08 - Statistical Inference

HCI/PSYCH 522
Iowa State University

February 15, 2022

Overview

- Inference
  - Population vs sample
  - Convenience sample
  - Random sample
  - Statistical inference
- Causality
  - Observational study
  - (Quasi-)experiment
  - Randomized experiment
  - Causal inference

Population

https://www.scribbr.com/methodology/population-vs-sample/

Definition

The population is the entire group that you want to draw conclusions about.

Examples:

- All graduate students at ISU
- All Mac M1 minis
- All interior design apps

Sample

https://www.scribbr.com/methodology/population-vs-sample/

Definition

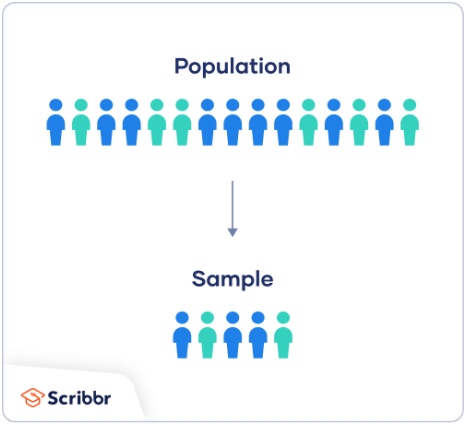
A sample is the specific group you will collect data from.

Examples:

- HCI students at ISU
- My Mac M1 mini
- All interior design apps on the Apple App Store

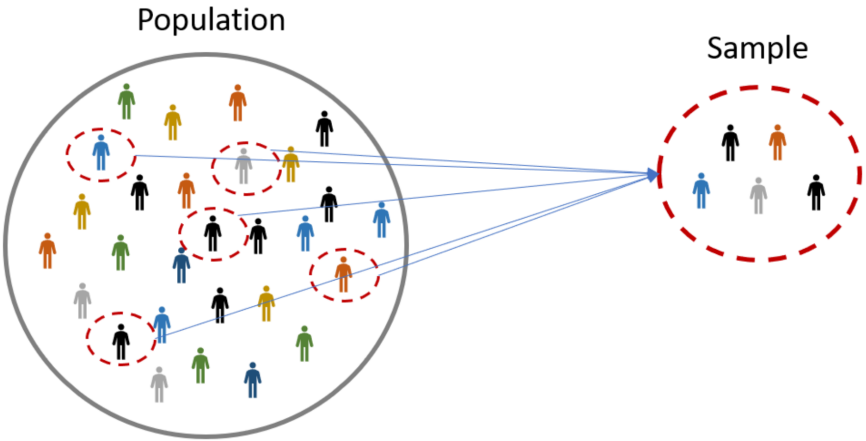
Population vs Sample

https://www.scribbr.com/methodology/population-vs-sample/



Population vs Sample

https://www.omniconvert.com/what-is/sample-size/



Parameters

https://en.wikipedia.org/wiki/Statistical_parameter

Definition

A parameter is any measur[able] quantity of a statistical population that summarizes or describes an aspect of the population.

Examples:

- Population mean
- Population standard deviation
- Population probability of success
- Population probability density function

Statistics

Definition

A statistic is a function of your data.

Examples:

- Numeric quantities
  - Sample mean
  - Sample standard deviation
  - Sample proportion of success
- Graphical statistics
  - Histogram
  - Scatterplot

Estimator

Definition

An estimator is a statistic that estimates a population parameter.

Examples:

- Sample mean estimates the population mean
- Sample standard deviation estimates the population mean
- Sample proportion of successes estimates the population probability of success
- Histogram estimates the probability density function

Sampling error

https://en.wikipedia.org/wiki/Sampling_error

Definition

Sampling error is the error caused by observing a sample instead of the whole population.

Examples:

- Sample mean minus population mean
- Sample proportion minus population probability

Representative sample

https://www.investopedia.com/terms/r/representative-sample.asp

Definition

A representative sample is a subset of a population that seeks to accurately reflect the characteristics of the larger group.

Random samples are probabilistically "guaranteed" to be representative.

Convenience sample

https://en.wikipedia.org/wiki/Convenience_sampling

Definition

Convenience sampling is a type of non-probability sampling that involves the sample being drawn from that part of the population that is easy to contact or reach.

