INTRODUCTION TO STATISTICS

GONZALO G. PERAZA MUES

AN INTRODUCTORY CASE

Is it better for children to start school at a younger or older age?

WHAT DOES DATA TELLS US?

- Older students fare better in tests applied at the end of the first year.
- Why?
- Can we gather more significant data?

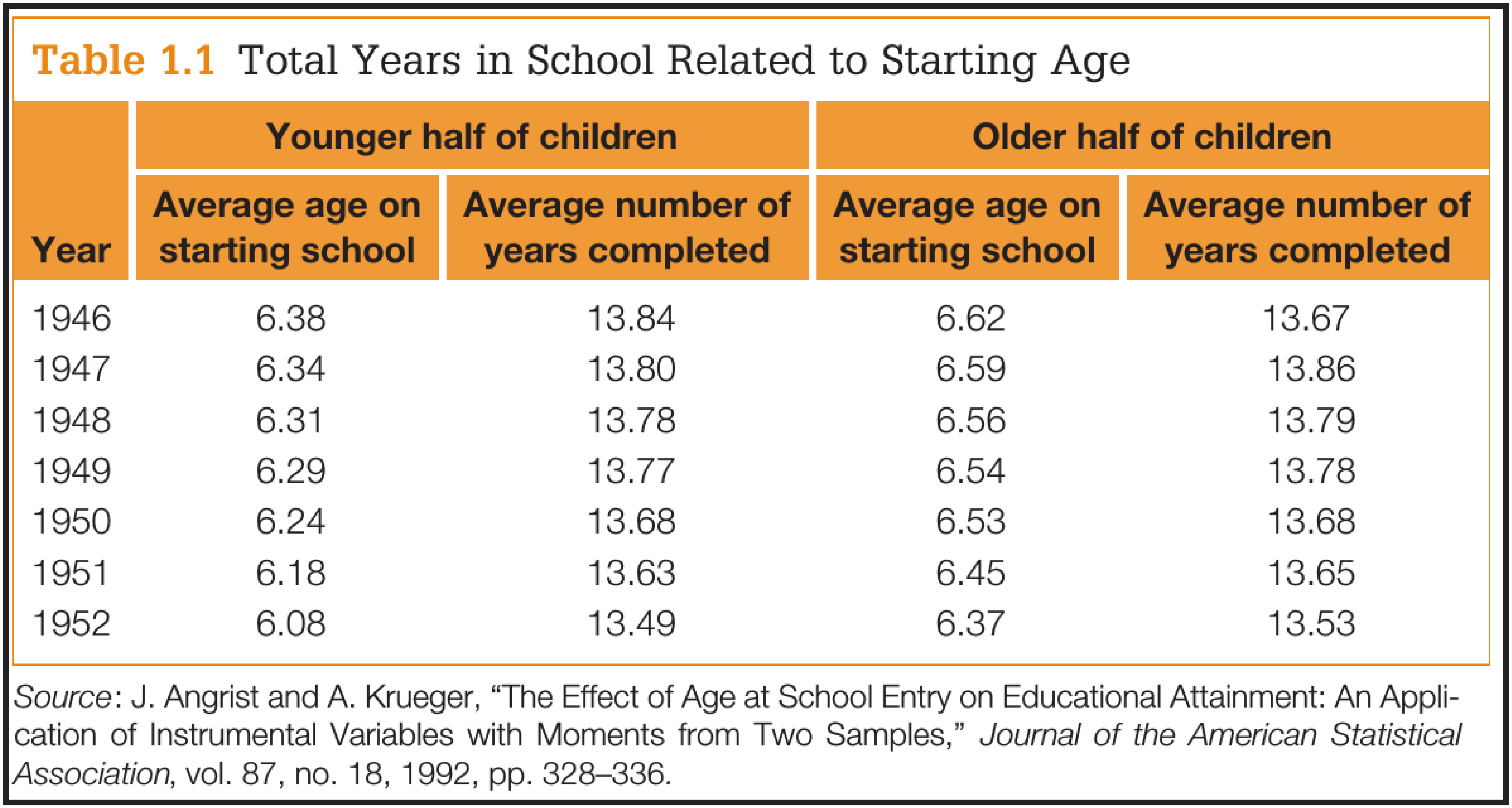


Table 1.1 Total Years in School Related to Starting Age

Year	Younger half of children		Older half of children	
	Average age on starting school	Average number of years completed	Average age on starting school	Average number of years completed
1946	6.38	13.84	6.62	13.67
1947	6.34	13.80	6.59	13.86
1948	6.31	13.78	6.56	13.79
1949	6.29	13.77	6.54	13.78
1950	6.24	13.68	6.53	13.68
1951	6.18	13.63	6.45	13.65
1952	6.08	13.49	6.37	13.53

Source: J. Angrist and A. Krueger, "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, vol. 87, no. 18, 1992, pp. 328–336.

DEFINITION OF STATISTICS

The art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.

DATA

- Collecting
- Organizing
- Analyzing
- Interpreting

THE RAW MATERIAL OF STATISTICS IS DATA

DATA?

DATA IN STATISTICS

In statistics, "Data" is conceptualized as having objects on which we observe or measure one or more characteristics.

Objects -> Individuals

Characteristics -> Variables

