Set02 - Data

STAT 401 (Engineering) - Iowa State University

January 13, 2017

Population vs sample

Definition

A population is any entire collection of objects we are interested in and would like to make statements about.

Definition

A sample is a group of units selected from the population.

Modified from `http://www.stats.gla.ac.uk/steps/glossary/basic_definitions.html`

Our typical goal as scientists and engineers is to take a sample to make a statement about the population.

Examples of populations

Examples of populations

- all labs in the world
- all iPhone 5s
- all diamonds bits for machine cutting
- all watersheds in Iowa
- ⋮

What are some examples of populations from your research?

Inference to this population

Definition

An inference is a conclusion that patterns in the data are present in some broader context.

Remark: A (statistically valid) inference to a population can be drawn from a random sample from that population, but not otherwise.

Definition

A simple random sample of size n from a population is a subset of the population consisting of n members selected in such a way that every subset of size n is afforded the same chance of being selected.

Using R to obtain a simple random sample

```r
# Get 10 random numbers from 1 to 100
sample(100, size = 10)
```

```
 [1] 19 70 57 17 91 90 13 78 44 51
```

```r
# Take a data set and extract 10 random rows
n = nrow(mydata)
mydataSRS = mydata[sample(n,10),]
```

```r
# To make it reproducible use `set.seed()`
seed = 20170112
set.seed(seed)
sample(100, size = 10)
```

```
 [1] 24 82  1 42 89 92 22 70 35 29
```

```r
sample(100, size = 10) # not the same so reset the seed
```

```
 [1] 81 76 55 53 26 28 43 52 49 34
```

```r
set.seed(seed)
sample(100, size = 10) # this is the same
```

```
 [1] 24 82  1 42 89 92 22 70 35 29
```

Randomized experiments vs observational studies

Definition

An experimental unit is the object which is actually studied by a researcher; the basic objects upon which measurements are taken.

Definition

An experiment is any process or study which results in the collection of data, the outcome of which is unknown. A randomized experiment is an experiment where the investigator controls the assignment of experimental units to groups and uses a **chance mechanism** to make the assignment. In an observational study, the group status of the subjects is not controlled by the investigator.

Remark: Statistical inference of cause-and-effect relationships can be drawn from randomized experiments, but not from observational studies.

Use R to assign treatments

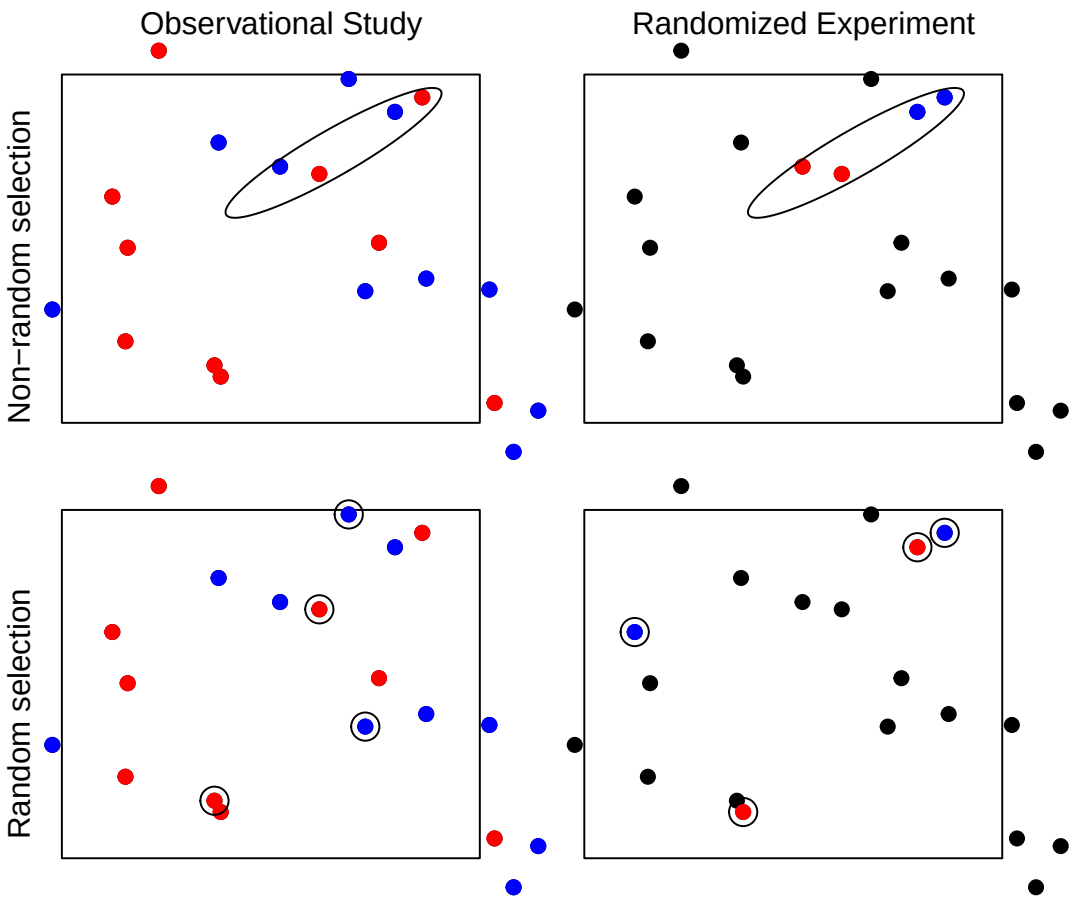
```
set.seed(20170113)
n <- 12
blocks     <- c("B1","B2")
treatments <- c("Trt1","Trt2")
data.frame(experimental_unit = paste0("EU",1:n),
           block             = rep(blocks, each = n/2)) %>%
  mutate(random              = sample(treatments, n, replace = TRUE),
         balanced            = sample(rep(treatments, each = n/2))) %>%
  group_by(block) %>%
  mutate(blocked             = sample(treatments, n/2, replace=TRUE),
         blocked_balanced    = sample(rep(treatments, each = n/4)))
```

```
Source: local data frame [12 x 6]
Groups: block [2]

   experimental_unit  block random balanced blocked blocked_balanced
               <fctr> <fctr>  <chr>    <chr>   <chr>            <chr>
1                EU1     B1   Trt1     Trt1    Trt1             Trt1
2                EU2     B1   Trt2     Trt2    Trt1             Trt2
3                EU3     B1   Trt2     Trt1    Trt1             Trt2
4                EU4     B1   Trt2     Trt1    Trt2             Trt1
5                EU5     B1   Trt1     Trt1    Trt1             Trt1
6                EU6     B1   Trt2     Trt1    Trt1             Trt2
7                EU7     B2   Trt2     Trt2    Trt2             Trt2
8                EU8     B2   Trt2     Trt1    Trt2             Trt1
9                EU9     B2   Trt1     Trt2    Trt2             Trt1
10              EU10     B2   Trt1     Trt2    Trt1             Trt1
11              EU11     B2   Trt2     Trt2    Trt2             Trt2
12              EU12     B2   Trt1     Trt2    Trt2             Trt2
```

