STAT 401A - Statistical Methods for Research Workers

Statistical Inference

Jarad Niemi (Dr. J)

Iowa State University

last updated: September 2, 2014

Population vs sample

Definition

A population is any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.

Definition

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

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

Examples of populations

Taken from `http://www.epa.gov/agriculture/ag101/demographics.html`

- "people living in the United States"
- individuals that "claim farming as an occupation"
- "farms"
- individuals who "actually live on farms"
- "small family farms"
- ⋮

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 An 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 software to obtain a simple random sample

SAS

```
PROC SURVEYSELECT DATA=mydata
    METHOD=srs N=100 OUT=mydataSRS;
RUN;
```

R

```
n = nrow(d)
mydataSRS = mydata[sample(n,100),]
```

Excel

```
RAND()
sort
```

Randomized experiments vs observational studies

Definition

An experimental unit is a person, animal, plant or thing which is actually studied by a researcher; the basic objects upon which the study or experiment is carried out.

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.

Remark In an observational study, the group status of the subjects is not controlled by the investigator.

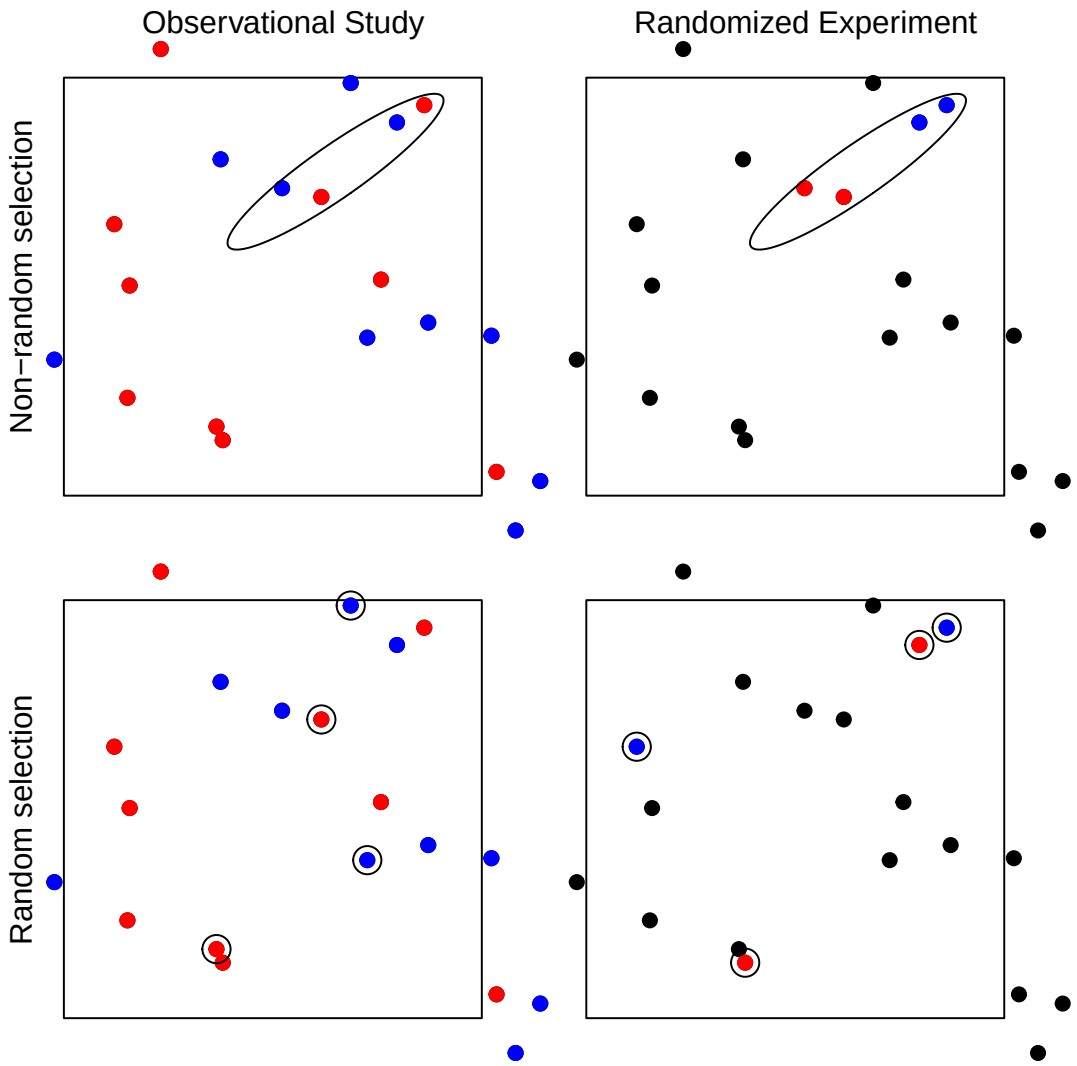
Remark Statistical inference of cause-and-effect relationships can be drawn from randomized experiments, but not from observational studies.

Chance mechanism

How do you assign experimental units to groups?

