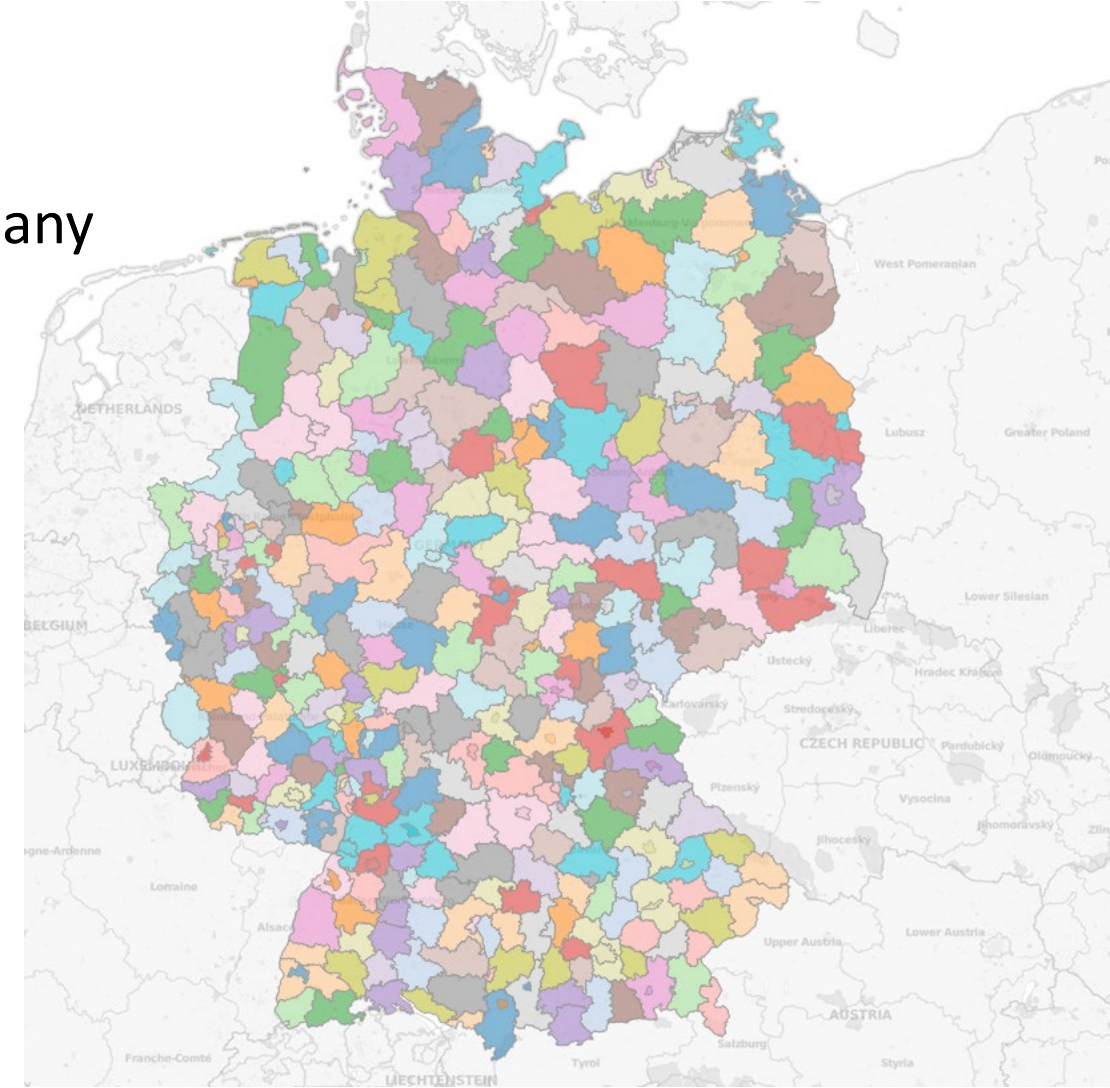
# Survey analysis
# week 6
# “ratio and regression estimation”

# Today

- Why ratio estimation?
- Class exercise ratio estimation
  - New example: coffees at UU
- Lecture ratio and regression estimation
- Class exercise regression estimation

# First rewind to cluster sampling….

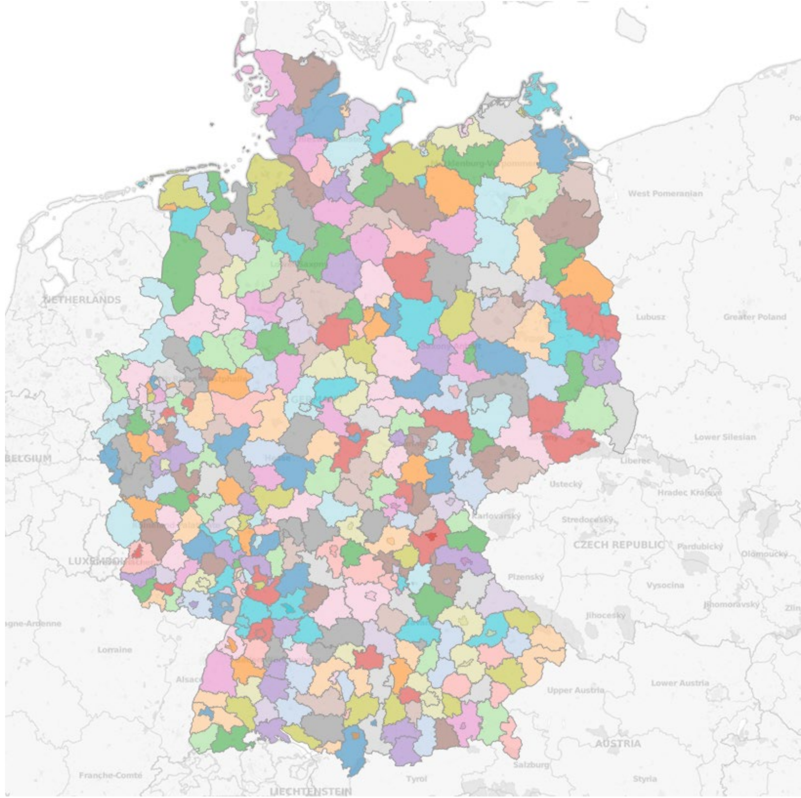
- You would like to estimate:
  - The number of “Whatsapp scams” in Germany

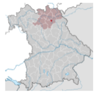
# Cluster sampling in Germany

- 411 Kreise (in 2022)
- Sampling frames only available at level of Kreise
- Select k clusters (50)
  - Stratify?
- Select households in clusters
  - Size=1500 per cluster

- Two-stage cluster samples
- Can we do better?



# Size of clusters is known

| Wappen | Stadt | Stadtkreis ID | BL | Regierungsbezirk[5] | UZ | Fläche in km² | tEW 1939 [6] | tEW 1950 [7] | tEW 1970 [8] | tEW 1990 [9] | tEW 2011 [10] | EW jetzt | Bev.-dichte (EW/km²) | Lagekarte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Aachen[1] (♁ 50° 47′ N, 6° 5′ O) | 05334 | NW | Köln | AC, MON | 160,85 | 162,2 | 129,8 | 175,5 | 239,2 | 260,5 | 249.070 (2021) | 1548 |  |
|  | Amberg (♁ 49° 27′ N, 11° 51′ O) | 09361 | BY | Oberpfalz | AM | 50,13 | 31,8 | 37,9 | 41,3 | 42,9 | 43,5 | 41.994 (2021) | 838 |  |
|  | Ansbach (♁ 49° 18′ N, 10° 34′ O) | 09561 | BY | Mittelfranken | AN | 99,91 | 26,0 | 33,2 | 33,2 | 37,6 | 40,3 | 41.662 (2021) | 417 |  |
|  | Aschaffenburg (♁ 49° 59′ N, 9° 9′ O) | 09661 | BY | Unterfranken | AB | 62,45 | 45,4 | 45,5 | 55,1 | 63,6 | 68,8 | 71.381 (2021) | 1143 |  |
|  | Augsburg (♁ 48° 22′ N, 10° 54′ O) | 09761 | BY | Schwaben | A | 146,85 | 185,4 | 185,2 | 213,2 | 254,3 | 266,6 | 296.478 (2021) | 2019 |  |
|  | Baden-Baden (♁ 48° 46′ N, 8° 14′ O) | 08211 | BW | Karlsruhe | BAD | 140,19 | 33,2 | 36,6 | 37,2 | 51,5 | 54,5 | 55.527 (2021) | 396 |  |
|  | Bamberg (♁ 49° 54′ N, 10° 54′ O) | 09461 | BY | Oberfranken | BA | 54,62 | 59,5 | 76,2 | 70,4 | 70,2 | 70,1 | 77.749 (2021) | 1423 |  |
|  | Bayreuth (♁ 49° 57′ N, 11° 35′ O) | 09462 | BY | Oberfranken | BT | 66,89 | 45,0 | 58,8 | 64,2 | 72,0 | 73,1 | 73.909 (2021) | 1105 |  |
|  | Berlin (♁ 52° 31′ N, 13° 24′ O) | 11000 | B | – | B | 891,69 | 4338,8 | 3336,0 | 3200,7 | 3420,6 | 3501,9 | *3.677.472* | *3948* |  |
|  | Bielefeld (♁ 52° 1′ N, 8° 32′ O) | 05711 | NW | Detmold | BI | 258,83 | 129,5 | 153,6 | 168,6 | 317,2 | 323,4 | 334.002 (2021) | 1290 |  |

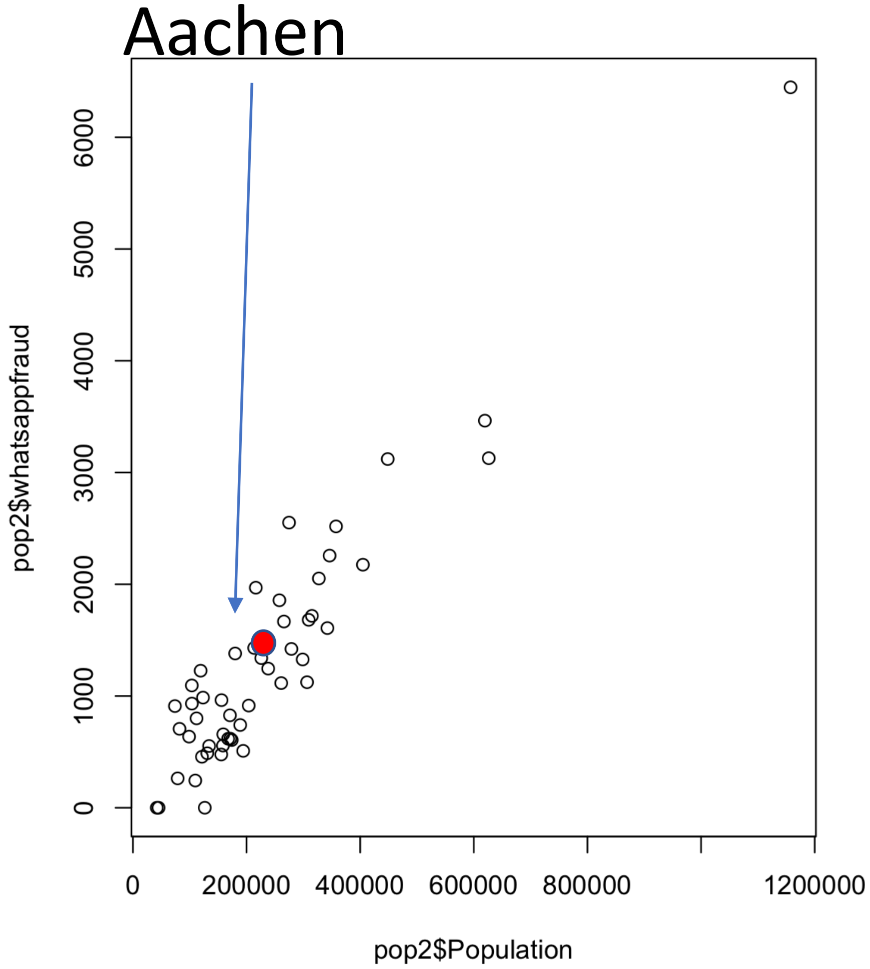
# Estimate at individual cluster level

- Imagine we select the city Aachen as one cluster
  - We draw an SRS of 2.000 households
  - Conduct the survey: 1.000 households participate
  - We find that 12 people were victim of Whatsapp fraud last year
  - What is the total number of Whatsapp frauds in Aachen?

