Hoff Ch. 2: Probability

Setup

- $\mathcal{Y}$ the sample space, e.g., $\mathcal{Y} = \{H, T\}$
- An <u>event</u> is a subset of $\mathcal{Y}$
- A <u>probability measure</u> is a function $\mathbb{P}$ that assigns probabilities to events

Axioms of probability

1. $0 = \mathbb{P}(\emptyset) \leq \mathbb{P}(H) \leq \mathbb{P}(\mathcal{Y}) = 1$

2. If $H_1 \cap H_2 = \emptyset$ then

$$\mathbb{P}(H_1 \cup H_2) = \mathbb{P}(H_1) + \mathbb{P}(H_2)$$

More generally, if $H_j \cap H_k = \emptyset \ \forall j \neq k$,

$$\mathbb{P}\left(\bigcup_{k=1}^{K} H_k\right) = \sum_{k=1}^{K} \mathbb{P}(H_k).$$

Axioms of conditional probability

1. $0 = \mathbb{P}(H^c | H) \leq \mathbb{P}(F | H) \leq \mathbb{P}(H | H) = 1.$

2. If $F \cap G = \emptyset$

$$\mathbb{P}(F \cup G | H) = \mathbb{P}(F | H) + \mathbb{P}(G | H)$$

3. $\mathbb{P}(F \cap G | H) = \mathbb{P}(F | G \cap H)\mathbb{P}(G | H)$

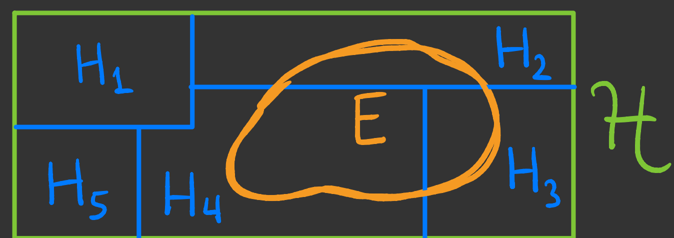
* Often people write

$$\mathbb{P}(F, G | H) = \mathbb{P}(F | G, H)\mathbb{P}(G | H).$$

Marginal probability

Let E be an event and $\{H_k\}_{k=1}^{K}$ a partition of event $\mathcal{H}$ with $\mathbb{P}(\mathcal{H}) = 1$.

$$\mathbb{P}(E) = \sum_{k=1}^{K} \mathbb{P}(E \cap H_k) = \sum_{k=1}^{K} \mathbb{P}(E \mid H_k)\,\mathbb{P}(H_k).$$



Bayes' rule

$$\mathbb{P}(H_j \mid E) = \frac{\mathbb{P}(H_j \cap E)}{\mathbb{P}(E)}$$

$$= \frac{\mathbb{P}(E \mid H_j)\,\mathbb{P}(H_j)}{\mathbb{P}(E)}$$

$$= \frac{\mathbb{P}(E \mid H_j)\,\mathbb{P}(H_j)}{\sum_{k=1}^{K} \mathbb{P}(E \mid H_k)\,\mathbb{P}(H_k)}$$

Independence

Events F and G conditionally independent given H if $P(F, G \mid H) = P(F \mid H)\, P(G \mid H)$.

$$P(F \mid G, H) = \frac{P(F, G \mid H)}{P(G \mid H)} = \frac{P(F \mid H)\, P(G \mid H)}{P(G \mid H)}$$

$$= P(F \mid H).$$

Random variables

- A <u>random variable</u> is an unknown numerical quantity about which we make probability statements.

- Random variables typically denoted by capital letters, e.g., Y.

- Y takes values in state space $\mathcal{Y}$

A more formal definition*

- A probability space is a triple $(\Omega, \mathcal{H}, \mathbb{P})$ where Ω is a nonempty set, $\mathcal{H}$ is a collection of events, i.e., subsets of Ω, and $\mathbb{P}$ is a probability measure, i.e., a function that assigns a probability $\mathbb{P}(H)$ to each event $H \in \mathcal{H}$.

* We will not use this level of formality in this course, but it is good to be aware of.

- A <u>random variable</u> is a function

$$Y : \Omega \longrightarrow \mathcal{Y}$$

- Given a subset $A \subseteq \mathcal{Y}$, the event

"$Y \in A$" is shorthand for the event

$$\{\omega \in \Omega : Y(\omega) \in A\} \subseteq \Omega.$$

$$\Rightarrow \mathbb{P}(Y \in A) = \mathbb{P}(\{\omega \in \Omega : Y(\omega) \in A\})$$

Discrete vs. continuous

- Y is <u>discrete</u> if its state space $\mathcal{Y}$ is countable, e.g., if $\mathcal{Y} \subseteq \mathbb{N} = \{1, 2, 3, \ldots\}$.

- Loosely speaking, Y is <u>continuous</u> if it can take an uncountable number of values, e.g., if $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = [0, 1]$

Probability density functions

- For discrete r.v., write

$$p(y) = \mathbb{P}(Y = y) = \mathbb{P}(Y \in \{y\}).$$

- Note that
  - $0 \leq p(y) \leq 1$
  - $1 = \mathbb{P}(\mathcal{Y}) = \sum_{y \in \mathcal{Y}} p(y).$

- For $A \subseteq \mathcal{Y}$, $\mathbb{P}(A) = \sum_{y \in A} p(y)$.