Graphical representation

Within the box is the population and colors are the treatment.



Statistical inference

	Observational Study	Randomized Experiment
Non-random Selection		Causal Inference
Random Selection	Inference to Population	Causal Inference to Population

- Random sampling $\rightarrow$ inference to population
- Random treatment assignment $\rightarrow$ causal inference

ZMapp therapy for Ebola

Current Ebola status: `http://www.cdc.gov/vhf/ebola/outbreaks/guinea/`

from: `http://en.wikipedia.org/wiki/ZMapp`

> *In 2014, Samaritan's Purse worked with the FDA and Mapp Biopharmaceutical to make the drug available to two of its health workers, who were infected by Ebola during their work in Liberia, under the Expanded access program. At the time, there were only a few doses of ZMapp in existence. According to news reports, Kent Brantly received the first dose of ZMapp nine days after falling ill. According to Samaritan's Purse, Brantly received a blood transfusion from a 14-year old boy who survived an Ebola virus infection before being treated with the ZMapp serum. Nancy Writebol, working alongside Brantly, was also treated with Zmapp. The condition of both health workers improved, especially in Brantly's case, before being transported back to the United States, to Emory University Hospital, specialized for Ebola treatment. Writebol and Brantly were released from hospital on August 21, 2014.*
>
> *A Roman Catholic priest, 75-year-old Miguel Pajares, was flown back to Spain from Monrovia on 7 August after being infected with Ebola. With the permission of Spains drug safety agency, he was given ZMapp. He died on August 12, two days after receiving the drug.*
>
> *The west African nation of Liberia, which has been affected by the 2014 outbreak, has secured enough ZMapp to treat three individual Liberians with the disease. One of the three to receive the drug, Dr. Abraham Borbor, a Liberian doctor and deputy chief physician at Liberia's largest hospital, died August 25th, 2014.*
>
> *William Pooley, a British male nurse who contracted Ebola while working in Sierra Leone, was also treated with ZMapp in August 2014.*

Question: Is ZMapp an effective therapeutic for the treatment of Ebola and prevention of death?

Uncertainty/Randomness

Definition

Uncertainty is a lack of certainty, a state of having limited knowledge where it is impossible to exactly describe current state or future outcome, (or the existence of more than one possible outcome).

Example

- The hardness measurement for a particular cubic boron nitride sample.
- Time until failure of a system component.
- The brightness of a light on an electrical circuit.

Probability and Statistics

We want to study physical processes that are not completely deterministic. Using probability and statistics to understand the random components of such processes can help us do this.

- Probability: mathematical theory for modeling *data* where outcomes occur randomly.
- Statistics: use data to make inferences about questions of interest

Because statistical inference makes use of probability models, probability is a foundation for statistics.