# Survey analysis
# week 39

# 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 39-~47
  - Model-based
    - Big data, any data?
    - Model all the data errors, but how?
    - Week 44-~50

# Class exercise

- 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

# The sampling distribution

- See file “simulation cards srs.R”
- Lets repeat the experiment 10.000 times!

# 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

# 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)

# So 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 1000), 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

# Computation SRSWOR (without)

1. Mean under Simple Random Sampling

$$\begin{aligned}\overline{y}_0 &= \frac{y}{n} = \frac{1}{n}\sum_{j=1}^{n} y_j \\ &= \frac{1}{n}[y_1 + y_2 + \ldots + y_n]\end{aligned}$$

2. Variance of the SRS mean estimate

$$\begin{aligned}var(\overline{y}_0) &= (1-f)\frac{s^2}{n} \\ s^2 &= \frac{1}{n-1}\sum_{j}^{n}(y_j - \overline{y})^2\end{aligned}$$

Correction 1: fpc

Correction 2: Divide by n-1

# How do we compute s.e.?

1. Mean under Simple Random Sampling (SRS):

$$\begin{aligned}\overline{y}_0 &= \frac{y}{n} = \frac{1}{n}\sum_{j=1}^{n} y_j \\ &= \frac{1}{n}[y_1 + y_2 + \ldots + y_n]\end{aligned}$$

2. Variance of the SRS mean estimate:

$$\begin{aligned}var(\overline{y}_0) &= (1-f)\frac{s^2}{n} \\ s^2 &= \frac{1}{n-1}\sum_{j}^{n}(y_j - \overline{y})^2\end{aligned}$$

3. S.e. of the SRS mean estimate:

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

n = sample size, s=standard deviation in sample

# Intermezzo 1: Fpc in practice

- Fpc = (1-n/N) or (N-n)/N
- Sampling is done without replacement
- 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 s.e.) using sample data
- Why?
  - Ideal: $\sum_j (y_j - \mu)^2$
  - In practice: $\sum_j (y_j - \bar{y})^2$

$$var(\bar{y}_0) = (1-f)\frac{s^2}{n}$$

$$s^2 = \frac{1}{n-1}\sum_j^n (y_j - \bar{y})^2$$

  - 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

# Why smaller?

- Sum of squares is too small when using a sample
- Why? Here is what we would like

$$\sum_j ((y_j - \bar{y}) + (\bar{y} - \mu))^2$$

- 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" on Blackboard

# Computation SRSWR (with)

1. Mean under Simple Random Sampling

$$\begin{aligned}\bar{y}_0 &= \frac{y}{n} = \frac{1}{n}\sum_{j=1}^{n} y_j \\ &= \frac{1}{n}[y_1 + y_2 + \ldots + y_n]\end{aligned}$$