- Number of individuals in selected households: 2123
  - 12/2123 = .56% of individuals experiences Whatsapp fraud
- Number of whatsapp fraud in AAchen= .0056 * 249070 = 1394

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aachen[1] (♁ 50° 47′ N, 6° 5′ O) | 05334 | NW | Köln | AC, MON | 160,85 | 162,2 | 129,8 | 175,5 | 239,2 | 260,5 | 249.070 (2021) | | 1548 |

# Estimate at population level




Mean pop size of cluster = 235509
Mean whatsapp fraud = 1307

Ratio = 180/ 1

Or .00555 of population

Germany: population is 83 Million
Whatapp fraud is 83M/180 = 461k

# Why ratio estimation?

- We know:
  - The size of each farm in the USA    Auxiliary information at level of farm
    - $N_h$ and $n_h$
- Estimate from a sample:
  - What crops they produce
  - What is their yield per acre (or total production)

- USA wheat production = wheat production per acre * total # acres of wheat

# Why ratio estimation?

- We know:
  - How many schools there are: # schools
    - $N_h$ (no. of clusters)

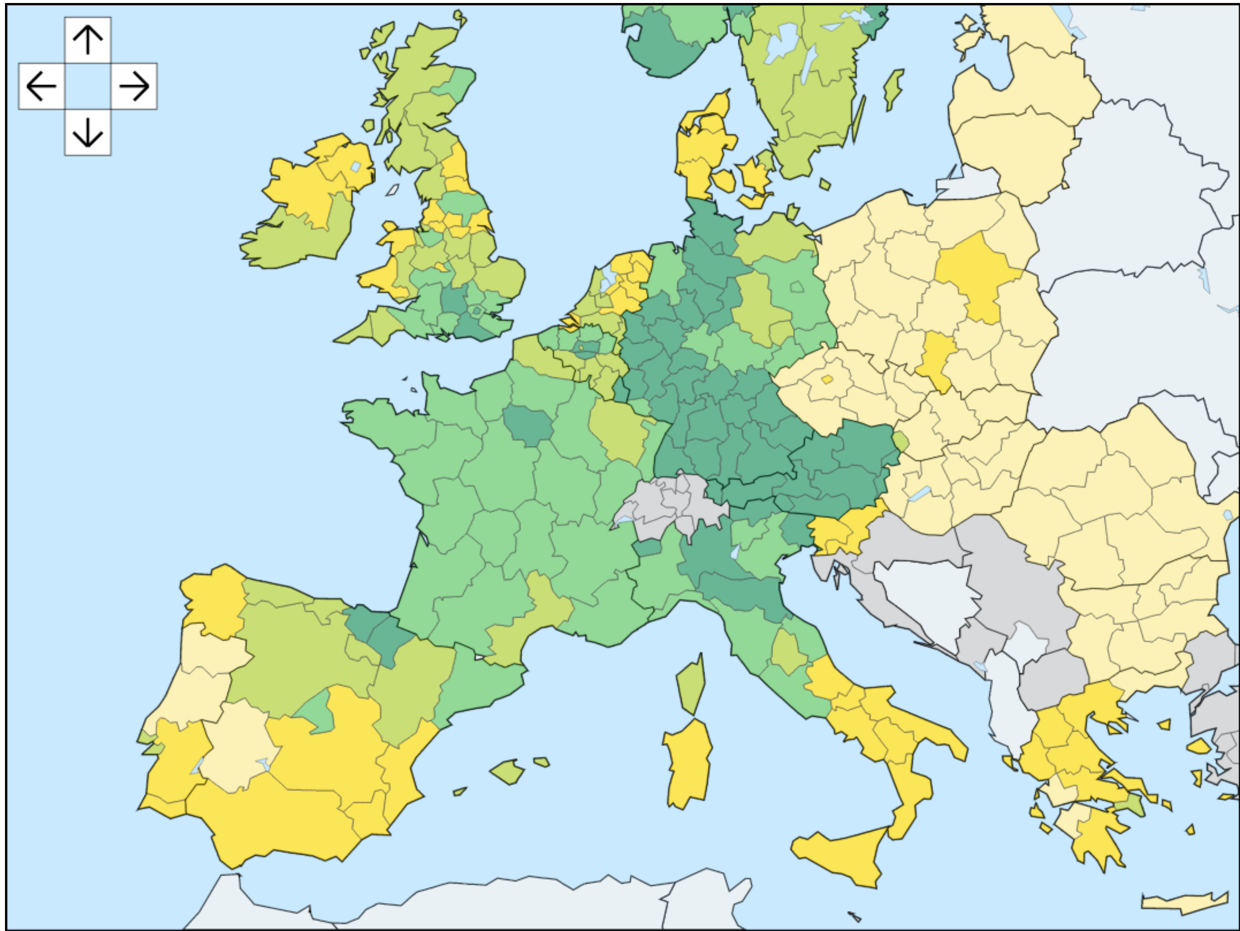
Auxiliary information at level of cluster

- Estimate from a sample:
  - The average number of children per school: $n_h$
  - the proportion with reading problems: p

- Total # children with learning diff = $n_h * N_h * p_{\text{children with reading problems or}}$

# Why so often in cluster samples?

- We often don't know much about individuals
- But we do know about the clusters
  - Public sources:
    - Population size
    - income, employment
    - Gender, age distribution
    - Etc.
- Is Y strongly correlated with these?
  - And a ratio variable?
  - Ratio estimation
- E.g. No. of births, marriages, death

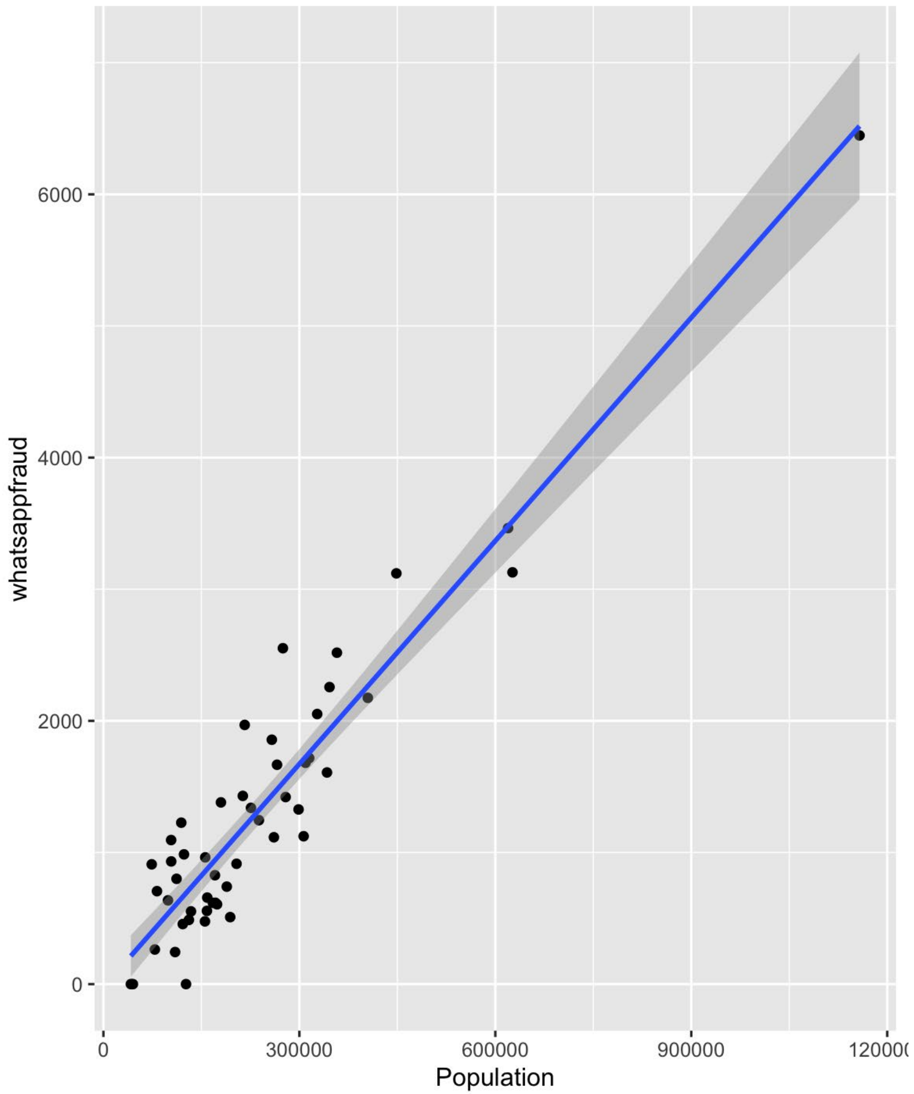


# Class exercise 1

- 25 minutes
- 4 questions…

# What is great in ratio estimation

- We can only sample some clusters
    - BUT: we know the size of each cluster
- Estimate fraud in some clusters
- The ratio $\frac{Population\ size}{Number\ of\ whatsapp\ fraud}$
- Allows us to estimate with great precision
    - We know quite a lot about the clusters we didn't  observe
    - Se = much lower than SRS
    - Design effect very small
    - We can lower sample size, and save $$$

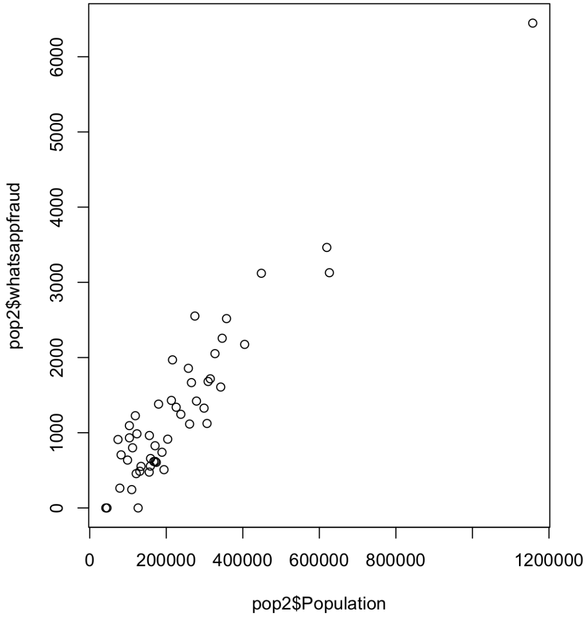


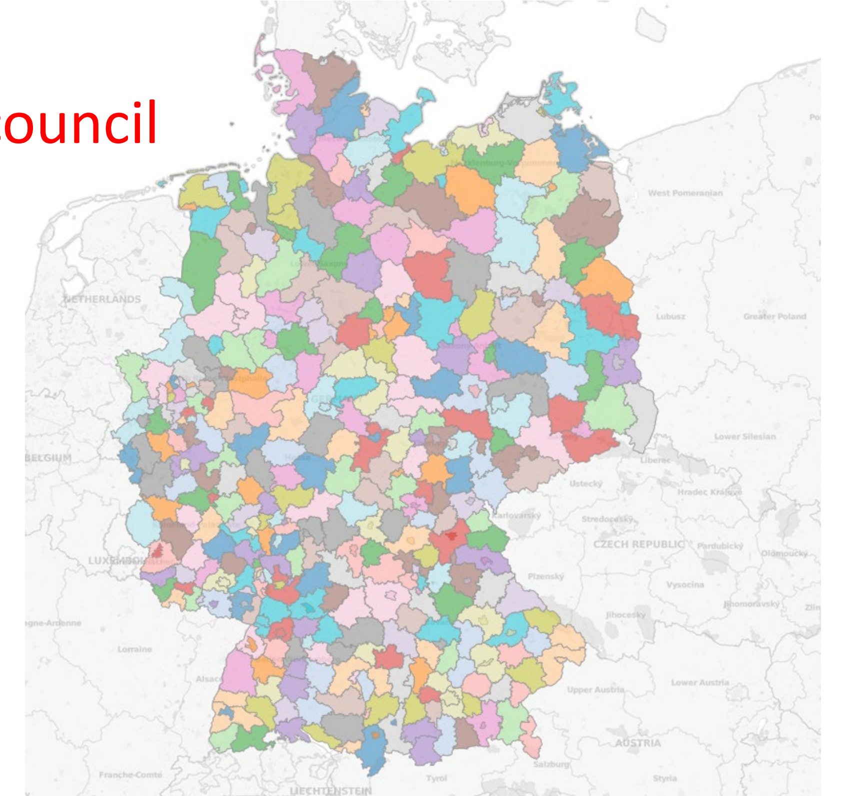
# What more is there?

- Estimating fraud in every cluster
  - Berlin: population 3.7 million. Whatapp fraud 3.7M / 180 = 20558
  - Ansbach: population 41k. Whatsapp fraud 41k/180 = 228

- We can now using a model to predict fraud in every council

| Wappen | Stadt | Stadtkreis ID | BL | Regierungsbezirk[5] | UZ | Fläche in km² | tEW 1939 [6] | tEW 1950 [7] | tEW 1970 [8] | tEW 1990 [9] | tEW 2011 [10] | EW jetzt | Bev.-dichte (EW/km²) | Lagekarte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Aachen[1] (♁ 50° 47′ N, 6° 5′ O) | 05334 | NW | Köln | AC, MON | 160,85 | 162,2 | 129,8 | 175,5 | 239,2 | 260,5 | 249.070 (2021) | 1548 | |
| | Amberg (♁ 49° 27′ N, 11° 51′ O) | 09361 | BY | Oberpfalz | AM | 50,13 | 31,8 | 37,9 | 41,3 | 42,9 | 43,5 | 41.994 (2021) | 838 | |
| | Ansbach (♁ 49° 18′ N, 10° 34′ O) | 09561 | BY | Mittelfranken | AN | 99,91 | 26,0 | 33,2 | 33,2 | 37,6 | 40,3 | 41.662 (2021) | 417 | |
| | Aschaffenburg (♁ 49° 59′ N, 9° 9′ O) | 09661 | BY | Unterfranken | AB | 62,45 | 45,4 | 45,5 | 55,1 | 63,6 | 68,8 | 71.381 (2021) | 1143 | |
| | Augsburg (♁ 48° 22′ N, 10° 54′ O) | 09761 | BY | Schwaben | A | 146,85 | 185,4 | 185,2 | 213,2 | 254,3 | 266,6 | 296.478 (2021) | 2019 | |
| | Baden-Baden (♁ 48° 46′ N, 8° 14′ O) | 08211 | BW | Karlsruhe | BAD | 140,19 | 33,2 | 36,6 | 37,2 | 51,5 | 54,5 | 55.527 (2021) | 396 | |
| | Bamberg (♁ 49° 54′ N, 10° 54′ O) | 09461 | BY | Oberfranken | BA | 54,62 | 59,5 | 76,2 | 70,4 | 70,2 | 70,1 | 77.749 (2021) | 1423 | |
| | Bayreuth (♁ 49° 57′ N, 11° 35′ O) | 09462 | BY | Oberfranken | BT | 66,89 | 45,0 | 58,8 | 64,2 | 72,0 | 73,1 | 73.909 (2021) | 1105 | |
| | Berlin (♁ 52° 31′ N, 13° 24′ O) | 11000 | B | – | B | 891,69 | 4338,8 | 3336,0 | 3200,7 | 3420,6 | 3501,9 | *3.677.472* | *3948* | |
| | Bielefeld (♁ 52° 1′ N, 8° 32′ O) | 05711 | NW | Detmold | BI | 258,83 | 129,5 | 153,6 | 168,6 | 317,2 | 323,4 | 334.002 (2021) | 1290 | |

