Chapter 3: Random Sampling, Probability, and the Binomial Distribution

Part I

Eric D. Nordmoe

Math 261
Department of Mathematics and Computer Science
Kalamazoo College

Spring 2009

Outline

- Some Data
- Probability and Random Sampling
- Properties of Probabilities
- Finding Probabilities in Trees
- Probability Rules

Some Data: The Value of Statistical Consulting

- Does the chance of publication depend on statistical help?

Cell Counts

Decision	Stats Help	No Stats Help	TOTAL
Accept	221	55	276
Reject	293	135	428
TOTAL	514	190	704

Statistical Questions:

- Are acceptance rates significantly different?
- Cause and effect?
- Generalizable?

Vitamin C and the Common Cold

- Data from a Canadian experiment studying Pauling's claim that Vitamin C helps prevent the common cold.
- Randomized double-blind experiment with 818 volunteers:

Cell Counts

	Cold	No Cold	Total
Placebo	335	76	411
Vitamin C	302	105	407
Total	637	181	818

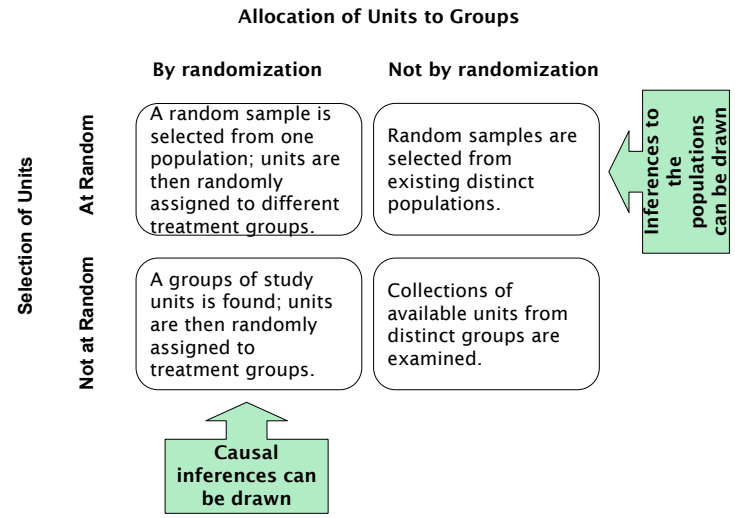
Statistical Questions:

- Are cold percents significantly different?
- Cause and effect?
- Generalizable?

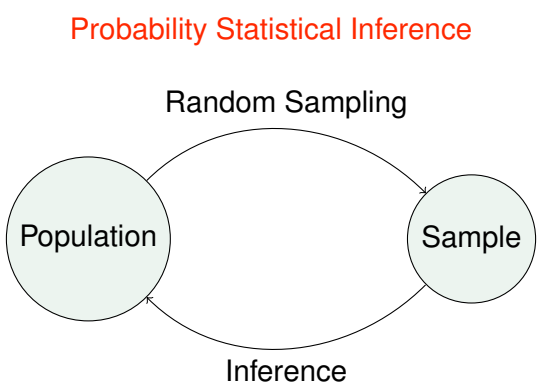
Why study probability?

- Some biological process are directly affected by chance
- The results of experiments are affected by chance.
- The designs of experiments and samples often incorporate chance.
- The conclusions of a statistical data analysis are typically stated in terms of probabilities.
- The models used in statistical inference involve probability.

Statistical Inference and Study Design

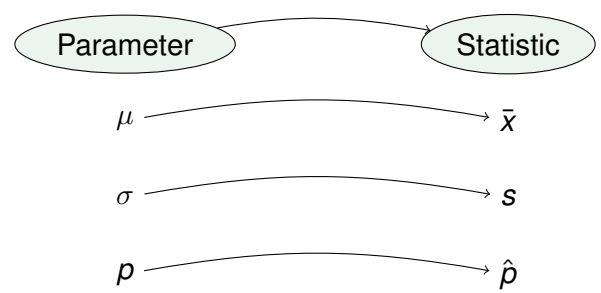


Random Sampling Model



Sampling: Parameters and Statistics

- Population: the group about which we wish to make inference.
- Sample: the subgroup of the population actually studied.
- A sample characteristic is called a statistic. A population characteristic is a parameter.



