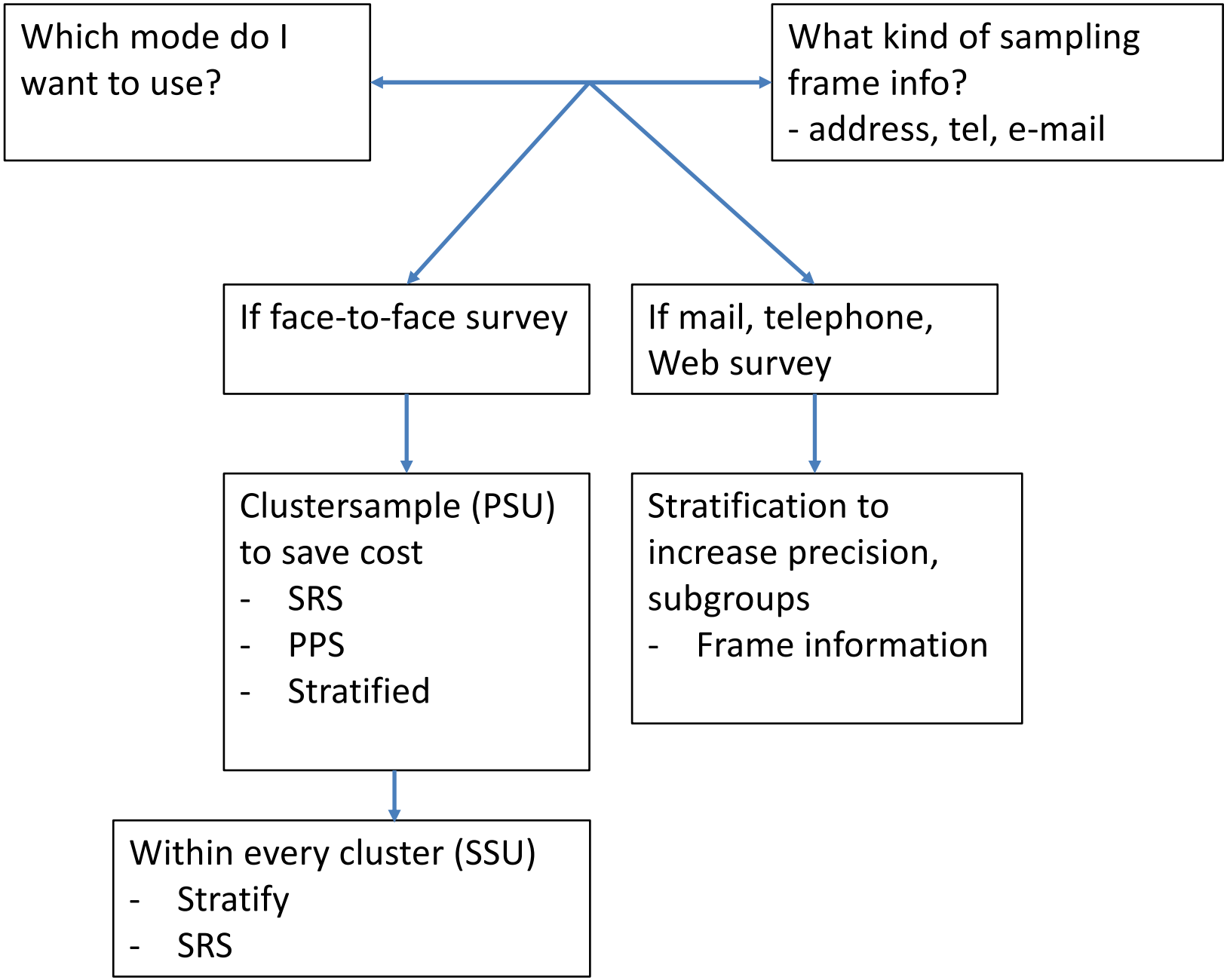
What kind of sampling frame info?

