Introduction to Probability and Statistics

Anastasiia Kim

January 22, 2020

General Information

Instructor (MWF 9.00-9.50 am): Anastasiia Kim
Email: anastasiiakim@unm.edu
Office Hours: MW TBD, SMLC 319

Tutors: Jared DiDomenico, Md Rashidul Hasan
Emails: jdidomen@unm.edu, mdhasan@unm.edu
Recitation/Tutoring Hours: MW 5 pm - 6 pm, TR 4 pm - 5 pm at DSH TBD

Course Outline

- Sample Spaces and Events
- Fundamentals of probability
- Discrete and continuous distributions
- Descriptive Statistics
- Parameter Estimation
- Confidence Intervals
- Hypothesis Testing

Books

Course syllabus, slides, and homeworks will be posted at:
https://anastasiiakim.github.io/teaching/stat345
Course books (not required):

- A First Course in Probability, by Sheldon Ross
- Statistical Inference, by George Casella and Roger L. Berger
- Introduction to Probability, by Joseph K. Blitzstein and Jessica Hwang
- Applied Statistics and Probability for Engineering, by Douglas C. Montgomery and George C. Runger
- https://www.probabilitycourse.com

Assessment

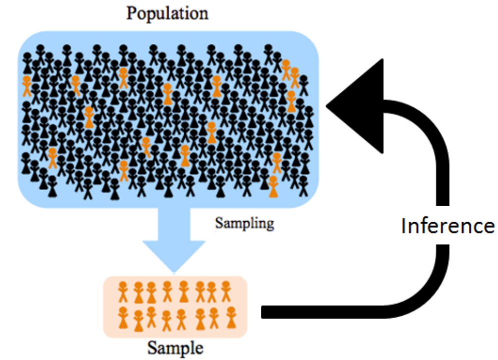
- Homeworks (50%):
  - Assigned biweekly. Expect around 7-8 homeworks
  - Students are encouraged to work together on homework problems, but they must turn in their own write-ups
  - Some homework assignments require the R statistical software (https://www.r-project.org)
- Midterm (25%)
- Final exam (25%)

Why study statistics?

- Statistics helps us make decisions and draw conclusions in the presence of variability
- Many decisions have to be made that involve uncertainties:
  - an economist wants to estimate the unemployment rate
  - an environmentalist tests whether new controls have resulted in a reduction in pollution
  - a biologist is interested in estimating the clutch size for a particular type of bird

Why study statistics?

- The sample along with inferential statistics allows us to draw conclusions about the population
- A group of individual persons, objects, or items from which samples are taken for statistical measurement constitutes a population



Misuse of Statistics

- Misleading data visualization
- Data fishing. When data mining is abused
  - If enough different variables are looked at, some will show correlations that occur solely by chance rather than representing a true relationship
  - If a selection bias is introduced when selecting the sub-sample from the data that previously showed no correlation can be altered to suggest a positive result
- Sampling bias (undercoverage, nonresponse, voluntary response, etc.)
  - Mall interviews will not contact a sample that is representative of the entire population
- Poor data quality
- False causality (Correlation does not imply causation!)
  - Children that watch a lot of TV are the most violent. Clearly, TV makes children more violent
  - Drinking tea increases diabetes by 40%
- Choosing incorrect methods
- Violating model assumptions

Why study probability?

- Probability theory is fundamentally important to inferential statistical analysis.
- Probability provides mathematical models for random phenomena and experiments, such as:
    - gambling
    - stock market
    - racing
    - clinical trials
    - weather forecasts
    - genetic mutations, etc.

Why study probability?

- The theory of probability has always been associated with gambling:
  - if a fair coin is tossed n times, the relative frequency of tails will be close to $1/2$
  - if a fair six-sided die is thrown n times, the relative frequency of getting 3 is likely to be $1/6$
  - If a card is drawn from a shuffled deck and then replaced, the deck is reshuffled, and the process is repeated n times, the relative frequency of hearts is likely to be very close to $1/4$
- The purpose of probability theory is to describe and predict such relative frequencies in terms of probabilities of events
- The probability of an event may be determined empirically or mathematically

The idea of probability

- A random experiment is an experiment that can result in different outcomes, even though it is repeated in the same manner every time
  - ex. Five tosses of a coin constitute a single experiment
- The probability of any outcome of a random experiment is the proportion of times the outcome would occur in a very long series of repetitions

Sample spaces and Events

- Every probabilistic model involves an experiment that will produce exactly one out of several possible outcomes
- An event (E) is a collection of possible outcomes
- The set of ALL possible outcomes is called the Sample Space (S)
- The events in S must be mutually exclusive

Discrete and continuous sample spaces

- S is discrete if it consists of a finite or countable infinite set of outcomes
  - Toss three fair coins. What is the probability of exactly one Tails (T)?
  - The sample space S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
  - The event of getting exactly one Tail is E = {HHT, HTH, THH} and probability is $3/8$
- S is continuous if it contains an interval of real numbers
  - Experiment: note the time of arrival past the departure time of the last train. If T is the interval between two consecutive trains, then the sample space for the experiment is the interval $S = [0, T] = \{x : 0 < x \leq T\}$

