

---

---

Introductory Statistics Lectures  
Statistics: The foundations

---

---

ANTHONY TANBAKUCHI  
DEPARTMENT OF MATHEMATICS  
PIMA COMMUNITY COLLEGE

REDISTRIBUTION OF THIS MATERIAL IS PROHIBITED  
WITHOUT WRITTEN PERMISSION OF THE AUTHOR

© 2009

(Compile date: Tue May 19 14:47:40 2009)

## Contents

<b>1 Statistics: The foundations</b>	<b>1</b>	<b>1.4 Types of data</b>	<b>6</b>
1.1 Course Overview	1	1.5 Design of experiments	8
1.2 Sampling populations	2	1.6 Rounding	9
1.3 Basic definitions	3	1.7 Summary	10

## 1 Statistics: The foundations

### 1.1 Course Overview

#### What is statistics?

Statistics allow us to determine what **reliable information** can be obtained from **data**. It also helps us to design experiments and studies so that they can gather meaningful data. Statistics is frequently used to guard scientists from being misled by false impressions.

In studying STDs, an epidemiologist may use statistics to:

- Design a study to determine what are the primary risk factors for contracting a certain STD.
- Analyze study data to determine if females have a higher risk for contracting HIV than males.
- Design an experiment using mice to determine if cervical cancer can be induced with a sexually transmitted virus.
- Conduct a census to determine the extent of a STD outbreak in a city.
- Design an experiment to determine if a new STD treatment is more effective than the current treatment.

#### Where is statistics used?

**Government** consumer price index, trends in the economy, determine if there is a recession, employment patterns, population trends, determine if you should be audited, ...

**Politics** determine which political candidate is more popular, learn what constituents are mainly concerned about, ...

**Scientific research** clinical trials to investigate new drugs, estimation of fundamental constants, psychological tests, ...

**Business** predict demand for a product, check quality of manufactured goods, manage investment portfolios, calculate insurance rates, ...

**Quantum physics** quantum particles are fundamentally described by probability.

Nearly every field uses statistics!

### Outline of course


#### Six primary components of class

- Fundamentals: what are statistics, parameters, experiments?
- Descriptive statistics: how can we describe data?
- Probability: what is the probability of ...
- Inferential statistics: inferences about populations from samples.
- Hypothesis testing: using statistics to prove claims
- Regression: modeling and prediction

## 1.2 Sampling populations

*Example 1.* The number of sexual partners an individual has provides useful information to epidemiologists. An epidemiologist would like to estimate what the mean (average) number of sexual partners is for US college students.

*Question 1.* What is the population we are interested in studying?



*Question 2.* How could we determine the true population mean?



*Question 3.* What are the drawbacks to a census?



### 1.3 Basic definitions

#### Fundamental terminology

**data** collected observations.

ex. number of sexual partners

$h = \{3, 2, 4, 0, 6, 2, 8, 7, 3, 11, \dots\}$

**statistics methods** for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting and making conclusions based on the data.

**population** the **complete** set of all data to be studied.

ex. US college students

**census** collection of data from **every** member of the population.

**sample** a **sub-collection** of members selected from a population

ex. students at this institution are a **sample** of US college students

#### Two key statistical terms concerning data

PARAMETER.

numerical measure describing a characteristic of a **population**.

ex. mean number of sexual partners for all US college students

DEFINITION 1.1

STATISTIC.

numerical measure describing a characteristic of a **sample**.

ex. mean number of sexual partners for a sample of students from this institution.

DEFINITION 1.2

The **parameter** is the **truth** we desire to find but doing a census is not always possible or reasonable. Therefore, we use a **statistic** to **estimate** the parameter.

#### Case Example II

*Example 2.* An epidemiologist would like to determine the mean (average) number of sexual partners for our class. A census is not possible, so the epidemiologist samples 10 students in the class.

*Question 4.* If the epidemiologist can only sample 10 students, how should those students be picked?

**Proper sampling techniques (VERY IMPORTANT)**

Sampling entails selecting a representative subset of a population.

DEFINITION 1.3

RANDOM SAMPLE.

**each member** has an equal chance of being selected.

- “Random” sample is selected based on some characteristic.
- Hidden factors in the random sampling may **bias** results.
- Not ideal.

DEFINITION 1.4

SIMPLE RANDOM SAMPLE (KEY METHOD).

**every possible sample** of the same size has an equal chance of being selected.

- Method of selecting sample is completely random. You don’t use any characteristic to make random selection.
- Ideal.