Same as without replacement

2. Variance of the SRS mean estimate

$$var(\bar{y}_0) = \frac{s^2}{n}$$

No fpc

$$s^2 = \frac{1}{n-1}\sum_{j}^{n}(y_j - \bar{y})^2$$

# 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
- 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:

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

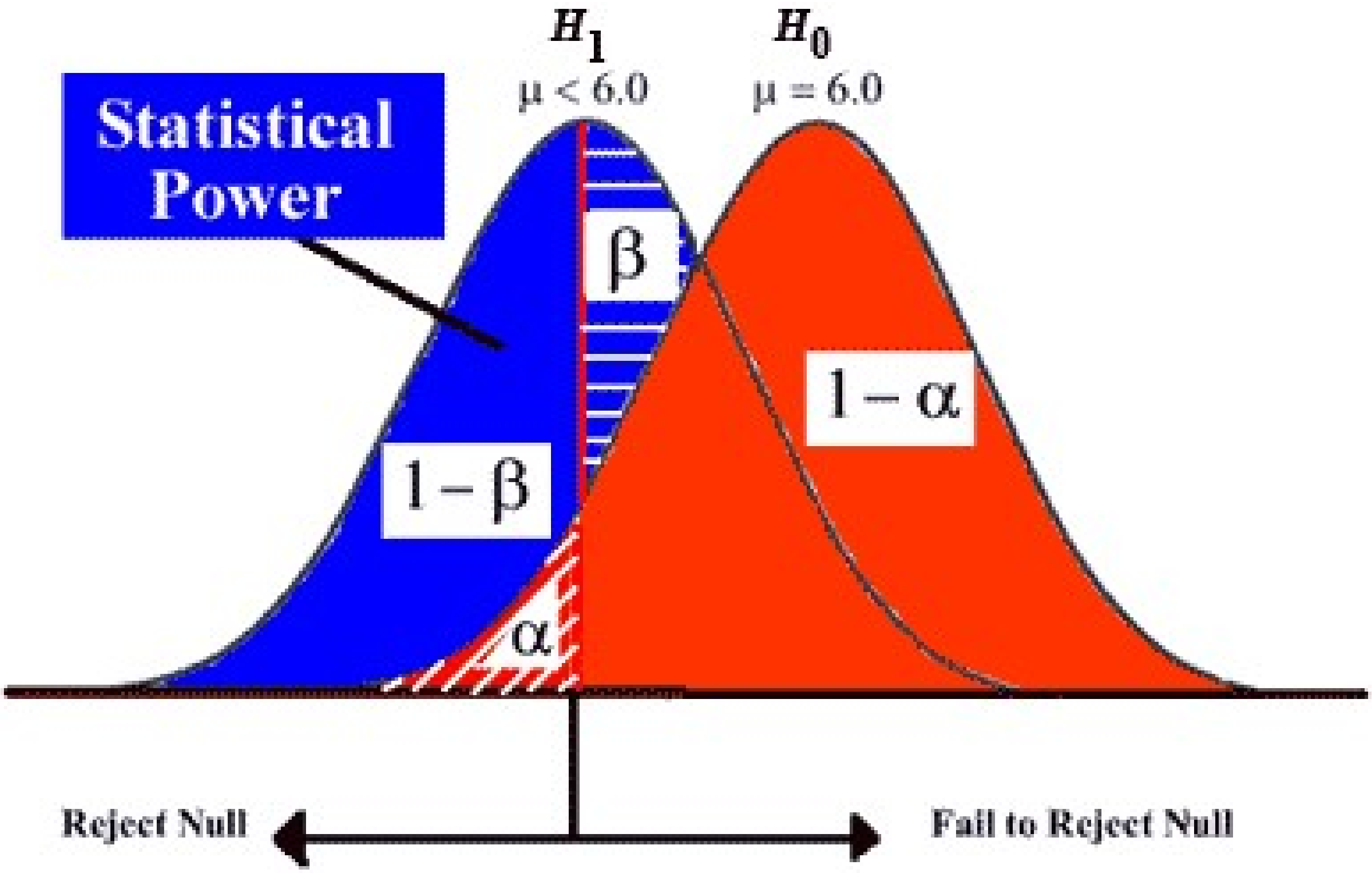
# How large should my sample be?

- #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)

  - Leads to POWER (beta).

