Unit 1: Introduction to data
Lecture 1: Data Collection + Observational studies & experiments

Statistics 101

Dr. Çetinkaya-Rundel

January 12, 2015

1 Announcements

2 Main points

1. Use a sample to make inferences about the population
2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
3. Sampling schemes can suffer from a variety of biases

Housekeeping

1 Announcements

2 Main points

1. Use a sample to make inferences about the population
2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
3. Sampling schemes can suffer from a variety of biases

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

1. Use a sample to make inferences about the population

- Our ultimate goal is to make inferences about populations
- However populations are difficult or impossible to access
- Therefore we use a sample from that population, and use *statistics* from that sample to make inferences about the unknown population *parameters*
  - We want to know how many offspring female lemurs have, on average
  - It's not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center
  - We use the sample mean from these data as an estimate for the unknown population mean
- The better (more *representative*) sample we have, the more reliable our estimates and more accurate our inferences will be

Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

Sampling is natural

- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's *exploratory analysis*
- If you generalize and conclude that your entire soup needs salt, that's an *inference*
- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population)

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling: Clusters* are not necessarily homogenous. First randomly sample a few clusters, then randomly sample from within them
  - e.g.
  - Usually preferred for economical reasons

Demo: *http://bl.ocks.org/avimoondra*

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling: Clusters* are not necessarily homogenous. First randomly sample a few clusters, then randomly sample from within them
  - e.g.
  - Usually preferred for economical reasons

Demo: *http://bl.ocks.org/avimoondra*

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling: Clusters* are not necessarily homogenous. First randomly sample a few clusters, then randomly sample from within them
  - e.g.
  - Usually preferred for economical reasons

Demo: *http://bl.ocks.org/avimoondra*

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling: Clusters* are not necessarily homogenous. First randomly sample a few clusters, then randomly sample from within them
  - e.g.
  - Usually preferred for economical reasons

Demo: *http://bl.ocks.org/avimoondra*

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling:* *Clusters* are not necessarily homogenous. First randomly sample a few clusters, then randomly sample from within them
  - e.g.
  - Usually preferred for economical reasons

Demo: *http://bl.ocks.org/avimoondra*

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling: Clusters* are not necessarily homogenous. First randomly sample a few clusters, then randomly sample from within them
  - e.g.
  - Usually preferred for economical reasons

Demo: *http://bl.ocks.org/avimoondra*

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling: Clusters* are not necessarily homogenous. First randomly sample a few clusters, then randomly sample from within them
  - e.g.
  - Usually preferred for economical reasons

Demo: *http://bl.ocks.org/avimoondra*

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

- *Simple random sampling:* Randomly select cases from the population, each case is equally likely to be selected
- *Stratified sampling:* First divide the population into homogenous *strata*, then randomly sample from each stratum
  - e.g. Stratify to control for socio-economic status
- *Cluster sampling: Clusters* are not necessarily homogenous. First randomly sample a few clusters, then randomly sample from within them
  - e.g.
  - Usually preferred for economical reasons

Demo: *http://bl.ocks.org/avimoondra*

Clicker question

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Which approach would likely be the least effective?

(a) Simple random sampling

(b) Stratified sampling, where each cluster is a neighborhood

(c) Cluster sampling, where each cluster is a neighborhood

Clicker question

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Which approach would likely be the least effective?

(a) Simple random sampling

(b) Stratified sampling, where each cluster is a neighborhood

(c) *Cluster sampling, where each cluster is a neighborhood*

3. Sampling schemes can suffer from a variety of biases

- *Non-response:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population.
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample.

3. Sampling schemes can suffer from a variety of biases

- *Non-response:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population.
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample.

3. Sampling schemes can suffer from a variety of biases

- *Non-response:* If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue since such a sample will also not be representative of the population.
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample.

Clicker question

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

II. Overall, the school district has strong support from parents to move forward with the policy approval.

III. It is possible that majority of the parents of high school students disagree with the policy change.

IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

Clicker question

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

II. Overall, the school district has strong support from parents to move forward with the policy approval.

III. It is possible that majority of the parents of high school students disagree with the policy change.

IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) *I and III* (d) III and IV (e) Only IV