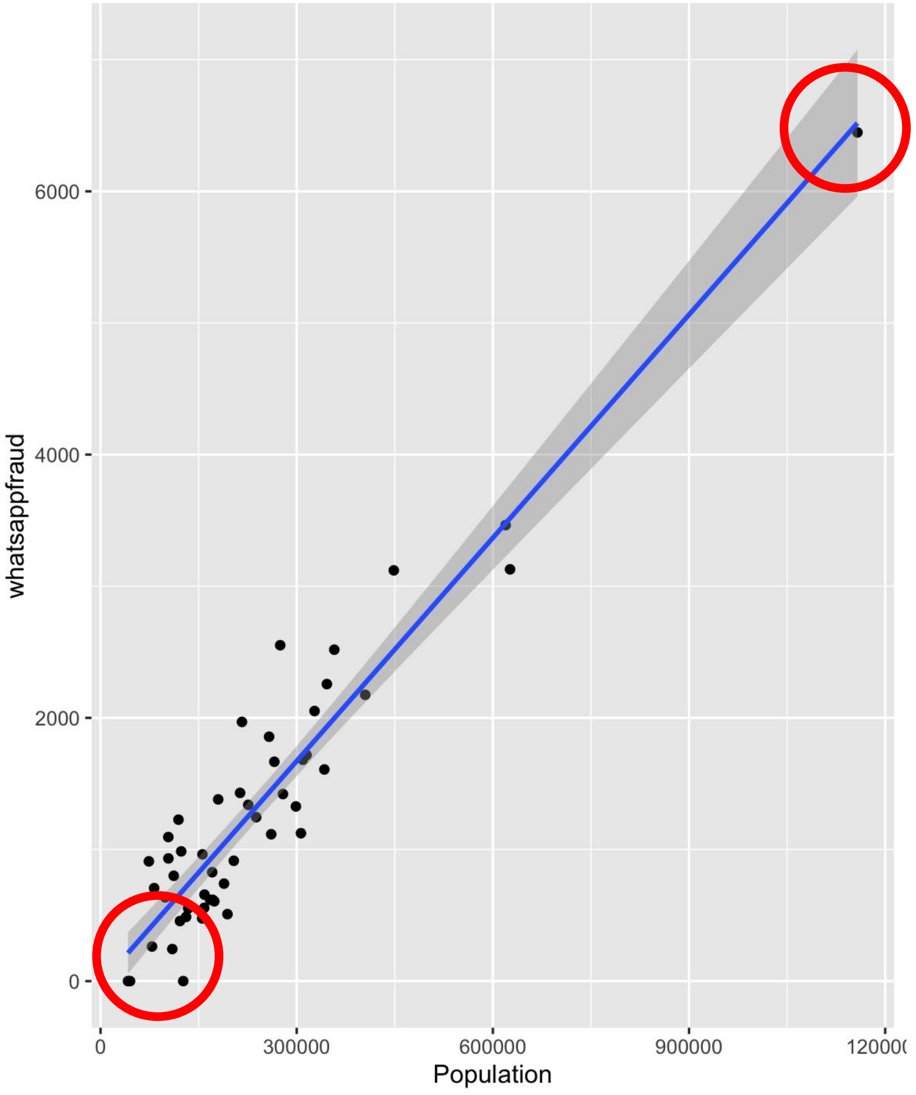


# What is a problem in ratio estimation

Question 4: class exercise

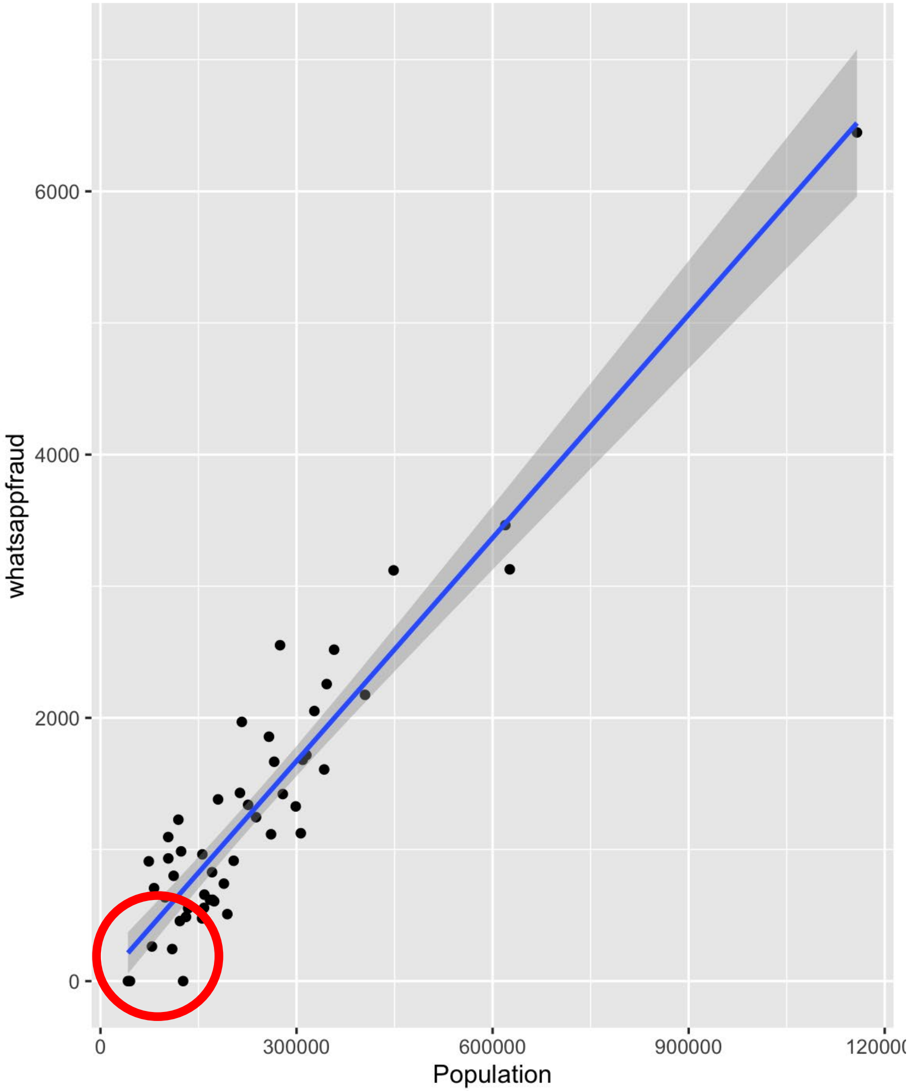
- There may be bias!
  - Outlier clusters
    - Large cities drive results
    - Whatsapp fraud may be local
  - Population size = 0 doesnt happen, but whatsapp fraud = 0 does!
    - The origin does not really exist



# What about making the model more complex?

- What about including other covariates?
  - Urban/rural
  - Average income of cluster
  - State of the council
  - Etc.

- We build a regression model
  - More covariates
  - Why not an intercept?



# Model-based estimation

- Using a survey from some clusters….
- We try to predict Fraud in other clusters
- And the sum of all predictions is the total

| Wappen | Stadt | Stadtkreis ID | BL | Regierungsbezirk[5] | UZ | Fläche in km² | tEW 1939 [6] | tEW 1950 [7] | tEW 1970 [8] | tEW 1990 [9] | tEW 2011 [10] | EW jetzt | Bev.-dichte (EW/km²) | Lagekarte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Aachen[1] (50° 47′ N, 6° 5′ O) | 05334 | NW | Köln | AC, MON | 160,85 | 162,2 | 129,8 | 175,5 | 239,2 | 260,5 | 249.070 (2021) | 1548 | |
| | Amberg (49° 27′ N, 11° 51′ O) | 09361 | BY | Oberpfalz | AM | 50,13 | 31,8 | 37,9 | 41,3 | 42,9 | 43,5 | 41.994 (2021) | 838 | |
| | Ansbach (49° 18′ N, 10° 34′ O) | 09561 | BY | Mittelfranken | AN | 99,91 | 26,0 | 33,2 | 33,2 | 37,6 | 40,3 | 41.662 (2021) | 417 | |
| | Aschaffenburg (49° 59′ N, 9° 9′ O) | 09661 | BY | Unterfranken | AB | 62,45 | 45,4 | 45,5 | 55,1 | 63,6 | 68,8 | 71.381 (2021) | 1143 | |
| | Augsburg (48° 22′ N, 10° 54′ O) | 09761 | BY | Schwaben | A | 146,85 | 185,4 | 185,2 | 213,2 | 254,3 | 266,6 | 296.478 (2021) | 2019 | |
| | Baden-Baden (48° 46′ N, 8° 14′ O) | 08211 | BW | Karlsruhe | BAD | 140,19 | 33,2 | 36,6 | 37,2 | 51,5 | 54,5 | 55.527 (2021) | 396 | |
| | Bamberg (49° 54′ N, 10° 54′ O) | 09461 | BY | Oberfranken | BA | 54,62 | 59,5 | 76,2 | 70,4 | 70,2 | 70,1 | 77.749 (2021) | 1423 | |
| | Bayreuth (49° 57′ N, 11° 35′ O) | 09462 | BY | Oberfranken | BT | 66,89 | 45,0 | 58,8 | 64,2 | 72,0 | 73,1 | 73.909 (2021) | 1105 | |
| | Berlin (52° 31′ N, 13° 24′ O) | 11000 | B | – | B | 891,69 | 4338,8 | 3336,0 | 3200,7 | 3420,6 | 3501,9 | *3.677.472* | *3948* | |
| | Bielefeld (52° 1′ N, 8° 32′ O) | 05711 | NW | Detmold | BI | 258,83 | 129,5 | 153,6 | 168,6 | 317,2 | 323,4 | 334.002 (2021) | 1290 | |

# Class exercise 2

- Regression estimation in practice
- 30 minutes

# Design-based versus model-based

**Variance of the estimator:**

**Design-based:**
Average squared deviation of the estimate and the expected value, averaged over all possible samples under the sampling design
(i.e. we repeat the sampling procedure 10000 times, and estimate variance in the total)

**Model-based**
Average squared deviation of the estimate and the expected value, averaged over all possible samples under the model
(i.e. we assume the model is correct, and sample 10000 times new observations, fit the regression line, and estimate variance in total)

# When ratio vs. regression?

Ratio

- Size of area/no. of buildings -> people in a certain area
- Turnover per company/no. of peppers -> total pepper production

Often, good frame information, and a meaningful 0

Regression

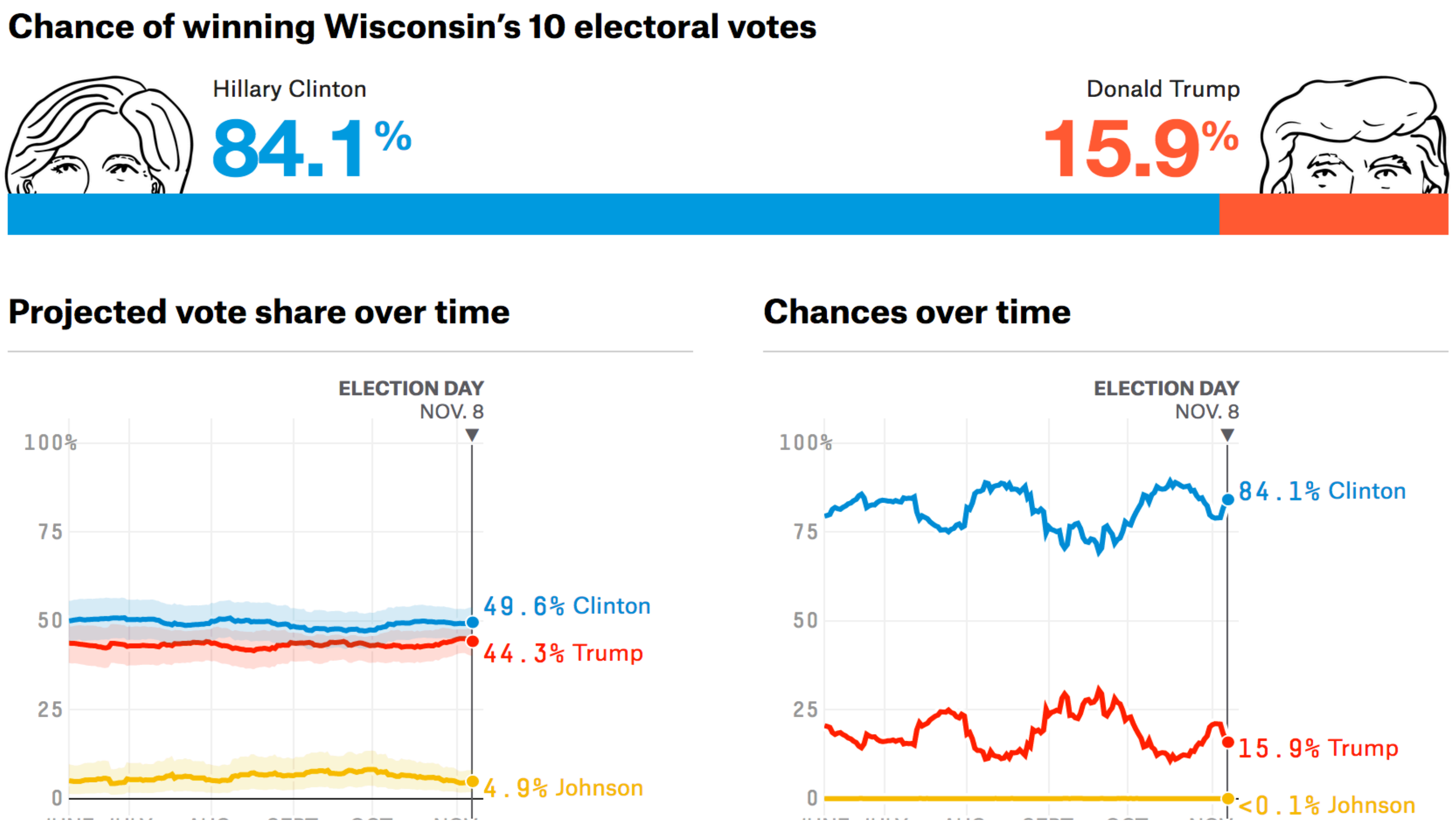
- Happiness <- grades:gender:income:sociallife
- Vote <- race:age:gender:education
- Fraud <-population:urban:incomes

Often, little good frame information, no meaningful 0

# Implicationss of going model-based

- Sampling is not so important!
  - We just get data, and as long as we are confident that our model is correct **in the population**, we are fine…
- We need a good (regression) model for Y
- We need to worry about sample <-> population
  - On a more conceptual level, not about inclusion probabilities
  - Sample should capture variation
  - Selection bias, nonresponse
- From now on: more focus on model-based inference
  - Nonresponse model -> weights
  - Missing data model -> imputation
  - Selection bias model -> ???