Remark Use a computer, e.g. SURVEYSELECT in SAS or `sample` in R, to assign experimental units to groups!

Graphical representation



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?

Scientific hypotheses

http://en.wikipedia.org/wiki/Null_hypothesis

Definition

The null hypothesis is a general statement or default position that there is no relationship between two measured phenomena.

Definition

The alternative hypothesis is a general statement or default position that there IS a relationship between two measured phenomena.

Examples of null hypotheses

- Hog feed makes no difference on average daily gain
- Fertilizer level has no effect on corn yield
- Prairie strips do not decrease nitrogen leaching
- Logging has no effect on bird populations

Pvalues

Definition

A statistic is a numerical quantity calculated from the data. A test statistic is a statistic used to measure the plausibility of an alternative hypothesis to a null hypothesis.

Definition

A pvalue is the probability of observing a test statistic as (extreme as) or more extreme than that observed if the null hypothesis is true.

Randomization pvalues

Remark In a randomized experiment if a treatment has no effect, we should see (on average) no difference in means (or other test statistics) between two groups with different treatments.

Consider the following experiment:

- Two fertilizers (A and B) are randomly assigned: A to 3 plots and B to 2 plots.
- The observed corn yield (bushels per acre) are A: 136, 146, 140 and B: 145, 139.
- The difference in means (A-B) is -1.33.
- Is this difference significant?

Remark Calculate a randomization pvalue by calculating the difference in means for every possible treatment assignment and calculate the proportion of times the difference in these means is as or more extreme (farther away from zero) than observed (-1.33).

```
library(combinat)
fertilizer = c("A","A","A","B","B")
yield = c(136,146,140,145,139)
rands = as.data.frame(matrix(unlist(unique(permn(fertilizer))),ncol=5,byrow=TRUE))
names(rands) = yield
rands$meanA = apply(rands, 1, function(x) mean(yield[x=='A']))
rands$meanB = apply(rands, 1, function(x) mean(yield[x=='B']))
rands$diffs = with(rands, meanA-meanB)
rands
```

```
   136 146 140 145 139 meanA meanB   diffs
1    A   A   A   B   B 140.7 142.0 -1.3333
2    A   A   B   A   B 142.3 139.5  2.8333
3    A   B   A   A   B 140.3 142.5 -2.1667
4    B   A   A   A   B 143.7 137.5  6.1667
5    B   A   A   B   A 141.7 140.5  1.1667
6    A   B   A   B   A 138.3 145.5 -7.1667
7    A   A   B   B   A 140.3 142.5 -2.1667
8    A   B   B   A   A 140.0 143.0 -3.0000
9    B   A   B   A   A 143.3 138.0  5.3333
10   B   B   A   A   A 141.3 141.0  0.3333
```

Calculate the proportion of diffs that have absolute value greater than $|-1.33|$.

```
truediff = mean(yield[fertilizer=="A"])-mean(yield[fertilizer=="B"])
mean(rands$diffs <= -abs(truediff) | rands$diffs >= abs(truediff))
```

```
[1] 0.8
```

Permutation pvalues

Remark In an observational study, if the group has no effect, we should see (on average) no difference in means (or other test statistics) between two groups.

Consider the following observational study:

- Five plots were sampled
  3 on the West side of a river and 2 on the East side of a river
- The observed corn yield (bushels per acre) are
  W: 136, 146, 140 and E: 145, 139.
- The difference in means (W-E) is -1.33.
- Is this difference significant?

Remark Calculate a permutation pvalue by calculating the difference in means for every possible permutation of observations and calculate the proportion of times the difference in these means is as or more extreme (farther away from zero) than observed (-1.33).

```
side = c("W","W","W","E","E")
perms = as.data.frame(matrix(unlist(permn(yield)), ncol=5))
names(perms) = side
perms$meanW = rowSums(perms[,1:3])/3
perms$meanE = rowSums(perms[,4:5])/2
perms$diffs = with(perms, meanW-meanE)
head(perms,10)
```

```
     W   W   W   E   E meanW meanE   diffs
1  136 139 140 140 145 138.3 142.5 -4.1667
2  146 145 136 146 139 142.3 142.5 -0.1667
3  140 136 145 139 146 140.3 142.5 -2.1667
4  145 140 139 145 140 141.3 142.5 -1.1667
5  139 146 146 136 136 143.7 136.0  7.6667
6  136 139 140 140 145 138.3 142.5 -4.1667
7  146 136 136 139 146 139.3 142.5 -3.1667
8  140 145 145 146 139 143.3 142.5  0.8333
9  139 140 146 145 140 141.7 142.5 -0.8333
10 145 146 139 136 136 143.3 136.0  7.3333
```

```
pvalue = mean(perms$diffs<=-1.33 | perms$diffs>=1.33)
pvalue
```

```
[1] 0.6333
```

```
hist(perms$diffs,20, main="Permutation distribution", xlab="Theoretical differences (W-E)")
abline(v=c(-1.33,1.33), col="red", lwd=2)
```