WHY VARIABLES

A characteristics that varies from one individual to another.

individuals

- observations
- subjects
- objects
- cases

variables

- characteristics
- attributes
- features
- traits

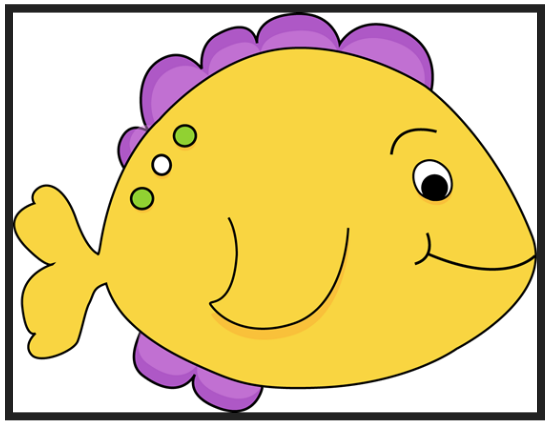
VARIABLES PLAY THE STARRING ROLE IN STATISTICAL STUDIES

Given an object (individual)

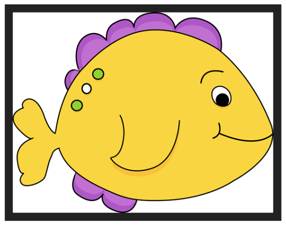
What variables (characteristics) can you think about?

- color
- flavor
- weight
- volume
- ripeness
- region

- calories
- sugars
- vitamins
- seeds
- sweetness
- skin opacity



What variables (characteristics) can you think about?



- length
- weight
- volume
- age
- # fins

- sex
- diet
- size
- main color
- speed

VARIABLES

- Qualitative
  - non numerical
  - information
- Quantitative
  - numerical
  - information

SOME QUALITATIVE VARIABLES

- Gender of newborn
  - male, female
- Icecream flavors
  - chocolate, vanilla, lemons
- Types of pasta
  - spaghetti, tortellini, ravioli

- Frequency of usage
  - never, sometimes, always
- Clothe sizes
  - XS, SM, MD, LG, XL
- Levels of spicyness
  - none, mild, moderate, high

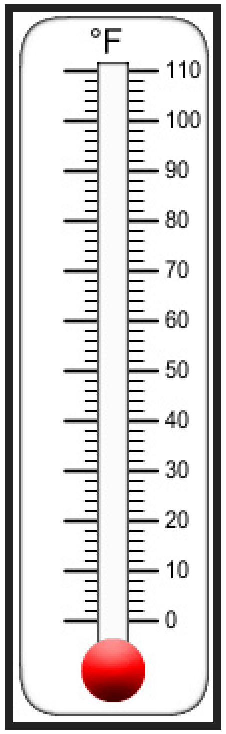
SOME QUANTITATIVE VARIABLES

- Temperature
  - -10, 0, 15, 45, 70, 100
- year number
  - 1776, 1810, 1905, 1999, 2005