# Model-based inference – an example




How did it end?

# Wisconsin – election outcome 2016

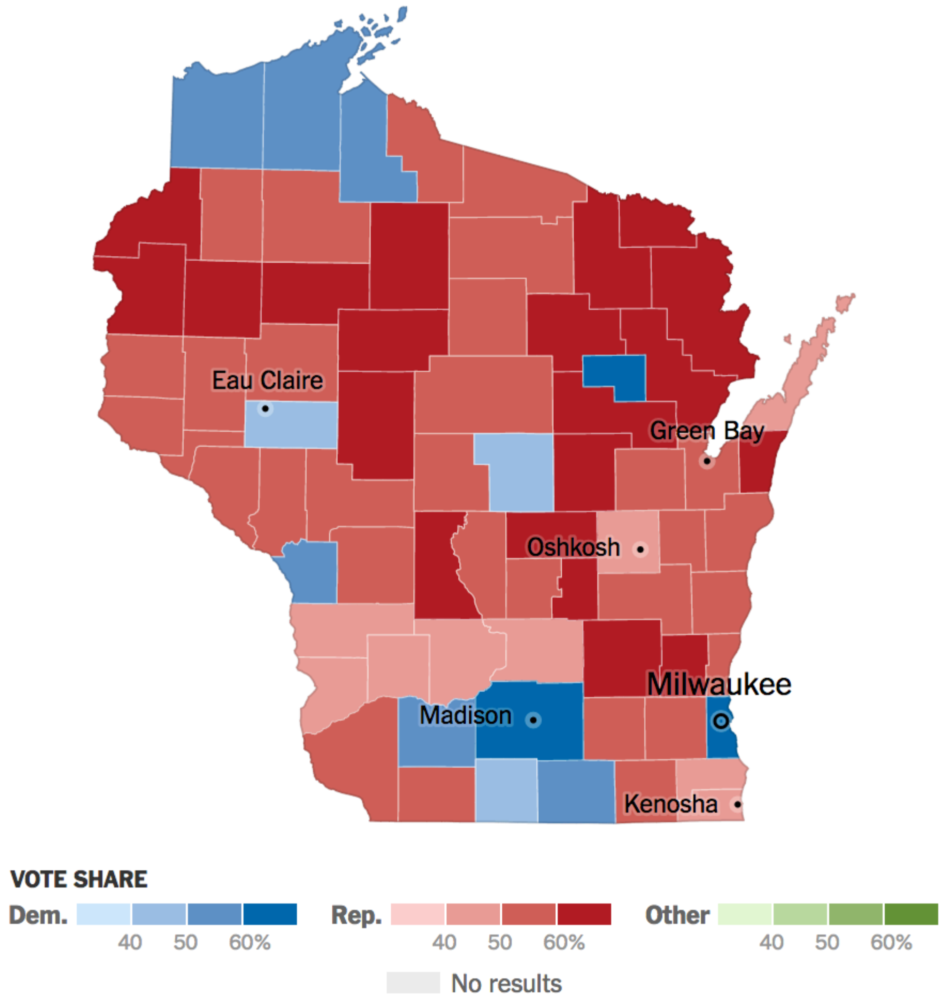
## President

| CANDIDATE | PARTY | VOTES | PCT. | E.V. |
|---|---|---|---|---|
| ✔ Donald J. Trump | Republican | 1,405,284 | 47.2% | 10 |
| Hillary Clinton | Democrat | 1,382,536 | 46.5 | — |
| Gary Johnson | Libertarian | 106,674 | 3.6 | — |
| Others | Independent | 35,150 | 1.2 | — |
| ▼ Others | | 46,506 | 1.6 | — |

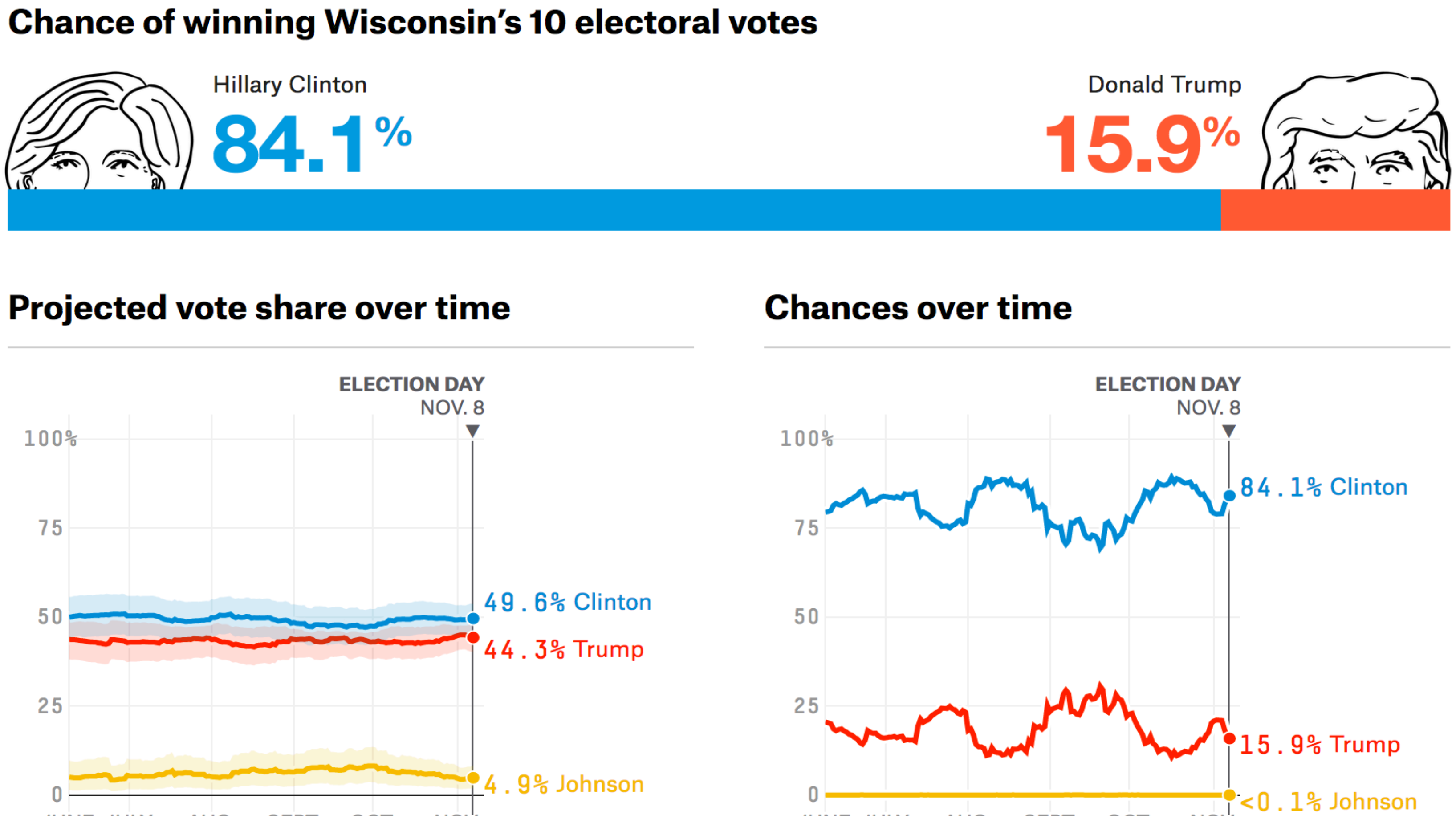
100% reporting (3,620 of 3,620 precincts)

President Map »

**Race Preview:** Wisconsin, a competitive state that leans Democratic, has 10 electoral votes. With a large population of white, working-class Democrats, it seemed promising for Mr. Trump, but he has struggled with Republican-leaning voters in the Milwaukee suburbs. Barack Obama won Wisconsin in 2012 by 6.9 percentage points.



# Model-based inference – an example




If this were an individual poll of n=10000

s.e. = $\sqrt{p(1-p)/n}$

= sqrt(.496(1-.504)/10000)
= .005

Clinton Vote CI:
[.4904 - .5096]

# How does political polling in the USA work?

| DATES | POLLSTER | GRADE | SAMPLE | WEIGHT | CLINTON | TRUMP | JOHNSON | LEADER | ADJUSTED LEADER |
|---|---|---|---|---|---|---|---|---|---|
| OCT. 26-31 | Marquette University | A | 1,255 LV | 3.79 | 46% | 40% | 4% | Clinton +6 | Clinton +5 |
| NOV. 1-2 | Remington | | 2,720 LV | 3.26 | 49% | 41% | | Clinton +8 | Clinton +9 |
| NOV. 1-2 | Clarity Campaign Labs | B | 1,129 LV | 2.99 | 47% | 43% | 4% | Clinton +4 | Clinton +5 |
| NOV. 3-6 | Gravis Marketing | B- | 1,184 RV | 2.84 | 47% | 44% | 3% | Clinton +3 | Clinton +4 |
| OCT. 31-NOV. 1 | Public Policy Polling | B+ | 891 LV | 2.81 | 48% | 41% | | Clinton +7 | Clinton +7 |
| NOV. 1-7 | SurveyMonkey | C- | 2,246 LV | 2.53 | 44% | 42% | 7% | Clinton +2 | Clinton +1 |
| OCT. 31-NOV. 1 | Loras College | B- | 500 LV | 1.62 | 44% | 38% | 7% | Clinton +6 | Clinton +5 |
| OCT. 27-28 | Emerson College | B | 400 LV | 1.23 | 48% | 42% | 9% | Clinton +6 | Clinton +7 |
| OCT. 13-16 | St. Norbert College | A- | 664 LV | 1.20 | 47% | 39% | 1% | Clinton +8 | Clinton +5 |
| NOV. 1-7 | Google Consumer Surveys | B | 914 LV | 1.03 | 43% | 31% | 4% | Clinton +12 | Clinton +12 |
| OCT. 15-18 | Monmouth University | A+ | 403 LV | 0.98 | 47% | 40% | 6% | Clinton +7 | Clinton +4 |
| OCT. 5-7 | YouGov | B | 993 LV | 0.93 | 43% | 39% | 4% | Clinton +4 | Clinton +2 |
| OCT. 24-NOV. 6 | Ipsos | A- | 625 LV | 0.92 | 46% | 40% | | Clinton +6 | Clinton +6 |
| OCT. 18-20 | McLaughlin & Associates | C- | 600 LV | 0.85 | 48% | 43% | 4% | Clinton +5 | Clinton +3 |
| OCT. 18-19 | Public Policy Polling | B+ | 804 LV | 0.73 | 50% | 38% | | Clinton +12 | Clinton +9 |

Multiple polls
Weighted by:

- Quality of organisation (grade)
- Recency

Results presented is aggregated total

# But forecasters do not stop there…

| | CLINTON | TRUMP | JOHNSON |
|---|---|---|---|
| **1. Polling average** | 46.4% | 40.5% | 4.9% |
| Adjust for likely voters | +0.1 | +0.2 | -0.1 |
| Adjust for convention bounce | -0.0 | +0.0 | +0.0 |
| Adjust for vice-presidential selection | -0.0 | +0.0 | +0.0 |
| Adjust for omitted third parties | -0.2 | -0.2 | +0.0 |
| Adjust for trend line | +0.3 | +1.0 | -0.7 |
| Adjust for house effects | -0.2 | -0.5 | +0.1 |
| **2. Adjusted polling average** | 46.4% | 41.0% | 4.2% |
| Allocate undecided and third-party voters | +3.3 | +3.3 | +0.5 |
| **3. Polls-based vote share** | 49.6% | 44.2% | 4.8% |
| Calculate demographic regression | 49.6% | 44.2% | 5.5% |
| **4. Polls- and demographics-based projection** Weighted average 91% polls-based, 9% demographics | 49.6% | 44.2% | 4.9% |
| Calculate fundamentals forecast | 47.6% | 46.3% | 4.9% |
| **5. Projected vote share for Nov. 8** Weighted average 99% polls/demographics, 1% fundamentals | 49.6% | 44.3% | 4.9% |