Examples:

- Students in HCI 522 representing all ISU graduate students
- ISU undergraduate students representing all people
- Interior design apps on the App Store representing all interior design apps

Random sample

Definition

A random sample is a sampling technique that uses random mechanism to include individuals in the sample.

Random mechanism examples:

- Rolling dice
- Lotteries
- Random number table
- Random number generation, e.g. RAND() in Excel

Simple random sample

https://www.investopedia.com/terms/s/simple-random-sample.asp

Definition

A simple random sample is a subset of the population in which each member of the subset has an equal probability of being chosen.

Simple random sample in R

```
n <- 10000 # enumerate all n individuals
sample(n, size = 10)

##  [1] 7085 7707  399 4713 8797 9625 2386 1954  164 6308
```

Alternatively

```
data.frame(individual = 1:n,
           random_number = runif(n)) %>% # RAND() in Excel
  arrange(random_number) %>%
  head(10) %>%
  pull(individual)

##  [1] 9233 6764 1195 8612 8273 4602 5213 5452 3906 7771
```

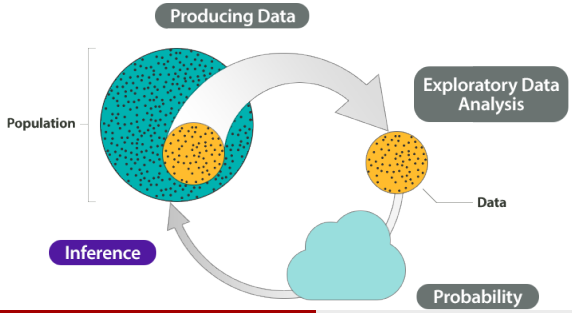
Statistical inference

https://online.stanford.edu/courses/stats200-introduction-statistical-inference

Definition

Statistical inference is the process of using data [from a sample] to draw conclusions about a population.

https://courses.lumenlearning.com/wm-concepts-statistics/chapter/wim-linking-probability-to-statistical-inference/



Binomial example

Parameter: proportion of app users who use a chatbot
Statistic: 6 out of 20 randomly sampled users use the chatbot

```
y <- 6; n <- 20; a <- 1-0.95
qbeta(c(a/2,1-a/2), 1+y, 1+n-y) %>% round(2)

## [1] 0.15 0.52
```

This 95% credible interval is a statement about the population parameter.

Normal example

Parameter: mean EEG alertness level of air traffic controllers
Statistic: in a random sample of 30 air traffic controllers, normalized EEG alertness level had a mean of 73 and a standard deviation of 5

```
n <- 30; mn <- 73; sd <- 5; se <- sd/sqrt(n); a <- 1-0.95
(qt(c(a/2,1-a/2), df = n-1)*se + mn) %>% round(1)

## [1] 71.1 74.9
```

The following 95% credible interval is a statement about the population parameter.

Observational study

https://study.com/academy/lesson/observational-study-in-statistics-definition-examples.html

Definition

An observational study is a study in which the researcher simply observes the subjects without interfering.

Examples:

- Recording how long it takes an undergraduate student to register
- Recording usage of a twitter hashtag

(Quasi-)experiment

Definition

A (quasi-)experiment is a study where the researcher (non-randomly) assigns a treatment to an experimental unit, e.g. individual.

Examples:

- Live HCI 522 students must use R while asynchronous students must use Excel
- First 100 students to register have no chatbot, next 100 students to register have a chatbot

Randomized experiment

Definition

A randomized experiment is a study where the researcher assigns a treatment to an experimental unit, e.g. individual.

Examples:

- Randomly forcing HCI 522 students to use R or Excel
- Each student randomly gets a chatbot or not when registering for courses

Randomizing treatments in R

```
data.frame(subject = 1:10) %>%
  mutate(treatment = sample(c("A","B"), size = n(), replace = TRUE))
```

```
##    subject treatment
## 1        1         A
## 2        2         B
## 3        3         B
## 4        4         B
## 5        5         B
## 6        6         B
## 7        7         B
## 8        8         B
## 9        9         A
## 10      10         B
```

Randomizing treatments in R

```
data.frame(subject = 1:10) %>%
  mutate(treatment = sample(c("A","B","C"), size = n(), replace = TRUE, prob = c(2,3,5)/10))
```

```
##    subject treatment
## 1        1         B
## 2        2         C
## 3        3         A
## 4        4         A
## 5        5         C
## 6        6         B
## 7        7         B
## 8        8         C
## 9        9         B
## 10      10         A
```

Balanced

Definition

An experiment is balanced if there are the same number of experimental units for each treatment (or combination of treatments).