Find a sample space

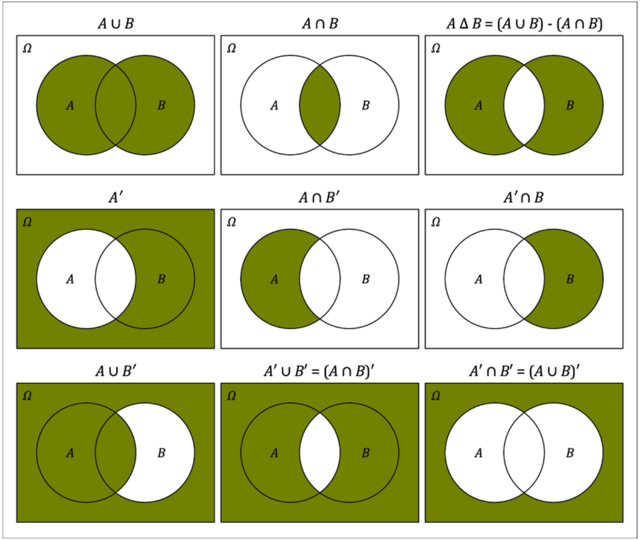
- If the experiment consists of flipping two fair coins
- If the outcome of an experiment is the order of finish in a race among the 5 horses
- If the experiment consists of measuaring the lifetime of a phone battery
- Consider an event E=$\{$sum of the faces of two independently thrown dice is 7$\}$. Find the probability of this event

What's wrong with this sample space?

- Roll a die
- $S = \{\text{Even number}\}$
- $S = \{(1 \text{ or } 3), (1 \text{ or } 4)\}$

Sets via Venn diagram

For any two events A and B of a sample space S



Sets

For any two events A and B of a sample space S

English	**Sets**
Events and occurrences	
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
something must happen	$s_{\text{actual}} \in S$
New events from old events	
A or B (inclusive)	$A \cup B$
A and B	$A \cap B$
not A	A^c
A or B, but not both	$(A \cap B^c) \cup (A^c \cap B)$
at least one of $A_1, \ldots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \cdots \cap A_n$
Relationships between events	
A implies B	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \emptyset$
$A_1, \ldots, A_n$ are a partition of S	$A_1 \cup \cdots \cup A_n = S, A_i \cap A_j = \emptyset$ for $i \neq j$

The basic principle of counting: multiplication rule

- Suppose that two experiments are to be performed. Then if experiment A can result in any one of m possible outcomes and if, for each outcome of experiment A, there are n possible outcomes of experiment B, then together there are mn possible outcomes of the two experiments.

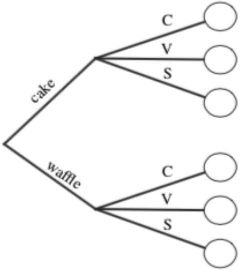


Figure 1: Tree diagram for choosing an ice cream cone. You can choose whether to have a cake cone or a waffle cone, and whether to have chocolate, vanilla, or strawberry as your flavor.

Example

Roll a die 3 times. What is the probability that you get different numbers?

- Identify the set of equally likely outcomes
- Compute the total number of outcomes and the number of good outcomes
- Compute the probability as #of good outcomes/total # of outcomes

History of probability

In 1654 the Flemish aristocrat Chevalier de Méré sent a letter to the mathematician Blaise Pascal:

- I used to bet even money that I would get at least one 6 in four rolls of a fair die. The probability of this is 4 times the probability of getting a 6 in a single die, i.e., $4/6 = 2/3$; clearly I had an advantage and indeed I was making money. Now I bet even money that within 24 rolls of two dice I get at least one double 6. This has the same advantage ($24/62 = 2/3$), but now I am losing money. Why?

History of probability

In 1654 the Flemish aristocrat Chevalier de Méré sent a letter to the mathematician Blaise Pascal:

- I used to bet even money that I would get at least one 6 in four rolls of a fair die. The probability of this is 4 times the probability of getting a 6 in a single die, i.e., $4/6 = 2/3$; clearly I had an advantage and indeed I was making money. Now I bet even money that within 24 rolls of two dice I get at least one double 6. This has the same advantage $(24/62 = 2/3)$, but now I am losing money. Why?
- de Méré's reasoning was faulty: if the number of rolls were 7 in the first game, the logic would give the nonsensical probability $7/6$.

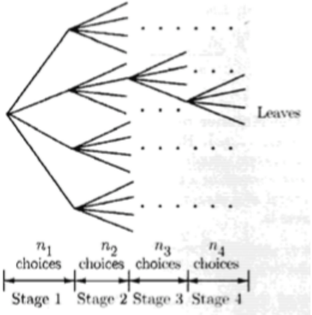
History of probability

How to compute probabilities for de Méré's games?

- Game 1: there are 4 rolls and he wins with at least one 6.
- Game 2: there are 24 rolls of two dice and he wins by at least one pair of 6's rolled.

Fundamental Theorem of counting

- If k experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes; and if, for each of these n_1 possible outcomes, there are n_2 possible outcomes of the second experiment; and if, for each of the possible outcomes of the first two experiments, there are n_3 possible outcomes of the third experiment; and if ..., then there is a total of $n_1 \cdot n_2 \cdot ... \cdot n_k$ possible outcomes of the k experiments.



Problems

- How many different 7-place license plates are possible if the first 3 places are to be occupied by letters and the final 4 by numbers? 175,760,000
- How many license plates would be possible if repetition among letters or numbers were prohibited? 78,624,000