Survey Data analysis
week 3

Simple Random Sampling

The big picture

- Inference
  - use a small dataset to say something about the world
  - Design based:
    - probability based sampling and inferenc
    - Estimate and correct for each TSE source
    - Weeks 3-~8
  - Model-based
    - Big data, any data?
    - Model all the data errors, but how?
    - Week 9 -~14

Take home exercise of week 2

- Deck of 52 cards
  - Spades, diamonds, clubs, hearts
  - Each suit: 13 cards
- How many cards of Spades?
  - When sample of size 10/40
  - When drawing with/without replacement

- Your results

Simple Random Sampling

- Every element on the sampling frame has an equal, non-zero probability of being selected into sample
  - Element: individuals/households/companies
  - Population: collection of elements
- Why/when use a SRS?

Simple random sampling: when?

- There is a sampling frame consisting of population elements
  - Bonus Q: what to do if we have no frame?
- No need for clustering
  - Depends on mode
    - Web/mail vs. face-to-face/telephone
- No need for stratification
  - Little is known about people on sampling frame
  - Known characteristics do not correlate with dependent variables

The sampling distribution

- See file “simulation_cards_srs.R”
- Idea:
  - Every sample will have a slightly different estimate
  - What matters is whether [the method] gives you a consistent estimate of the population in the long run
  - Simple Random Sampling is an asymptoticially unbiased estimator
- We can repeat the experiment 10.000 times!

Sampling with/without replacement

- When does it not matter?
  - Selecting 1 out of 52 cards

Sampling without replacement (SRSWOR)

- When does with/without not matter?
  - Selecting 1 out of 52 cards

- What happens when we select 2 cards WOR
  - Card 1:
    - 13/52 chance for Spades
  - Card 2:
    - 75% chance for 13/51
    - 25% chance for 12/51

- Expected value for 2 cards:
  - 0.25 + (.75*13/51+.25*12/51)=
  - 0.25 + .1912 + .0588 = .50 Spades

Sampling with replacement (SRSWR)

- When does it not matter?
  - Selecting 1 out of 52 cards

- What happens when we select 2 cards WR
  - Card 1:
    - 13/52 chance for Spades
  - Card 2:
    - 13/52

- Expected value for 2 cards:
  - 0.25 + 0.25 = 0.50

- SRS(WR) and SRSWOR are both unbiased estimators of population mean
  - Also of mode/median (the beauty of the central limit theorem)
  - We assume no other errors (coverage, nonresponse)

what's the fuss – variance of estimator

- Extreme case: select 52 of 52 cards
  - Expected value: 13 Spades in both
  - Variance SRSWOR estimator: 0
    - Repeating it a 1000 times -> always 13 spades
    - This method needs correction -> without it is biased
  - Variance SRS(WR) estimator: 9.48
    - Repeating it a 1000 times -> variation
- Difference in variance is larger when a larger proportion of population is sampled

Estimators

- If we repeat a study n times (say 10000 times), we can investigate:
  - Bias: is the mean/variance/etc. correctly estimated in the long run?
    - Do we get p=.25 for spades on average?
  - Variance of estimator (precision)
    - How much variation is there in the mean?
    - In reality we take just 1 sample!
  - Consistent: does it work across all situations?
    - Different kinds of data

- Mean Square Error = bias^2 + variance
- A good estimator often minimizes MSE

Computation SRSWOR (without)

1. Mean under Simple Random Sampling

$$\bar{y} = \frac{1}{n}\sum y_i$$

2. Variance of the SRS mean estimate

$$var(\bar{y}) = (1\text{-}\ f)\frac{s^2}{n}$$ Correction 1: fpc

$$s^2 = \frac{1}{n-1}\sum(\ y_i\ \text{-}\ \bar{y})^2$$ Correction 2: Divide by n-1

n = sample size, s^2 = variance in sample

How do we compute se in SRSWOR ?

1. Mean under Simple Random Sampling

$$\bar{y} = \frac{1}{n}\sum y_i$$

2. Variance of the SRS mean estimate

$$var(\bar{y}) = (1\text{- f})\frac{s^2}{n}$$

$$s^2 = \frac{1}{n-1}\sum(\, y_i \text{ - } \bar{y})^2$$

3. Standard error of the mean

$$\text{Se }(\bar{y}) = \sqrt{var(\bar{y})} = \sqrt{1-f}\,\frac{s}{\sqrt{n}}$$

Se = standard error, n = sample size, s=standard deviation in sample

Intermezzo 1: Fpc in practice

- Fpc = (1-f) = (1-n/N) or (N-n)/N
  - In SRSWOR, correction
- fpc approaches 1 when n/N small
  - when sample of 1.000 people in the Netherlands is drawn:
  - Fpc = 1 – 1.000/17.000.000 = 1 – 0,00058 = 0,99942

- When sampling fraction n/N < .05, ignore FPC
  - We assume a infinite population

Intermezzo 2: (n-1) or n?