**Improper sampling results in useless data!**

Assume our class is composed of students sitting in rows of 10.

Determine if the following methods are random samples or simple random samples. If it is a random sample, indicate how it could be biased.

*Question 5.* The epidemiologist sampling our class picks 10 students by randomly choosing a row.

*Question 6.* The epidemiologist picks the lowest 10 social security numbers.

*Question 7.* The epidemiologist puts the student’s names on cards, completely shuffles them, and randomly chooses 10.

**Online Anonymous Survey**

Students, complete anonymous survey on class website!

Ten students from our class are randomly selected using a computerized **simple random sample**.

*Question 8.* Sample 1: What is the mean number of sexual partners?

Let's repeat this experiment 2 more times:

*Question 9.* Sample 2: What is the mean number of sexual partners?

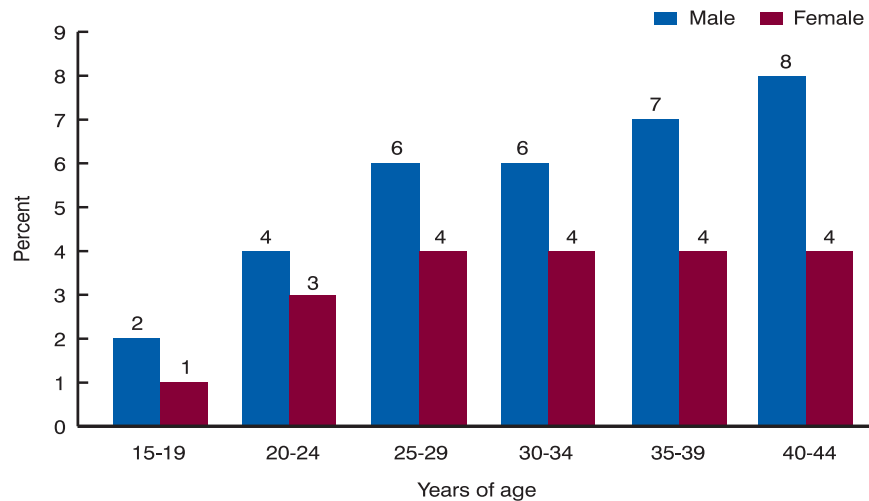
*Question 10.* Sample 3: What is the mean number of sexual partners?

*Question 11.* Are the sample means equal for the 3 samples? Why?

*Question 12.* If the sample mean varies each time, how useful is the sample?

*Question 13.* The epidemiologist now conducts a census of the class to determine the true population mean and to see how close the sample estimates were.

How does our class compare to the US population?



NOTE: Counts vaginal intercourse, and oral and anal sex with opposite-sex partners.  
SOURCE: Tables 10 and 11.

Figure 1: Median number of opposite-sex sexual partners in lifetime, by age and sex. “Advance data from vital and health statistics”, Mosher et., al, number 362, Sep. 15, 2005, US CDC.

### Two key types of “errors”

#### DEFINITION 1.5

SAMPLING “ERROR” (NOISE).

errors due to **chance** sample fluctuations.

ex. Repeatedly sampling 5 students from the class and recording their mean number of sexual partners. Each sample will have a slightly different mean due to random variation in the sample.

- A “natural” error. Like noise. NOT caused by anything you do!

#### DEFINITION 1.6

NON-SAMPLING ERROR.

errors due to data entry error, analysis error, ...

ex. I incorrectly calculate a statistic using the wrong formula.

**Make sure you can differentiate sampling error from non-sampling error!**

*Question 14.* What type of error caused the mean number of sexual partners (the statistic) for 10 randomly chosen students to vary from sample to sample?

## 1.4 Types of data

### Key concepts: Types of data

- Understand the difference between: qualitative / quantitative

- Understand the 4 levels of measurement: (1) nominal, (2) ordinal, (3) interval, (4) ratio.

**Two key types of data**

QUALITATIVE DATA: CATEGORICAL/ATTRIBUTE. DEFINITION 1.7  
 categories represented by some non-numerical characteristic.  
 ex. gender, hair color, car type, shoe type

QUANTITATIVE DATA: NUMERICAL. DEFINITION 1.8  
 numbers representing **measurements**  
 ex. weight, height, temperature, speed, income  
 Two kinds:  
**discrete countable** data, finite number of values.  
 ex. number of students, number of siblings  
**continuous** continuous scale with infinite values.  
 ex. height, weight, temperature

**Four levels of data measurement**

Listed in **increasing** amount of information.

