Statistical Methods and Data Analysis I
Lecture 10: Sampling Fundamentals

Oleg Goldshmidt
oleg.goldshmidt@post.idc.ac.il

Arison School of Business
Interdisciplinary Center (IDC)
Herzliya, Israel

April 02, 2019

Population vs. Sample: Consider an Election...

Population vs. Sample: Consider an Election...

- An election is essentially a *sample of the entire population* — a *census*.

Population vs. Sample: Consider an Election...

- An election is essentially a *sample of the entire population* — a *census*.
- Taking a *census* (for information, not for election)
  - difficult and expensive!
  - there always are people who are hard to locate or measure
  - and they may share certain characteristics that distinguish them from the general population
  - populations change...

Population vs. Sample: Consider an Election...

- An election is essentially a *sample of the entire population* — a *census*.
- Taking a *census* (for information, not for election)
  - difficult and expensive!
  - there always are people who are hard to locate or measure
  - and they may share certain characteristics that distinguish them from the general population
  - populations change...
- We want to get results (predictions, guidance for action, feedback) without all the difficulties of taking a *census*
  - poll a *sample*!

Population vs. Sample: Consider an Election...

- An election is essentially a *sample of the entire population* — a *census*.
- Taking a *census* (for information, not for election)
  - difficult and expensive!
  - there always are people who are hard to locate or measure
  - and they may share certain characteristics that distinguish them from the general population
  - populations change...
- We want to get results (predictions, guidance for action, feedback) without all the difficulties of taking a *census*
  - poll a *sample*!
- Sampling is **natural**! Like tasting the dish you are cooking...

Population vs. Sample: Consider an Election...

- An election is essentially a *sample of the entire population* — a *census*.
- Taking a *census* (for information, not for election)
  - difficult and expensive!
  - there always are people who are hard to locate or measure
  - and they may share certain characteristics that distinguish them from the general population
  - populations change...
- We want to get results (predictions, guidance for action, feedback) without all the difficulties of taking a *census*
  - poll a *sample*!
- Sampling is **natural**! Like tasting the dish you are cooking...
  - Is it salty enough? We will call it *exploratory analysis*.

Population vs. Sample: Consider an Election...

- An election is essentially a *sample of the entire population* — a *census*.
- Taking a *census* (for information, not for election)
  - difficult and expensive!
  - there always are people who are hard to locate or measure
  - and they may share certain characteristics that distinguish them from the general population
  - populations change...
- We want to get results (predictions, guidance for action, feedback) without all the difficulties of taking a *census*
  - poll a *sample*!
- Sampling is **natural**! Like tasting the dish you are cooking...
  - Is it salty enough? We will call it *exploratory analysis*.
  - No — *the whole soup needs more salt*! This is *inference* or *induction*.

Population vs. Sample: Consider an Election...

- An election is essentially a *sample of the entire population* — a *census*.
- Taking a *census* (for information, not for election)
  - difficult and expensive!
  - there always are people who are hard to locate or measure
  - and they may share certain characteristics that distinguish them from the general population
  - populations change...
- We want to get results (predictions, guidance for action, feedback) without all the difficulties of taking a *census*
  - poll a *sample*!
- Sampling is **natural**! Like tasting the dish you are cooking...
  - Is it salty enough? We will call it *exploratory analysis*.
  - No — *the whole soup needs more salt*! This is *inference* or *induction*.
  - What if the salt is at the bottom of the pot? Is your sample *representative*?

Population vs. Sample: Consider an Election...

- An election is essentially a *sample of the entire population* — a *census*.
- Taking a *census* (for information, not for election)
  - difficult and expensive!
  - there always are people who are hard to locate or measure
  - and they may share certain characteristics that distinguish them from the general population
  - populations change...
- We want to get results (predictions, guidance for action, feedback) without all the difficulties of taking a *census*
  - poll a *sample*!
- Sampling is **natural**! Like tasting the dish you are cooking...
  - Is it salty enough? We will call it *exploratory analysis*.
  - No — *the whole soup needs more salt*! This is *inference* or *induction*.
  - What if the salt is at the bottom of the pot? Is your sample *representative*?
  - Stir well (*randomize*) before tasting (*sampling*)!

Randomization

Randomization

- Sampling *people* is **very difficult** (compared to sampling *soup*):

Randomization

- Sampling *people* is **very difficult** (compared to sampling *soup*):
  - let's walk up to people in the streets — you are likely to select those who look civil and are neatly dressed...

Randomization

- Sampling *people* is **very difficult** (compared to sampling *soup*):
  - let's walk up to people in the streets — you are likely to select those who look civil and are neatly dressed...
  - parliamentarians should be weary of letters from their constituents — biased towards organized pressure groups, self-selecting activists, busybodies will an excess of free time...