- address, tel, e-mail


Which mode do I want to use?
What kind of sampling frame info?
- address, tel, e-mail
If mail, telephone, Web survey
Stratification to increase precision, subgroups
- Frame information


Which mode do I want to use?
What kind of sampling frame info?
- address, tel, e-mail
If face-to-face survey
Clustersample (PSU) to save cost
- SRS
- PPS
- Stratified


Which mode do I want to use?
What kind of sampling frame info?
- address, tel, e-mail
If face-to-face survey
Clustersample (PSU) to save cost
- SRS
- PPS
- Stratified
Within every cluster (SSU)
- Stratify
- SRS


Which mode do I want to use?
What kind of sampling frame info?
- address, tel, e-mail
If face-to-face survey
If mail, telephone, Web survey
Clustersample (PSU) to save cost
- SRS
- PPS
- Stratified
Stratification to increase precision, subgroups
- Frame information
Within every cluster (SSU)
- Stratify
- SRS

Class exercises

1. Other statistics
2. Hurvitz-Thompson estimator
   - Design weights
   - Inclusion robabilities
3. Stratified cluster samples

Extra slides

Not discussed in class, but in case you want to know the end of the story of the "student" sample…

Horvitz-Thompson estimation

- We discussed SRS, stratified and cluster sampling
  - With and without replacement
  - Equal + unequal probabilities
  - All with slighlty different formulas
- Horvitz and Thompson (1952) designed a general framework for inference for random (probability surveys)

  - For mean: $\overline{y}_w = \frac{\sum w_i y_i}{\sum w_i}$

HT-estimation – a unifying framework…

HT-estimation works for all design-based sampling methods

- SRS equal probabilities: $\boldsymbol{\pi}_i$ = equal
- Stratified: $\boldsymbol{\pi}_i$ depends on strata selection
- One-stage cluster: $\boldsymbol{\pi}_i$ depends on cluster selection
- Two-stage (and more complex): cluster and within-cluster

All you need is $\boldsymbol{\pi}_i$, for every individual on your sampling frame