**Adjustments for:**

- Likely voters
  - Not all people are likely to go and vote
- Omitted third parties
  - Not all polls ask for all parties
- Adjust for trend line
  - A smoothing adjustment to avoid large fluctuations
- House effects
  - Some pollsters are known to have a bias

# But forecasters do not stop there…

| | CLINTON | TRUMP | JOHNSON |
|---|---|---|---|
| **1. Polling average** | 46.4% | 40.5% | 4.9% |
| Adjust for likely voters | +0.1 | +0.2 | -0.1 |
| Adjust for convention bounce | -0.0 | +0.0 | +0.0 |
| Adjust for vice-presidential selection | -0.0 | +0.0 | +0.0 |
| Adjust for omitted third parties | -0.2 | -0.2 | +0.0 |
| Adjust for trend line | +0.3 | +1.0 | -0.7 |
| Adjust for house effects | -0.2 | -0.5 | +0.1 |
| **2. Adjusted polling average** | 46.4% | 41.0% | 4.2% |
| Allocate undecided and third-party voters | +3.3 | +3.3 | +0.5 |
| **3. Polls-based vote share** | 49.6% | 44.2% | 4.8% |
| Calculate demographic regression | 49.6% | 44.2% | 5.5% |
| **4. Polls- and demographics-based projection**<br>Weighted average 91% polls-based, 9% demographics | 49.6% | 44.2% | 4.9% |
| Calculate fundamentals forecast | 47.6% | 46.3% | 4.9% |
| **5. Projected vote share for Nov. 8**<br>Weighted average 99% polls/demographics, 1% fundamentals | 49.6% | 44.3% | 4.9% |

**Adjustments for:**

- Undecideds
  - Assumption about how “don’t know” answers will vote

# But forecasters do not stop there...

| | CLINTON | TRUMP | JOHNSON |
|---|---|---|---|
| **1. Polling average** | 46.4% | 40.5% | 4.9% |
| Adjust for likely voters | +0.1 | +0.2 | -0.1 |
| Adjust for convention bounce | -0.0 | +0.0 | +0.0 |
| Adjust for vice-presidential selection | -0.0 | +0.0 | +0.0 |
| Adjust for omitted third parties | -0.2 | -0.2 | +0.0 |
| Adjust for trend line | +0.3 | +1.0 | -0.7 |
| Adjust for house effects | -0.2 | -0.5 | +0.1 |
| **2. Adjusted polling average** | 46.4% | 41.0% | 4.2% |
| Allocate undecided and third-party voters | +3.3 | +3.3 | +0.5 |
| **3. Polls-based vote share** | 49.6% | 44.2% | 4.8% |
| Calculate demographic regression | 49.6% | 44.2% | 5.5% |
| **4. Polls- and demographics-based projection** Weighted average 91% polls-based, 9% demographics | 49.6% | 44.2% | 4.9% |
| Calculate fundamentals forecast | 47.6% | 46.3% | 4.9% |
| **5. Projected vote share for Nov. 8** Weighted average 99% polls/demographics, 1% fundamentals | 49.6% | 44.3% | 4.9% |

**Demographic regression**

Use data from other states:

1. Fit a model with demographics
   (ethnicity, age, college degree, income)
2. What is predicted vote in Wisconsin?
3. Mix the poll outcome with model-based outcome

# Why were the polls wrong?

- It wasn't all the modeling.....
  - Polls only: 46 vs. 40 – result: 46.5 vs. 47.2
  - + modeling: 49 vs. 44

AAPOR report (Kennedy et al, 2017) – week 1

- Shy Trump vote
- Low turnout
- Late swing to Trump
- Failure to correct for overrepresentation of highly educated

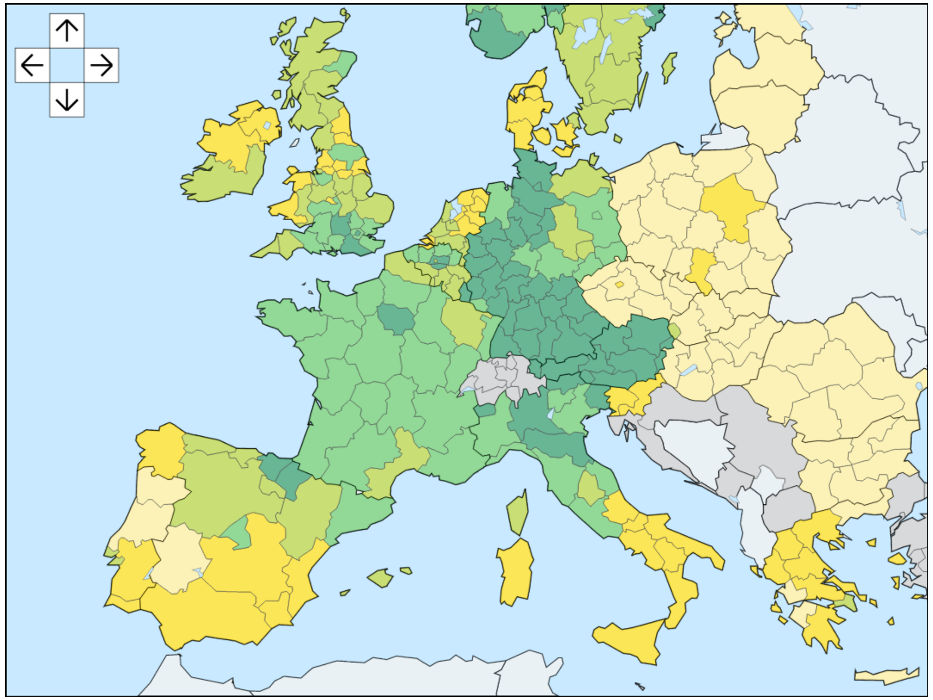
# Why were the polls wrong?

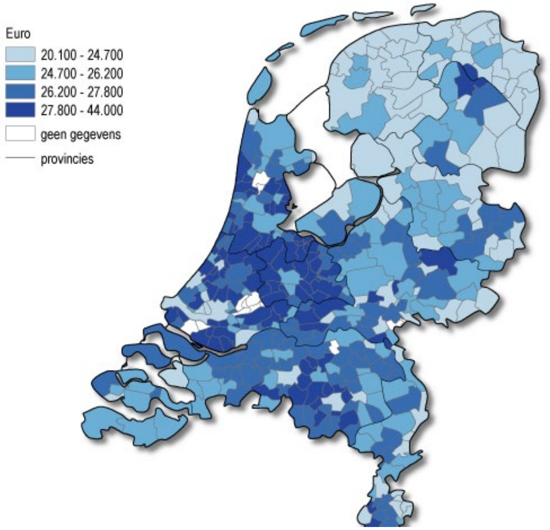
- Model based estimation depends on quality of model!
  - In design based, we can estimate error
  - In model-based -> much more difficult

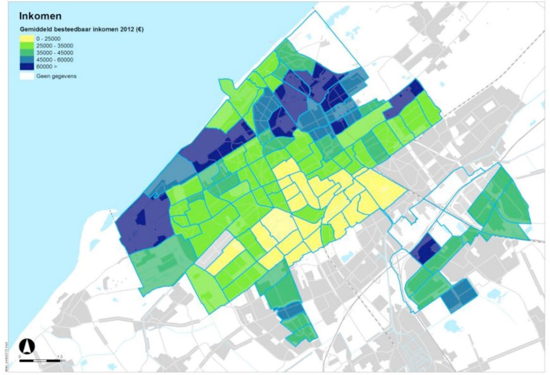
- Why not do design-based inference?
  - Costs
  - Time
  - Problems with coverage, nonresponse
  - -> still needs modeling
  - There are too many people who want to do a a poll
    - 100s in Wisconsin alone

# What is a cluster?

What size should be a cluster be?



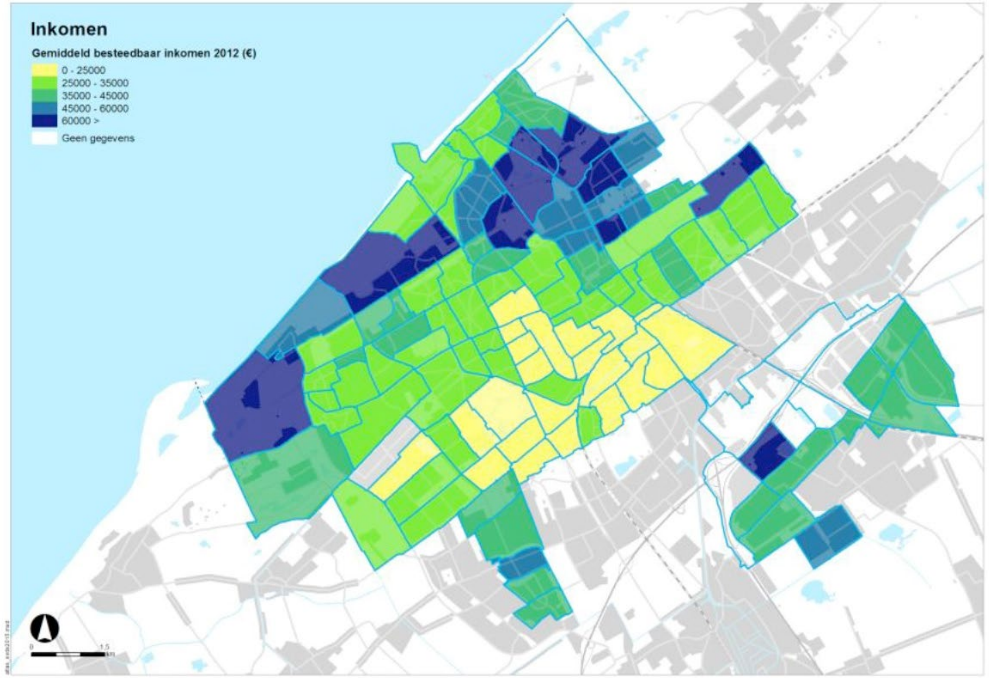


# Small Area Estimation

- Desire for detailed statistics at low geographical level.
- Would result in 1000s clusters in Netherlands, even more in Europe
- Solution: Small area estimation
  - Analogue to coffee machines example
  - There are 100s of machines at UU
  - Build an elaborate model with many auxiliary variables

- Predict Y in every cluster by using a model

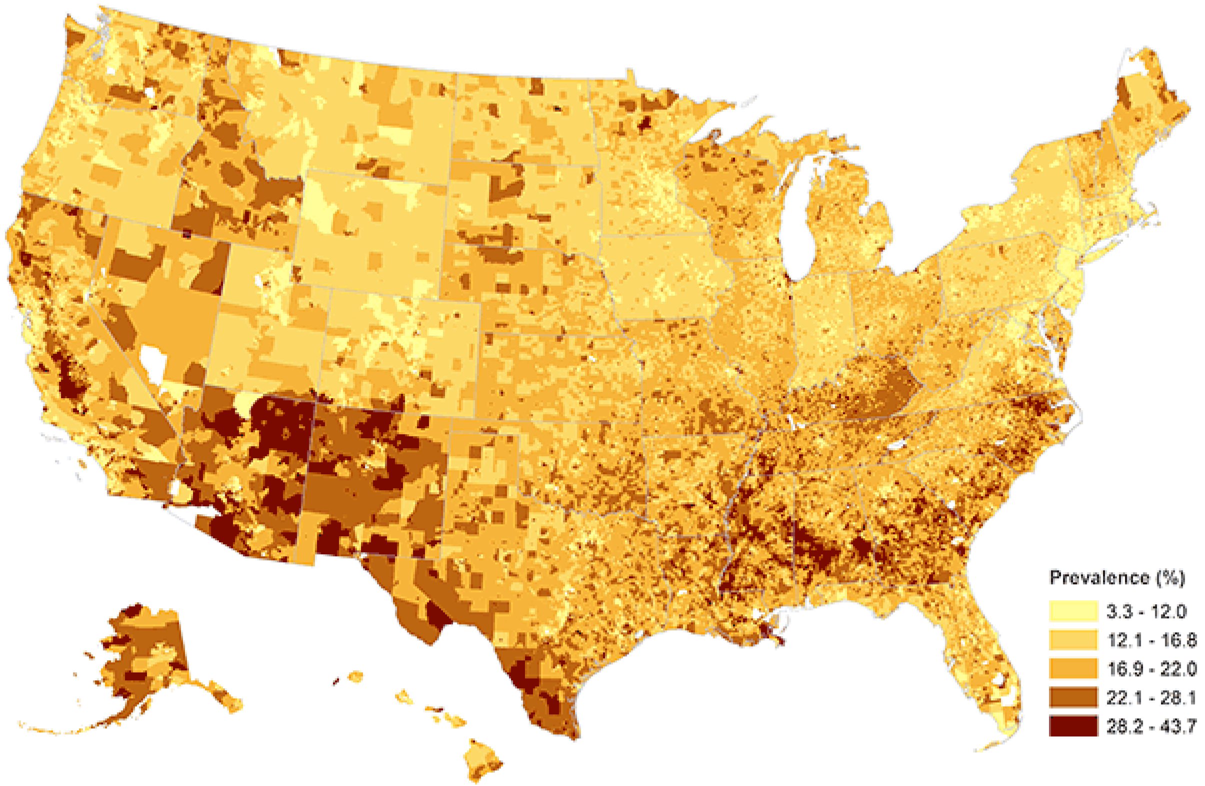


# Example

Childhood obesity

Used 91,642 completed interviews from NCSH survey:

- Model for every county:
- NSCH child obesity status (yes or no) = sex + age + race (individual level)
- + median household income + lifestyle classifications + urbanization levels (zip-code level)
- + median household income + urban-rural (county-level)
- + random effects (state- and county levels)



# Next week(s)

- Next week: class free
  - Finish regression exercise
  - Catch up on reading
- In two weeks: nonresponse

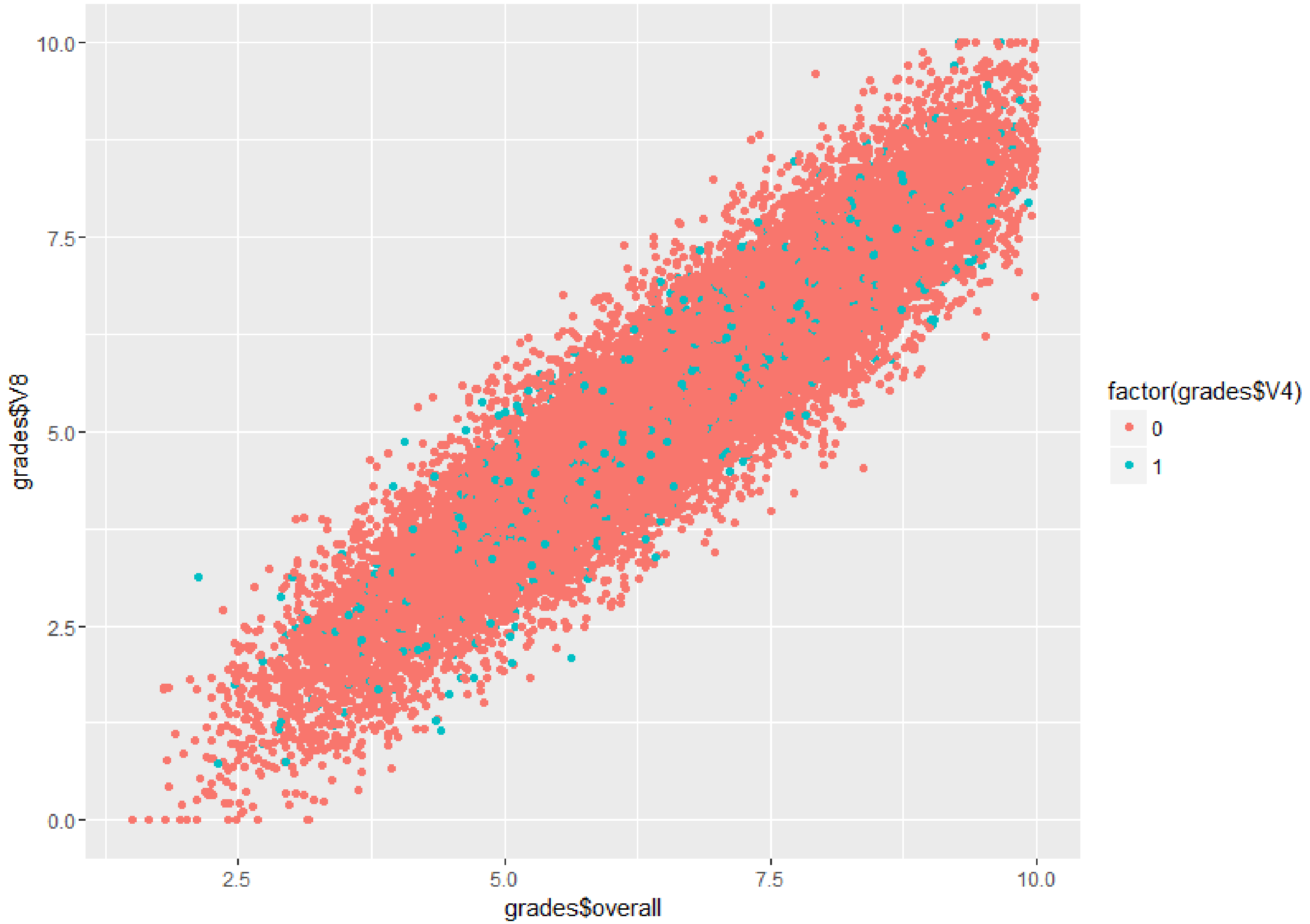
  Readings: several articles
- During class-free week: assignment 1 (!)

# Extra slides

- What goes right and wrong?

# Model based sampling

- Let's bring student happiness in!

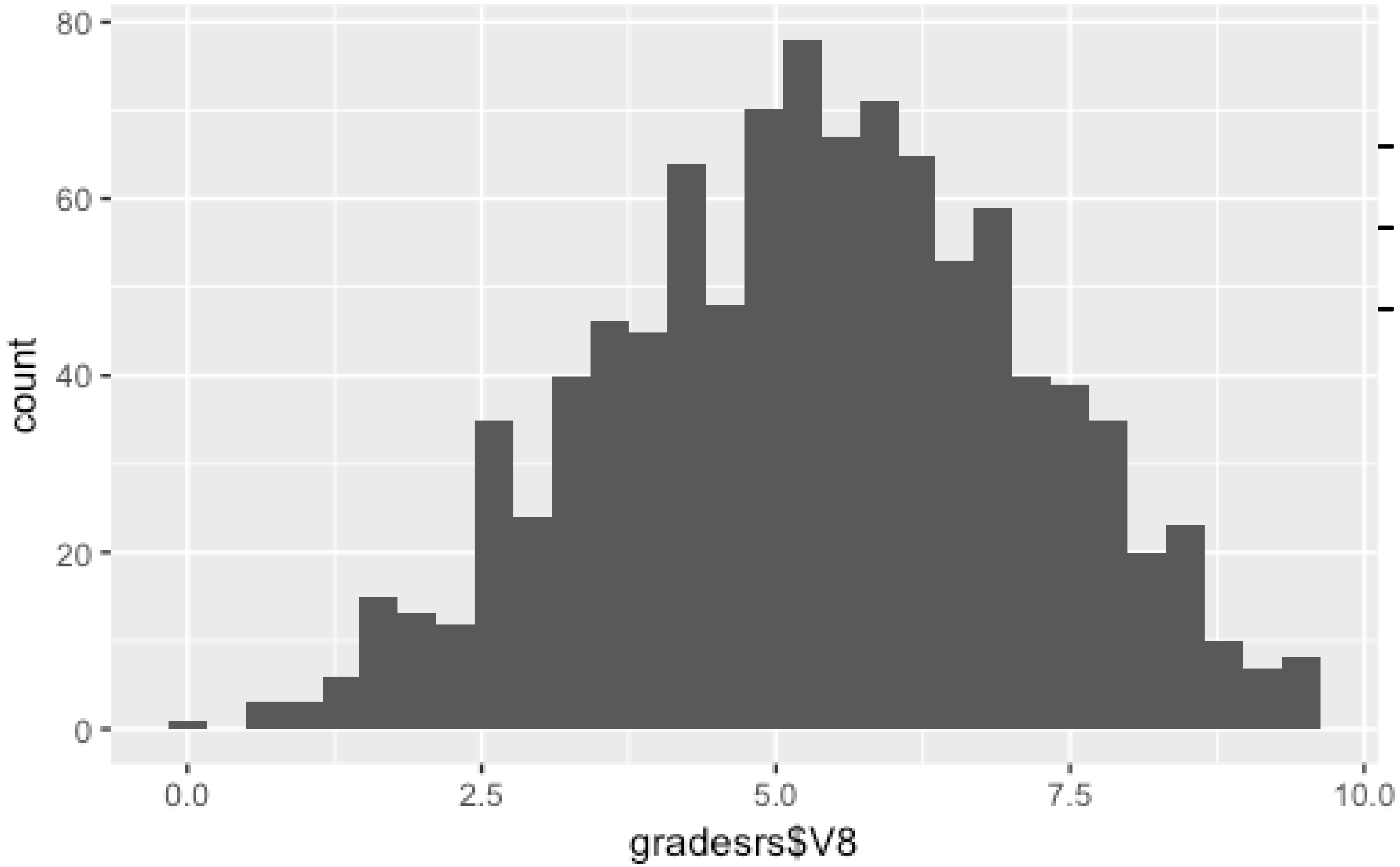



Population data:

- N=20000
- X = grades
- Y = student happiness (also 0-10 scale)

- Mean happiness = 5.37

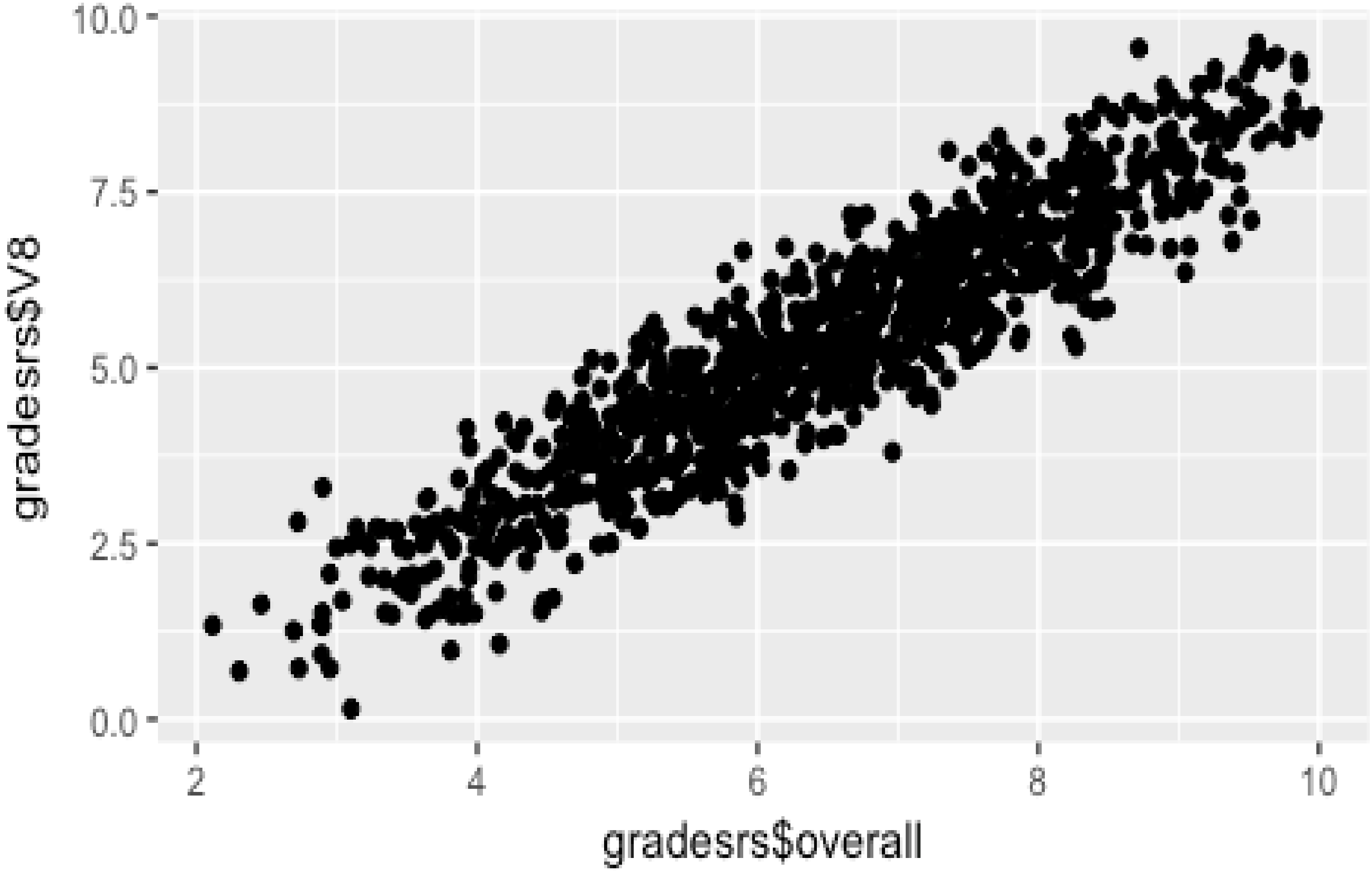
# Simple Random Sampling




- Simple Random Sample
- n=1000
- **Happiness (on X):**
  - Mean = 5.32
  - S.e = .05477

  - CI: [5.21; 5.43]

# Ratio estimation under SRS




Svyratio(~happiness, ~grades, design = ratio.design)

