SDSC2102
Statistical Methods and Data Analysis

Topic 1. Basic Probability and Statistics Theory

Probability vs. Statistics

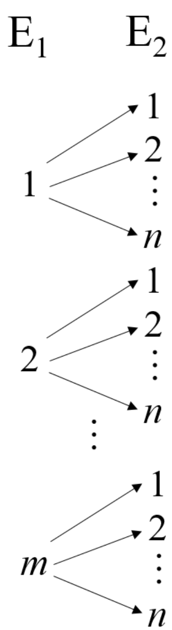
- **Probability: A method for representing random events or outcomes.**

- **Statistics: Techniques for collecting and analyzing data, so as to make decisions about the population that generated the data.**
  - Note: “data” is plural and “datum” is singular.

Probability

Basic Principle of Counting

- **Multiplication Rule:** Two experiments: If Experiment 1 has m outcomes, and out of each outcome of Experiment 1, Experiment 2 has n outcomes, then together there are $m \times n$ outcomes.



$$\begin{array}{cccc} (1,1) & (1,2) & \dots & (1,n) \\ (2,1) & (2,2) & \dots & (2,n) \\ & \vdots & & \\ (m,1) & (m,2) & \dots & (m,n) \end{array}$$

Example: Roll a die twice

$m = 6, n = 6$

$$\begin{array}{cccc} (1,1) & (1,2) & \dots & (1,6) \\ (2,1) & (2,2) & \dots & (2,6) \\ & \vdots & & \\ (6,1) & (6,2) & \dots & (6,6) \end{array}$$

Basic Principle of Counting

Define

- Factorial: $n! = n(n-1)(n-2)\ldots(1)$

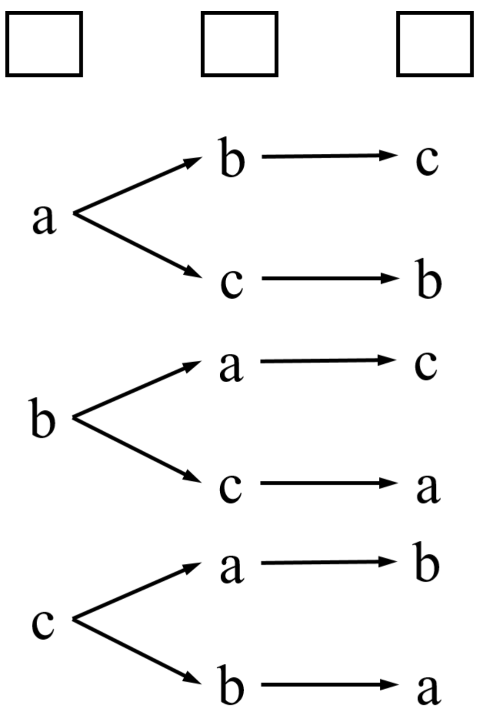
  For example, $5! = 5\times4\times3\times2\times1 = 120$

  - $n! = n(n-1)! = n(n-1)(n-2)!$ etc.
  - $1! = 1$ and $0! = 1$

Permutations

- # ways to arrange n distinct objects:
  $n(n-1)(n-2)\ldots(1) = n!$ permutations

Example: Number of ways to order letters {a, b, c}



$$3! = 3 \times 2 \times 1 = 6$$

Combinations

- # ways to select *r* objects from a set of *n*:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Note: In selection of objects from a larger set, different orderings of the objects are only counted as one combination.

Example: Number of ways to select 3 from{a, b, c, d, e}

$\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \ldots$

$$\binom{5}{3} = \frac{5!}{3! \times 2!} = \frac{5 \times 4 \times 3 \times 2!}{3! \times 2!} = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} = 10$$

Sample Space and Events

- Random experiment: the task that results in a random outcome.
  - Toss a coin / Roll a die
  - See how long a light bulb lasts
  - Count how many people enter the post office between 2:00 and 3:00 pm tomorrow
- Sample space (S) = the set of all possible outcomes.
- An event consists of one or more outcomes
  - A = {odd number on a die}
  - A = {light bulb lasts < 200 hours}
  - A = {at least 3 people enter the post office between 2:00 and 3:00 pm tomorrow}

Discrete Sample Spaces

- Tossing a coin: $S = \{\text{H}, \text{T}\}$
- Tossing a die: $S = \{1, 2, 3, 4, 5, 6\}$
- Tossing 3 coins:

