
Survey analysis
week 1
“the inference wars”
© Peter Lugtig

Outline for today

- Introduction
- Surveys and samples
- The first inference war:
  - 1948~1960
- The second inference war:
  - 2005~2020
- Election polling
- Class exercise

But first: Speed-dating!

- Form 2 rows of 10 people

- Ask 1 question (1 min max)
- Get 1 in return (1 min max)

- Row facing the Screen: shift one place on signal

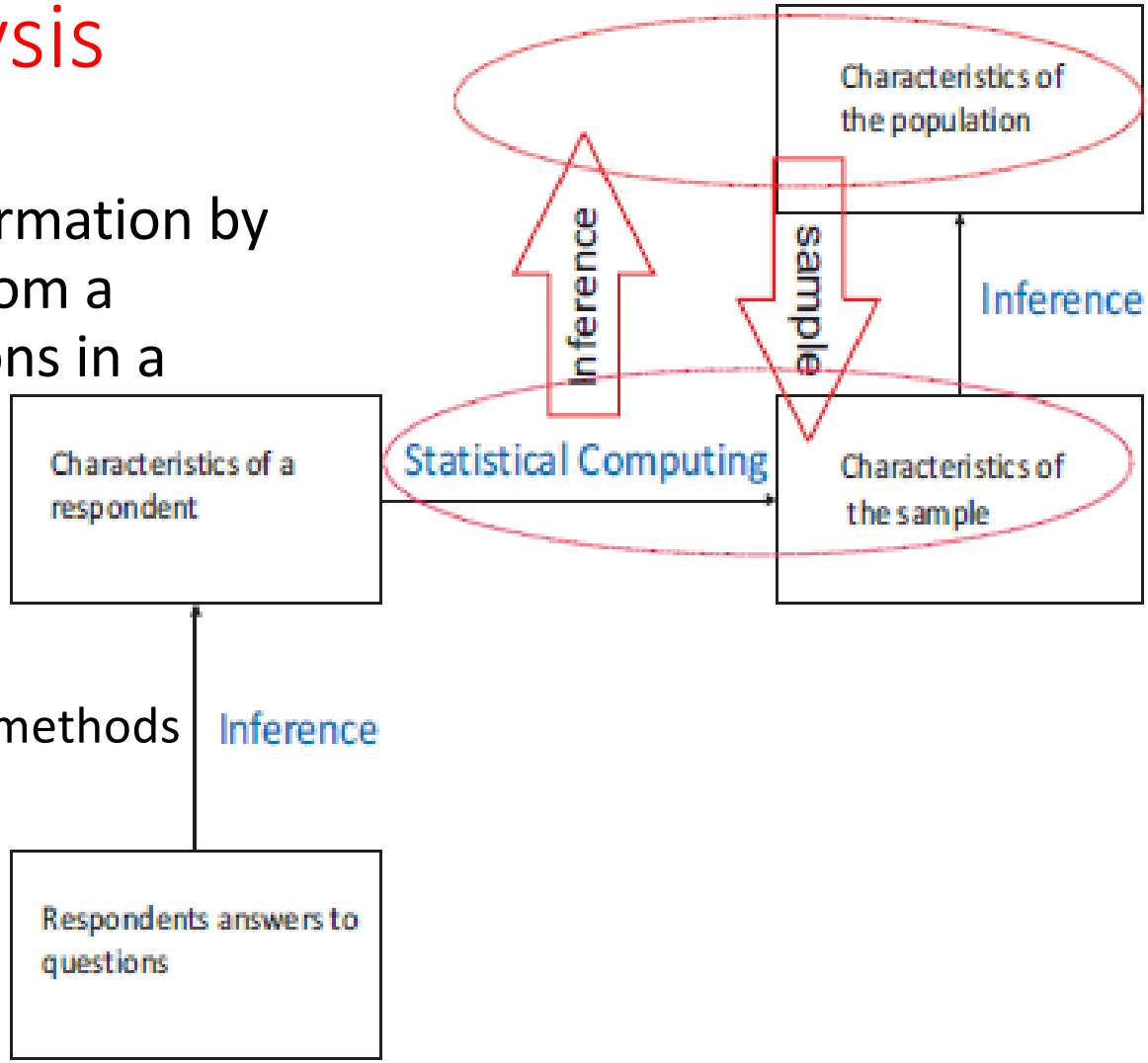
- https://www.youtube.com/watch?v=2dAorgAB0I4

Survey data analysis

- A method to acquire information by asking people selected from a predefined group questions in a structured manner
  - Dominant method of collecting data in the social and behavorial sciences
  - Learn about the methodology of conducting the survey

Survey data analysis

- A method to acquire information by asking people selected from a predefined group questions in a structured manner
- Survey data analysis
  - Sampling
  - Inference
  - Doing this for all kinds of methods



Why focus on Sampling?