Our recurring example

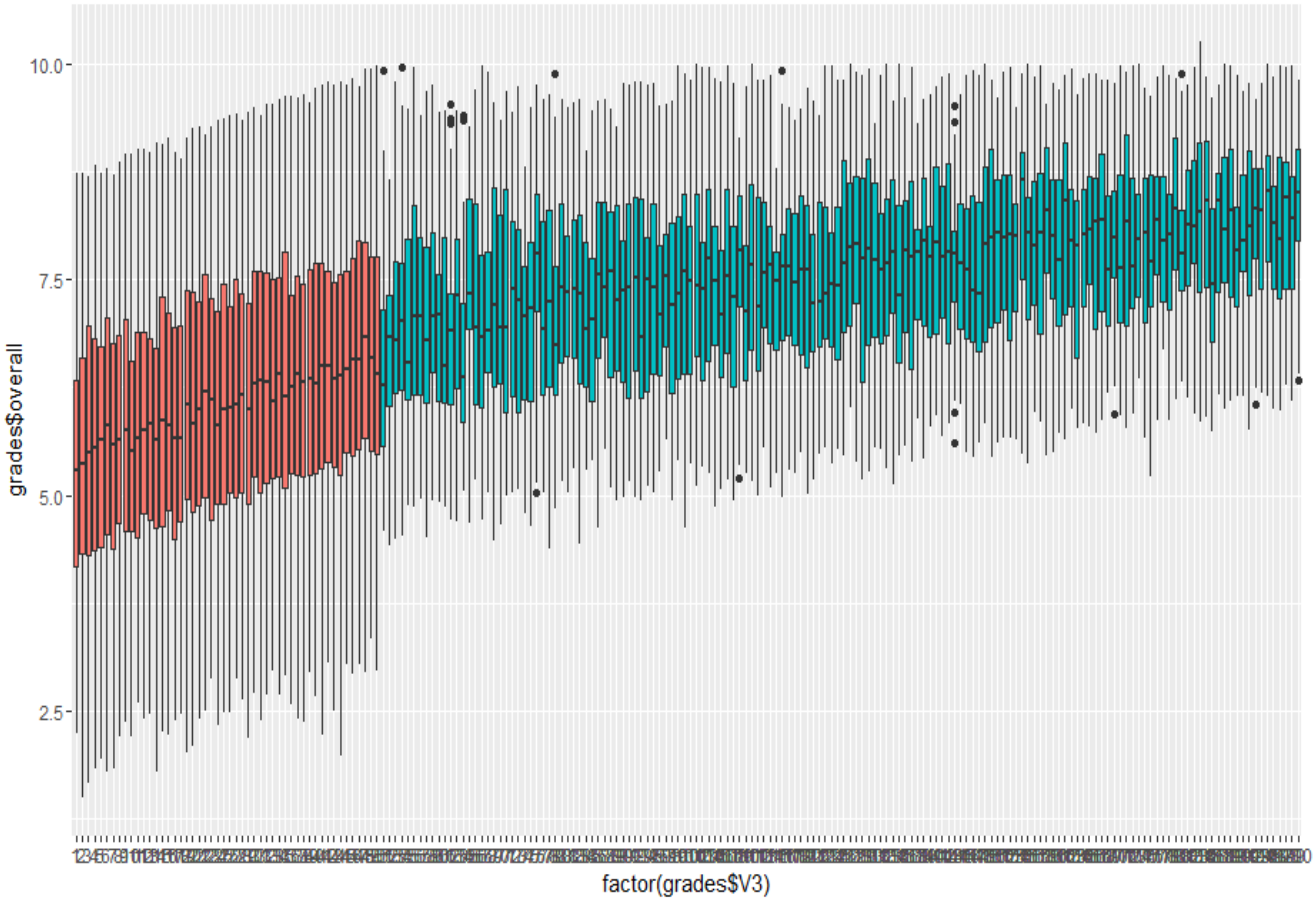
- We would like to do a survey among all students at Utrecht University
  - Population = 20.000
  - RQ: Interested in differences in **grades** and **student happiness** between programmes
  - approx. 49 BA programmes and 150 MA programmes
  - Limited budget (cannot do census) for about n=1000

- This week:

What if we combine clustering and stratification?

Example – 150 programmes (Ba/MA)