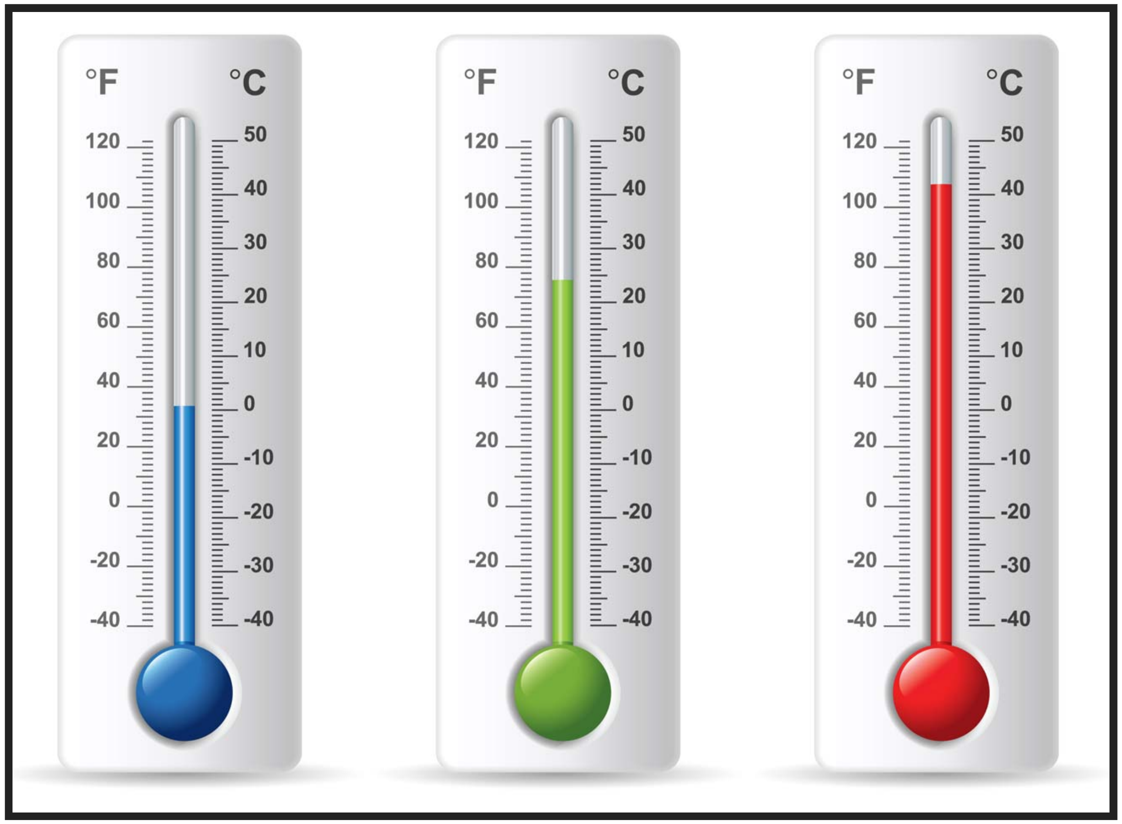
- age
  - 19yrs, 22yrs, 25yrs, 24yrs
- income
  - $0, $99.99, $2350, $1000
- weight
  - 3kg, 6.5kg, 16.3kg

HOW WOULD YOU CHANGE A QUANTITATIVE VARIABLES INTO A QUALITATIVE ONE?

Converting temperature into a qualitative variables



Cold	Mild	Hot
< 5 ^{o}C	5 - 25	^{o}C > 25 ^{o}C



HOW WOULD YOU CHANGE A QUALITATIVE VARIABLES INTO A QUANTITATIVE ONE?