Balanced treatments in R

```
data.frame(subject = 1:10) %>%
  mutate(treatment = sample(rep(c("A","B"), times = 5), size = n()))
```

```
##    subject treatment
## 1        1         B
## 2        2         B
## 3        3         B
## 4        4         A
## 5        5         A
## 6        6         A
## 7        7         B
## 8        8         B
## 9        9         A
## 10      10         A
```

Randomizing treatments in R

```
data.frame(subject = 1:10) %>%
  mutate(treatment = sample(rep(c("A","B","C"), times = c(2,3,5)), size = n()))
```

```
##    subject treatment
## 1        1         C
## 2        2         B
## 3        3         B
## 4        4         C
## 5        5         A
## 6        6         C
## 7        7         C
## 8        8         B
## 9        9         C
## 10      10         A
```

Causal inference

https://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i4418.xml

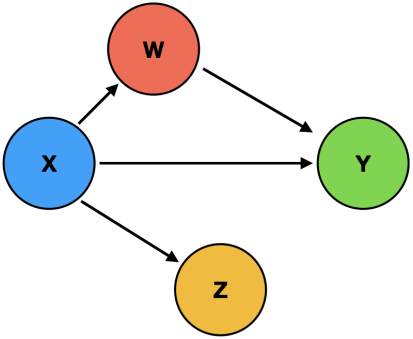
Definition

Causal inference refers to the process of drawing a conclusion that a specific treatment (i.e., intervention) was the "cause" of the effect (or outcome) that was observed.

Causal inferences can only be drawn from randomized experiments.