simulated data

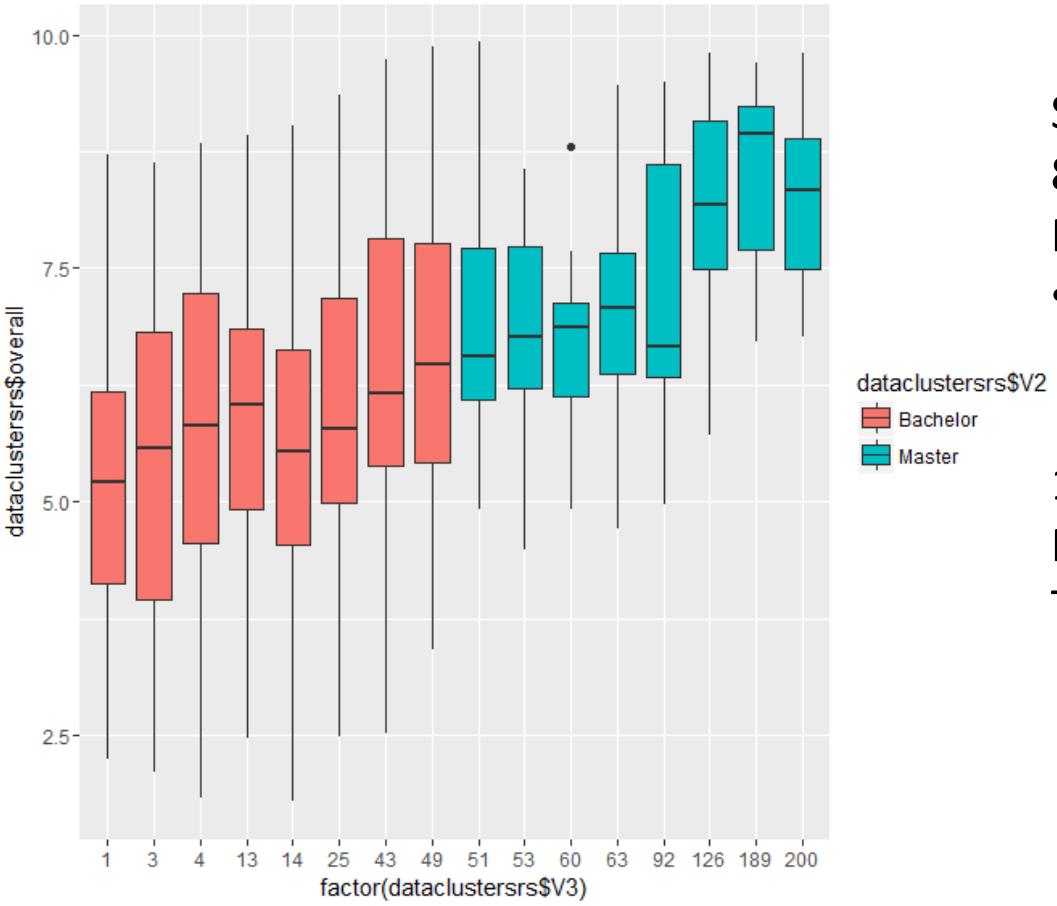


- Student grades (y)
- 200 programmes (x)
  - 50 BA, n=280 each
  - 150 MA, n=40 each