- Bessel's correction for variance: Divide by n-1 when you calculate variances (or se) using sample data
- Why?
  - Ideal: $\sum(y_i - \mu)^2$ $\quad s^2 = \frac{1}{n-1} \sum(y_i - \mu)^2$
  - In practice: $\sum(y_i - \bar{y})^2$ $\quad s^2 = \frac{1}{n-1} \sum(y_i - \bar{y})^2$

  - The sample mean is always a bit biased
  - the sum of squares is smaller than it should be
- Divide by n-1 in denominator to adjust

μ = population mean

Why smaller?

- Sum of squares is too small when using a sample
- Why? Here is what we would like
  - $\sum(y_i - \overline{y})$+ ($\overline{y}$ - μ)^2

- Divide by n-1 in denominator to adjust
  - dividing by n-1 works for variance, but biased for s! (sqrt(s^2))
  - When you would resample many times
    - Not the smallest MSE with many types of data
    - often sqrt(1.5) used instead of n-1 in larger samples
- Just remember: use n-1 for variance estimate of mean
  - Want to know more? See “bessels correction.r”

A real example

- I 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

- 5 minutes: how do we do this?

Example: possible solution

- Cheap: e-mail
  – Sample all students? A census
- Can do complicated stratification to ensure enough students from every programme
  – 200 + programmes…
- Simple random sampling (SRS)
  – Risk of small n for some programmes.
  – Let’s work out how SRS works
  – And talk about sample size

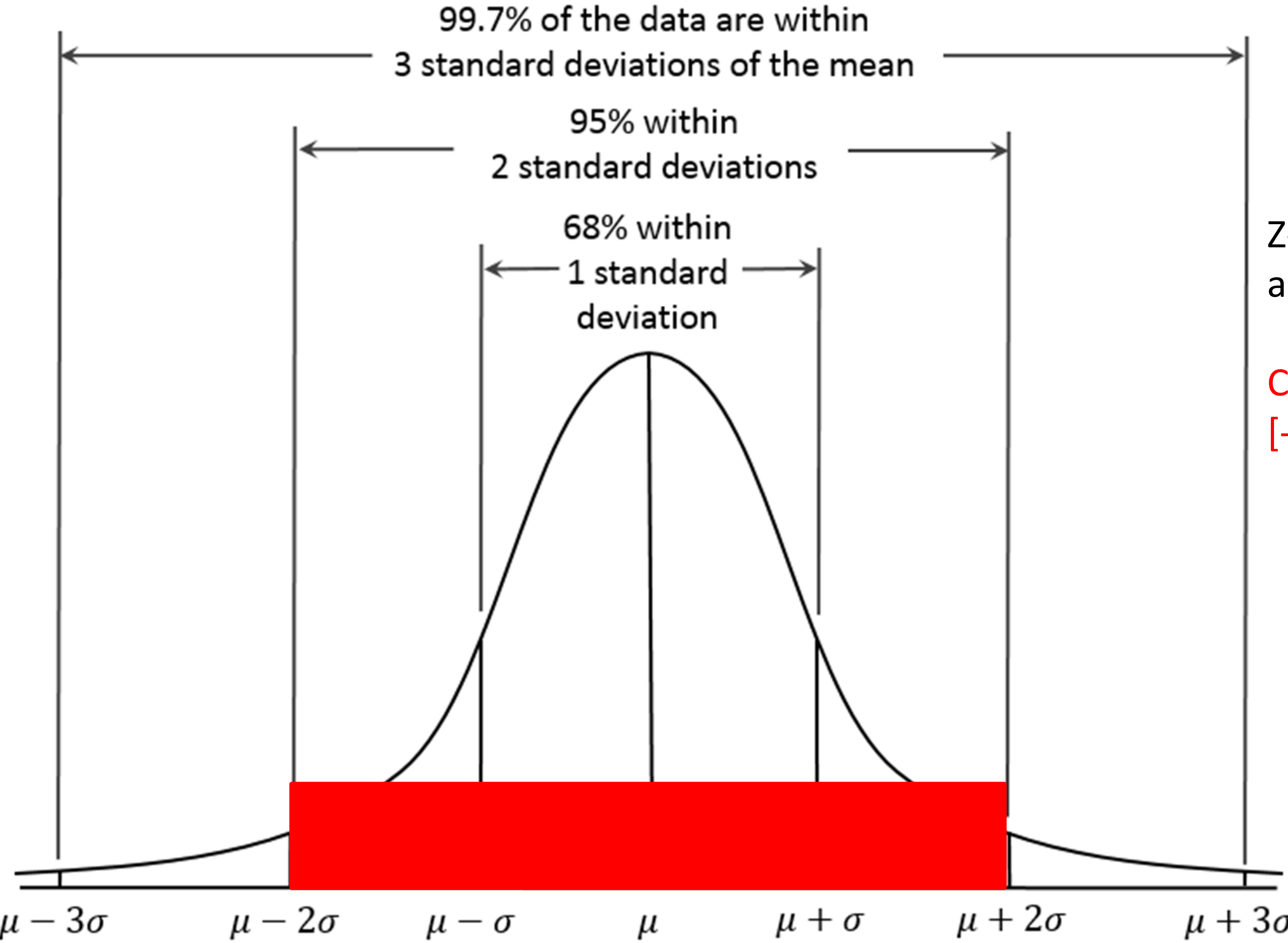
Why is standard error useful?