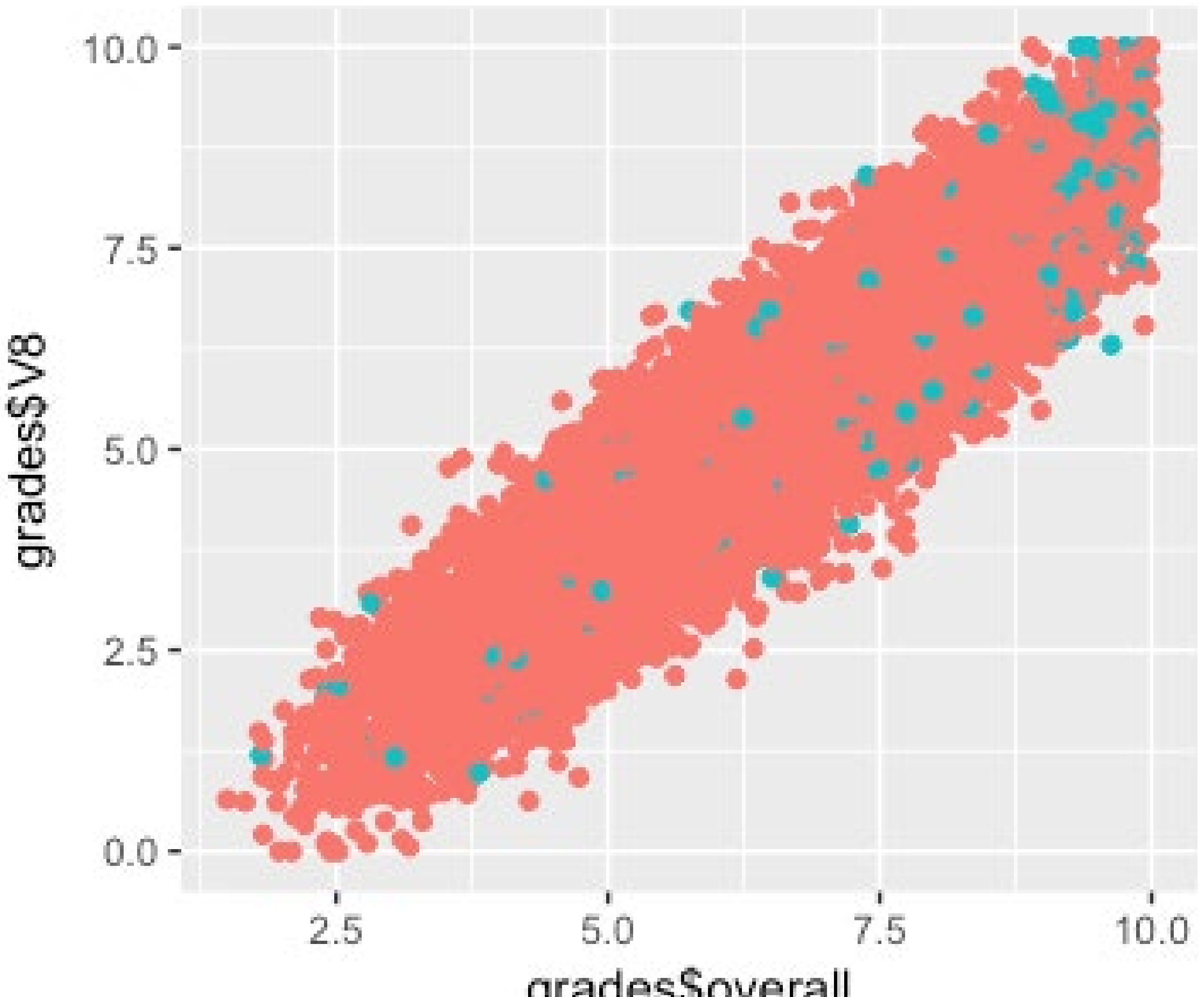
- B= .8231
- s.e. = .0024
- Predicted mean= 5.34

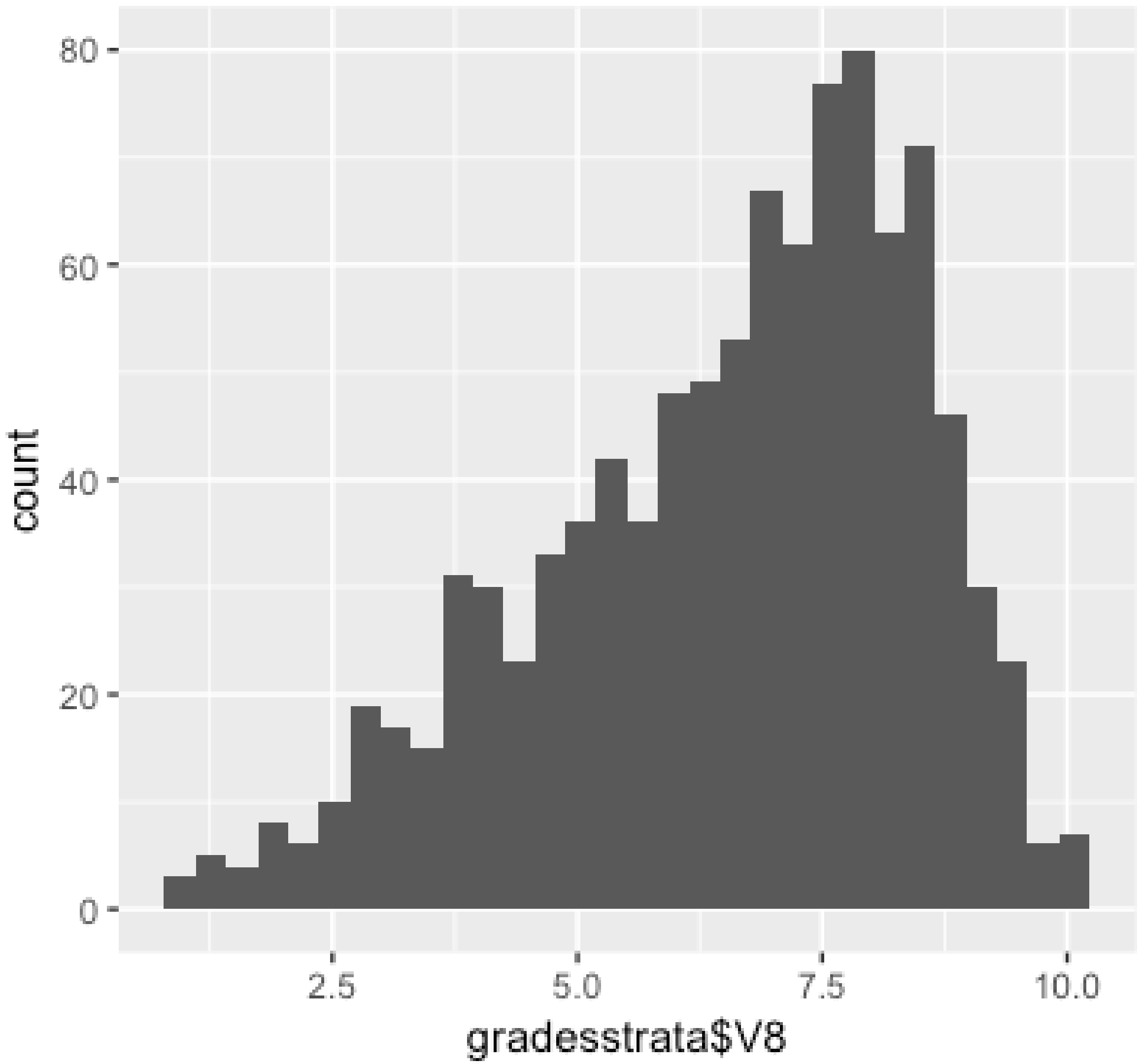
Or:
summary(lm(happiness~0+grades,data =gradesrs, subset=(V4==1)))

Mean(data$fittedvalues )

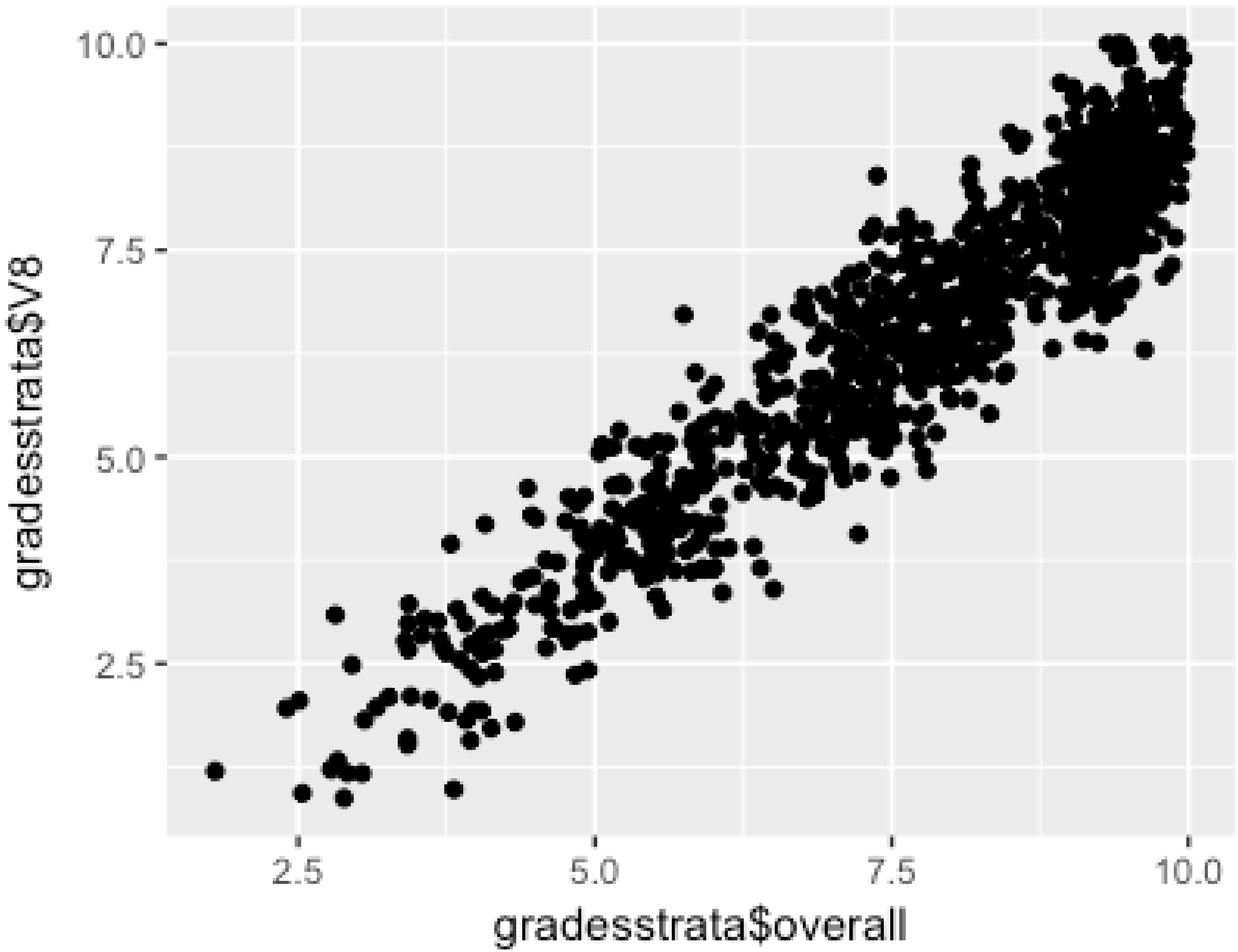
- B = .83
- s.e. =.0036
- Predicted mean = 5.42