- Basis of all inferential statistics
  - Standard error in t-test, Anova, regression.
- Sampling errors can be estimated
  - Design based inference
  - (model-based inference later in course)
- Power analysis, efficient design of studies
- Important in understanding other techniques (e.g. Bootstrapping)

- There are not so many people who know about this
  - A specific skill ‘survey statistics’

Links to other courses

- Multivariate Statistics
- Fundamentals of Statistics
- Computation inference with R

- **Survey data analysis:**
  - **Links/overlap**
    - **Apply the general linear model under correct inferential design**
    - **Apply fundamental knowledge about estimators**
    - **Apply R skills to analyse a real-life dataset and data problem**
  - **Specific goals: inference, data collection, missing data**

Sample

- NL: steekproef
- DE: stichprobe
- FR: Échantillon
- ES: muestra

The first inference war (1936-1952)

The 1936 Literary Digest survey

- Correct result since 1916
- Sample: 2.4 million people(!)
- Opt-in sample

The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

Those figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased THE LITERARY DIGEST?" A telephone message only the day before these lines were written: "Was the Republican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased THE LITERARY DIGEST?" "Is the Pope of Rome a stockholder of THE LITERARY DIGEST?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in THE LITERARY DIGEST straw vote. I consider this conclusive evidence as to the choice of the people of this country for a change in the National Government. THE LITERARY DIGEST poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of the voters from

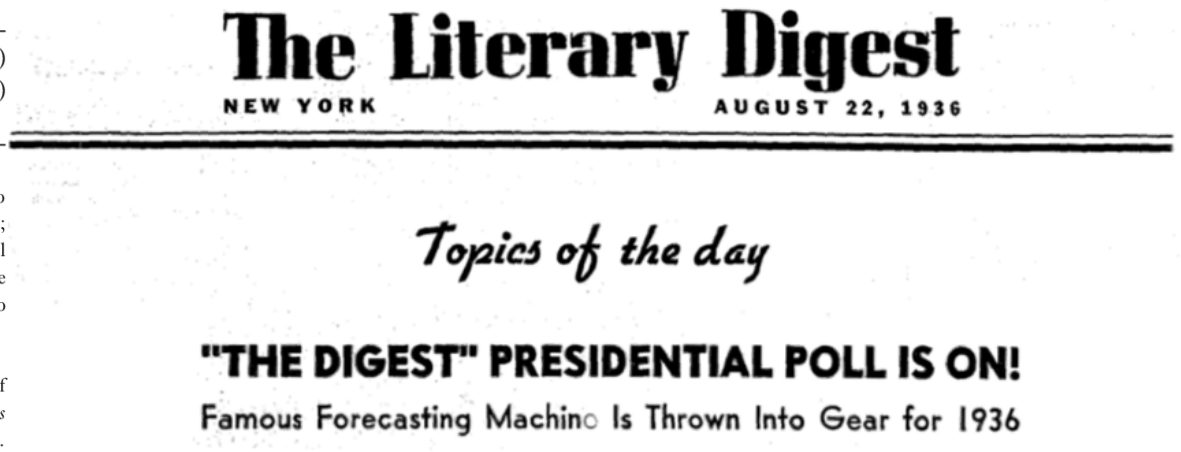
Candidate	*Digest* poll respondent: Yes		No		Total	
Roosevelt	**42.9**	(48.5)	*60.9*	(66.1)	*56.6*	(54.8)
Landon	**57.1**	(51.5)	*39.1*	(33.9)	*43.4*	(45.2)
Total[a]	**23.8**	(63.7)	**76.2**	(36.3)	764[b]	

Source: Roper Center for Public Opinion Research 2003b.

Note: This table indicates voting preference *at the time of* the *Digest* poll; that is, it accounts for those who remembered changing their minds ($n = 26$); those who claimed not to have changed their minds ($n = 433$); those who said that they did not remember ($n = 13$); and those who gave no answer ($n = 292$). The original unweighted results are in parentheses ($n_{11} = 236$; $n_{12} = 183$; $n_{21} = 251$; $n_{22} = 94$). The weighted results are italicized ($n_{11} = 78$; $n_{12} = 355$; $n_{21} = 104$; $n_{22} = 228$). (Weighted cell frequencies do not sum to total due to rounding.) The bolded numbers are known (population) values.