  $S = \{$ HHH,
  HHT, HTH, THH,
  HTT, THT, TTH,
  TTT $\}$
- Count # people that enter the post office:
  $S = \{0, 1, 2, 3, 4, 5, \ldots\}$

Continuous Sample Spaces

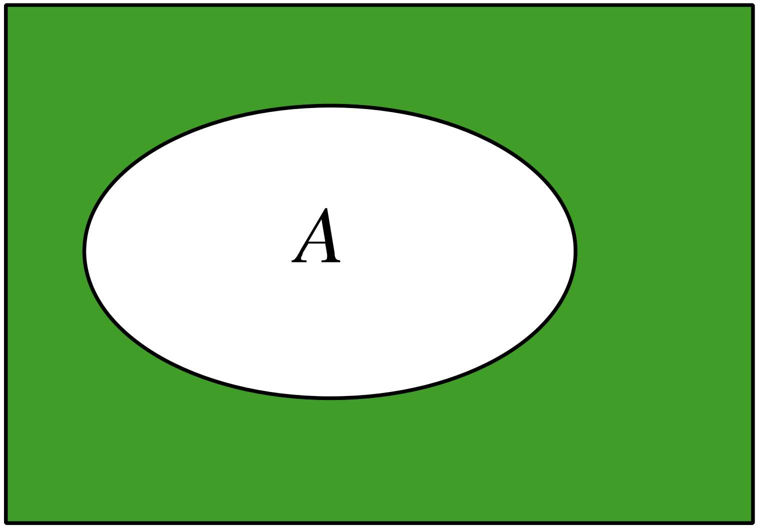
- Values between 0 and 130:
  $S = \{x \mid 0 < x < 130\}$
- Lifetime of a light bulb:
  $S = \{x \mid x \geq 0\}$
- Points on a circle of radius 2:
  $S = \{(x, y) \mid x^2 + y^2 = 4\}$
- Points inside a circle of radius 2:
  $S = \{(x, y) \mid x^2 + y^2 < 4\}$

Events

- An event is a subset of a sample space.
  - Toss 3 coins: A = {exactly two heads}
    = {HHT, HTH, THH}
  - Light bulb: A = {lasts $<$ 200 hours}
    $= \{x \mid 0 \leq x < 200\}$
  - Post office: A = {$\geq$ 3 people enter}
    $= \{3, 4, 5, \ldots.\}$

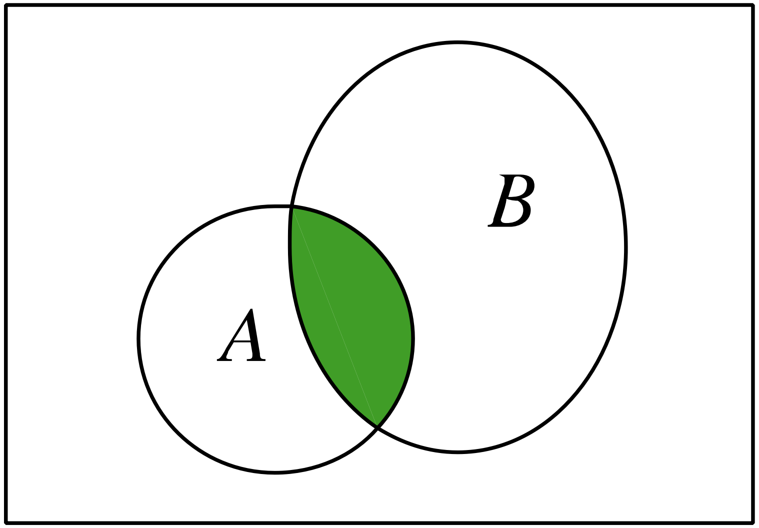
Set Notation / Venn Diagrams

- A^C is the complement of $A = S - A$
  - Light bulb: $A = \{x \mid 0 \le x < 200\}$
    $A^c = \{x \mid 200 \le x < \infty\} = \{x \mid x \ge 200\}$
  - Toss 3 coins: A = {HHT, HTH, THH}
    A^c = {HHH, HTT, THT, TTH, TTT}
  - Venn Diagram: S = entire rectangle



