

Introduction to statistics

Yonghyun Kwon

Department of Mathematics, Korea Military Academy

Course Objectives

Objectives:

- Understand the basic theories and principles of statistics.
- Be able to analyze real data using R program.

Eligible Cadets and Course Credits:

- First-year cadets (3 credits / 3 hours per week)

Grading:

- Quizzes (30%) - Four Quizzes (90%), Attitude in class (10%)
- Midterm Exam (30%)
- Final Exam (40%)



Important Course Guidelines

- Bring the textbook and the **printed lecture notes** to class.
 - Textbook: Openintro Statistics(4th ed.); pdf available online.
- All the course materials are available on **LMS**.
- **Evening study sessions** will be conducted
 - for those who do not meet the desired level (scoring below 60 on quizzes or midterm) and volunteers.
- Grading on **attitude**:
 - Deductions for poor classroom behavior.
 - Positive behavior can offset deductions.
 - We go over the assignments every Tuesday in class.
- **Section leader's** responsibilities:
 - Clean the blackboard, and perform other assigned duties.



Chapter 1: Introduction to data

Yonghyun Kwon

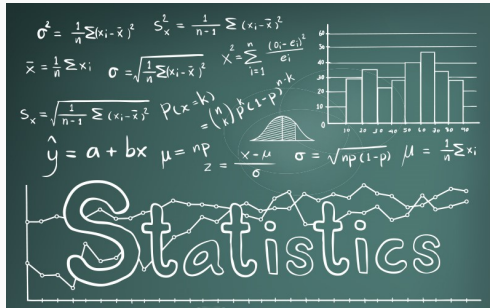
Department of Mathematics, Korea Military Academy

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

The slides may be copied, edited, and/or shared via the CC BY-SA license.

Some images may be included under fair use guidelines (educational purposes).

What is statistics?



collect -> analyze -> conclude

= the whole course

- **Statistics** is the study of how best to collect, analyze, and draw conclusions from data.
- **Data** is collected from various sources like field notes, surveys, and experiments.

Classroom survey

A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey, and the corresponding variables the data from the responses were stored in:

6 Variables

- gender: What is your gender?
- intro_extra: Do you consider yourself introverted or extraverted?
- sleep: How many hours do you sleep at night, on average?
- bedtime: What time do you usually go to bed?
- countries: How many countries have you visited?
- dread: On a scale of 1-5, how much do you dread being here?

each question -> one column



Data matrix

Data collected on students in a statistics class on a variety of variables:

variable = a column



Stu.	gender	intro_extra	...	dread
1	male	extravert	...	3
2	female	extravert	...	2
3	female	introvert	...	4
4	female	extravert	...	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	...	3



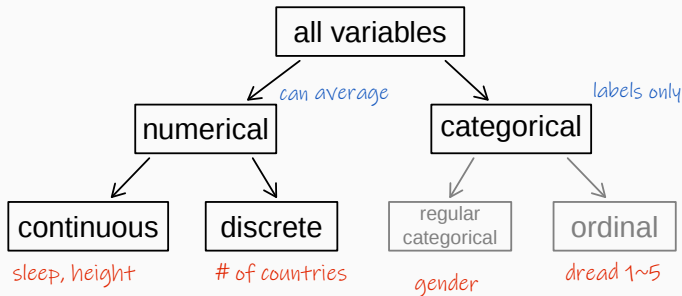
observation

= a row (one student)

$n = 86$



Types of variables



Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2	0	2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- gender: *categorical*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: *numerical, discrete*
- dread: *categorical, ordinal*

5.5 is not whole
-> continuous

'10-12' is a range,
and it has an order

What type of variable is a telephone area code?

(a) numerical, continuous

(b) numerical, discrete

(c) categorical *just a label*

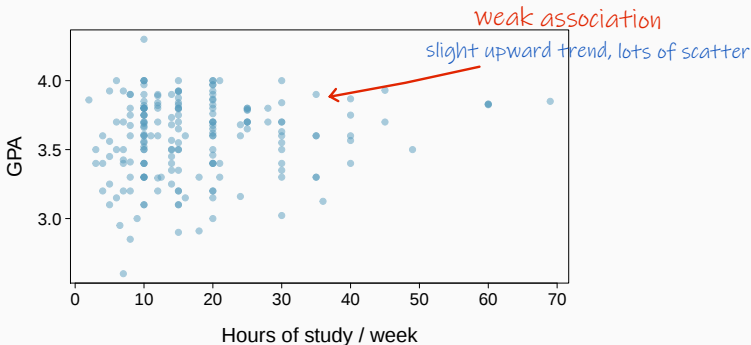
(d) categorical, ordinal *no natural order*

area codes 02, 031, 051 ... averaging them means nothing



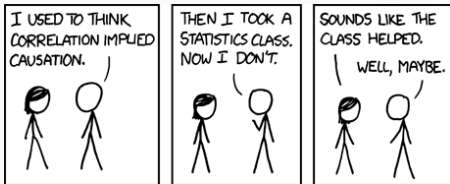
Relationships among variables

Does there appear to be a relationship between GPA and number of hours students study per week?



Association vs. causation

- When two variables show some connection with one another, they are called *associated* variables.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be *independent*. *associated $\neq$ one causes the other*
- In general, association does not imply causation, and causation can only be inferred from a randomized experiment.



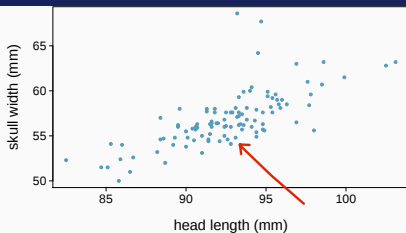
only a randomized experiment can show cause

<http://xkcd.com/552/>



Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



goes up together -> positive

- (a) There is no relationship between head length and skull width, i.e. the variables are independent.
- (b) Head length and skull width are positively associated.**
- (c) Skull width and head length are negatively associated.
- (d) A longer head causes the skull to be wider.
- (e) A wider skull causes the head to be longer.

*cause cannot be read
off a scatterplot*

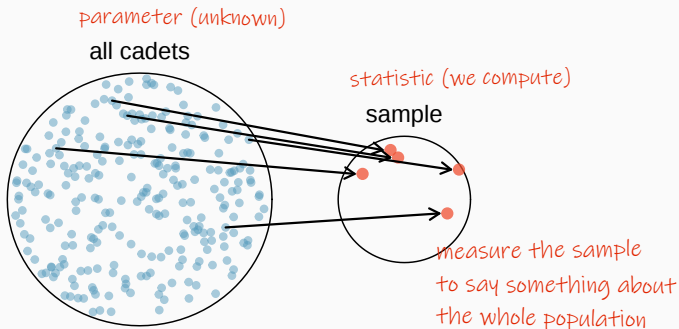


Populations and samples

Population: a set of similar items or events which is of interest.

Sample: a subset of the population.

Example: Estimate the average height of KMA cadets by collecting a sample of cadets.



Populations and samples(cont.)

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS



David De Looze/Getty Images

<http://well.blogs.nytimes.com/2012/08/29/>

finding-your-ideal-running-form

Research question: Can people become better, more efficient runners on their own, merely by running?

Population of interest:

all people who run

Sample: Group of adult women

Population to which results can be generalized: adult women

- only if the sample was random