Randomization

- Sampling *people* is **very difficult** (compared to sampling *soup*):
  - let's walk up to people in the streets — you are likely to select those who look civil and are neatly dressed...
  - parliamentarians should be weary of letters from their constituents — biased towards organized pressure groups, self-selecting activists, busybodies will an excess of free time...
- Let's choose a *simple random sample* of 1000 voters from the register.

Randomization

- Sampling *people* is **very difficult** (compared to sampling *soup*):
  - let's walk up to people in the streets — you are likely to select those who look civil and are neatly dressed...
  - parliamentarians should be weary of letters from their constituents — biased towards organized pressure groups, self-selecting activists, busybodies will an excess of free time...
- Let's choose a *simple random sample* of 1000 voters from the register.
  - Size matters: small samples will not necessarily reflect the population due to dumb luck of the draw.

Randomization

- Sampling *people* is **very difficult** (compared to sampling *soup*):
  - let's walk up to people in the streets — you are likely to select those who look civil and are neatly dressed...
  - parliamentarians should be weary of letters from their constituents — biased towards organized pressure groups, self-selecting activists, busybodies will an excess of free time...
- Let's choose a *simple random sample* of 1000 voters from the register.
  - Size matters: small samples will not necessarily reflect the population due to dumb luck of the draw.
  - Large samples will yield more reliable estimates *(under many assumptions discussed below)*

Randomization

- Sampling *people* is **very difficult** (compared to sampling *soup*):
  - let's walk up to people in the streets — you are likely to select those who look civil and are neatly dressed...
  - parliamentarians should be weary of letters from their constituents — biased towards organized pressure groups, self-selecting activists, busybodies will an excess of free time...
- Let's choose a *simple random sample* of 1000 voters from the register.
  - Size matters: small samples will not necessarily reflect the population due to dumb luck of the draw.
  - Large samples will yield more reliable estimates *(under many assumptions discussed below)*
    - with P_p and P_s denoting the population and sample proportions (e.g., of republican voters in a presidential election) we can construct a *confidence interval*: $\boxed{P_p = P_s \pm \text{"sampling allowance"}}$
    - How tight is the confidence interval? Depends on how sure we want to be (e.g., *"with probability 95%"*), on the *measured* sample proportion P_s, and on the sample size.

Randomization

- Sampling *people* is **very difficult** (compared to sampling *soup*):
  - let's walk up to people in the streets — you are likely to select those who look civil and are neatly dressed...
  - parliamentarians should be weary of letters from their constituents — biased towards organized pressure groups, self-selecting activists, busybodies will an excess of free time...
- Let's choose a *simple random sample* of 1000 voters from the register.
  - Size matters: small samples will not necessarily reflect the population due to dumb luck of the draw.
  - Large samples will yield more reliable estimates *(under many assumptions discussed below)*
    - with P_p and P_s denoting the population and sample proportions (e.g., of republican voters in a presidential election) we can construct a *confidence interval*: $\boxed{P_p = P_s \pm \text{"sampling allowance"}}$
    - How tight is the confidence interval? Depends on how sure we want to be (e.g., *"with probability 95%"*), on the *measured* sample proportion P_s, and on the sample size.
    - Details to follow...

(Some) Types of Sampling Biases

(Some) Types of Sampling Biases

- *Convenience sampling bias*
  - respondents who are easier to reach are more likely to be included in a sample

(Some) Types of Sampling Biases

- *Convenience sampling bias*
  - respondents who are easier to reach are more likely to be included in a sample
  - e.g., those who have phones (some people do not have land lines, others have poor cell coverage, quite a few have unlisted numbers, etc.)

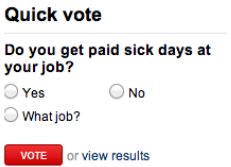
(Some) Types of Sampling Biases

- *Convenience sampling bias*
  - respondents who are easier to reach are more likely to be included in a sample
  - e.g., those who have phones (some people do not have land lines, others have poor cell coverage, quite a few have unlisted numbers, etc.)
- *Voluntary responders*, a.k.a. *self-selection bias*:
  - a sample containing a high percentage of people who volunteer to respond may be biased.

(Some) Types of Sampling Biases

- *Convenience sampling bias*
  - respondents who are easier to reach are more likely to be included in a sample
  - e.g., those who have phones (some people do not have land lines, others have poor cell coverage, quite a few have unlisted numbers, etc.)
- *Voluntary responders*, a.k.a. *self-selection bias*:
  - a sample containing a high percentage of people who volunteer to respond may be biased.

