

# STA 360/602L: MODULE 1.2

## PROBABILITY REVIEW

DR. OLANREWAJU MICHAEL AKANDE

# OUTLINE

- Random variables
- Joint distributions
- Independence
- Exchangeability

# DISCRETE RANDOM VARIABLES

- A **random variable** is **discrete** if the set of all possible outcomes is **countable**.
- The **probability mass function (pmf)** of a discrete random variable  $Y$ ,  $p(y)$  describes the probability associated with each possible value of  $Y$ .
- $p(y)$  has the following properties:
  1.  $0 \leq p(y) \leq 1$  for all values  $y \in \mathcal{Y}$ .
  2.  $\sum_{y \in \mathcal{Y}} p(y) = 1$ .
- Most distributions are often characterized by some parameter (or set/vector of parameters)  $\theta$ .
- So, to make this clear, we will often write the pmf instead as  $p(y|\theta)$ .

# BERNOULLI DISTRIBUTION

- The **Bernoulli distribution** can be used to describe an experiment with two outcomes, such as
  - Flipping a coin (heads or tails);
  - Vote turnout (vote or not); and
  - The outcome of a basketball game (win or loss).
- In all cases, we can represent this as a binary random variable where the probability of "success" is  $\theta$  and the probability of "failure" is  $1 - \theta$ .
- We usually write this as:  $Y \sim \text{Bernoulli}(\theta)$ , where  $\theta \in [0, 1]$ .
- It follows that

$$p(y|\theta) = \Pr(Y = y|\theta) = \theta^y(1 - \theta)^{1-y}; \quad y = 0, 1.$$

- What is the mean of this distribution? What is the variance?

# BINOMIAL DISTRIBUTION

- The **binomial distribution** describes the number of successes from  $n$  independent Bernoulli trials.
- That is,  $Y =$  number of "successes" in  $n$  independent trials and  $\theta$  is the probability of success per trial.
- We usually write this as:  $Y \sim \text{Bin}(n, \theta)$ , where  $\theta \in [0, 1]$ .

- The pmf is

$$p(y|\theta) = \Pr(Y = y|\theta, n) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}; \quad y = 0, 1, \dots, n.$$

- **Example:**  $Y =$  number of individuals with type I diabetes out of a sample of  $n$  surveyed.
- Binomial likelihoods are commonly used in collecting data on proportions.
- What is the mean of this distribution? What is the variance?

# POISSON DISTRIBUTION

- $Y \sim \text{Po}(\theta)$  denotes that  $Y$  is a **Poisson random variable**.
- The Poisson distribution is commonly used to model count data consisting of the number of events in a given time interval.
- The Poisson distribution is parameterized by  $\theta$  and the pmf is given by

$$p(y|\theta) = \Pr[Y = y|\theta] = \frac{\theta^y e^{-\theta}}{y!}; \quad y = 0, 1, 2, \dots; \quad \theta > 0.$$

- Similar to binomial but with no limit on the total number of counts.
- **What is the mean of this distribution? What is the variance?**

# GENERAL DISCRETE DISTRIBUTIONS

- Useful to consider general discrete distributions having an arbitrary form.
- Suppose  $Y \in \{y_1^*, \dots, y_k^*\}$ . Then define  $\Pr(Y = y_h^*) = \pi_h$  for each  $h = 1, \dots, k$ . That is,

$$p(y|\pi) = \Pr[Y = y|\pi] = \prod_h \pi_h^{1[Y=y_h^*]}; \quad y \in y_1^*, \dots, y_k^*$$

where  $\pi = (\pi_1, \dots, \pi_k)$ .

- $(y_1^*, \dots, y_k^*)$  are "atoms" representing possible values for  $Y$ .
- For example, these may be words in a dictionary or values for education as a categorical variable. Useful for text data, categorical observations, etc.
- Can also write as  $Y \sim \sum_{h=1}^k \pi_h \delta_{y_h^*}$ , where  $\delta_{y_h^*}$  denotes a unit mass at  $y_h^*$ .
- Often called the **categorical distribution** or **generalized Bernoulli distribution**. Also, see the **multinomial distribution**.

# CONTINUOUS RANDOM VARIABLES

- The **probability density function (pdf)**,  $p(y)$  or  $f(y)$ , of a continuous random variable  $Y$  has slightly different properties:
  1.  $0 \leq f(y)$  for all  $y \in \mathcal{Y}$ .
  2.  $\int_{y \in \mathbb{R}} f(y) dy = 1$ .
- The pdf for a continuous random variable is not necessarily less than 1.
- Also,  $f(y)$  is NOT the probability of value  $y$ .
- However, if  $f(y_1) > f(y_2)$ , we say informally that  $y_1$  has a "higher probability" than  $y_2$ .
- As we did in the discrete case, we will also often write the pdf instead as  $f(y|\theta)$  or  $p(y|\theta)$  to make the conditioning obvious.

# UNIFORM DENSITY

- The simplest example of a continuous density is the **uniform density**.
- $Y \sim \text{Unif}(a, b)$  denotes density is uniform in interval  $(a, b)$ .
- The pdf is simply

$$f(y|a, b) = \frac{1}{b - a}; \quad y \in (a, b).$$

- The cdf is

$$F(y) = \Pr(Y \leq y) = \int_a^y \frac{1}{b - a} dz = \frac{y - a}{b - a}$$

- The mean (expectation) is

$$\frac{a + b}{2}$$

- What is the variance? Also, can you prove the formula for the mean?