Switching to quantitative variables

Frequency of usage

- never
- rarely
- sometimes
- often
- always

Quantifying

- times/day
- # days
- # weeks
- months
- years

Switching to quantitative variables

Icecream flavors

- chocolate
- vanilla
- lemon

Quantifying

- sugar content
- milk content
- pH

MORE ABOUT QUANTITATIVE VARIABLES

Continuous

Discrete

DISCRETE QUANTITATIVE VARIABLES

Takes only a finite number of values or a countable number of values.

DISCRETE QUANTITATIVE VARIABLES

- Number of days in a year
- Number of laps you can swim in 5 minutes
- Number of goals in the FIFA World Cup

CONTINUOUS QUANTITATIVE VARIABLE

Takes on any of the countless number of values in a line interval

CONTINUOUS QUANTITATIVE VARIABLE

- Time to run 100 meters
- Distance while running 30 minutes
- Size of a text file

TWO TYPES OF STATISTICS

- Descriptive statistics: The part of statistics concerned with the description and summarization of data.
- Inferential statistics: The part of statistics concerned with the drawing of conclusions from data.

DATA COLLECTION

- Existing data sets: government, institutions, etc. Examples:
  - INEGI: http://www.inegi.org.mx/
  - World bank open data: http://data.worldbank.org/
- Collecting new data from experiments. Examples:
  - Surveys
  - Measurements

EXAMPLE

Design an experiment to test a new drug.

VARIABLES OF INTEREST

UNKNOWN VARIABLES

REDUCING INFLUENCE OF UNKNOWNS

- Control group
- Random sampling

HOW SURE ARE WE THAT THE RESULTS ARE REAL?

- We need to take into account the influence of chance.

THINK ABOUT A COIN

- 7 heads out of 10 flips
- 47 heads out of 50 flips

PROBABILITY MODELS

- Assumptions about the chances (probabilities).

POPULATION

The total collection of all the elements that we are interested in.

SAMPLE

- A subgroup of the population that we study in detail.
- Must be representative about the population.
- It is always an approximation.
- How do we choose a sample?

RANDOM SAMPLING

A sample of k members of a population is said to be a random sampling if the members are chosen in such a way that all possible choices of the k members are equally likely.

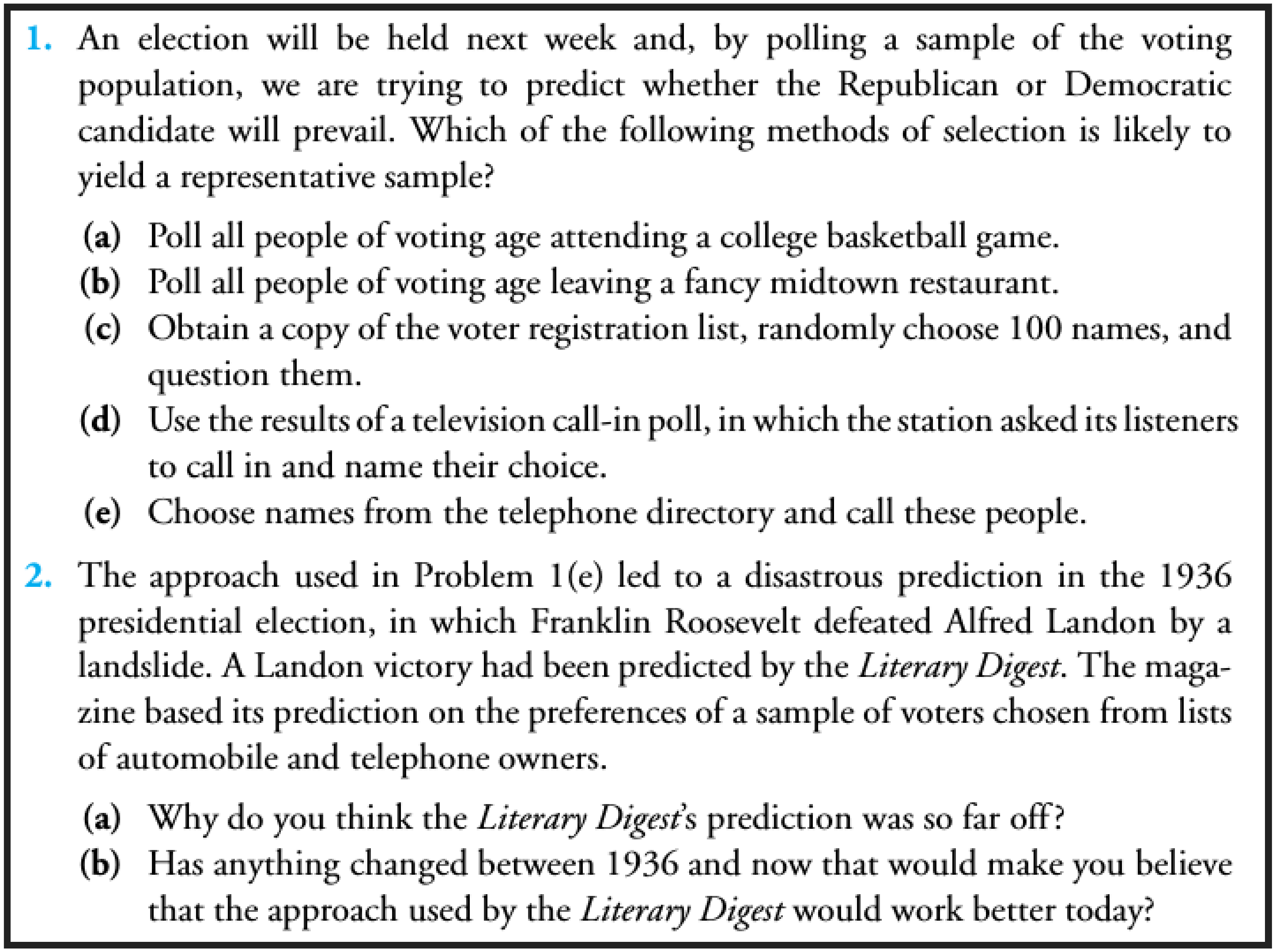
CONCLUSION FROM THE DATA ARE ALWAYS EXACT FOR THE SAMPLE, BUT NEVER FOR THE POPULATION.