[a] Row percentages.

[b] Total sample size. It is made up of AIPO poll respondents who report having received a *Digest* ballot. Of those, *respondents* are individuals who claim to have returned their straw ballots ($n = 487$), and *nonrespondents* are individuals who said that they did not return theirs ($n = 246$) or did not remember returning it ($n = 31$).



The Literary Digest

NEW YORK — AUGUST 22, 1936

Topics of the day

"THE DIGEST" PRESIDENTIAL POLL IS ON!

Famous Forecasting Machine Is Thrown Into Gear for 1936

The 1936 Literary Digest survey

- Sample: 2.4 million people(!)
- Opt-in sample
- The literary Digest

	Digest poll respondent		
Candidate	Yes	No	Total
Roosevelt	**42.9**		
Landon	**57.1**		
Total[a]	**23.8**		

Source: Roper Center for Public Opinion Research 2003b.

Note: This table indicates voting preference *at the time of* the *Digest* poll; that is, it accounts for those who remembered changing their minds ($n = 26$); those who claimed not to have changed their minds ($n = 433$); those who said that they did not remember ($n = 13$); and those who gave no answer ($n = 292$). The original unweighted results are in parentheses ($n_{11} = 236$; $n_{12} = 183$; $n_{21} = 251$; $n_{22} = 94$). The weighted results are italicized ($n_{11} = 78$; $n_{12} = 355$; $n_{21} = 104$; $n_{22} = 228$). (Weighted cell frequencies do not sum to total due to rounding.) The bolded numbers are known (population) values.

[a]Row percentages.

[b]Total sample size. It is made up of AIPO poll respondents who report having received a *Digest* ballot. Of those, *respondents* are individuals who claim to have returned their straw ballots ($n = 487$), and *nonrespondents* are individuals who said that they did not return theirs ($n = 246$) or did not remember returning it ($n = 31$).

1936-1948: Quota Sampling

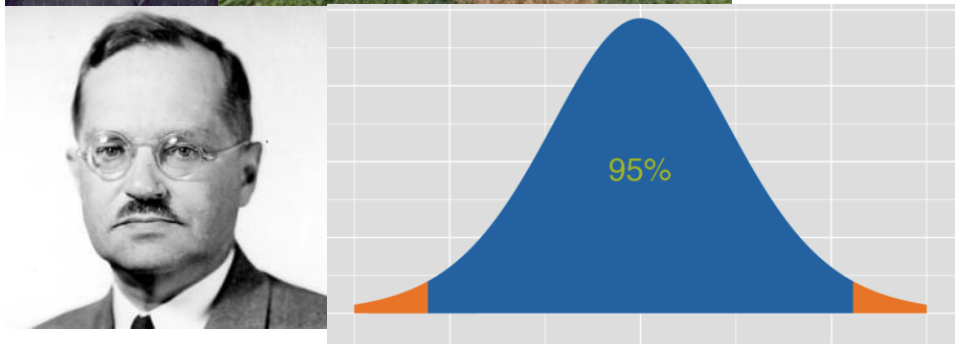
- Opt-in sample, but:
  - Record characteristics that are important for voring behavior.
  - We want a sample of 1000 respondents
  - Quota:
    - White, male, aged 35-54, blue collar worker — 5% of population -> 50 in sample
    - Black, female, aged 18-35, in education — 2% of population -> 20 in sample
    - ….
  - Choice of variables for quota is important!
- Why does this go wrong in election polling? -------->

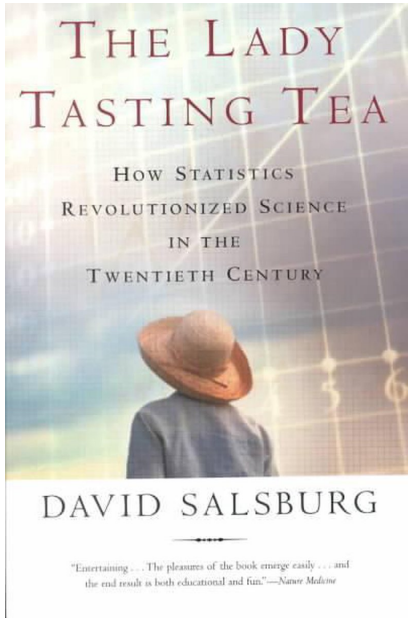
1948: Truman vs. Dewey

- **Gallup**, Roper
- Quota sample
- Prediction: +5 for Dewey
- Result: +5 for Truman
  - Problem with quotas
    - Gender, age, race
  - Late undecideds