- Gives indication of both
  - Uncertainty due to sampling error
  - Uncertainty in estimation (e.g. ML estimation)

- Used to construct confidence Interval:

$$[\overline{y} - z_{\alpha/2} SE(\overline{y}), \overline{y} + z_{\alpha/2} SE(\overline{y})]$$

s.e. and confidence intervals



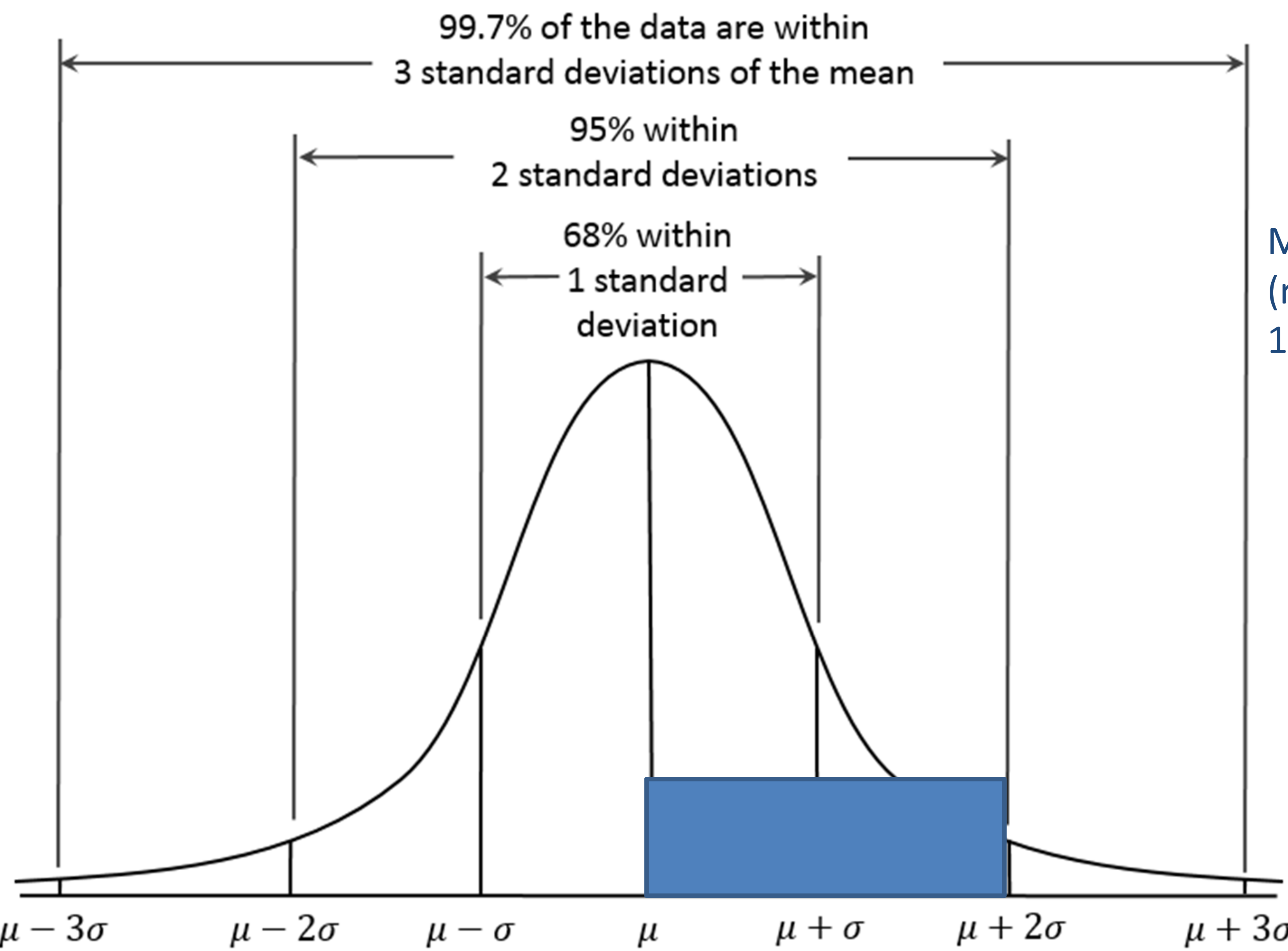
With alpha=.05

Two-sided test

Z-value for alpha=.05: 1.96

Confidence interval [-1.96*se, +1.96*se]

Margin of error

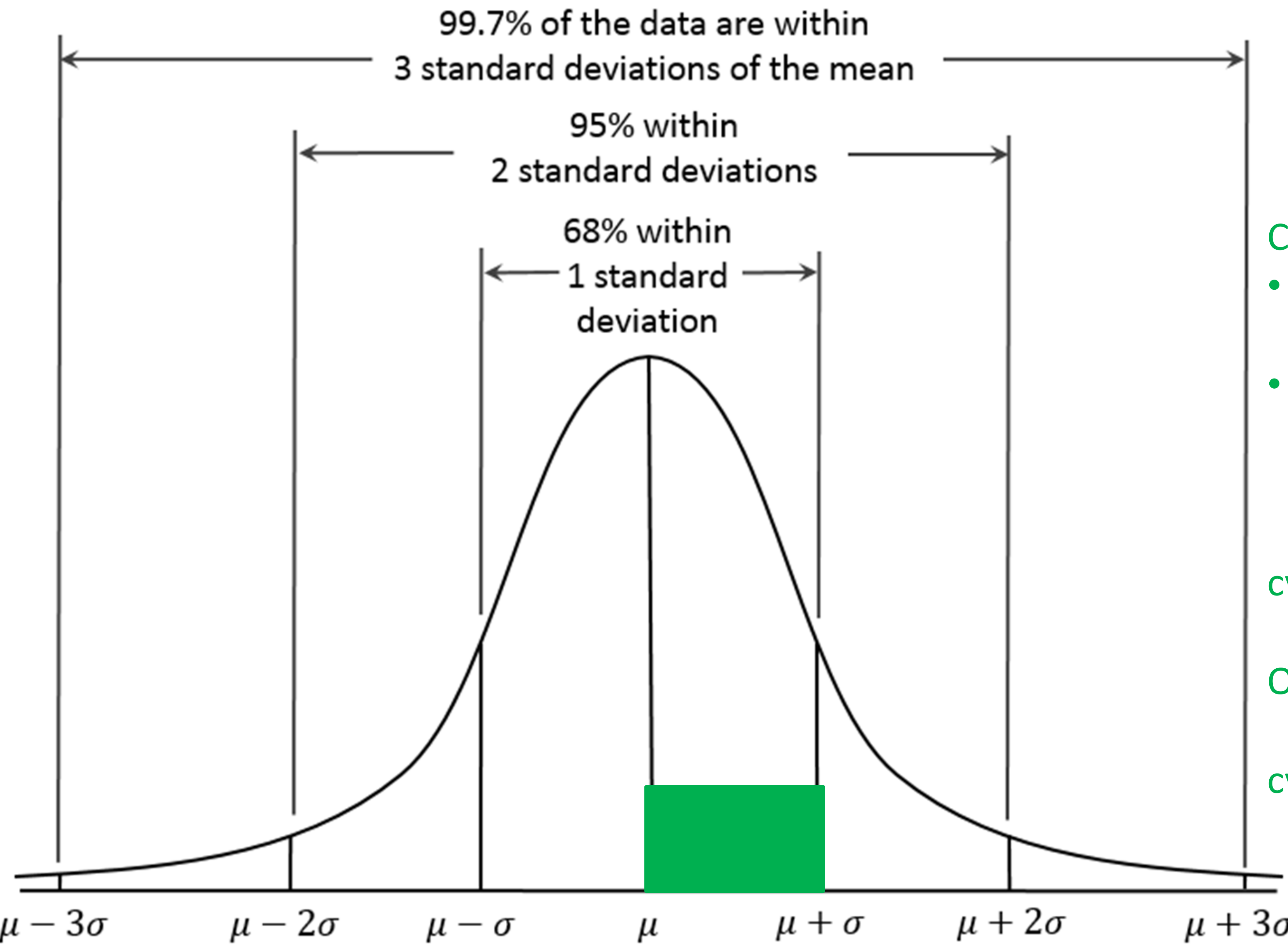


With alpha=.05

Two-sided test

Margin of error
(radius of CI)
1.96 * s.e.