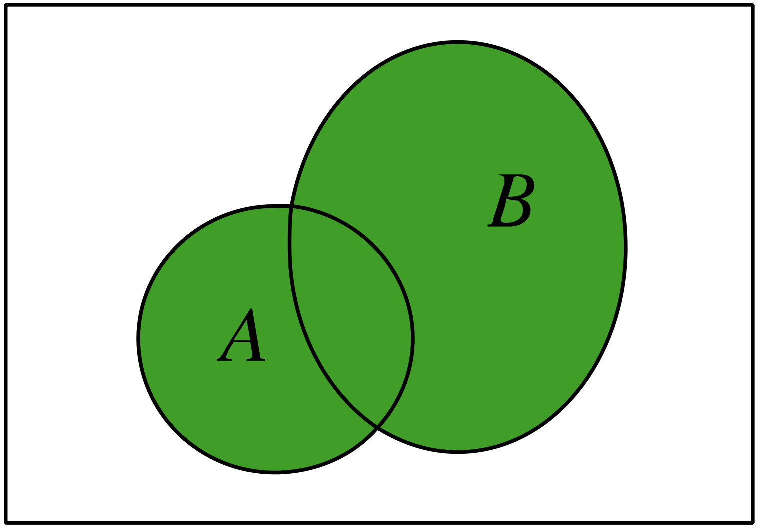
Set Notation / Venn Diagrams

- $A \cap B$ is the intersection of A and B
  - Events for which both A and B occur
  - Light bulb: $A = \{x \mid 0 \le x < 200\}$, $B = \{x \mid x > 50\}$ $\Rightarrow A \cap B = \{x \mid 50 < x < 200\}$
  - Toss 3 coins: A = {HHT, HTH, THH}, B = {HHH, THH, THT} $\Rightarrow A \cap B$ = {THH}



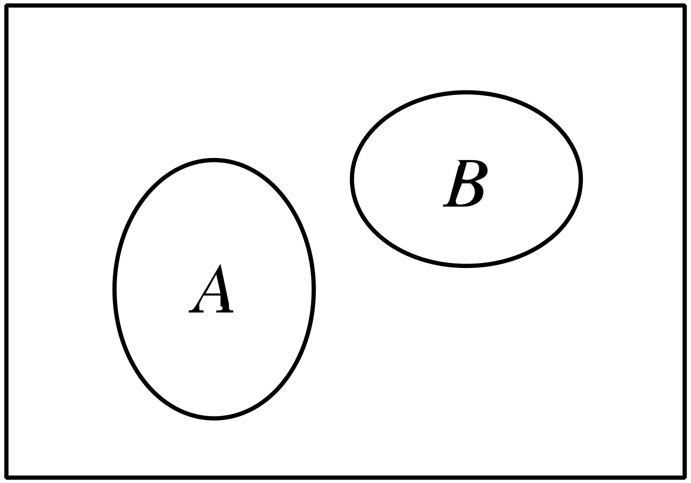
Set Notation / Venn Diagrams

- $A \cup B$ is the union of A and B
  - Events for which either A or B (or both) occur
  - Light bulb: $A = \{x \mid 0 \leq x < 200\}$,
    $B = \{x \mid 5 \leq x < 500\} \Rightarrow A \cup B = \{x \mid 0 \leq x < 500\}$
  - Toss 3 coins: $A = \{\text{HHT, THH}\}$, $B = \{\text{HHH, THH}\}$
    $\Rightarrow A \cup B = \{\text{HHH, HHT, THH}\}$



Set Notation / Venn Diagrams

- **Mutually exclusive (or disjoint) events**
  - Events that have nothing in common
  - $A \cap B = \varnothing$ (empty set)



Set Rules for Events

$A \cap B = B \cap A$ $\qquad$ $A \cup B = B \cup A$

$A \cap A = A$ $\qquad$ $A \cup A = A$

$A \cap S = A$ $\qquad$ $A \cup S = S$

$A \cap \varnothing = \varnothing$ $\qquad$ $A \cup \varnothing = A$

$A \cap A^c = \varnothing$ $\qquad$ $A \cup A^c = S$

$$A \cap (B \cap C) = (A \cap B) \cap C$$

$$A \cup (B \cup C) = (A \cup B) \cup C$$

$(A \cap B)^c = A^c \cup B^c$ $\qquad$ $(A \cup B)^c = A^c \cap B^c$

What Is Probability?