Developments in statistics 1900-1930

- Gosset (Student-T)
- Fisher
- Neyman





Sampling and the central limit theorem

- Galton board

- Central limit theorem:
  - https://gallery.shinyapps.io/CLT_mean/

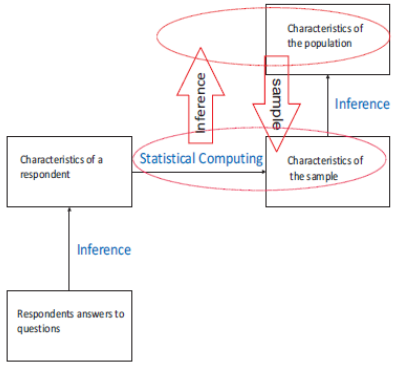
Sampling and the central limit theorem

- Galton board
- Central limit theorem:
  - https://gallery.shinyapps.io/CLT_mean/
  - The distribution of means that are the result from repeatedly sampling from any population distribution will result in a normal distribution
- Law of large numbers: the confidence interval of an estimate from a sample will become smaller the larger the sample size is.

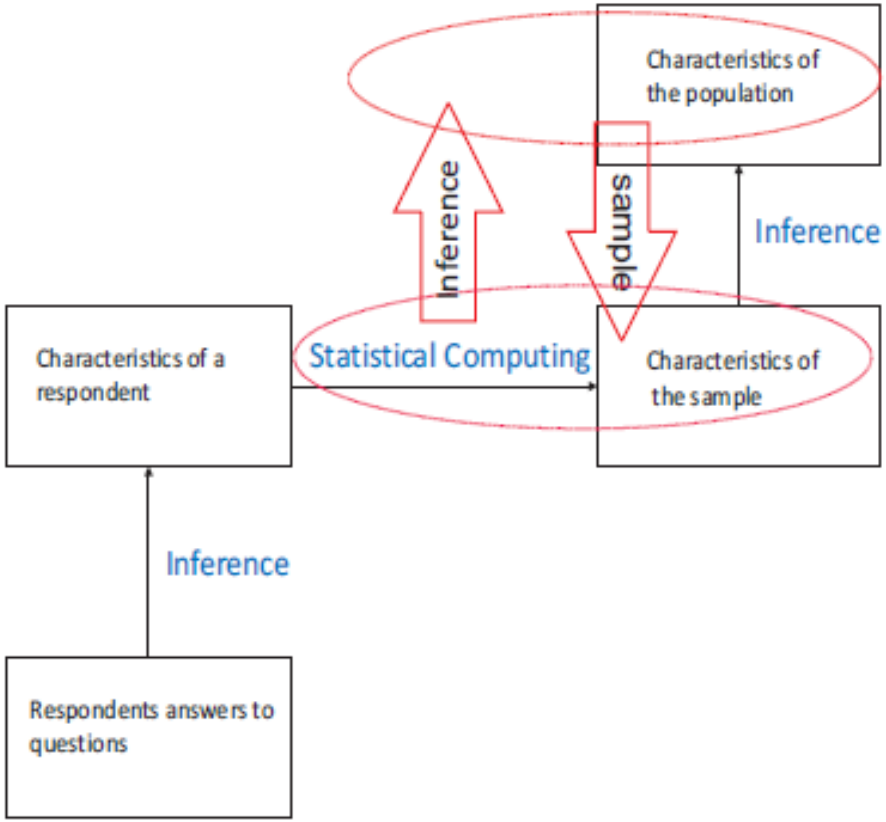
1952 onwards: Michigan election studies

Idea: a random sample from the population

- 1. Population
- 2. Frame for population elements
  - Addresses, phone-numbers, etc.
- 3. Sample from frame
- 4. Response from the sample