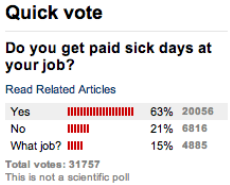
From cnn.com, Jan 14, 2012



(Some) Types of Sampling Biases

- *Convenience sampling bias*
  - respondents who are easier to reach are more likely to be included in a sample
  - e.g., those who have phones (some people do not have land lines, others have poor cell coverage, quite a few have unlisted numbers, etc.)
- *Voluntary responders*, a.k.a. *self-selection bias*:
  - a sample containing a high percentage of people who volunteer to respond may be biased.

From cnn.com, Jan 14, 2012



(Some) Types of Sampling Biases

- *Convenience sampling bias*
  - respondents who are easier to reach are more likely to be included in a sample
  - e.g., those who have phones (some people do not have land lines, others have poor cell coverage, quite a few have unlisted numbers, etc.)
- *Voluntary responders*, a.k.a. *self-selection bias*:
  - a sample containing a high percentage of people who volunteer to respond may be biased.

From cnn.com, Jan 14, 2012

- *Non-response bias* (also a type of *self-selection bias*):
  - If only a fraction choose to respond the result may be *not representative*.

(Some) Types of Sampling Biases

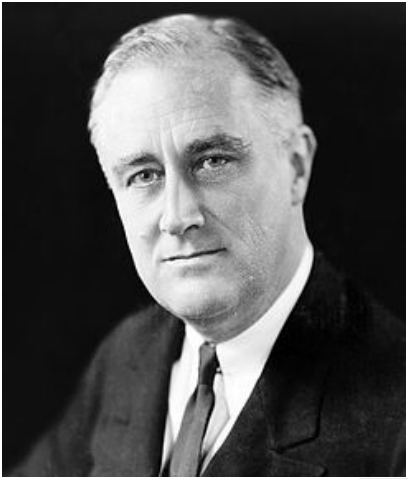
- *Convenience sampling bias*
  - respondents who are easier to reach are more likely to be included in a sample
  - e.g., those who have phones (some people do not have land lines, others have poor cell coverage, quite a few have unlisted numbers, etc.)
- *Voluntary responders*, a.k.a. *self-selection bias*:
  - a sample containing a high percentage of people who volunteer to respond may be biased.

From cnn.com, Jan 14, 2012

- *Non-response bias* (also a type of *self-selection bias*):
  - If only a fraction choose to respond the result may be *not representative*.
- *Some respondents may lie...*

Election prediction in the past: 1936

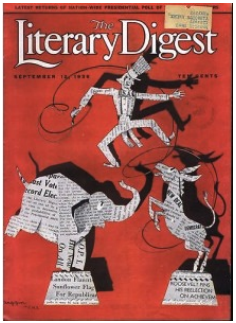
In 1936, Alfred Mossman ("Alf") Landon sought the Republican presidential nomination opposing the re-election of (the Democratic President) Franklin Delano Roosevelt.



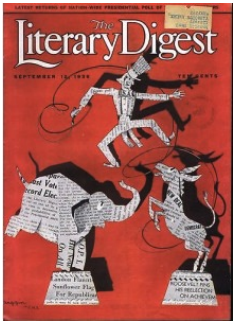
The *Literary Digest* Poll

The Poll:

- The *Literary Digest* polled about *10 million* Americans
- About *2.4 million* responded.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.



The *Literary Digest* Poll

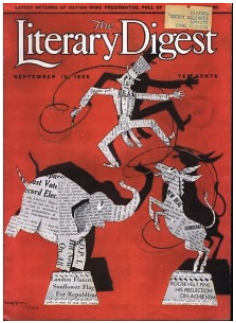


The Poll:

- The *Literary Digest* polled about *10 million* Americans
- About *2.4 million* responded.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.

The Results:

The *Literary Digest* Poll



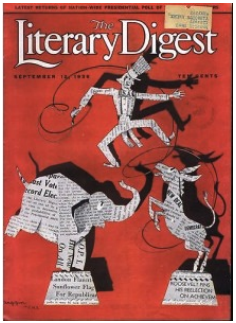
The Poll:

- The *Literary Digest* polled about *10 million* Americans
- About *2.4 million* responded.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.

The Results:

- FDR won, with 62% of the votes. Landon got 37%.

The *Literary Digest* Poll



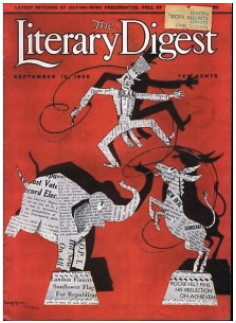
The Poll:

- The *Literary Digest* polled about *10 million* Americans
- About *2.4 million* responded.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.

