Survey analysis
week 42
“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

- Your dataset and svydesign settings

Why ratio estimation?

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

Auxiliary information at level of farm

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

Why ratio estimation?

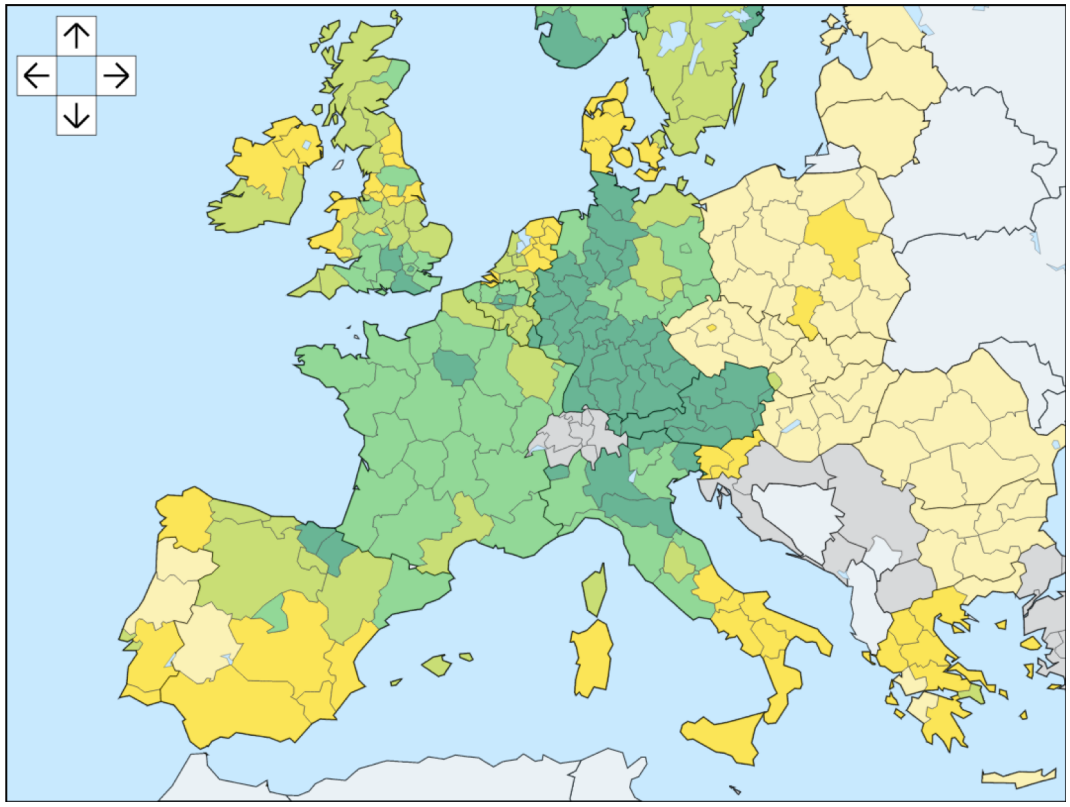
- We know:
  - How many schools there are: # schools
    - N_h (no. of clusters)
- Estimate from a sample:
  - The average number of children per school: n_h
  - the proportion with reading problems: p

Auxiliary information at level of cluster

- 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 short questions…

A more complex example:

estimation in 2-stage cluster sampling

- More realistic
- Time coffee machines in Ruppert take to make coffee

	1	2	3	4	5	6	7
Machine 1							
Machine 2							
Machine 3							
Machine 4							

- We can at most sample 2 machines, and get 10 coffees
- How can we estimate the mean time it will take someone to get a coffee?
  - 1 minute in pairs

Example: 2 clusters (m) of size (n) 2

Some data

- Time coffee machines in Ruppert take to make coffee

	1	2	3	4	5	6	7
Machine 1	20	20	19	21	20		
Machine 2	22	23	22	22	21		
Machine 3							
Machine 4							

- What is estimate of mean(time) for someone to get a coffee?