Design based inference



Elements of design-based inference

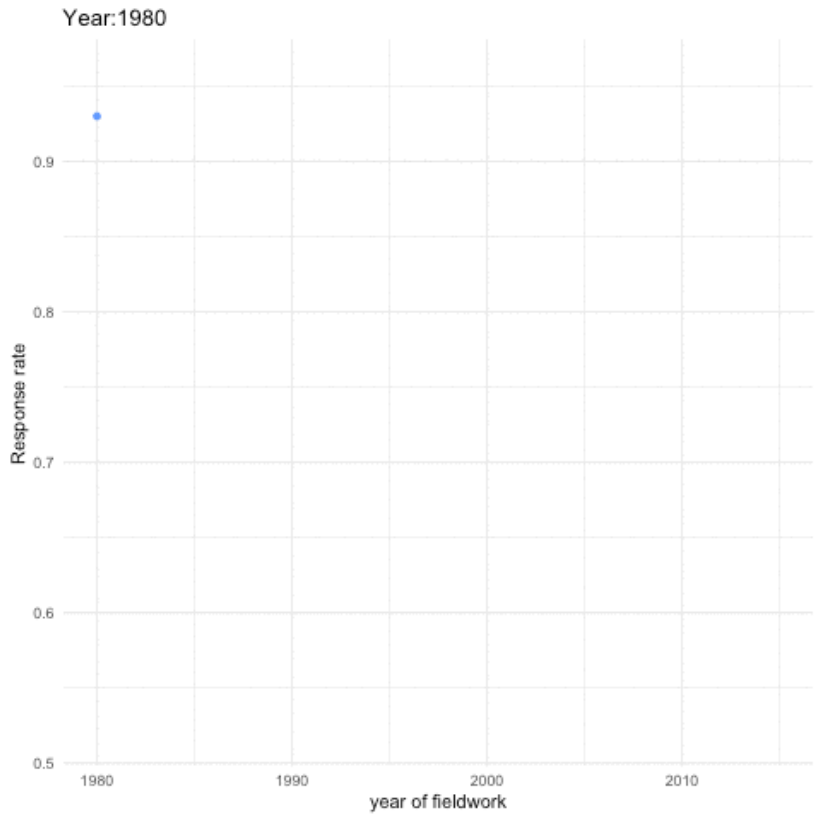
- Sampling methods: clustering, stratification
  - Corrections for unequal sampling methods
  - Hurvitz-Thomposon estimator (1948, 1952) (week 4)
- Corrections for missing data: weighting, imputation
  - Nonresponse (week 8-9,11-12)
  - Errors on frame (weeks 8,9)
- Estimation of uncertainty (errors) (week 3-7)
- Variance estimators (week 3-7)

- 1952-now: development of a framework for inference: **design-based**

So, what led to the 2nd inference war?

Problems with surveys

- Lists are getting worse (telephone esp)
- Falling response rates
- Costs!



So, what led to the 2nd inference war?

- Lists are getting worse
- Falling response rates
- Costs!

- The Internet
- Growing need for information
- Quick and cheap valued more than slow and expensive

What is the 2nd inference war about?

- Design-based inference:
  - We have worked 60 years to work out sampling theory and survey practice
  - Methods are unbiased and consistent (week 39)
  - Using opt-in or quota samples is unscientific

Vs.

- **Model-based inference:**
  - Response rates are too low
  - It is just not feasible anymore to do expensive surveys
    - And a lot of other data is just there!
  - You need to model nonresponse anyway
  - Lets model the whole selection process

Election polling

- Need for fast results
- Can’t be too expensive
  - Internet panels: quota samples from pre-recruited panel members
  - Telephone: random digit dialing (USA)
  - Telephone: registers (phone book)
  - IVR vs. in-person calls
- 1000s of polls per election cycle
- Polls are generally accurate (see article by Kennedy et al)
  - But recent high profile misses: US election, British EU-referendum
  - https://utrecht-university.shinyapps.io/SDA_shinyelectionbias/

Inference peace remains as well: design-based dominant

Class exercise

- The 2016 U.S. elections
  - News stations believed Clinton would easily win
  - Trump won the election (although lost the popular vote)
  - Was there a polling miss?

- www.fivethirtyeight.com has a database with about1,600 polls conducted before the election
  - Sample size, company, date conducted, state conducted, reputation of pollster, whole pulation vs. likely voters, raw and adjusted % for Trump and Clinton

Class exercise

- Go to https://utrecht-university.shinyapps.io/SDA_shinyelectionbias/

Five groups (4 people each- 20 minutes):

1. Was Trump underestimated? (or was he particularly in the swing states)?
2. Did the quality of the pollster matter? (what is quality?)
3. Was there a difference between sampling likely voters and registered voters?
4. Are larger polls better?
5. Is there a difference between raw and adjusted (modeled) poll estimates?

Class exercise (1)

Was Trump underestimated? in the swing states? (Michigan, Wisconsin, Pensylvania)

Class exercise (2)

Did the quality of the pollster matter?

Class exercise (3)

Was there a difference between sampling likely voters and registered voters?

Class exercise (4)

Are larger polls better?

Class exercise (5)

Is there a difference between raw and modeled polls?

Next week: Total survey error

- We move into design-based surveys
- Read articles
- Complete take-home exercise and bring to class
  - THE 1- Your adopted survey.