- R-code is available on Blackboard

- Population mean: 6.52

Stratified cluster sample

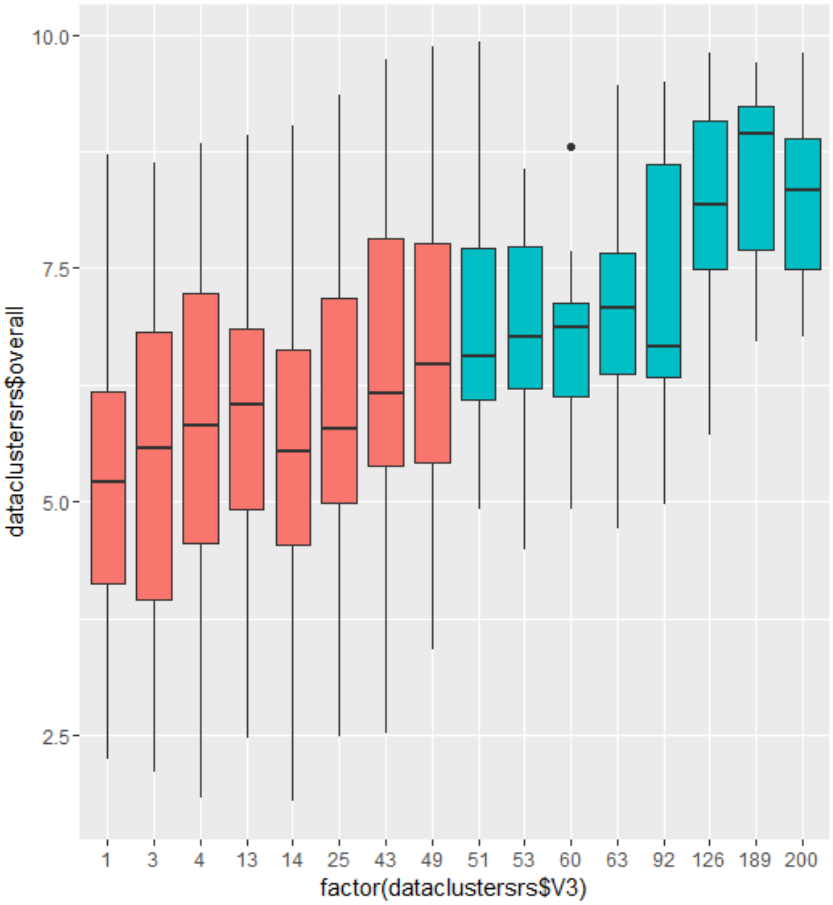


Stratify on programme (2)
8 clusters in each (can also vary)
Random sample per cluster PPS:

- sample with p=.4

16 clusters
For BA:
Total n=1000 out of population 20000

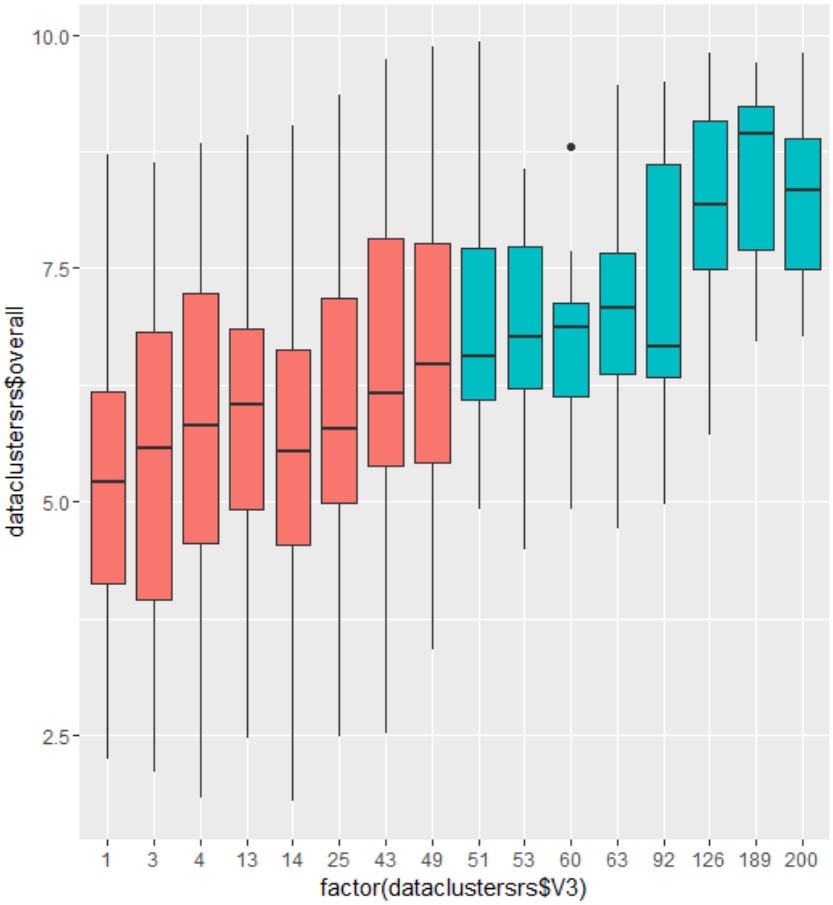
Variance estimation



- How do we calculate variances.
- Alternative: Horvitz-Thompson estimator
  - Stage 1: stratify
  - Stage 1: cluster
  - Stage 2: Select individuals

- Weights:
- Stage 2: per cluster:
  - Wt|s,master = 15 out of 40 -> 2.5
  - Wt|s,Bachelor = 112 out of 280 -> 2.5

