

# 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](mailto:anastasiiakim@unm.edu)

Office Hours: MW TBD, SMLC 319

Tutors: Jared DiDomenico, Md Rashidul Hasan

Emails: [jdidomen@unm.edu](mailto:jdidomen@unm.edu), [mdhasan@unm.edu](mailto: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

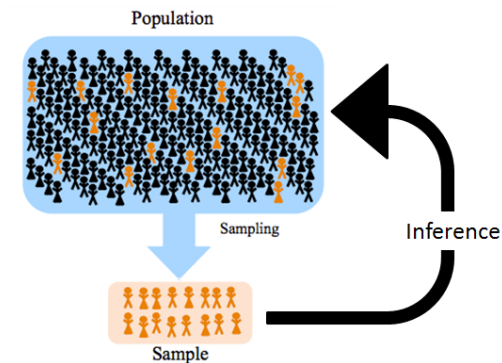
- ▶ Homeworks (50%):
  - ▶ Assigned biweekly. Expect around 7-8 homeworks
  - ▶ Students are encouraged to work together on homework problems, but they must turn in their own write-ups
  - ▶ Some homework assignments require the R statistical software (<https://www.r-project.org>)
- ▶ Midterm (25%)
-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

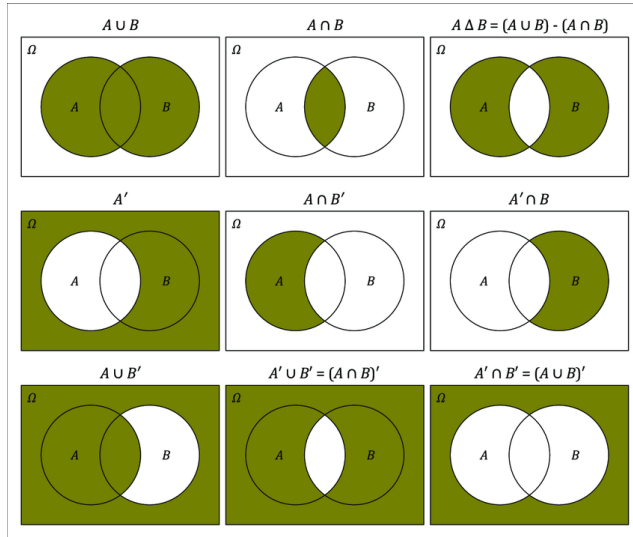
- ▶ If the experiment consists of flipping two fair coins
- ▶ If the outcome of an experiment is the order of finish in a race among the 5 horses
- ▶ If the experiment consists of measuring the lifetime of a phone battery
- ▶ Consider an event  $E = \{\text{sum of the faces of two independently thrown dice is 7}\}$ .  
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
<i>Events and occurrences</i>	
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	$s_{\text{actual}} \in S$
<i>New events from old events</i>	
$A$ 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 $A_1, \dots, A_n$	$A_1 \cup \dots \cup A_n$
all of $A_1, \dots, A_n$	$A_1 \cap \dots \cap A_n$
<i>Relationships between events</i>	
$A$ implies $B$	$A \subseteq B$
$A$ and $B$ are mutually exclusive	$A \cap B = \emptyset$
$A_1, \dots, A_n$ are a partition of $S$	$A_1 \cup \dots \cup A_n = S, A_i \cap A_j = \emptyset \text{ for } i \neq j$