The Results:

- FDR won, with 62% of the votes. Landon got 37%.
- The magazine was completely discredited because of the poll, and was soon discontinued.

The *Literary Digest* Poll



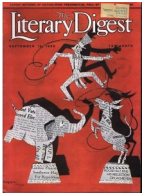
The Poll:

- The *Literary Digest* polled about *10 million* Americans
- About *2.4 million* responded.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.

The Results:

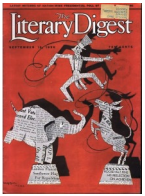
- FDR won, with 62% of the votes. Landon got 37%.
- The magazine was completely discredited because of the poll, and was soon discontinued.
- Hardly anyone remembers Landon's name anymore...

The *Literary Digest* Poll: What Went Wrong?



- *Convenience sampling:* the magazine had surveyed
  - its own readers,
  - registered automobile owners, and
  - registered telephone users.

The *Literary Digest* Poll: What Went Wrong?



- *Convenience sampling:* the magazine had surveyed
  - its own readers,
  - registered automobile owners, and
  - registered telephone users.
- *Non-representative sample:*
  - The participants had incomes well above the national average of the day.
  - During the Great Depression era high income individuals were far more likely to support Republicans than a truly *representative* American voter of the time.
- *Self-selection bias:*
  - The 24% who *responded* were even more biased towards the Republican party.
- But the sample was **HUGE**...
  - Doesn't help if it is biased...
  - If your soup is *not* well stirred it does not matter how large a spoon you've got, but if it *is* well stirred a small spoon will suffice.

Why Sample?

So, sampling may easily lead to biases and errors. Wouldn't it be better to do a *census*?

Why Sample?

So, sampling may easily lead to biases and errors. Wouldn't it be better to do a *census*?

- Limited resources: neither time nor money is available for a census:
  - Often true not just for political polls but also for market surveys.
  - Testing new equipment: is it better than the old one? We can't test the whole *population* of that the new production line will produce. Instead, we will do a sample run and *infer* the decision (to buy or not to buy) from that sample.

Why Sample?

So, sampling may easily lead to biases and errors. Wouldn't it be better to do a *census*?

- Limited resources: neither time nor money is available for a census:
  - Often true not just for political polls but also for market surveys.
  - Testing new equipment: is it better than the old one? We can't test the whole *population* of that the new production line will produce. Instead, we will do a sample run and *infer* the decision (to buy or not to buy) from that sample.
- Scarcity: sometime only a sample is available.
  - Heredity vs. environmental factor studies are usually based on observation of identical twins.
  - Do rats carry diseases? How would you identify **all rats**?

Why Sample?

So, sampling may easily lead to biases and errors. Wouldn't it be better to do a *census*?

- Limited resources: neither time nor money is available for a census:
  - Often true not just for political polls but also for market surveys.
  - Testing new equipment: is it better than the old one? We can't test the whole *population* of that the new production line will produce. Instead, we will do a sample run and *infer* the decision (to buy or not to buy) from that sample.
- Scarcity: sometime only a sample is available.
  - Heredity vs. environmental factor studies are usually based on observation of identical twins.
  - Do rats carry diseases? How would you identify **all rats**?
- Destructive experiments.
  - Suppose we want to know how durable our product is...
  - We only give a small blood sample for testing...

Why Sample?

So, sampling may easily lead to biases and errors. Wouldn't it be better to do a *census*?

- Limited resources: neither time nor money is available for a census:
  - Often true not just for political polls but also for market surveys.
  - Testing new equipment: is it better than the old one? We can't test the whole *population* of that the new production line will produce. Instead, we will do a sample run and *infer* the decision (to buy or not to buy) from that sample.
- Scarcity: sometime only a sample is available.
  - Heredity vs. environmental factor studies are usually based on observation of identical twins.
  - Do rats carry diseases? How would you identify **all rats**?
- Destructive experiments.
  - Suppose we want to know how durable our product is...
  - We only give a small blood sample for testing...
- A sample may be **more accurate** than a census!
  - E.g., large, but inadequately trained personnel.
  - US Census Bureau: Decennial Census (cf. West Wing S1E6 "Mr. Willis of Ohio") is **less** accurate than the monthly Current Population Survey.

Sampling Frame: Definition and Examples

Sampling frame:

A set or device from which a sample may be obtained.

- Electoral register
- Telephone directory
- Employment records
- Medical files
- Database records
- Manufacturing line
- Systematic measurements
- Events generated by a computer system

Sampling Frames: Requirements

- Must be representative
- Elements must be identifiable for both observation and analysis
- Full coverage
- No duplicates
- No foreign elements
- Must be up-to-date