➢ Interpretations of probabilities

- **Subjective:** Based on people's experience and general knowledge

- **Equally likely:** Outcomes have equal probability

$\Rightarrow$ Roll a six-sided die: $P(1) = \ldots = P(6) = 1/6$

$$P(A) = \frac{n_A}{N} = \frac{\#\text{ of outcomes in event } A}{\text{total } \#\text{ of outcomes in } S}$$

- **Limiting Relative Frequency:**

$$\frac{n_A}{n} = \frac{\#\text{ of times event } A \text{ occurs}}{\text{total } \#\text{ of trials}} \to P(A)$$

as $n \to \infty$

Probability Axioms

- $0 \leq P(A) \leq 1$
- $P(\varnothing) = 0$
- $P(S) = 1$
- If events $A_1, A_2, \ldots, A_k$ are disjoint, then

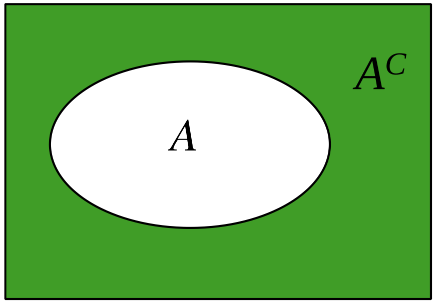
$$P(A_1 \cup A_2 \cup \ldots \cup A_k)$$
$$= P(A_1) + P(A_2) + \ldots + P(A_k)$$

- Roll a six-sided die:

$P(\text{even}) = P(2) + P(4) + P(6) = 3/6$
$P(\{2, 3, 4, 6\}) = P(\text{even}) + P(3) = 3/6 + 1/6 = 4/6$
$P(\{1, 2, 3, 4, 5, 6\}) = P(S) = 1$

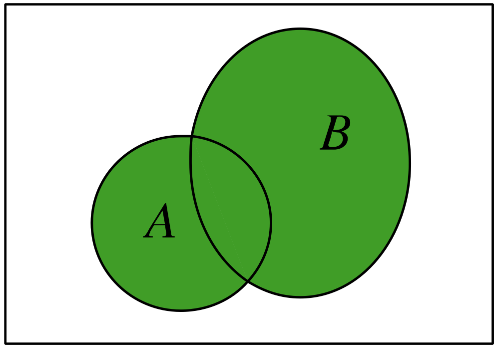
Additive Rules of Probability

- For event A: $P(A^C) = 1 - P(A)$



- For events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Class Problems

1. A pair of fair dice is rolled. What is the probability that the second die lands on a higher value than the first?

Define S: All possible outcomes of rolling two dice

E: Second die has higher value than the first die

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

$$P(E) = \frac{\#\ of\ Outcomes\ in\ E}{total\ \#\ of\ outcomes\ in\ S}$$

$$= \frac{15}{36}$$

Class Problems

2. There are 30 psychiatrists and 24 psychologists attending a certain conference. Three of these 54 people are randomly chosen to take part in a panel discussion. What is the probability that at least one psychologist is chosen?

Define A: At least one psychologist is chosen

A^C: No psychologist is chosen

(3 psychiatrists are chosen)

$$P(A) = 1 - P(A^C) = 1 - \frac{\binom{30}{3}}{\binom{54}{3}} = 1 - \frac{30 \times 29 \times 28}{54 \times 53 \times 52} = 0.84$$

Class Problems

3. Two cards are chosen at random from a deck of 52 playing cards. What is the probability that they

 a. are both aces

 b. have the same value?

(a) B_{A}: both cards are aces

$$P(B_{\mathrm{A}}) = \frac{\binom{4}{2}}{\binom{52}{2}} = \frac{4 \times 3}{52 \times 51}$$

Class Problems

(b) Let B_{A}, B_2, …, B_{K} be the event that both cards are aces, twos, …, Kings, respectively. These events are mutually exclusive (disjoint), and

$$P(B_{\mathrm{A}}) = P(B_2) = \cdots = P(B_{\mathrm{K}}) = \frac{4 \times 3}{52 \times 51}$$

Let E be the event that both cards have same value

$$E = B_{\mathrm{A}} \cup B_2 \ldots \cup B_{\mathrm{K}}$$

$$\begin{aligned} P(E) &= P(B_{\mathrm{A}} \cup B_2 \ldots \cup B_{\mathrm{K}}) \\ &= P(B_{\mathrm{A}}) + P(B_2) + \cdots + P(B_{\mathrm{K}}) = 13 \times \frac{4 \times 3}{52 \times 51} \end{aligned}$$