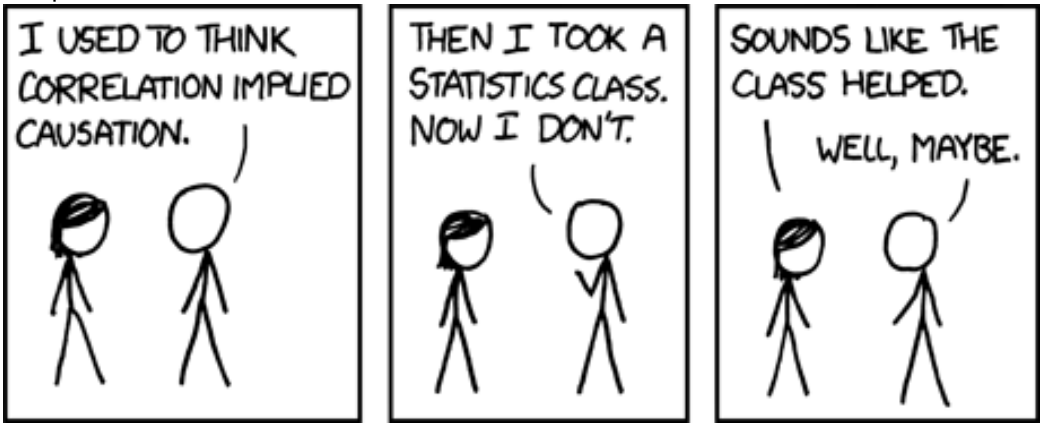
Causal inference

https://towardsdatascience.com/causal-inference-962ae97cefda



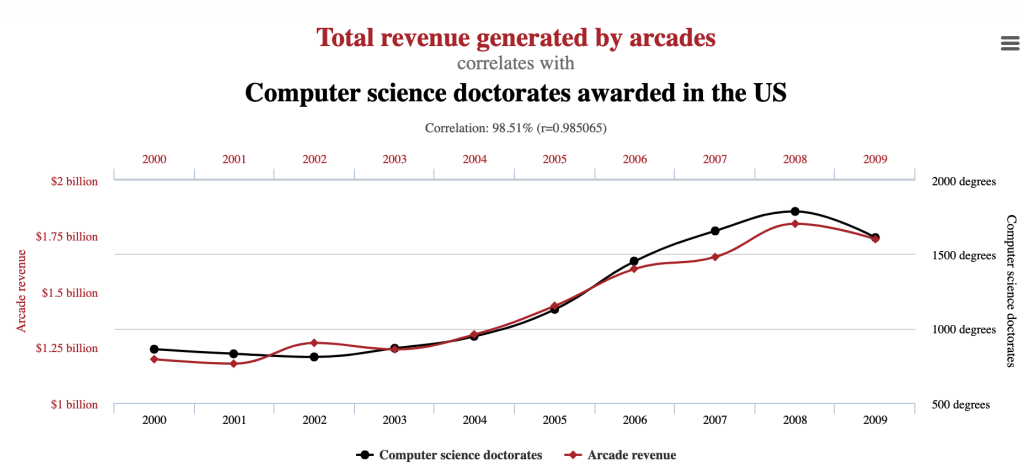
Correlations (not causation)

Correlations can be inferred to the population from an observational study based on a random sample.



Spurious correlations

https://www.tylervigen.com/spurious-correlations



Binomial comparison

Scientific question: effect of chatbot on probability to correctly register for courses
Experiment: 20 volunteer undergraduate students were randomly assigned a chatbot or no chatbot. Amongst the 10 in the chatbot group, all successfully registered. In the non-chatbot group 8/10 successfully registered.

```
n_reps <- 100000; a <- 1-0.95
theta_chatbot   <- rbeta(n_reps, shape1 = 1+10, shape2 = 1+10-10)
theta_nochatbot <- rbeta(n_reps, shape1 = 1+ 8, shape2 = 1+10- 8)
quantile(theta_chatbot - theta_nochatbot, probs = c(a/2, 1-a/2)) %>% round(2)

##  2.5% 97.5%
## -0.11  0.46
```

This 95% credible interval is a causal effect of the treatment (chatbot) but only for those students in this study.

Normal comparison

Scientific question: effect of chatbot on course registration time
Experiment: 40 randomly chosen ISU undergraduate students were randomly assigned a chatbot or no chatbot. Amongst the 21 in the chatbot group, the sample mean time was 2 minutes with a standard deviation of 1 minute. Amongst the 19 in the no chatbot group, the sample mean time was 1.5 minutes with a standard deviation of 0.75 minutes.

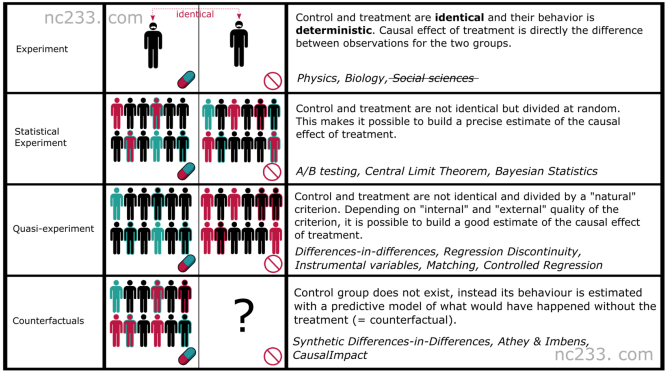
```
n_reps <- 100000; a <- 1-0.95
mu_chatbot   <- rt(n_reps, df = 21-1)*(   1/sqrt(21)) + 2
mu_nochatbot <- rt(n_reps, df = 19-1)*(0.75/sqrt(19)) + 1.5
quantile(mu_chatbot - mu_nochatbot, probs = c(a/2, 1-a/2)) %>% round(2)

##  2.5% 97.5%
## -0.08  1.08
```

This 95% credible interval is a causal effect of the treatment (chatbot) for all ISU undergraduate students.

Causal inference cheat sheet

https://nc233.com/2020/04/causal-inference-cheat-sheet-for-data-scientists/



Levels of evidence ladder for causal inference methods

Summary

- Samples
  - Population vs sample
  - Convenience sample
  - Random sample
  - Statistical inference
- Randomized experiment
  - Observational study
  - (Quasi-)experiment
  - Randomized experiment
  - Causal inference