**qualitative level :**

1. **nominal** Categories w/ **no order**. Data cannot be arranged in a meaningful order.  
 ex. hair color: brown, red, blond, grey, black
2. **ordinal** Categories w/ order but **differences are meaningless**  
 ex. car size: subcompact, compact, sedan, full-size

**quantitative level :**

3. **interval** Like ordinal but **differences between data points are meaningful** but **no natural zero**. An interval quantity can have a zero value, but it does not indicate “none”.  
 ex. temperature in Fahrenheit, years
  4. **ratio** interval level with a **natural zero** indicating none of the quantity. With a natural zero, **meaningful ratios** can be made (ie. value  $a$  is twice as much as value  $b$ ,  $a/b = 2$ .)  
 ex. temperature in Kelvin, weight
- Ratio measurements have the most information.

**Percents and ratios**

Make sure you understand percentages and can use them for calculations.

PERCENT. DEFINITION 1.9

$$\text{percentage \%} = \frac{\overbrace{\text{number satisfying criteria}}^{\text{ratio}}}{\text{total number}} \times 100\% \quad (1)$$

- **Always** convert to a ratio (decimal) when doing calculations!
- Percent: range 0%-100%

- Ratio: range 0-1.

*Question 15.* Convert 0.5% to decimal form. (Check: 0.005)

*Question 16.* If 5% of 200 students failed an exam, how many student failed? (Check: 10)

*Question 17.* Convert 0.995 to a percent. (Check 99.5%)

## 1.5 Design of experiments

### Sources of data

Make sure you understand the primary sources of data.

- Observational studies: (1) cross-sectional, (2) retrospective, (3) prospective.
- Experiments.

### Two distinct sources of data: observations & experimental

- 1. observational studies observe** and measure.  
ex. Measure heights of students in this class.
- 2. experiments apply a treatment** and measure effect.  
ex. Break class into two groups. Feed one group low protein, one group high protein food. Measure heights over time.

### Three primary types of observational studies

- 1. cross-sectional** collect data at **one point in time**.  
ex. Heights in class students today.



2. **retrospective (case-control)** examining **existing data**.  
ex. Heights of students from existing medical records.
3. **prospective (longitudinal/cohort)** collect data in the **future** from groups — **cohorts** — sharing common factors.  
ex. Follow male and female students in this class for next 50 years and measure their heights every 10 years.

### Other types of sampling

Know the other primary types of sampling: systematic, convenience, stratified, cluster.

**systematic** selecting some starting point and then selecting every k-th individual in the population.

ex. Selecting every other student in the class.

**convenience** use samples that are easy to get.

ex. Randomly selecting students by taking students in first row.

**stratified** divide population into strata (subgroups) based on a characteristic, then randomly sample from each stratum.

ex. Split class into males and females, randomly select 2 males and 2 females.

**cluster** divide population into clusters (regions), randomly select a few clusters, sample all members in chosen clusters.

ex. To estimate the mean height of Tucson residents, break the city up into city blocks, then randomly select 8 city blocks and measure the height of all residents in that block.

### Confounding

Make sure you understand what **confounding** is and how to prevent it.

CONFOUNDING.

when effects of different factors cannot be distinguished.

Must **plan** experiments **carefully** and **properly sample**.

DEFINITION 1.10

### How to prevent confounding in experiments

1. blinding
2. double-blinding
3. blocks (experimental units)
4. completely randomized experimental design
5. rigorously controlled design
6. large enough sample size
7. replication

## 1.6 Rounding

**Round final answers to 3 significant digits.**

Rounding Policy: avoid rounding during calculations. Final answers should be reported to 3 significant digits (Report first 3 **non-zero digits to the left**). If calculating by hand, intermediate steps should contain at least 5 significant digits.

Round the following values to 3 significant digits:

Question 18. 0.02566

Question 19. 0.0000045673

Question 20. 356,854,328.687

Question 21. 3.000

Question 22. 3

## 1.7 Summary

### Summary of main points

Make sure you do the reading and understand:

- Difference between statistic / parameter.
- Types of data.
- Methods of data collection: studies / experiments
- Sampling methods & errors: preventing bias.
- Preventing confounding.

### Sampling Error

The natural variation (noise) from random sample to random sample.

This class will spend a great deal of time studying sampling error to determine how “well” a statistic estimates a parameter.