Example: 2 clusters (m) of size (n) 2

Some data

	1	2	3	4	5	6	7
Machine 1	20	20	19	21	20		
Machine 2	22	23	22	22	21		
Machine 3							
Machine 4							

- What is estimate of Mean(time) for someone to get a coffee?
  - Equal cluster size -> SRS -> mean = 21
  - 4/2* 210. total=420. Mean = total/20 = 21

example: Why equal cluster sizes?

- What is the size of each cluster?

	1	2	3	4	5	6	7
Machine 1	20	21	21	20	20	19	19
Machine 2	22	23	21				
Machine 3							
Machine 4							

- What is estimate of Mean(time) for someone to get a coffee?
  - 1 minute in pairs

We are running to the limits of design based inference

	1	2	3	4	5	6	7
Machine 1	20	21	21	20	20	19	19
Machine 2	22	23	21				
Machine 3							
Machine 4							

- We can still compute mean -> 20 + 22 = 21!
  - There are effects on standard errors of course…
- Possible solutions:
  - Ratio estimation not possible, there is no ratio…
  - sample from every cluster instead of just 2 (i.e. do SRS!)
  - get Mi (ssu in each PSU).
  - Or switch to model-based estimation

Example: model based inference

- Time coffee machines in Ruppert take to make coffee

	1	2	3	4	5	6	7
Machine 1	20	21	21	20	20	19	19
Machine 2	22	23	21				
Machine 3							
Machine 4							

- What can our model be like?
  - 2 minutes (in pairs)

Example: model based inference

- Time coffee machines in Ruppert take to make coffee

	1	2	3	4	5	6	7
Machine 1	20	21	21	20	20	19	19
Machine 2	22	23	21				
Machine 3							
Machine 4							

- What can our model be like?
  - Assumptions about machine use (size of M_i)
  - Use a model with auxiliary data

How to design a model?

- Model-based approach:
  - Get extra information
  - Imagine, we know that with each year machines get older, they become slower
  - Build a model (regression-based estimation)

Example: model assisted inference

- Time coffee machines in Ruppert take to make coffee

	Age	1	2	3	4	5	6	7
Machine 1	1	20	21	21	20	20	19	19
Machine 2	3	22	23	21				
Machine 3	3							
Machine 4	5							

- What can our model be like?
  - Assumptions about machine use (size of M_i)
  - Use a model with auxiliary data (this is up to the modeler)
    - E.g. $T_i = M_i * \mu + B_1 * AGE_i * M_i$
- What is the mean? (2 minutes)

Example: model-assisted inference

- Time coffee machines in Ruppert take to make coffee

	Age	1	2	3	4	5	6	7
Machine 1	1	20	21	21	20	20	19	19
Machine 2	3	22	23	21				
Machine 3	3							
Machine 4	5							

- Use a model with auxiliary data (this is up to the modeler)
  - Still assumptions about M_i
  - Mean: M_1: 20, M_2: 22, M_3: 22, M_4: 24
    - Overall: 22

- Age may not even be available for machine 3 and 4 (tricky!)

Why the two approaches?
2-stage cluster sample with unequal cluster sizes

- Design-based inference
  - You need to know M_i, N_i
    - You often don't
  - Alternative is inference using the information you have
- Model-based inference is easier
  - You use the information you have to design a model for the information you don't have

- Ratio estimation is a particular technique based on regression
  - It is model assisted: can be used under design-based models
  - e.g. Assume that the ratio t_i/m_i = the same for every M_i
  - e.g. Assume that variances are equal across clusters

What to do if you don’t have a list…

- Question:

“How many double espresso’s are being drunk at UU in a year?”

