Intro to Probability and Statistics

Disclaimer: Work in progress. Portions of these written materials are incomplete.

Why Study Probability and Statistics?

- The world is full of uncertainty and we want to understand it to make **predictions** and **decisions**.
- Probability and Statistics provide a formal language and set of tools for dealing with uncertainty.

What is Probability?

Frequentist definition:

*“Probabilities can be found by a **repeatable** objective process (and are thus ideally **devoid of opinion**).”*

Example: roll a die multiple times to compute the probability of occurrence of 6.

What is Probability?

Bayesian definition:

"Probability is a reasonable expectation representing a ***state of knowledge*** *or a quantification of a* ***personal belief.****"*

Example: Amy currently resides in a country that has the reputation for great food. What is the probability that Amy is in France?

What is Probability?

A mathematical description of a random phenomenon → **Probability distribution.**

Example: toss a fair coin several times, count the average of heads and subtract 0.5 each time, record this number x. Repeat multiple times. The curve on the right side describes the true phenomenon behind the *behavior* of x.

We know the true phenomenon governing x (the curve), we can answer questions about x, e.g. probability that x is less than 1.

What is Statistics?

- The field of study that uses the **language of probability** and **data** to answer questions about the true phenomena underlying the world.
- Data = observations from a true phenomenon. Use this data to try to discover what the curve describing the phenomenon is

Formalizing Probability: Ingredients

- ❖ **Experiment:** a procedure that can be repeated an infinite number of times, has a well-defined set of possible outcomes each of which occurs by chance.

- ❖ **Example:** a coin toss. Other examples?

Formalizing Probability: Ingredients

- **Sample Space:** the set of all possible outcomes of an experiment. Call it S.

- **Example:** S = {H, T} for a coin toss experiment. How about for a die roll?

Formalizing Probability: Ingredients

- **Event:** any collection of possible outcomes of an experiment, i.e. any subset of S, including S itself.
- **Example:** Possible event E = {H,H,T} for a coin toss. How about for a die roll?

Formalizing Probability: Set Theory

- **Union:** the set of outcomes that belong to either A or B

$$A \cup B = \{x | x \in A \text{ or } x \in B\}$$

Formalizing Probability: Set Theory

- **Intersection:** the set of outcomes that belong to A and B

$$A \cap B = \{x | x \in A \text{ and } x \in B\}$$

Formalizing Probability: Set Theory

- **Complementation:**
  the set of outcomes
  that are not in A

$$A^c = \{x | x \notin A\}$$

Formalizing Probability: Ingredients

- **Sigma Algebra:** a collection of events $\mathcal{F}$ that satisfies the properties:
  - Contains the empty set: $\emptyset \in \mathcal{F}$
  - Contains the complement of all its elements: $\text{If } A \in \mathcal{F} \text{ then } A^c \in \mathcal{F}$
  - Closed under countable unions: $\text{If } A_1, A_2, \cdots \in \mathcal{F} \text{ then } \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Formalizing Probability: Probability Function

- A function from $\mathcal{F}$ to [0, 1] that satisfies the properties:
  - Positivity: $P(A) \geq 0 \text{ for all } A \in \mathcal{F}$
  - Totality: $P(S) = 1$
  - Additivity under disjoint union:

$$\text{If } A_1, A_2, \cdots \in \mathcal{F} \text{ are pairwise disjoint, then } P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

These are referred to as the ***axioms*** of probability

Formalizing Probability: Some Rules

If P is a probability function and A and B are in $\mathcal{F}$ then:

$$P(\emptyset) = 0$$
$$P(A) \leq 1$$
$$P(A^c) = 1 - P(A)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A) \leq P(B) \text{ if } A \subset B$$

These are some of the most useful rules of probability

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Independence

If two events A and B are independent then:

$$P(A \cap B) = P(A) \cdot P(B)$$

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

If two events are independent, knowing the occurrence of one doesn't tell me anything about the probability of occurrence of the other.