# Oversample students who get good grades



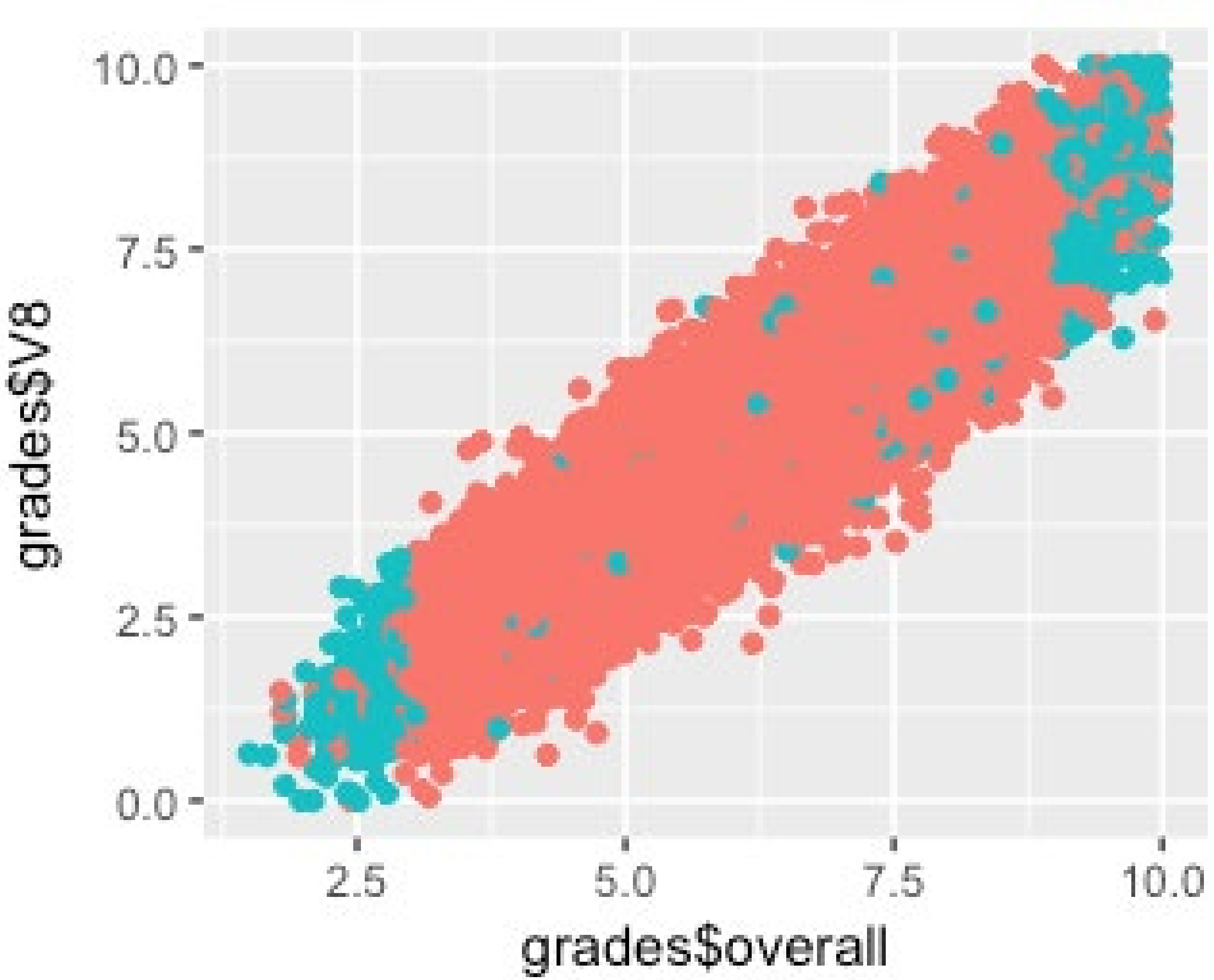



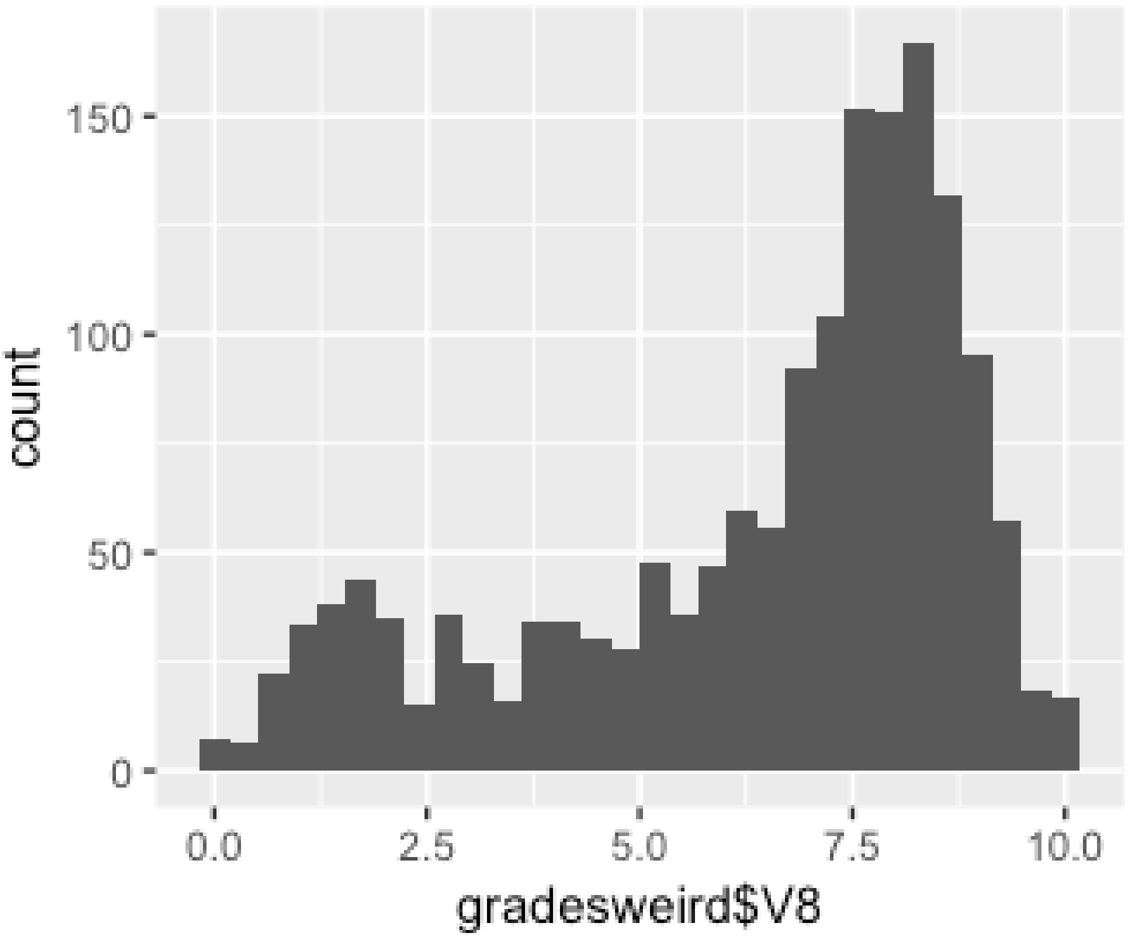
# Oversampling students with good grades



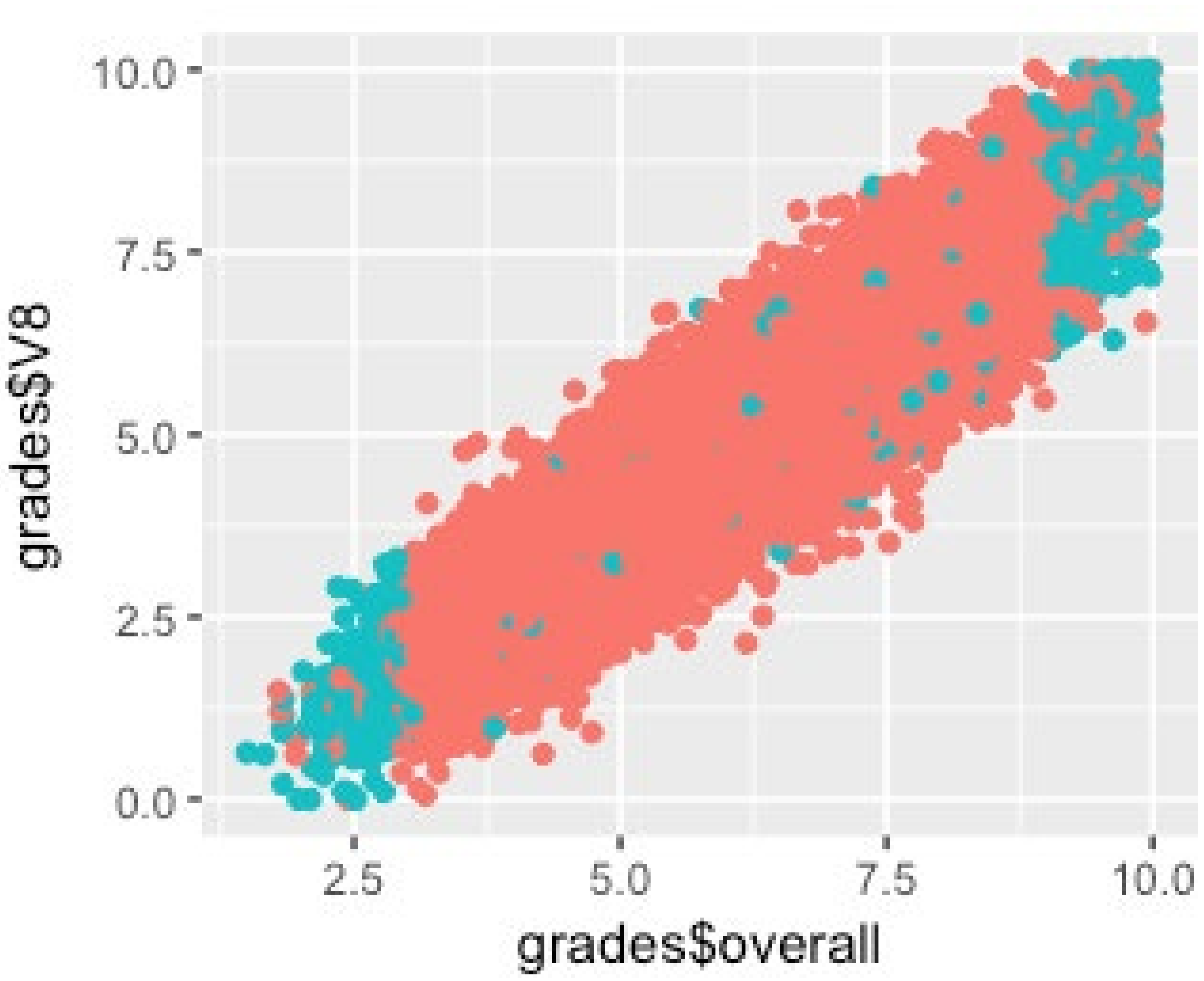

- Design based:
  - Mean = 5.33
  - S.e. = .0337
- Ratio estimation
  - B=.87
  - S.e. = .0026
  - Mean= 5.335
- Regression estimation
  - B = .83
  - S.e. = .0036
  - Mean=5.45

# Truly model based – extreme cases



# Truly model based – regression

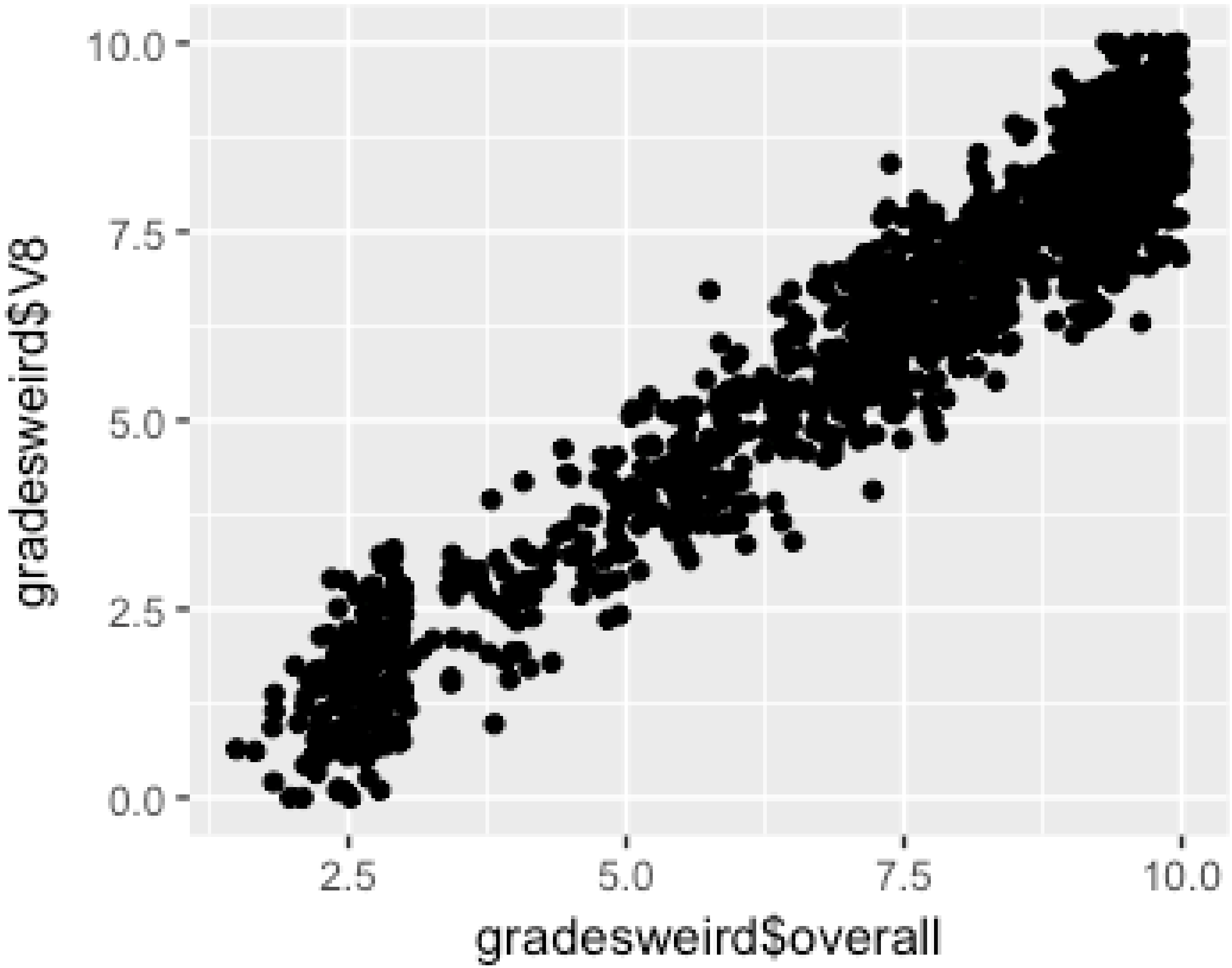



- Regression model:

  Happiness <- grades + programme

  (+ age, gender, etc.)

# Truly model based




- Design based
  - Mean = 5.33
  - S.e. = .0335
- Ratio estimation
  - B=.87
  - S.e. = .0026
  - Mean= 5.33
- Regression estimation
  - B = .87
  - S.e. = .0027
  - Mean=6.28

# What works?

| | Type of sample | Mean | Precision | Mean square error |
|---|---|---|---|---|
| Design Based | SRS | 5.32 | .0548 | $.05^2$ + .05 = .0525 |
| | Oversample good students | 5.33 | .0337 | $.04^2$ +  .03= .0353 |
| | Extreme cases | 5.33 | 0.335 | $.04^2$  + .03 = .0351 |
| Ratio-estimation | SRS | 5.34 | .0027 | $.03^2$ + .0027  = .0036 |
| | Oversample good students | 5.335 | .0026 | $.035^2$ + .0026 = .0037 |
| | Extreme cases | 5.335 | .0026 | $035^2$ + .0026 = .0037 |
| Regression estimation | SRS | 5.42 | .0036 | $.05^2$ + .0036 = .0061 |
| | Oversample good students | 5.45 | .0027 | $.08^2$ + .0027= .0091 |
| | Extreme cases | 6.28 | .0036 | $.93^2$ + .0036 = .8685 |

Notes: Population mean = 5.37. MSE = bias + $se^2$