What kind of sampling design? Choose:

- Design based
- Ratio-estimation
- Model-based

What to do if you don’t have a list…

- Question: “How many double espresso’s are being drunk at UU in a year?”
  - Do a random sample, and ask a filter question
    - E.g. “Do you ever drink from the coffee machine?”
    - stratify -> oversample those more likely to drink coffee (e.g. employees)
  - Do an area sample
    - Estimate the ‘double espresso per person-day/ area’ -> ratio estimation
  - Use model-based estimation
    - Somehow find coffee drinkers
    - Find auxiliary variables that correlate with Y
      - Sleep per night, stress, hours at UU, having a coffee card.

When ratio vs. regression?

Ratio

- ‘double espresso per person-day/area’ -> total no. espressos
- 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

Often, little good frame information, no meaningful 0

Class exercise 2

- Regression estimation in practice
- 30 minutes

Design-based versus model-based

Model based

- Residual analysis is important (to get correct SE)
- Valid results when model fits data (model also applies for observations not observed)
- Observations are usually not weighted (e.g., in linear regression)

Design based

- Residual analysis is not important (SE are design-based)
- Valid results regardless of fit model
- Inclusion probabilities (weights) will influence the estimates

Design-based versus model-based

Definition of variance of estimator

Model based

- Average squared deviation of the estimate and the expected value, *averaged over all possible samples under the model*.