# α?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

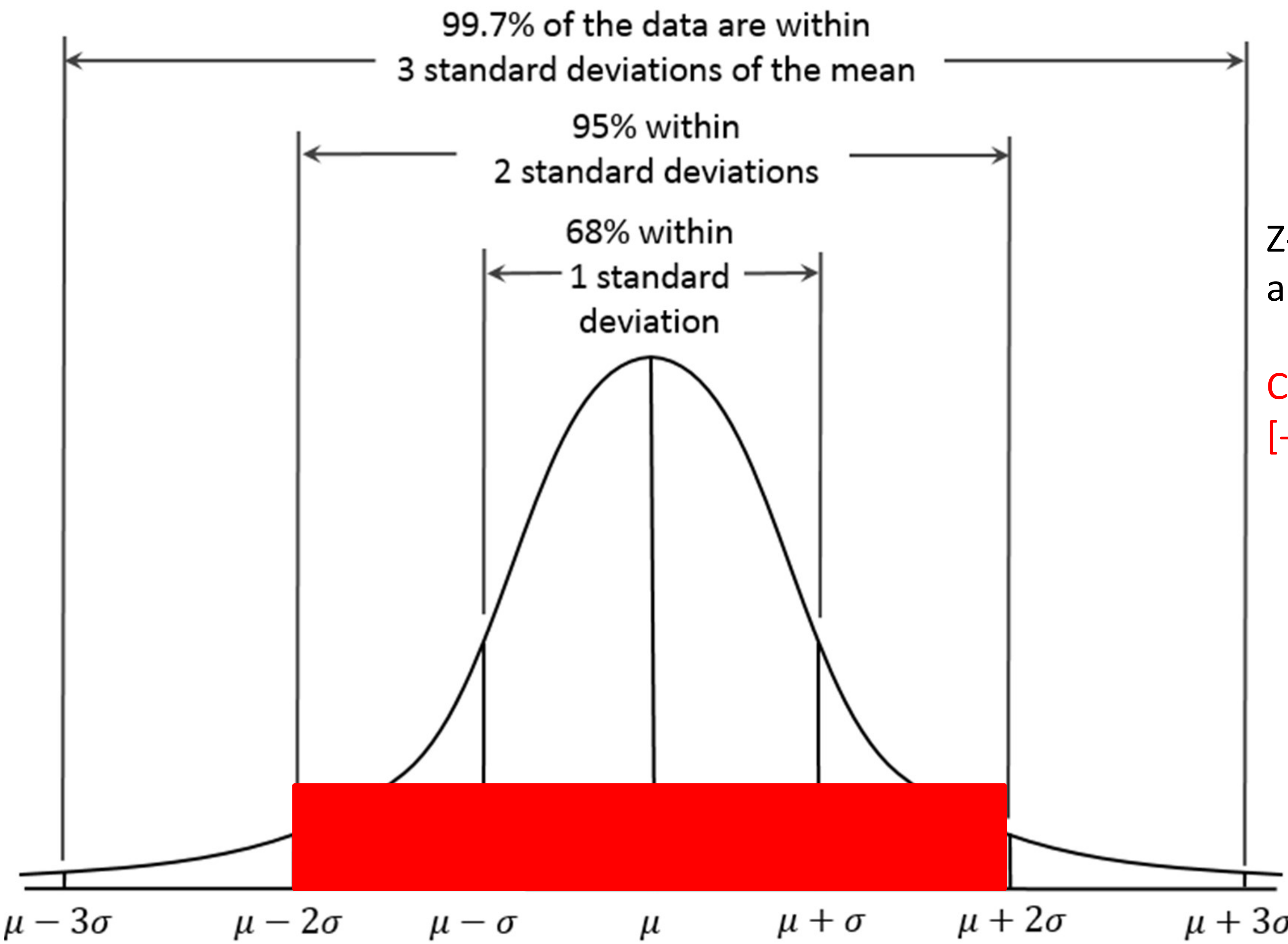


# How large should my sample be?

- $\alpha$ =.05
- Standard error?
  - Estimate relative error instead
  - Coefficient of variation