- For continuous r.v., start with <u>cumulative distribution function</u> (cdf):

$$F(y) = \mathbb{P}(Y \leq y).$$

  - (Radon-Nikodym theorem) If F is absolutely continuous, then there exists a function

$p : \mathcal{Y} \longrightarrow [0, \infty)$ such that

$$F(a) = \int_{-\infty}^{a} p(y)\,dy.$$

- p is called the <u>probability density function</u> (pdf) of Y.

- Note that
  - $F(-\infty) = 0$ and $F(\infty) = 1$

- $$\frac{d}{da} F(a) = \frac{d}{da} \int_{-\infty}^{a} p(y)\, dy = p(a),$$

i.e., p is the derivative of F.

Mean and variance

- The <u>mean/expectation</u> of Y is
  - $\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} y\, p(y)$ (discrete r.v.)
  - $\mathbb{E}[Y] = \int_{\mathcal{Y}} y\, p(y)\, dy$ (continuous r.v.)
- The <u>variance</u> of Y is

$$\mathrm{Var}(Y) = \mathbb{E}\left[(Y - \mathbb{E}(Y))^2\right]$$

$$= \mathbb{E}\left[Y^2 - 2Y\mathbb{E}(Y) + \mathbb{E}(Y)^2\right]$$

$$= \mathbb{E}[Y^2] - 2\,\mathbb{E}[Y]^2 + \mathbb{E}[Y]^2$$

$$= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2.$$

$$\boxed{\text{End lecture 2}}$$

Joint distributions

- These are just pdfs in the case where $Y = (Y_1, \dots, Y_n)$ is a multivariate r.v. taking values in the product space

$$\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n = \bigotimes_{i=1}^{n} \mathcal{Y}_i$$

- Let $y = (y_1, \dots, y_n)$.

- Discrete:

$$p(y) = \mathbb{P}(Y = y) = \mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n)$$

- Continuous:

$$F(a) = \mathbb{P}(Y_1 \leq a_1, \dots, Y_n \leq a_n)$$

$$= \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \cdots \int_{-\infty}^{a_n} p(y_1, \dots, y_n)\, dy_n \cdots dy_1$$

- <u>Marginal density</u> of y_j:

$$p(y_j) = \int_{\mathcal{Y}_{-j}} p(y)\, dy_1 \cdots dy_{j-1}\, dy_{j+1} \cdots dy_n$$

where $\mathcal{Y}_{-j} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{j-1} \times \mathcal{Y}_{j+1} \times \cdots \times \mathcal{Y}_n$

  - Integrate out/average over all coordinates except y_j.

- <u>Conditional density</u> of Y_j given $Y_{-j} = (Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_n)$:

$$p(y_j \mid y_{-j}) = \frac{p(y_1, \dots, y_n)}{p(y_j)}.$$

  - Probability of Y_j once we've observed all other variables.

Bayes rule (parameter estimation)

sampling distribution
aka likelihood

prior

$$p(\theta \mid y) = \frac{\overbrace{p(y \mid \theta)}^{\text{likelihood}}\;\overbrace{p(\theta)}^{\text{prior}}}{p(y)}$$

posterior
distribution/density

normalizing constant
(marginal of y)

Independent random variables

- $Y_1, \ldots, Y_n$ are <u>independent</u> if

$$\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(Y_i \in A_i)$$

for all events $(A_1, \ldots, A_n) \subseteq \bigotimes_{i=1}^{n} \mathcal{Y}_i$.

- <u>Conditionally independent</u> given θ if

$$\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n \mid \theta) = \prod_{i=1}^{n} \mathbb{P}(Y_i \in A_i \mid \theta).$$

- If $Y_1, \dots, Y_n$ are conditionally independent given θ, with pdfs $p_{Y_i}(y_i \mid \theta)$, then

$$p(y_1, \dots, y_n \mid \theta) = \prod_{i=1}^{n} p_{Y_i}(y_i \mid \theta).$$

- If, in addition, the Y_i are <u>identically distributed</u>, i.e., $p_{Y_i} = p \;\; \forall i$, then

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

Exchangeability and de Finetti

- $Y_1, \ldots, Y_n$ are <u>exchangeable</u> if

$$p(y_1, \ldots, y_n) = p(y_{\pi(1)}, \ldots, y_{\pi(n)})$$

for all permutations π of $\{1, \ldots, n\}$, i.e., subscripts don't matter.

- $Y_1, \ldots, Y_n \mid \theta$ iid $\Longrightarrow$ exchangeable

- (de Finetti's theorem) Essentially the converse, i.e., exchangeable $\Longrightarrow \exists\, \theta \sim p(\theta)$ such that

$$p(y_1, \ldots, y_n) = \int_{\Theta} \prod_{i=1}^{n} p(y_i \mid \theta)\, p(\theta)\, d\theta .$$

See Hoff, Section 2.8, for details.

- Provides foundation for using parameters θ and priors $p(\theta)$

End Ch. 2