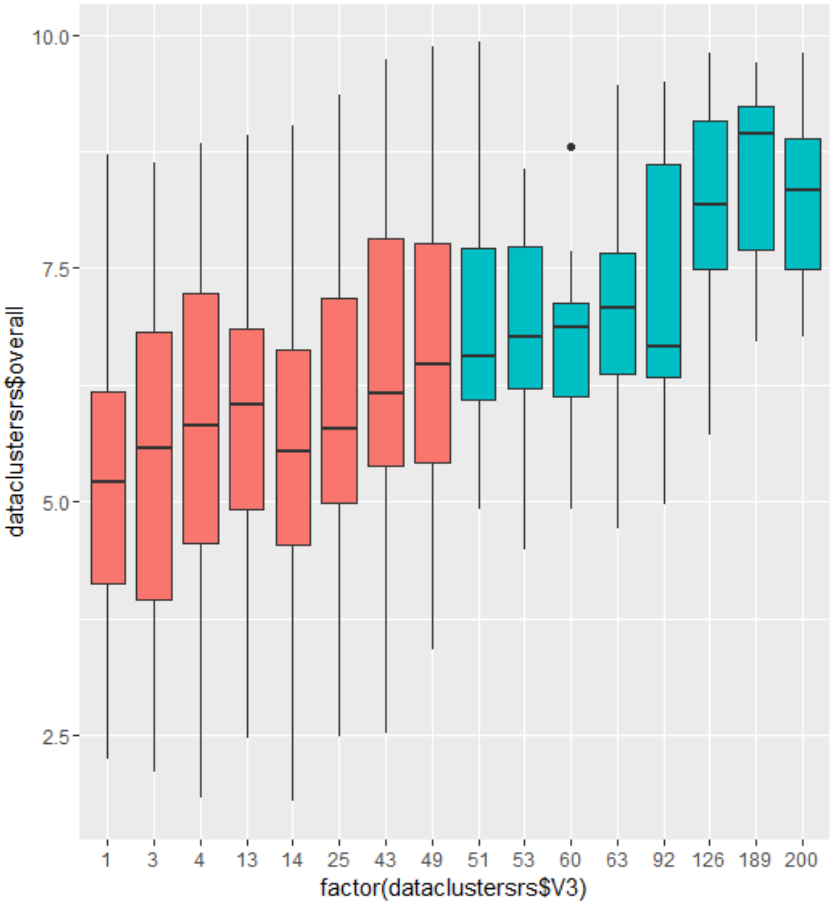
Variance estimation using weights



- Weights:
- Stage 2: per cluster:
  - Wt|s,master = 15 out of 40 -> 2.5
  - Wt|s,Bachelor = 112 out of 280 -> 2.5

- Stage 1: clusters out of strata
  - Wt|s,master = 8 out of 150 -> 18.75
  - Wt|s,Bachelor = 8 out of 50 -> 6,25