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

# Power and confidence intervals

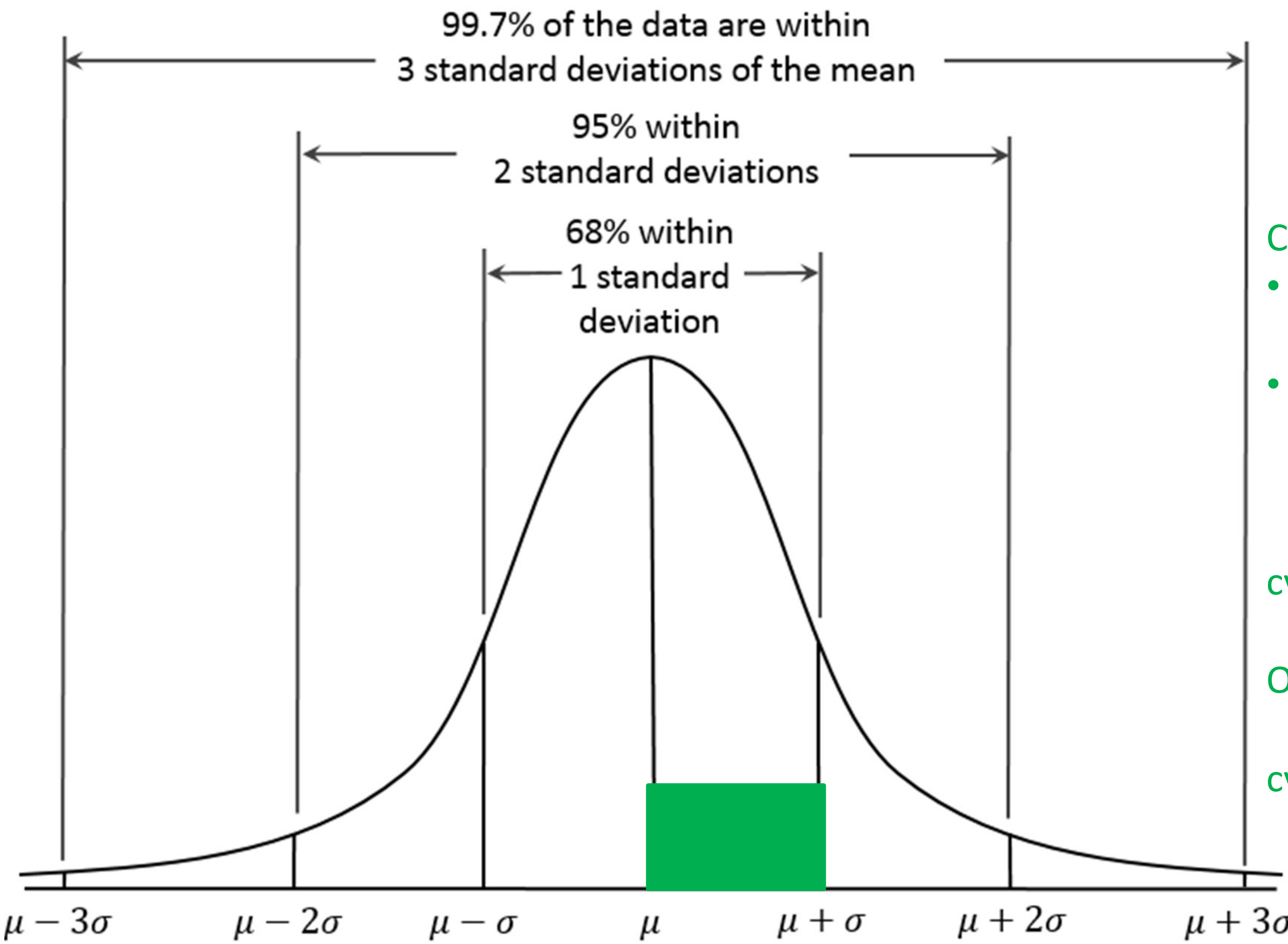



With alpha=.05
Two-sided test

Z-value for alpha=.05: 1.96

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

# Coefficient of variation




With alpha=.05

Two-sided test

Coefficient Variation

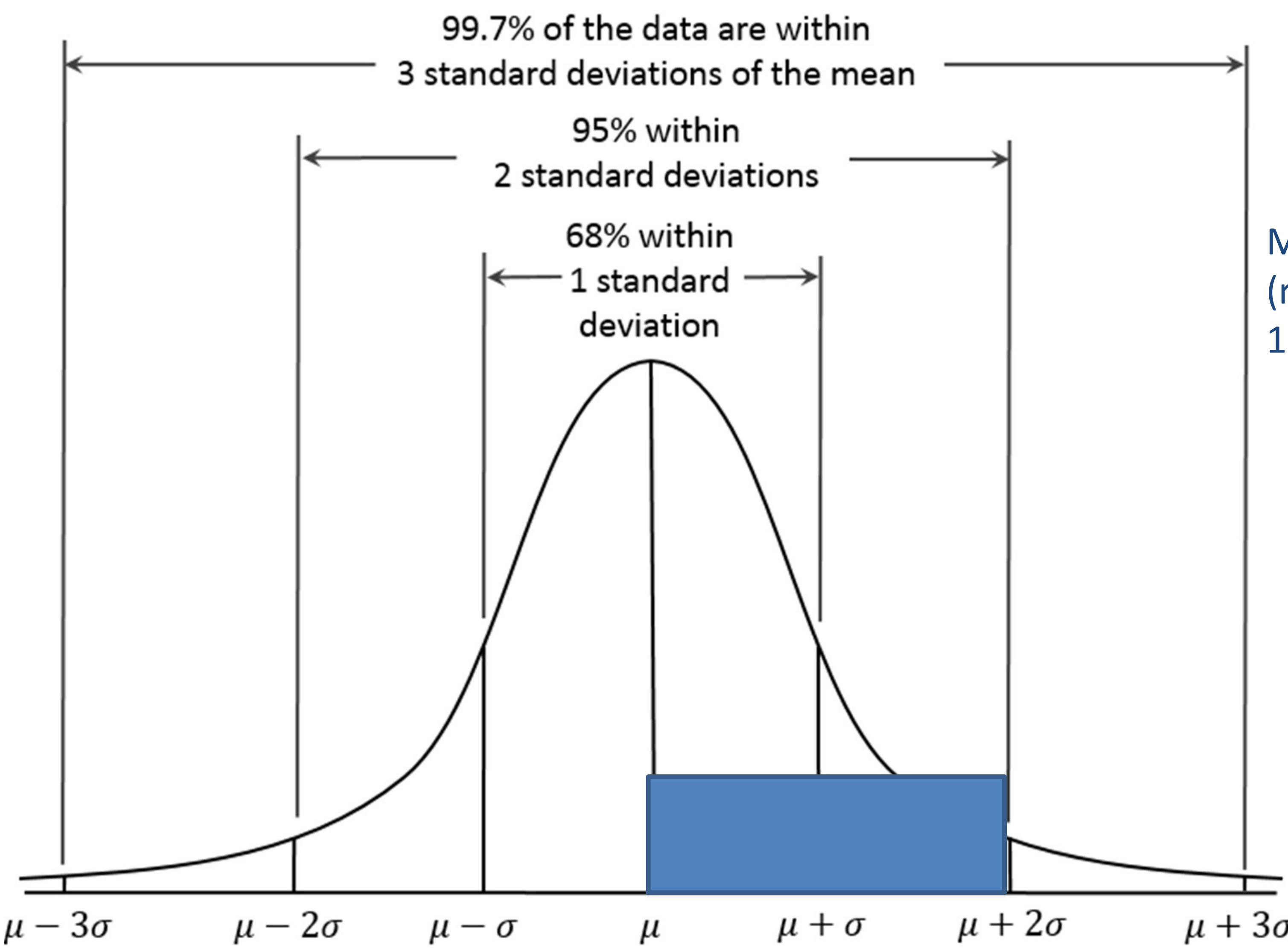
- relative standard error
- "units of the population average" (Stuart)

cv = $\sigma/\mu$

Or

cv = se/ $\bar{y}$

# Margin of error




With alpha=.05

Two-sided test

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

# 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 = sqrt(4)

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)

# Same exercise (if you have time)

## What if?

- Alpha = .005 (the "new") level proposed by Benjamin et al (2017)
- Margin of error = 5%?

# Solution alpha = .005? MoE 5%?

**1.** c.v = .05 /2.58 = .0193

(MoE = ½ CI, Z-value: 2.58)

**2.** standard error:

.0193 = x / 7. = 0.13566/7

**2.** Compute n under SRSWOR:

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

0.13566 = sqrt(1-f)* (2/sqrt(n))

