CS39000-DM0 Class Notes

Jennifer Neville, Sebastian Moreno

2 Background on Probability and Statistics

These are basic definitions, concepts, and equations that should have been covered in your earlier discrete math and probability courses.

Definition 2.1. *Sample space ($\mathbf{S}$)*
Set of all possible outcomes of an experiment (e.g., $\mathbf{S} = \{O_1, O_2, ..., O_s\}$).

> **Example.** Rolling one 6-sided die $\mathbf{S} = \{1, 2, 3, 4, 5, 6\}$

Definition 2.2. *Event*
Any subset of outcomes (e.g., $A = \{O_i, O_j, O_k\}$) contained in the sample space $\mathbf{S}$.

> **Example.** Odd numbers from rolling one 6-sided die $\mathbf{A} = \{1, 3, 5\}$

Definition 2.3. *Mutually exclusivity*
When events A and B have no outcomes in common (e.g., $A \cap B = \emptyset$).

> **Example.** Let A and B be the odd and even outcomes respectively, from rolling one 6-sided die, then A and B are mutually exclusive.

Definition 2.4. *Axioms of probability*
For a sample space $\mathbf{S}$ with possible events $\mathbf{A_S}$, a function that associates real values with each event A is called a **probability function** if the following properties are satisfied:

1. $0 \leq P(A) \leq 1$ for every A.
2. $P(\mathbf{S}) = 1$
3. $P(A_1 \vee A_2 \vee ... \vee A_s \in \mathbf{S}) = P(A_1) + P(A_2) + ... + P(A_n)$
   if $A_1, A_2, ..., A_n$ are pairwise mutually exclusive events

Properties of probability functions (i.e., implications of axioms):

- $P(A) = 1 - P(\neg A)$.
- $P(true) = 1$
- $P(false) = 0$
- If A and B are mutually exclusive then $P(A \wedge B) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

How to calculate probabilities

When the various outcomes of an experiment are *equally likely*, the task of computing probability reduces to counting:

1. Let $N := |\mathbf{S}|$ be the size of sample space (i.e., number of simple outcomes)
2. Let $N(A) := |\mathbf{A}|$ be the number of simple outcomes contained in the event A
3. Then $P(A) = \frac{N(A)}{N}$

   Example. Roll two 6-sided dice. What is the probability that the sum is 8?
   $|\mathbf{S}| = 6 \times 6$; Event $A = \{2, 6\}, \{3, 5\}, \{4, 4\}, \{5, 3\}, \{6, 2\}$
   $P(sum = 8) = \frac{|\mathbf{S}|}{|A|} = \frac{5}{36}$

Definition 2.5. *Permutation*
An **ordered** sequence of size k taken from a set of n distinct objects **without** replacement. The number of permutations of size k that can be constructed from n objects is:

$$P_{k,n} = \frac{n!}{(n-k)!}$$

If you are choosing an ordered sequence of k objects **with** replacement instead, there are n^k possibilities.

> **Example.** An urn contains ten balls, numbered from 0 to 9. Two balls are drawn at random. How many different **ordered** sequences can we draw?
> n=10; k=2; then we can draw $\dfrac{10!}{(10-8)!} = 90$.
>
> What happens if once we see a ball, we return it to the urn (i.e., the two draws are with replacement)?
> n=10; k=2; then we can draw $10^2 = 100$ (numbers from 00 to 99).

Definition 2.6. *Combination*
An **unordered** sequence of size k taken from a set of n distinct objects **without** replacement. The number of combinations of size k that can be constructed from n objects is:

$$C_{k,n} = \frac{P_{k,n}}{k!} = \frac{n!}{(n-k)!k!}$$

If you are choosing an unordered sequence of k objects **with** replacement instead, there are $C_{k,n+k-1}$ possibilities.

> **Example.** An urn contains ten balls, numbered from 0 to 9. Two balls are drawn at random. How many different **unordered** sequences can we draw?
> n=10; k=2; then we can draw $\dfrac{10!}{(10-8)!2!} = 45$.

What happens if once we see a ball, we return it to the urn (i.e., the two draws are with replacement)?

n=10; k=2; then we can draw $\dfrac{(10+2-1)!}{(10+2-1-2)!2!} = 55$

(previous result plus $00, 11, 22, 33, \cdots, 99$).

Definition 2.7. *Random variable (RV)*
Mapping from a measurement (i.e., property) of objects to a variable that can take on a set of possible different values.

You can think of a r.v. X as a measurement of interest in the context of an experiment. Each time the experiment is run, an outcome $O \in \mathbf{S}$ occurs and a value x is measured and associated with the outcome O. The r.v. X then consists of all possible values x that can occur as a result of the experiment. Note that the reference to $\mathbf{S}$ is suppressed, often because the sample space is hidden or unknown.