STRATIFIED RANDOM SAMPLING

RECOMMENDED READING

How to Lie with Statistics

Darrell Huff

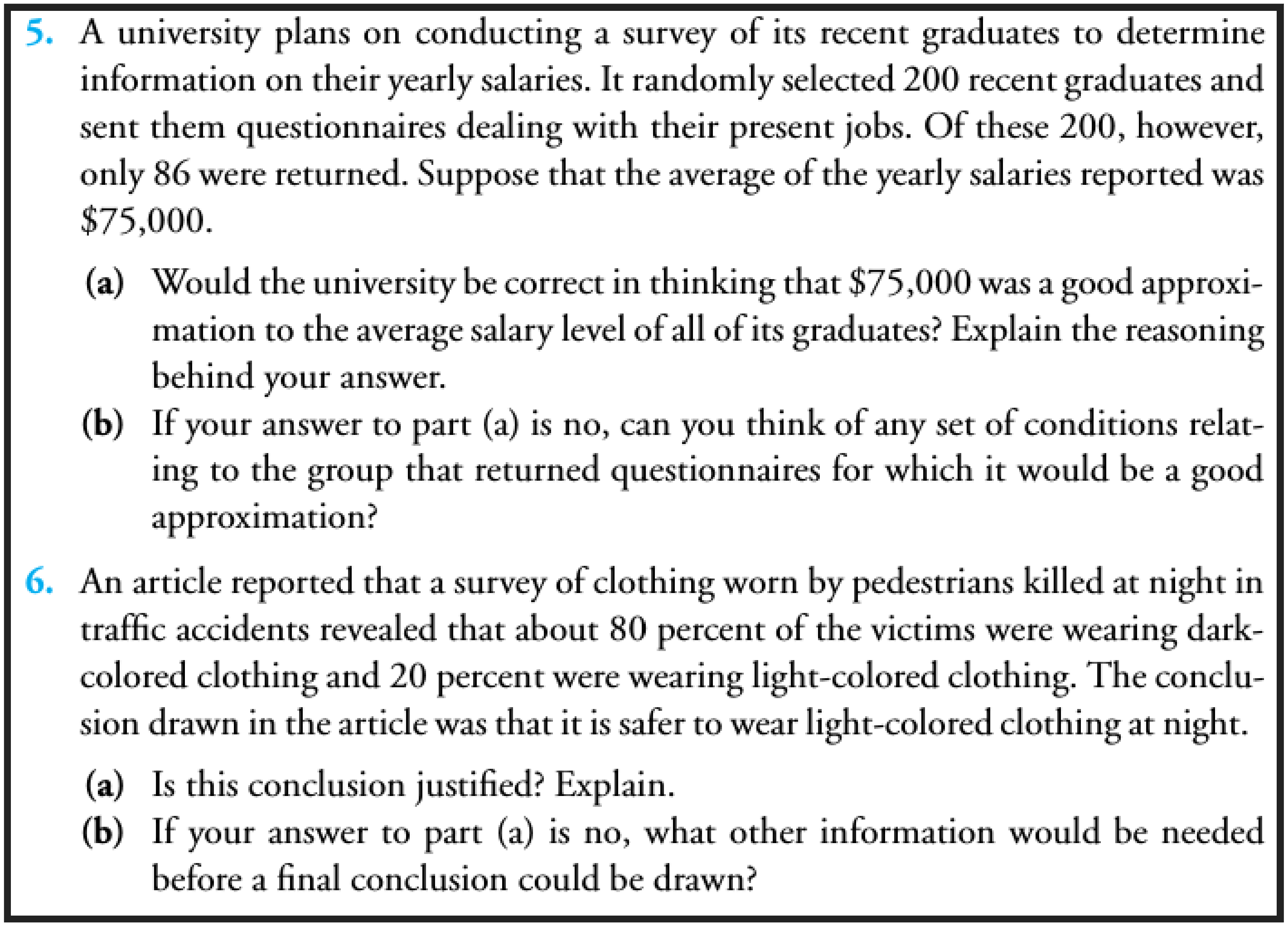


1. An election will be held next week and, by polling a sample of the voting population, we are trying to predict whether the Republican or Democratic candidate will prevail. Which of the following methods of selection is likely to yield a representative sample?

   (a) Poll all people of voting age attending a college basketball game.
   (b) Poll all people of voting age leaving a fancy midtown restaurant.
   (c) Obtain a copy of the voter registration list, randomly choose 100 names, and question them.
   (d) Use the results of a television call-in poll, in which the station asked its listeners to call in and name their choice.
   (e) Choose names from the telephone directory and call these people.

2. The approach used in Problem 1(e) led to a disastrous prediction in the 1936 presidential election, in which Franklin Roosevelt defeated Alfred Landon by a landslide. A Landon victory had been predicted by the *Literary Digest*. The magazine based its prediction on the preferences of a sample of voters chosen from lists of automobile and telephone owners.

   (a) Why do you think the *Literary Digest*'s prediction was so far off?
   (b) Has anything changed between 1936 and now that would make you believe that the approach used by the *Literary Digest* would work better today?



5. A university plans on conducting a survey of its recent graduates to determine information on their yearly salaries. It randomly selected 200 recent graduates and sent them questionnaires dealing with their present jobs. Of these 200, however, only 86 were returned. Suppose that the average of the yearly salaries reported was $75,000.

   (a) Would the university be correct in thinking that $75,000 was a good approximation to the average salary level of all of its graduates? Explain the reasoning behind your answer.

   (b) If your answer to part (a) is no, can you think of any set of conditions relating to the group that returned questionnaires for which it would be a good approximation?

6. An article reported that a survey of clothing worn by pedestrians killed at night in traffic accidents revealed that about 80 percent of the victims were wearing dark-colored clothing and 20 percent were wearing light-colored clothing. The conclusion drawn in the article was that it is safer to wear light-colored clothing at night.

   (a) Is this conclusion justified? Explain.

   (b) If your answer to part (a) is no, what other information would be needed before a final conclusion could be drawn?