Coefficient of variation



With alpha=.05

Two-sided test

Coefficient Variation

- relative standard error
- “units of the population average” (Stuart)

cv = σ/μ

Or

cv = se/ $\bar{y}$

Short exercise

- What is mean grade of students at Utrecht University (1-10-scale)
- Population = 20.000 students

- Best guesses for means and Variance?
  - Mean: 7.0
  - variance: 4

- Take a sample of n=200 (don’t worry about fpc (as n/N = .01)

- What is the standard error?
- What is the margin of error?
- What is coefficient of variation?

Solution:

1. standard error:

$$se = \sqrt{1-f}\,\frac{s}{\sqrt{n}}$$

$$se = \sqrt{1-f}\,\frac{2}{\sqrt{200}} = .14$$

2. Margin of error

MoE = 2 s.e = .28

3. Coefficient of variation

Cv = $\frac{se}{\bar{y}}$ = $\frac{.14}{\bar{7}}$ = .02 (2% of the mean)

What if we need to be more precise?

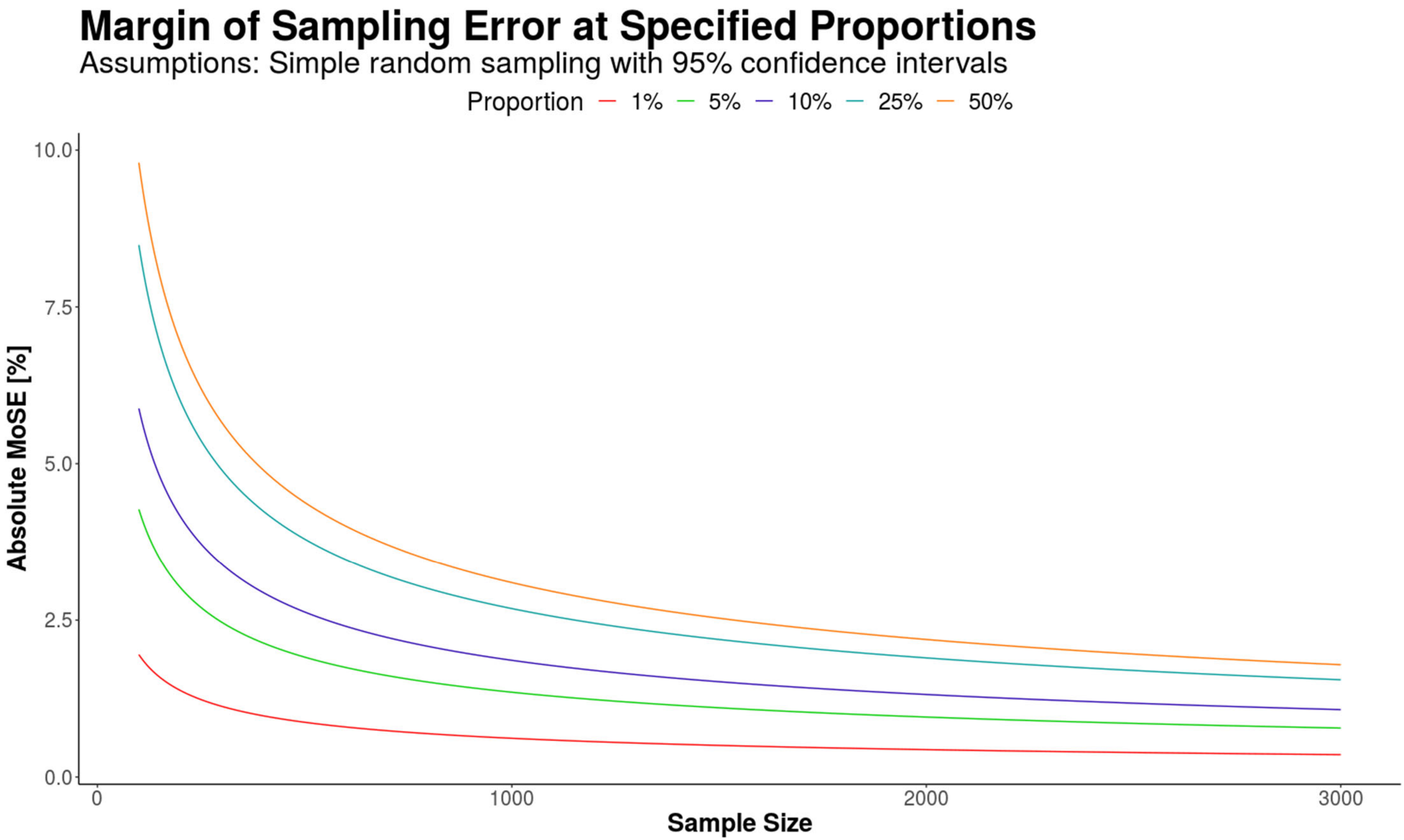
- How can confidence interval change?
  - Variance in sample/population
  - Required precision of Confidence Interval
    - Alpha
  - Size of sample (n)