Definition 2.7.1. *Discrete random variable*
A random variable with a finite set of possible values.

Example. Let X be the sum of the roll of two 6-sided dice.
X is a *discrete* random variable with possible values $X = \{2, \cdots, 12\}$.

Definition 2.7.2. *Continuous random variable*
A random variable with an infinite set of possible values.

Example. Let X be the output of a random number generator between 0 and 1. X is a *continuous* random variable with possible values $X = [0, 1]$.

Definition 2.8. *Probability distribution*
Probability mass function (for discrete random variables) or probability density function (for continuous random variables) specifies the probability of observing each possible value of a random variable.

Example. Let the random variable X represent the sum of the roll of two 6-sided dice, then its probability mass function is:

x	2	3	4	5	6	7	8	9	10	11	12
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Definition 2.9. *Joint probability distribution*
For a set of random variables, gives the probability of every possible combination of values for those random variables.

Example. Let W_1 be a discrete random variable over the possible weathers $W_1 = \{sunny, rainy, cloudy, snow\}$, and let W_2 be a discrete random variable over a possible weather warning $W_2 = \{true, false\}$

$W_2 \backslash W_1$	sunny	rainy	cloudy	snow
$true$	0.005	0.080	0.020	0.020
$false$	0.415	0.120	0.310	0.030

Definition 2.10. *Conditional (or posterior) probability*
Gives the probability of a set of random variables (e.g., A) given some evidence about the values of another set of random variables (e.g., B).

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \quad if\ P(B) > 0$$

> **Example.** Based on the previous joint probability distribution $P(W_1, W_2)$, what is the probability that there will be a weather warning given that is snowing?
>
> $$P(W_2 = true | W_1 = snow) = \frac{P(W_2 = true \wedge W_1 = snow)}{P(W_1 = snow)} = \frac{0.020}{0.020 + 0.030} = 0.400$$

Definition 2.11. *Mathematical rules of probability*

- Product rule:
$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

- Chain rule (via successive application of the product rule):
$$\begin{aligned} P(X_1, ..., X_n) &= P(X_n|X_1, ..., X_{n-1})P(X_1, ..., X_{n-1}) \\ &= P(X_n|X_1, ..., X_{n-1})P(X_{n-1}|X_1, ..., X_{n-2})P(X_1, ..., X_{n-2}) \\ &= ... \\ &= \prod_{i=1}^{n} P(X_i|X_1, ..., X_{i-1}) \end{aligned}$$

- Bayes rule (via product rule and definition of conditional probability):
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Definition 2.12. *Marginal (or unconditional) probability*
The probability that an event will occur regardless of conditioning events.

$$P(A) = \sum_{b \in B} P(A, b) = \sum_{b \in B} P(A|b)P(b)$$

Definition 2.13. *Independence*
Two events A and B are independent iff:

$$\begin{aligned} P(A|B) &= P(A) \qquad or \\ P(B|A) &= P(B) \qquad or \\ P(A, B) &= P(A)P(B) \end{aligned}$$

Example. Based on the previous joint probability distribution $P(W_1, W_2)$, are the events "weather warning" and "cloudy" independent?

$$\begin{aligned} P(W_1 = cloudy) &= P(W_1 = cloudy, W_2 = true) + P(W_1 = cloudy, W_2 = false) \\ &= 0.02 + 0.31 = 0.33 \\ P(W_2 = true) &= P(W_1 = sunny, W_2 = true + \cdots + W_1 = snow, W_2 = true) \\ &= 0.005 + 0.08 + 0.02 + 0.02 = 0.125 \end{aligned}$$

$$P(W_1 = cloudy \wedge W_2 = true) = 0.02 \neq 0.04125 = P(W_1 = cloudy) \cdot P(W_2 = true)$$

The events are *not* independent.

Definition 2.14. *Conditional independence*
Two variables A and B are conditionally independent given Z iff for all values of A, B, Z:

$$P(A, B|Z) = P(A|Z)P(B|Z)$$

Note: independence does not imply conditional independence or vice versa.

Definition 2.15. *Expected values*
The expectation of a random variable X is a measure of *location* and is defined as:

$$\text{Discrete: } E[X] = \sum_{x \in X} x \cdot p(x)$$

$$\text{Continuous: } E[X] = \int_x x \cdot p(x)\, dx$$

Definition 2.15.1. *Properties of expectation*

$$\text{Function of a rv: } E[h(X)] = \sum_{x \in X} h(x) \cdot p(x)$$

$$\text{Change in location: } E[X + b] = E[X] + b$$

$$\text{Scaling by constant: } E[aX] = a \cdot E[X]$$

$$\text{Sum of two rvs: } E[X + Y] = E[X] + E[Y]$$

Note that this expression holds even when the random variables X and Y are *dependent*. This is referred to as *linearity* of expectation.