Sampling Methods

- Probability sampling: each member of the sampling frame has a (theoretically) known chance of being selected for the sample.
- Non-probability sampling: any method that does not use probability sampling

Non-probability Sampling Methods

- Voluntary response: subjects self-select by responding to a general appeal.
- Convenience sampling: subjects sampled are those easiest to reach.

Probability Sampling Methods

- Simple random sample (SRS) a sample of size n drawn such that every possible sample of size n has the same chance of being drawn.

  Example 1 Drawn n names from a hat.

  Example 2 Randomly sort the population and take the first n cases.
- Stratified random sampling a sample obtained by first dividing the population into strata (identifiable homogenous groups) and then drawing separate SRSs from the strata. The separate SRSs are combined to form the full stratified sample.

  Example 1 Separate SRS of each sex

  Example 2 Separate SRS of juveniles and adults of a species

Sampling Error and Sampling Bias

- The discrepancy between the sample and the population is called chance error due to sampling or sampling error.
  - Such discrepancies are to be expected and are not, in fact, errors.
- Systematic differences between the sample and the population is called bias
  - Sampling bias is the systematic tendency for cases with certain values of the response variable to be selected more readily than others.

Definitions of Probability

The Subjective (Personal) Definition

The probability that an event occurs is determined by (expert) judgment about how likely the event is to occur.

The Relative Frequency Definition

The probability that an event occurs is the proportion of times an event would occur in an indefinitely long series of repetitions of a chance operation.

Probability on the Web

- Odds of Dying
- Eye Color

Some Formal Definitions

- A Random experiment is a process whose outcomes are determined by chance.
- The Outcome space of an experiment is the set of all possible outcomes of an experiment:
  - $\Omega = \{E_1, E_2, \ldots, E_k\}$
- An event is a subset of the outcome space.
- The probability of an event E is written

$$\Pr\{E\}$$

A Simple Example

Rolling a Die

- The Experiment is observing the outcome of the roll of a single die.
- Outcome space: $\Omega = \{1, 2, \ldots, 6\}$
  - $E_1 = 1$, $E_2 = 2$, ..., $E_6 = 6$
- The events:
  - A even roll
  - B roll less than a 3
  - C roll a 4
- Probabilities for a fair die with equally likely outcomes:

$$\Pr\{E_i\} = \frac{1}{6}$$

Basic Probability Rules

Rule 1 Probabilities are between 0 and 1.

- $0 \leq \Pr\{E\} \leq 1$.

Rule 2 The sum of the probabilities of all possible events equals 1.

- $\Pr\{E_1\} + \Pr\{E_2\} + \cdots + \Pr\{E_k\} = 1$

Rule 3 The probability of the complement of E^C (event E does not happen) is

- $\Pr\{E^C\} = 1 - \Pr\{E\}$

Example

The blood type distribution of the U.S. is 44% O, 42%A, 10%B, and 4% AB.

The Experiment

Type the blood of a randomly selected U.S. resident.

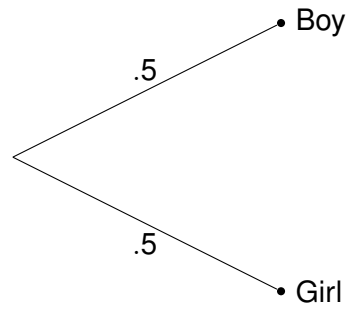
Some Probabilities

- $\Pr\{A\} = .42$
- $\Pr\{O\} + \Pr\{A\} + \Pr\{B\} + \Pr\{AB\} = .44 + .42 + .1 + .04 = 1$
- $\Pr\{O^C\} = 1 - .44 = .42 + .10 + .04 = .56$