- What if:
  - n=400, alpha = .05

Solution:

- n=400
- $se = \sqrt{1-f}\,\frac{2}{\sqrt{400}} = .1$
- Or $.1 * \sqrt{1 - \frac{400}{20.000}} = .1 * .98 = .98$ if we include fpc (n/N=.04)
  - Standard error becomes $\frac{.1}{.14} = .71$ times as wide when we double the sample size

MoE and sample size



Break

What if we need to be more precise?

- How can confidence interval change?
  - Variance in sample/population
  - Required precision of Confidence Interval
    - Alpha
  - Size of sample (n)

- What if:
  - n=400, alpha = .005 (multiple testing issue from MSSBBS02)

Solution

- Alpha = .005
  - Confidence interval: +- 2.58 se
  - Confidence interval becomes $\frac{2.58*2}{1.96*2}$ = 1.32 times wider

- Because we lower alpha, we get a wider CI

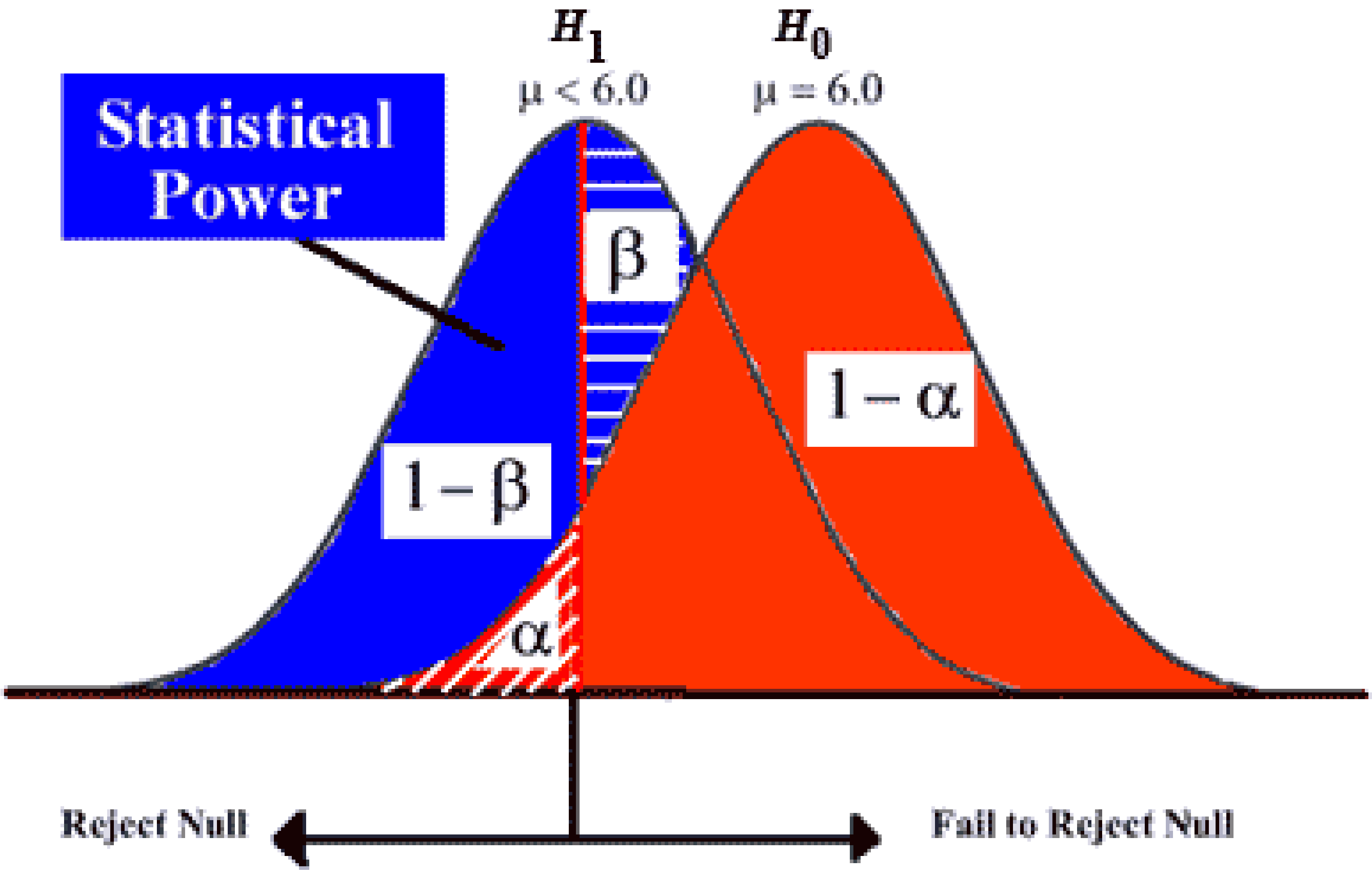
How large should my sample be?