Bayes Rule

Let's partition the sample space S into pairwise disjoint events A1, A2, ... whose total union is S. For any set B and any i = 1, 2, ...

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}$$

Random Variable

- A random variable is a function from a sample space S to the real numbers.
- You can think of it as a quantity of interest from an experiment.
- Example: the number of heads in a coin toss experiment. Other examples?

Random Variable

- A random variable X has a probability distribution, which defines the probability assigned to each of the values X can take.
- A random variable can be **discrete** or **continuous**.

Cumulative Distribution Function (CDF)

- ❖ Every random variable is associated with a cumulative distribution function, denoted by $F_X(x)$

$$F_X(x) = P(X \leq x) \text{ for all } x$$

Cumulative Distribution Function (CDF)

- ❖ Every random variable is associated with a cumulative distribution function, denoted by $F_X(x)$

$$F_X(x) = P(X \leq x) \text{ for all } x$$

- ❖ **Example:** Toss a coin until a head appears. Let p = probability of a head on any given toss. Define random variable X = number of tosses needed until a head appears. Then for any x = 1, 2, ...

$$P(X = x) = (1-p)^{x-1}p$$

$$F_X(x) = P(X \leq x) = \sum_{i=1}^{x} P(X = i) = \sum_{i=1}^{x} (1-p)^{i-1}p$$

Probability Density Function (PDF)

- The probability mass function of a discrete random variable X is

$$f_X(x) = P(X = x) \text{ for all } x$$

You can think of it as the size of the jump in the CDF curve at x

- The probability density function of a continuous random variable X is $f_X(\cdot)$ such that:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \text{ for all } x$$

Probability Density Function: Examples

You can compute probabilities using the probability density function, e.g. probability that X is between a and b.

Expectation

- The expectation or expected value of a random variable X is a quantity that summarizes its distribution. This is the ***typical*** value of X.

- The expectation of a discrete random variable X is

$$E(X) = \sum_{x} x \cdot P(X = x)$$

- The expectation of a continuous random variable is

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) \; dx$$

Measure of central tendency

Moments

- The expectation is the first moment of a distribution. There are higher-order moments:

$$\mu_n = E(X-\mu)^n \text{ where } \mu = E(X)$$

- The variance is the second moment, n = 2. Measures the degree of spread around the mean.

Joint Distribution

- Describes the co-occurence of two or more random variables.

- If Y is another random variable, we can talk about its joint distribution with X,

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

- If X and Y are independent, then

$$P(X,Y) = P(X)P(Y)$$

Marginal Distribution

- Consider the previous setting and assume X and Y are discrete. The marginal distribution of X is

$$P(X) = \sum_{y} P(X, Y = y)$$

It is the average distribution of X conditional on all possible values of y. The weighting function is the probability distribution of Y.

Random Variables in Probabilistic Modeling

- ❖ Data
- ❖ Hidden Variables:
  - ➢ Latent variables
  - ➢ Parameters
  - ➢ Noise variables

Bayesian Inference

- ❖ Data: $\mathbf{x}$
- ❖ Hidden structure: $\mathbf{z}$ (***latent variables***)
- ❖ Model: $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) \cdot p(\mathbf{x}|\mathbf{z})$ (*Joint* = *Prior* x "*Likelihood*")
- ❖ **Discovery** of the underlying structure through ***posterior inference***:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{\int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}}$$

Variational Inference

Finding Topics Underlying a Corpus of Documents

Republican	Yankees	Wine	Court	Company	Music
Bush	Game	Restaurant	Judge	Million	Dance
Campaign	Baseball	Food	Case	Stock	Songs
Senator	Season	Dishes	Justice	Shares	Opera
Democrats	Mets	Restaurants	Trial	Billion	Concert

Topics found on a corpus of 1.8 Million articles from *The New York Times* defining a vocabulary of more than 200,000 words.