Variance estimation – constructing weights

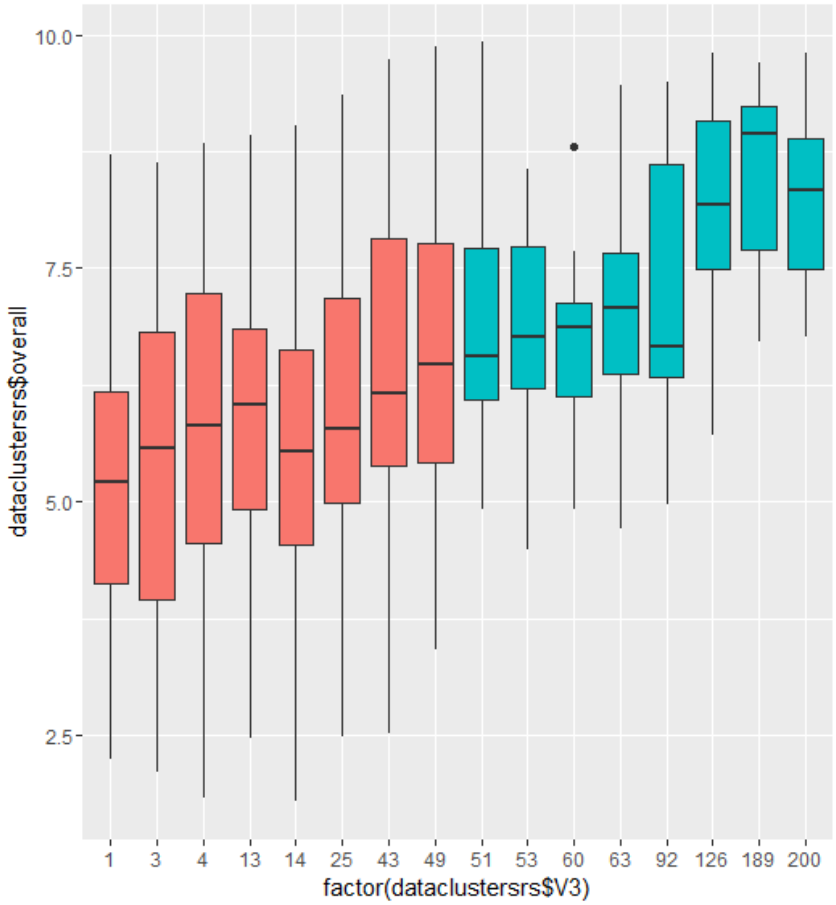


- Weights:
- Stage 2: per cluster:
  - Wt|s,master = 15 out of 40 -> 2.5
  - Wt|s,Bachelor = 112 out of 280 -> 2.5

- Stage 1: clusters out of population
  - Wt|s,master = 8 out of 150 -> 18.75
  - Wt|s,Bachelor = 8 out of 50 -> 6,25

- Total weight
  - Wt|s,master = 2.5 * 18.75 -> 46.875
  - Wt|s,Bachelor = 2.5 * 6.25 -> 18.75
- Rescaled weight
  - Wt|s,master = 46.875/mean(Wt) = 2,42
  - Wt|s,Bachelor = 18.75/mean(Wt)= 0,81

Variance estimation in R – identical results




```
clus2a <- svydesign(ids=~cluster+id, strata=~programme,
          #   weights=~weights, fpc = ~fpc1+fpc2, data=dataclustersrs)
```

```
Clus2b <- svydesign(ids=~cluster+id,  #strata=~dataclustersrs$V2,
          weights=~weights,#  fpc = ~fpc1+ffpc2,data=dataclustersrs)
```