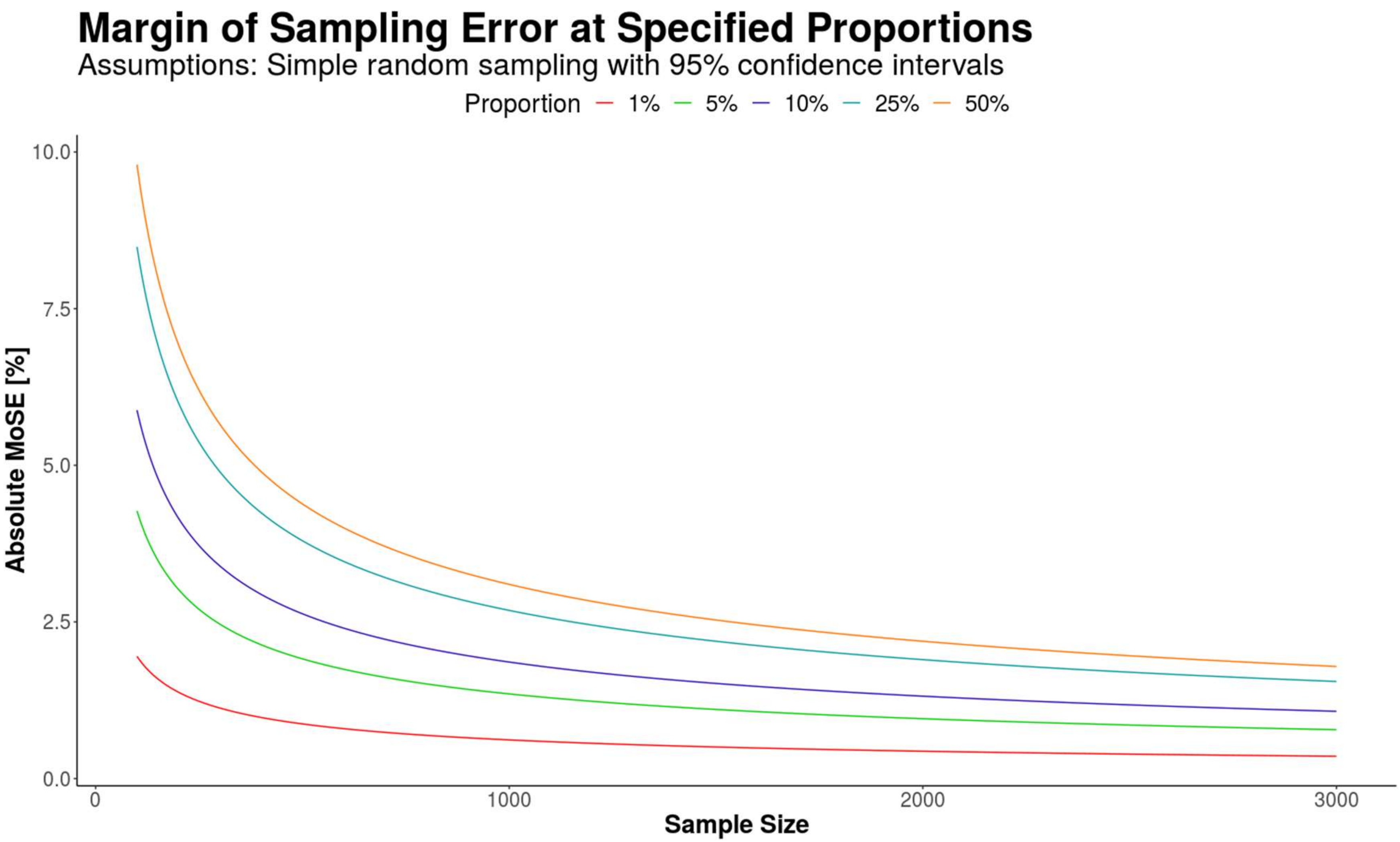
2/0.13566 = sqrt(n)/sqrt(1-f).

N=217.35

We may ignore fpc because sampling fraction <5% (.01)

Otherwise: 2/.1356/sqrt(.99) = sqrt(n) = $14.90^2$ = 219.52 (or 220)