Class Problems

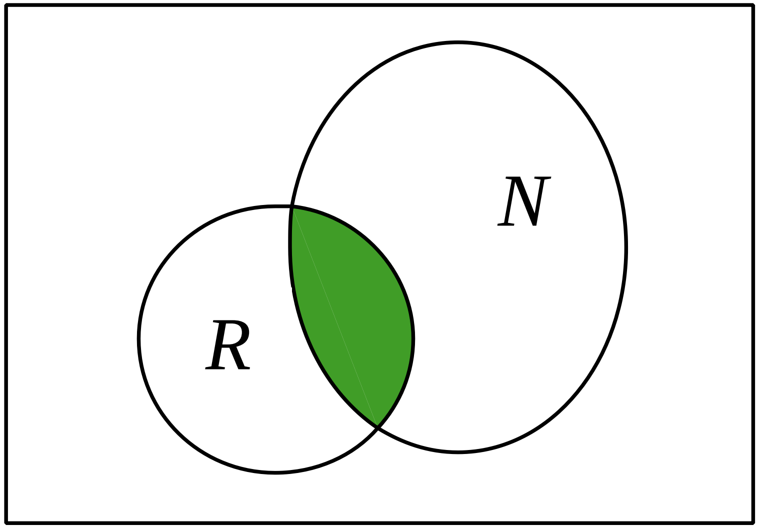
4. 60% of the students at a certain school wear neither a ring nor a necklace. 20% wear a ring and 30% wear a necklace. If one of the students is chosen randomly, what is the probability that this student is wearing:

   a. a ring or a necklace

   b. a ring and a necklace

Class Problems

R: A student wears a ring
N: A student wears a necklace
$P\left((R \cup N)^c\right) = 0.6$
$P(R) = 0.2$
$P(N) = 0.3$
Find $P(R \cup N)$, $P(R \cap N)$



a. $P(R \cup N) = 1 - P\left((R \cup N)^c\right) = 1 - 0.6 = 0.4$

b. $P(R \cap N) = P(R) + P(N) - P(R \cup N)$
$= 0.2 + 0.3 - 0.4$
$= 0.1$

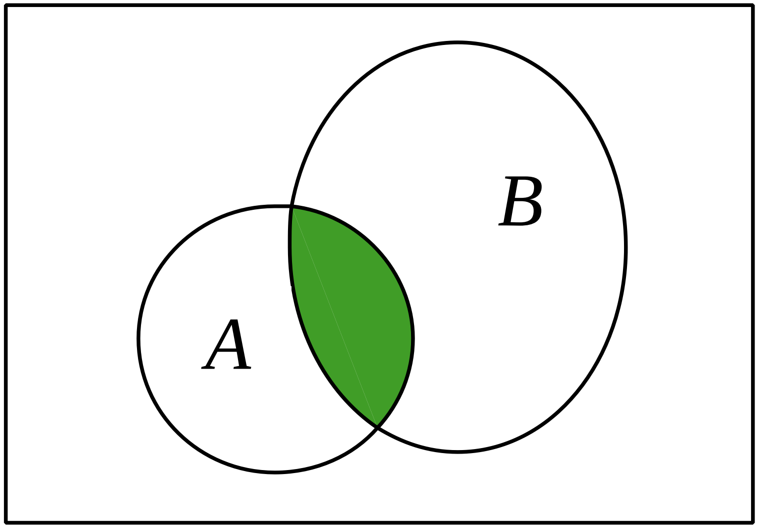
Class Problems

5. A total of 28% of American males smoke cigarettes, 7% smoke cigars, and 5% smoke both cigarettes and cigars.

a. What percentage of males smokes neither cigars nor cigarettes?
b. What percentage of males smokes cigars but not cigarettes?

Class Problems

A: A male smokes cigar
B: A male smokes cigarettes
$P(B) = 0.28$
$P(A) = 0.07$
$P(A \cap B) = 0.05$
Find $P\left((A \cup B)^C\right)$, $P(A \cap B^C)$



a. $P\left((A \cup B)^C\right) = 1 - P(A \cup B) = 1 - 0.3 = 0.7$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.07 + 0.28 - 0.05 = 0.3$$

b. $P(A \cap B^C) = P(A) - P(A \cap B) = 0.07 - 0.05 = 0.02$