$$\text{Conditional expectation: } E[X|Y = y] = \sum_{x \in X} x \cdot P(X = x|Y = y)$$

Example. Based on the previous rv X that represents the sum of the roll of two 6-sided dice, what is its expected value?

$$E[X] = \sum_{x=2}^{12} x \cdot p(x) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + ... + 12 \cdot \frac{1}{36} = 7$$

Definition 2.16. *Variance*
The variance of a random variable X is a measure of *dispersion* and is defined as:

$$\begin{aligned} Var(X) &= E[(x - E[X])^2] \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

$$\text{Standard deviation: } \sigma = \sqrt{Var(X)}$$

Definition 2.16.1. *Properties of variance*

$$\text{Function of a rv: } Var(h(X)) = \sum_{x \in X} (h(x) - E[h(x)])^2 \cdot p(x)$$

$$\text{Change in location: } Var(X + b) = Var(X)$$

$$\text{Scaling by constant: } Var(aX) = a^2 \cdot Var(X)$$

$$\text{Sum of two rvs: } Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Note that in contrast to expectation, this expression is linear only if X and Y are uncorrelated or independent.

$$\text{Conditional variance: } Var(X|Y = y) = E[(X - E[X|Y = y])^2 | Y = y]$$

Definition 2.16.2. *Covariance*
The covariance between two random variable X and Y is a measure of *relation* between the two variables and is defined as:

$$\begin{aligned} Cov(X, Y) &= E[(x - E[X])(y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

> **Example.** Based on the previous rv X that represents the sum of the roll of two 6-sided dice, what is its expected value?
>
> $$Var(X) = E[X^2] - (E[X])^2 = \sum_{x=2}^{12} x^2 \cdot p(x) - (7)^2 = 4 \cdot \frac{1}{36} + 9 \cdot \frac{2}{36} + ... + 144 \cdot \frac{1}{36} - 7^2 = \frac{5}{36}$$

Definition 2.17. *Common probability distributions for rvs*

- **Bernoulli**: Binary rv that takes value 1 with success probability p and value 0 with probability $1 - p$.

  Let $X \sim Bernoulli(p)$, then the probability distribution of X is

  $$P(X = x; p) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

  $$E[X] = p \qquad\qquad V[X] = p(1 - p)$$

- **Binomial**: Describes the number of successful outcomes (i.e., 1s) in n independent $Bernoulli(p)$ trials.

  Let $X \sim Bin(n, p)$, then the probability distribution of X is

$$P(X = x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x \in \{0, 1, \cdots, n\}$$
$$E[X] = np \qquad\qquad V[X] = np(1-p)$$

- **Multinomial**: Generalization of binomial with n trials to case where each trial has k possible outcomes, and outcome i has probability p_i of occurring.

  Let $X = (X_1, X_2, \cdots, X_k) \sim Mult(n, p_1, p_2, \cdots, p_k)$ such that $\sum_{i=1}^{k} p_i = 1$, then the probability distribution of X is

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = \begin{cases} \dfrac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} & \text{if } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise} \end{cases}$$
$$E[X_i] = np_i \qquad\qquad V[X_i] = np_i(1-p_i)$$

- **Poisson**: Expresses the probability of a given number of events occurring in a fixed interval of time, if the events occur with a known average rate (λ) and the events are *independent.*

  Let $X \sim Poisson(\lambda)$, then the probability distribution of X is

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$
$$E[X] = \lambda \qquad\qquad V[X] = \lambda$$

- **Normal (Gaussian)**: Very commonly occurring distribution, sometimes informally called the *bell curve*, which is continuous, symmetric about its mean, and is non-zero over the entire real line.

  Let X be a normal distribution with mean μ and variance σ^2 (i.e., $X \sim N(\mu, \sigma)$), then the probability distribution of X is

$$P(X = x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$E[X] = \mu \qquad\qquad V[X] = \sigma^2$$

Definition 2.18. *Multivariate random variable*
A multivariate rv $\mathbf{X} = \{X_1, X_2, ..., X_p\}$ is a list of p random variables that are grouped together, often because they refer to different properties of an individual entity (e.g., height, weight, age of a person).

Definition 2.18.1. *Properties of multivariate rvs*

$$\text{Joint density: } P(\mathbf{X}) = P(X_1, X_2, ..., X_p)$$

$$\text{Marginal density of a subset: } P(X_i) = \sum_{\mathbf{x} \in \mathbf{X} - X_i} P(X_1 = x_1, X_2 = x_2, ..., X_p = x_p)$$

$$\text{Conditional density of a subset: } P(X_i | \mathbf{X} - X_i) = \frac{P(X_1, X_2, ..., X_p)}{P(X_1, ..., X_{i-1}, X_{i+1}, ..., X_p)}$$