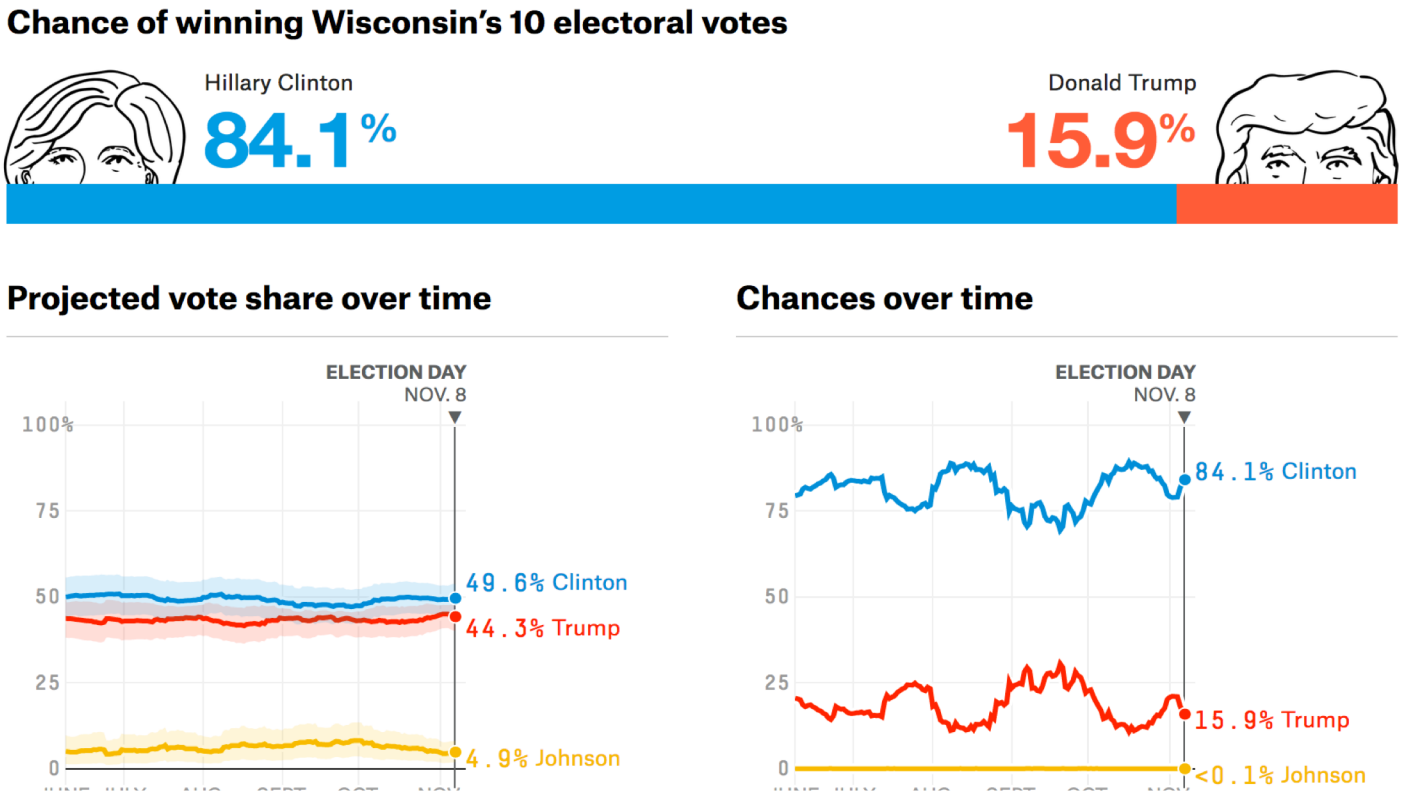
Design based

- Average squared deviation of the estimate and the expected value, *averaged over all possible samples under the design*.

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 – Presidential election 2016