Finding Probabilities in the Trees

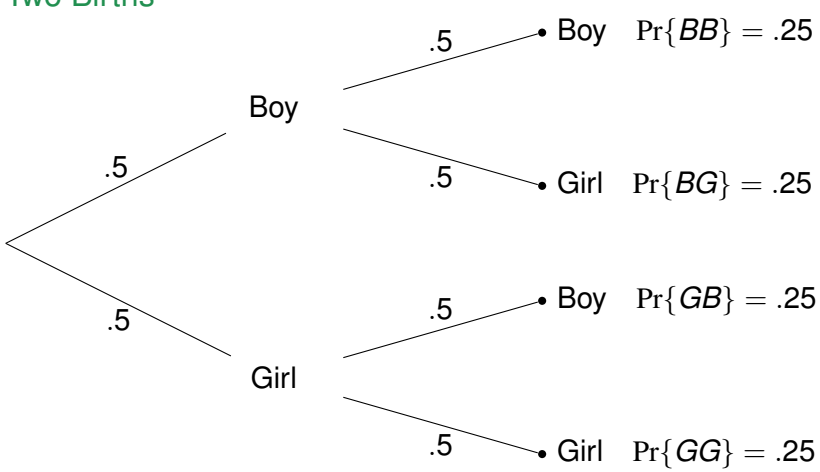
In more complex situations, probability trees are helpful for breaking a problem into parts.

Births



Finding Probabilities in the Trees

Two Births



Multiply probabilities to get the probability of a path.

Using Trees for Medical Testing

Testing for Hypertension

Suppose

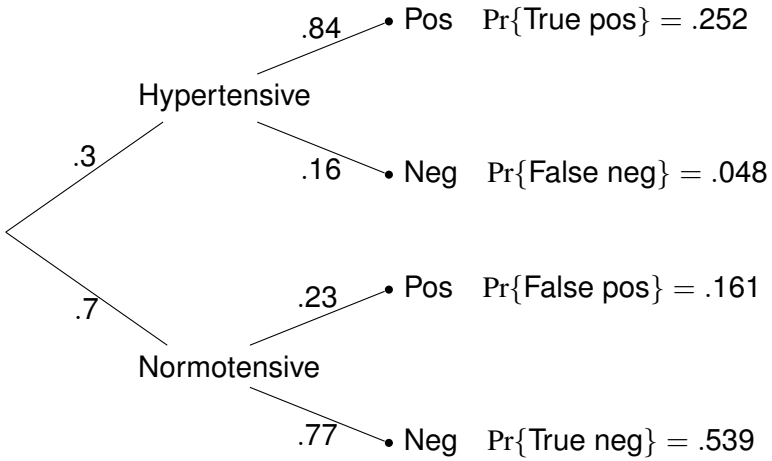
- 30% of the U.S. adult population is hypertensive
- 84% of hypertensives test positive on an automatic blood-pressure machine
- 23% of normotensives test positive

Questions

1. What is the chance that a randomly selected adult tests positive?
2. What percent of those testing positive are hypertensive?

Finding Probabilities in the Trees

Testing for Hypertension



Finding the Probabilities

Question 1 What is the chance that a randomly selected adult tests positive?

$$\begin{aligned}\Pr\{\text{Test pos}\} &= \Pr\{\text{True pos}\} + \Pr\{\text{False pos}\} \\ &= .252 + .161 = \boxed{.413}\end{aligned}$$

Question 2: What proportion of those testing positive are hypertensive?

$$\begin{aligned}\Pr\{\text{Hyper} \mid \text{Pos}\} &= \frac{\Pr\{\text{True pos}\}}{\Pr\{\text{Test pos}\}} \\ &= \frac{.252}{.252 + .161} = \boxed{.610}\end{aligned}$$

Exercise

Testing for HIV

Enzyme Immunoassay (EIA) tests are used to screen blood specimens for the presence of antibodies to HIV. Below are approximate probabilities of positive and negative test results:

	Test Result	
	+	-
Antibodies present	0.9985	0.0015
Antibodies absent	0.0060	0.9940

Questions

Suppose 1% of the population carries HIV antibodies in their blood:

1. What is the chance that a randomly selected adult tests positive for HIV?
2. What percent of those testing positive actually have the antibodies?

Sample Mutants

Judgement Sampling

- Use judgement to choose a representative sample of size 10 from the population of 100. Note the number of mutants and estimate the proportion.

SPSS Random Sample

- Draw a random sample of 10 from the population of 100.
- Record the number of mutants in your sample.
- Repeat for a total of 10 samples.
- Record your results in the format on p.77 of the text.