# BETA DENSITY

- The uniform density can be used as a prior for a probability if  $(a, b) \subset (0, 1)$ .
- However, it is very inflexible clearly.

## Why?

- An alternative for  $y \in \mathcal{Y}$  is the **beta density**, written as  $Y \sim \text{Beta}(a, b)$ , with

$$f(y|a, b) = \frac{1}{B(a, b)} y^{a-1} (1-y)^{b-1}; \quad y \in (0, 1), \quad a > 0, \quad b > 0.$$

where  $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ .  $\Gamma(n) = (n-1)!$  for any positive integer  $n$ .

- As we have already seen, the beta density is quite flexible in characterizing a broad variety of densities on  $(0, 1)$ .
- **Beta(1,1) is the same as Unif(0,1). Workout the pdfs to convince yourself!**

# GAMMA DENSITY

- The **gamma density** will be useful as a prior for parameters that are strictly positive.
- For random variables  $Y \sim \text{Ga}(a, b)$ , we have the pdf

$$f(y|a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}; \quad y \in (0, \infty), \quad a > 0, \quad b > 0.$$

- Properties:

$$\mathbb{E}[Y] = \frac{a}{b}; \quad \mathbb{V}[Y] = \frac{a}{b^2}.$$

- **Note:** there are multiple parameterizations of the gamma distribution. We will rely on this version in this course.
- Under this parameterization,  $a$  is known as the shape parameter, while  $b$  is known as the rate parameter.
- Under this parameterization, if  $Y \sim \text{Ga}(1, \theta)$ , then  $Y \sim \text{Exp}(\theta)$ , that is, the **exponential distribution**.

# CONTINUOUS JOINT DISTRIBUTIONS

- Suppose we have two random variables  $\theta = (\theta_1, \theta_2)$ .
- Their **joint distribution function** is

$$\Pr(\theta_1 \leq a, \theta_2 \leq b) = \int_{-\infty}^a \int_{-\infty}^b f(\theta_1, \theta_2) d\theta_1 d\theta_2,$$

where  $f(\theta_1, \theta_2)$  is the **joint pdf**.

- The **marginal** density of  $\theta_1$  can be obtained by

$$f(\theta_1) = \int_{-\infty}^{\infty} f(\theta_1, \theta_2) d\theta_2,$$

which is referred to as marginalizing out  $\theta_2$ .

- We will be doing a lot of "marginalizations", so take note!

# FACTORIZING JOINT DENSITIES AND INDEPENDENCE

- The joint density  $f(\theta_1, \theta_2)$  can be factorized as

$$f(\theta_1, \theta_2) = f(\theta_1|\theta_2)f(\theta_2), \quad \text{or} \quad f(\theta_1, \theta_2) = f(\theta_2|\theta_1)f(\theta_1).$$

- For independent random variables, the joint density equals the product of the marginals:

$$f(\theta_1, \theta_2) = f(\theta_1)f(\theta_2).$$

- This implies that  $f(\theta_2|\theta_1) = f(\theta_2)$  and  $f(\theta_1|\theta_2) = f(\theta_1)$  under independence.

- These relationships extend automatically to  $\theta = (\theta_1, \dots, \theta_p)$ . That is,

$$f(\theta_1, \dots, \theta_p) = \prod_{j=1}^p f(\theta_j),$$

under mutual independence of the elements of the  $\theta$  vector.

# CONDITIONAL INDEPENDENCE

- Suppose  $y_i \stackrel{iid}{\sim} f(y_i|\theta)$  for  $i = 1, \dots, n$ .
- Data  $\{y_i\}$  are independent & identically distributed draws from distribution  $f(y_i|\theta)$ .
- The data are said to be **conditionally independent** given  $\theta$  if

$$f(y_1, \dots, y_n|\theta) = \prod_{i=1}^n f(y_i|\theta).$$

- $f(y_1, \dots, y_n|\theta)$  is also the likelihood function  $L(\theta|y)$  of the data.
- The **marginal likelihood** of the data is

$$L(y) = f(y_1, \dots, y_n) = \int_{\Theta} f(y_1, \dots, y_n|\theta)p(\theta)d\theta = \int_{\Theta} L(\theta|y)p(\theta)d\theta.$$

- Here,  $L(y)$  can not be written as a product of densities as in  $\prod_{i=1}^n f(y_i)$ ; we lose independence when we marginalize out  $\theta$ .

# EXCHANGEABILITY

- In marginalizing out  $\theta$ , the observations  $\{y_i\}$  are not marginally independent.
- $\{y_i\}$  are **exchangeable** if  $f(y_1, \dots, y_n) = f(y_{\pi_1}, \dots, y_{\pi_n})$ , for all permutations  $\pi$  of  $\{1, \dots, n\}$ .
- **de Finetti's Theorem:** Suppose  $\{y_i\}$  are exchangeable under above definition for any  $n$ . Then

$$f(y_1, \dots, y_n) = \int_{\Theta} \left[ \prod_{i=1}^n f(y_i|\theta) \right] p(\theta) d\theta.$$

for some  $\theta$ , prior distribution  $p(\theta)$  and sampling model  $f(y_i|\theta)$ .

- Simply put, de Finetti's Theorem states that exchangeable observations are conditionally independent relative to some parameter.
- de Finetti's Theorem is critical in providing a motivation for using parameters and for putting priors on parameters.

# WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!