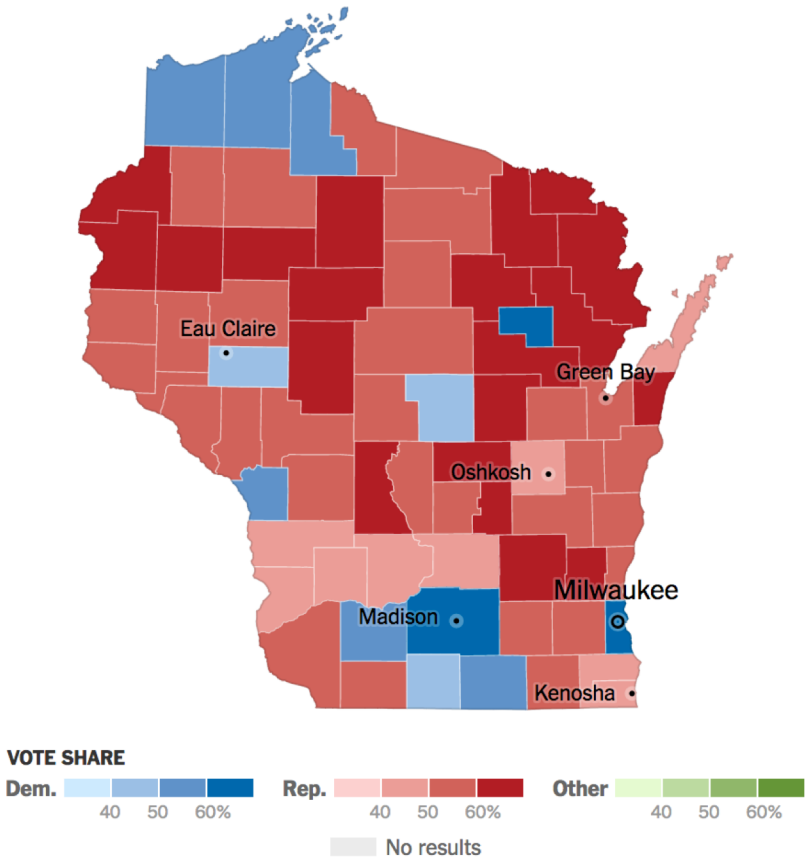
President

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	—

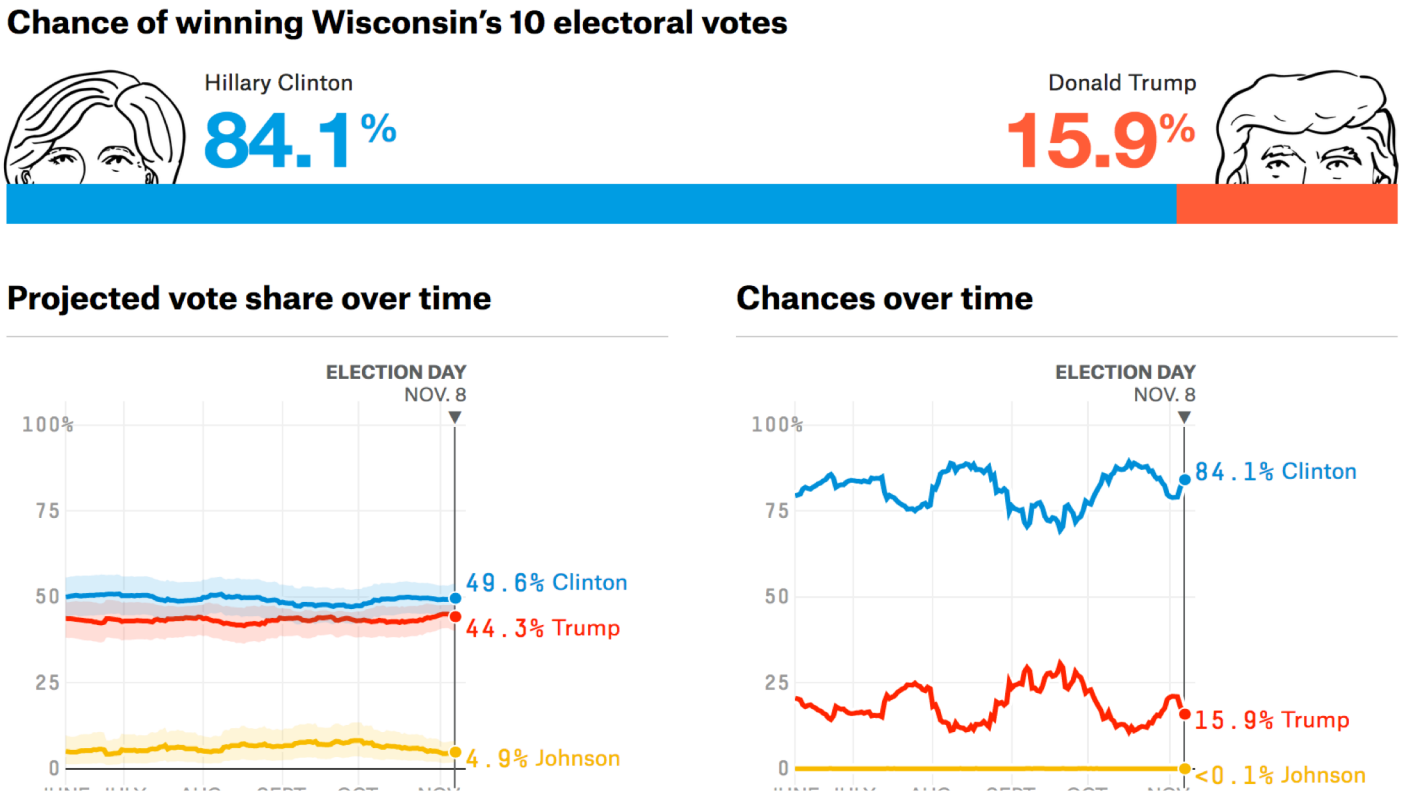
100% reporting (3,620 of 3,620 precincts)

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-.596)/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 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:

- 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

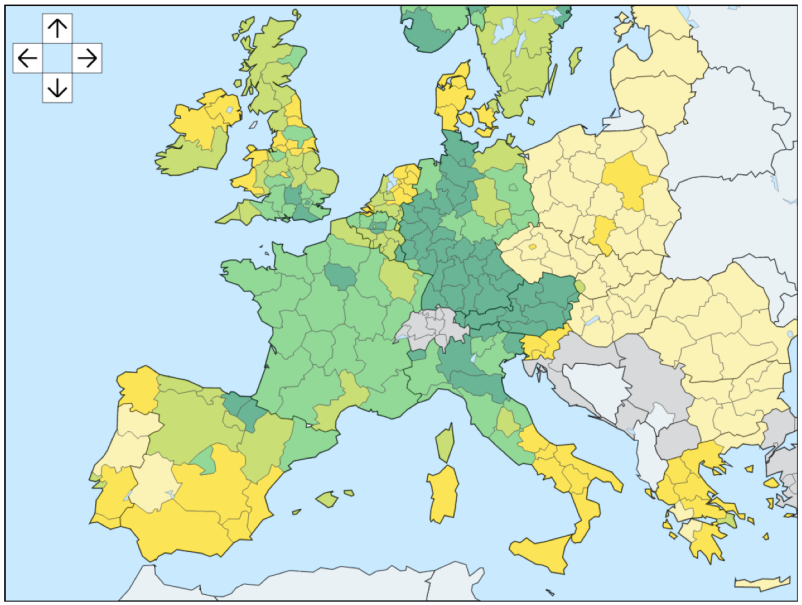
Why were the polls wrong?

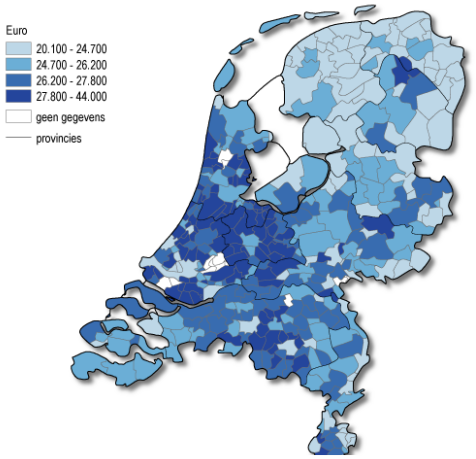
- Model based estimation depends on quality of model!
  - In design based, we can estimate error
  - In model-based -> much more difficult

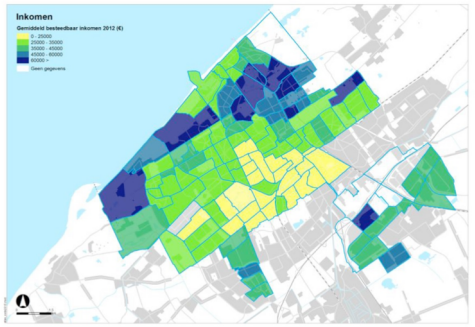
- Why not do design-based inference?
  - Costs
  - Time
  - Problems with coverage, nonresponse
  - -> still needs modeling
  - There are too many people who want to do a a poll
    - 100s in Wisconsin alone

What is a cluster?

What size should be a cluster be?



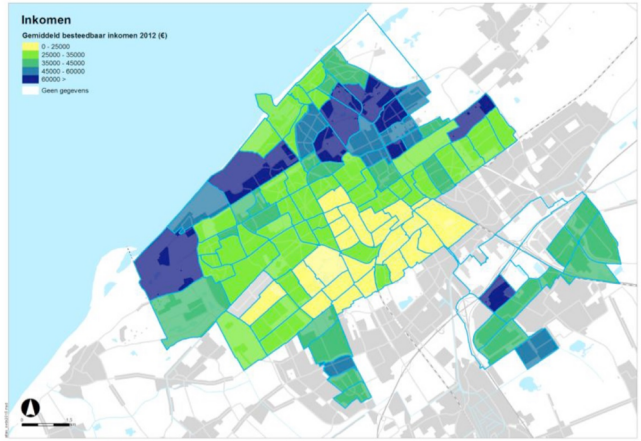




Small Area Estimation

- 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)

Your data and svysettings

- How do I know whether it works
- How can I check what my weights mean?
- What should ~fpc be? (esp. with use of weights)

Next week(s)

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

  Readings: several articles
- In two weeks: assignment 1 (!)