Mean: 6.04

s.e. = .15275

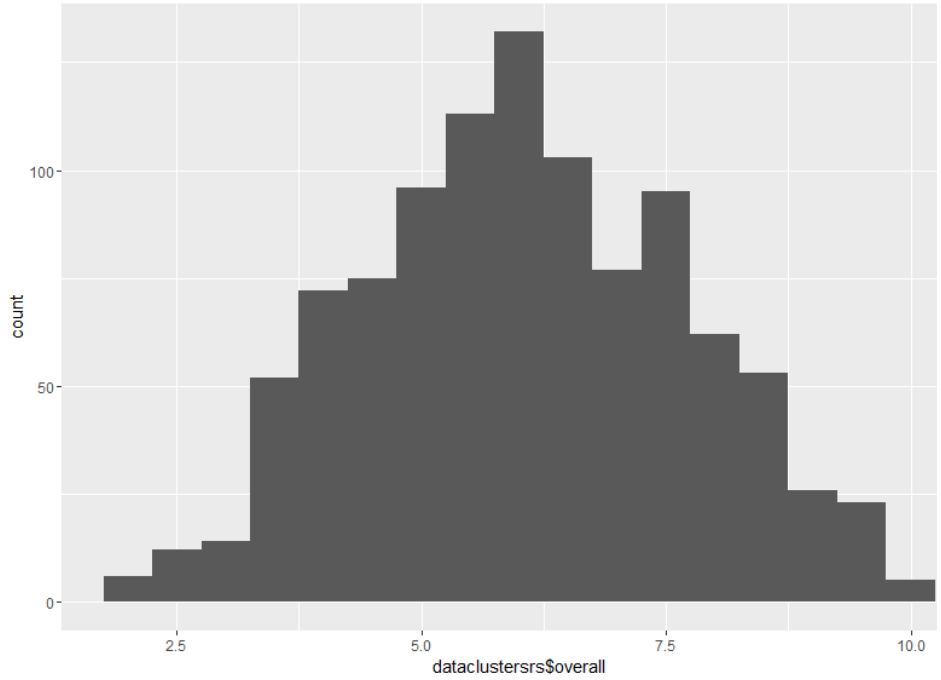
Deff= 8.86

Weights

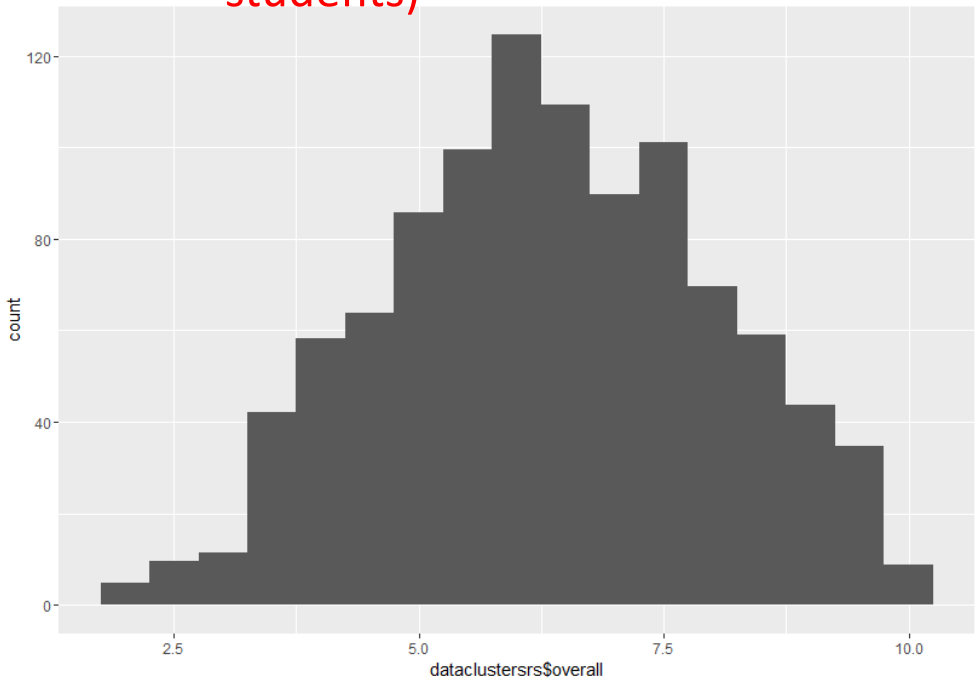
- The study doesn't stop at sampling
  - nonresponse weights (see week 44,45)
- Variance in weights indication of difference with perfect SRS design without nonresponse
  - In SRS -> Wi=1, Var(weights)=0.
  - In our design -> Var(weights)=.27
  - Likely in our design with NR: Var(weights) >.27
    - Variance inflation
- Can trim weights if they are large (rescaled weights >3 or 5)
  - Bias becomes larger
  - Variance lower -> precision higher
  - Goal is to Minimize Mean Square Error ($bias^2$ + variance)

Weighted graphs (using ggplot2)

- Without weights



- With weights
  - Heavier mass in upper tail (high weights for MA students)



Next weeks:

- Next week:
  - In two weeks: class-free week
- In two weeks:
  - Last week about sampling: model assisted estimation
    - Design based ---------- model-based
    - Ratio and regression estimation
  - Stuart 71-90
  - Finish class exercises today
  - Take home exercise:
    - Specify your survey design in R
  - Assignment 1 online
  - Deadline: 21 October 17:00