- #1. question in statistical consultation
- Depends on:
  - Statistic of interest (here: mean)
  - Variance in sample/population
  - Required precision of Confidence Interval
    - Alpha, standard error
  - Size of sample/population (n/N)

  - Often, we want to test for size of difference between groups (see take home exercise) and therefore Power also plays a role

α? Power (β)?

- Type I error (α) is to reject H0 while H0 is true
- Type II error (β) is to accept H0 while H1 is true



How large should my sample be?

- α =.05
- Standard error?
  - Estimate relative error instead
  - Coefficient of variation

$$cv(\overline{y}) = \frac{se(\overline{y})}{\overline{y}}$$

Class exercise

- What is mean grade of students at Utrecht University (1-10-scale) under SRS?
  - Population = 20.000 students
- Best guesses for means and Variance?
  - Mean: 7.0
  - variance: 4
- I want to be precise: s.e. restricted to 2% (cv=.02)
  - Implies CI of [-1.96 *2 ; 1.96*2] = 7.84%, and
  - Margin of error [1.96*2] = 3.92%
- Alpha = .05
- How large should sample be?

$$cv(\bar{y}) = \frac{se(\bar{y})}{\bar{y}} \qquad se(\bar{y}_0) = \sqrt{var(\bar{y}_0)} = \sqrt{(1-f)}\frac{s}{\sqrt{n}}$$

Solution:

1. standard error: $cv(\bar{y}) = \frac{se(\bar{y})}{\bar{y}}$

.02 = x / 7 = .14/7

2. Compute n under SRSWOR:

$$se(\bar{y}_0) = \sqrt{var(\bar{y}_0)} = \sqrt{(1-f)}\frac{s}{\sqrt{n}}$$

.14 = sqrt(1-f)* (2/sqrt(n))

2/.14 = sqrt(n)/sqrt(1-f) = 14.286^2 /sqrt(1-f).

n=204.08 (or 205)

- We may ignore fpc because sampling fraction <5%
- Or: f = 1-(205/20.000) = 1-.01 = .99
- 2/.14/(sqrt(.99) = sqrt(n) = 14.43^2 = 206.14 (or 207)

Estimator

- Equal selection probabilities (SRS):
  - Unbiased estimator of mean, variance in population
    - Can use sample size to manipulate precision
  - Also of regression (OLS), other estimates
  - So, with an SRS, you can use the data as is
  - When there are *no coverage and nonresponse errors*
- Unequal selection probabilities
  - You can't use the data as is…

Why unequal probabilities?

- Because you want to make your design cheaper or more efficient
  - Stratification and clustering
  - Next time
- Because there are issues with your list
  - Why?

Coverage and sampling issues in SRS

- Your list may have double entries
  - E.g. students enrolled in multiple programmes
  - Sometimes you don't know in advance
    - E.g. multiple telephone nos in RDD designs
- You have a list of clusters, but want individuals
  - Addresses -> individuals
  - How many people live at this address?

- Problem -> take unequal selection probabilities into account

What to do:

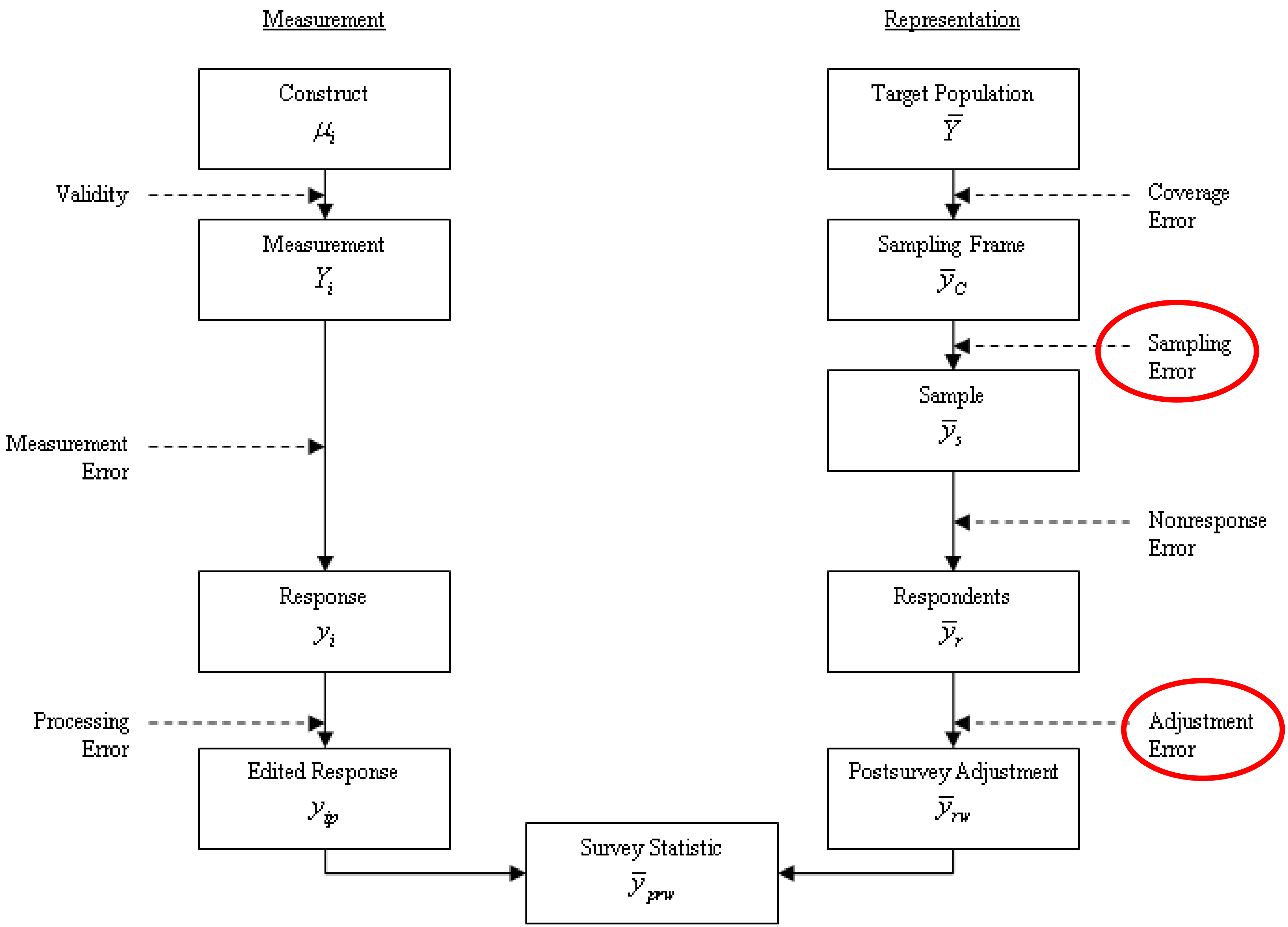
1. Estimate individual selection probabilities: π_i
2. Weight cases by the inverse of their selection probabilities: $w_i = \frac{1}{\pi_i}$
3. In computing statistics, every case is weighted in analysis:

- e.g. $\overline{y}_w = \frac{\sum w_i y_i}{\sum w_i}$

Weights

- In SRS, w_i = 1 for all.
- Sampling or design weights correct for unequal selection probabilities in sampling
  - Correction for bias in sampling
  - Nonresponse weights also exist (week 9)
- Point estimates are weighted
  - Means, B,
- Variances more complex
  - Next week

Weights



Measurement
Representation
Construct
μ_i
Target Population
$\bar{Y}$
Validity
Coverage
Error
Measurement
Y_i
Sampling Frame
$\bar{y}_C$
Sampling
Error
Sample
$\bar{y}_s$
Measurement
Error
Nonresponse
Error
Response
y_i
Respondents
$\bar{y}_r$
Processing
Error
Adjustment
Error
Edited Response
y_{ip}
Postsurvey Adjustment
$\bar{y}_{rw}$
Survey Statistic
$\bar{y}_{prw}$

Next week

- Take home exercise week 3
  - Draw SRS samples (once more)
  - Work with Svydesign in R
  - Work with design weights
  - Compute sample sizes (power analysis)
- **Next time (in 2 weeks):**
  - We will discuss sampling designs with explicit unequal selection probabilities (stratification and clustering)
  - Read Stuart