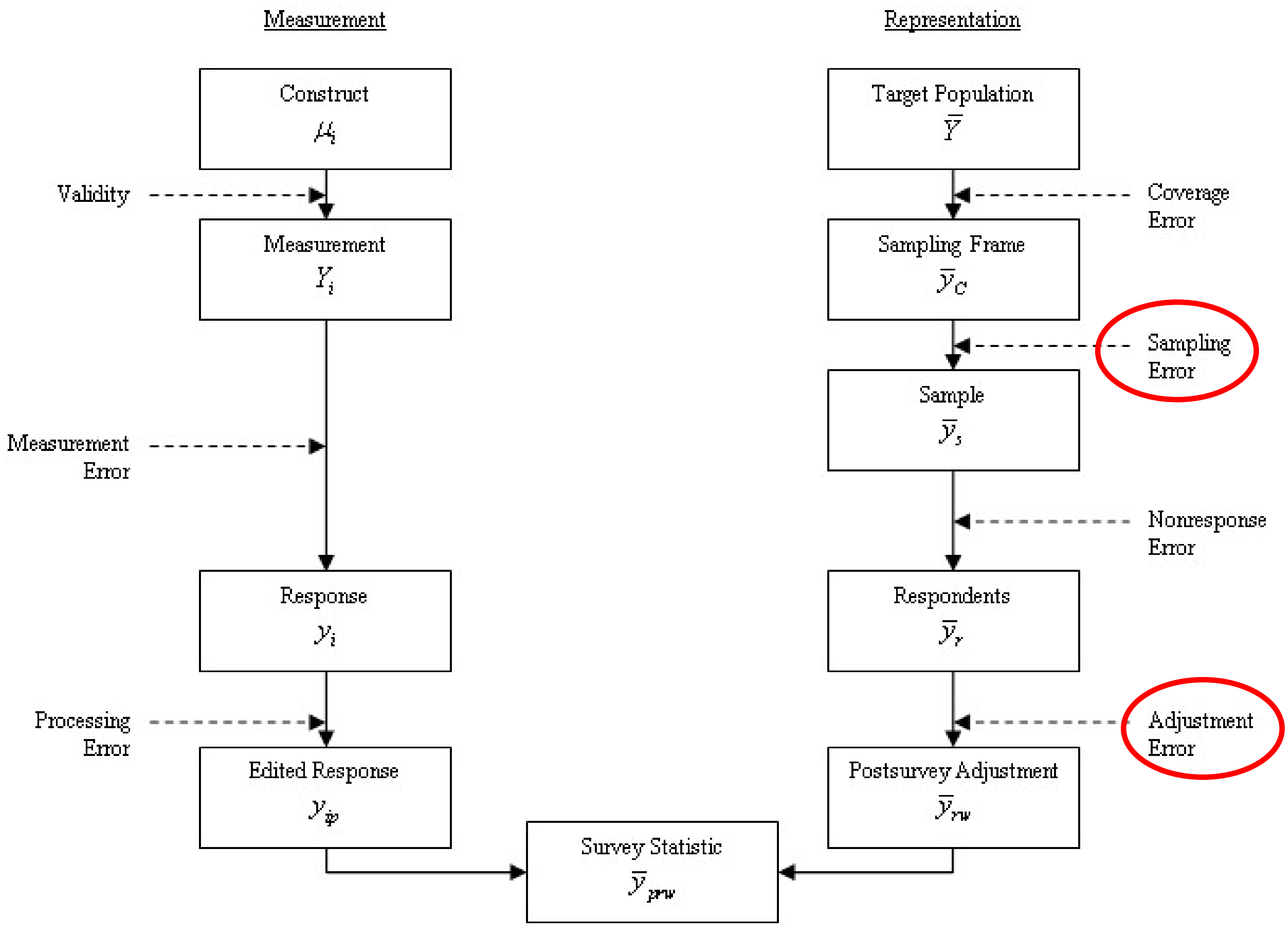
# MoE and sample size



# Estimator

- Equal selection probabilities (SRS):
  - Unbiased estimator of mean, variance in population
  - Also of regression (OLS), other estimates
  - When there are *no coverage and nonresponse errors*
- Unequal selection probabilities
  - All formulas shown so far do not work
  - Next week…

# Weights



Measurement
Representation
Construct
$\mu_i$
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
$\bar{y}_{prw}$

# Weights

- Goal of weights is to correct for:
  - Differences in sample selection probabilities: design weights
    - why?
  - Coverage and nonresponse error corrections
    - why? Think back to election example
  - Every case gets weight $W_i$
  - Typically: inverse of selection probability: $1/\pi_i$
    - Sample weight (design), nonresponse weight

# Weights: under SRS

- Mean $\bar{y}_w = \dfrac{\sum w_i \, y_i}{\sum w_i}$
- Under SRS:
  - $W_i$ = N/n, for every i for sample total
  - $W_i$ = 1; for sample estimates
- For stratification, clustering:
  - Corrects bias in means by assigning different weights
  - Variance: more complex (later weeks)
    - Variation in weights add to total variance in weighted statistics!

# Next week

- Take home exercise week 39
  - Draw SRS samples (once more)
  - Work with Svydesign (new!)
  - Work with design weights (new!)
- **Next week:**
  - We will discuss sampling designs with explicit unequal selection probabilities (stratification and clustering)
  - Read Stuart