

# Mathematical Preliminaries

---

Max Turgeon

DATA 2010—Tools and Techniques in Data Science

# Lecture Objectives

- Review basic definitions and properties from probability theory
- Review simple concepts from descriptive statistics

# Motivation

- Probability theory allows us to reason coherently about **uncertainty**
  - E.g. Will it rain tomorrow? Who will be the next Prime Minister?  
Does the prediction change when a new survey comes in?
- Even when studying deterministic systems, we rarely have complete information.
- In practice, a simple model with some uncertainty is more useful than a complex model with little uncertainty.

# Probability

---

# Basic definitions

- **Probability** is a framework for understanding the *likelihood* of certain events to occur.
- Probabilities have to satisfy certain properties:
  - $p \geq 0$  and  $p \leq 1$
  - $\sum_{s \in S} p_s = 1$
- Several *interpretations* are possible:
  - Relative frequency of an event over time (i.e. as we repeat the sampling process)
  - Degree of belief in the uncertainty of an event

# Independence i

- **Independence** tells us about the relationship between two events.
- *Conceptually*: Knowing something about event  $A$  does not affect the likelihood of event  $B$ .
- *Mathematically*: Two events  $A, B$  are **independent** if and only if the joint probability is given by

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

## Independence ii

- This can also be defined in terms of **information content**:

$$I(A) = -\log P(A).$$

- If  $I(A) = 0$  (i.e.  $P(A) = 1$ ), the event  $A$  contains no information.
- As  $I(A) \rightarrow \infty$ , the event  $A$  becomes rarer, and so its occurrence contains a lot of information.

### Definition

Two events  $A, B$  are independent if and only if the information content of the combined event is given by

$$I(A \cap B) = I(A) + I(B).$$

# Conditional Probabilities

- Let  $A, B$  be two events. The **conditional probability** of  $A$  given  $B$  is defined as

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

- If we assume  $A, B$  are independent, we get

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



# Prosecutor's fallacy

- **Careful!** The order of events is important. In general, we have

$$P(A \mid B) \neq P(B \mid A).$$

- Assuming these two quantities are equal is known as the **prosecutor's fallacy**. For example:

- $A$  = Defendant's DNA matches sample at crime scene
- $B$  = Defendant is guilty

- $P(A \mid \neg B)$  is probably very small! But you can't conclude that

$$P(B \mid A) = 1 - P(\neg B \mid A)$$

is very large.

## Exercise

Suppose that 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what is the probability that she also likes jelly?

Define:

- $A$  = Person likes peanut butter
- $B$  = Person likes jelly

We are interested in  $P(B \mid A)$ . We can use the definition of conditional probability.

$$\begin{aligned}P(B \mid A) &= \frac{P(A \cap B)}{P(A)} \\&= \frac{0.78}{0.8} \\&= 0.975.\end{aligned}$$

# Bayes Theorem

- Bayes' theorem is an important tool that allows us to **update prior beliefs**.
- Let  $A, B$  be two events. Bayes theorem says

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}.$$

- **Note:** If  $A, B$  are independent, we simply get  $P(B \mid A) = P(B)$ .

## Proof

We can quickly prove Bayes Theorem using the definition of conditional probabilities. Note that we can write  $P(A \cap B)$  in two ways:

$$\begin{aligned}P(A \cap B) &= P(A \mid B)P(B), \\P(A \cap B) &= P(B \mid A)P(A).\end{aligned}$$

Therefore, we have

$$P(B \mid A)P(A) = P(A \mid B)P(B).$$

Dividing both sides by  $P(A)$  finishes the proof.



## Application: Positive Predictive Value $i$

- Suppose a new test has been developed to diagnose PACG (a form of glaucoma).
  - PACG affects around 1% of the population.
- The test has a sensitivity of 90% and specificity of 95%
  - **Sensitivity:** Probability to receive a positive test result when the patient has the disease.
  - **Specificity:** Probability to receive a negative test result when the patient does not have the disease.
- If a patient receives a positive test result, what is the probability they have PACG?

## Application: Positive Predictive Value ii

- Define the following:
  - $A$  = Patient receives a positive test result.
  - $B$  = Patient has PACG.
- Therefore, we want to compute  $P(B \mid A)$ .
- Note that:
  - $P(B) = 0.01$
  - $P(A|B) = 0.9$  (This is the sensitivity)
  - $P(\neg A|\neg B) = 0.95$  (This is the specificity)
- We could use Bayes theorem, but we are missing  $P(A)$ ...



## Application: Positive Predictive Value iii

- We can use the **Law of Total Probability**: a patient either has PACG or does not have it.
  - In other words,  $B$  and  $\neg B$  form a *partition* of the sample space.
- LTP tells us that

$$P(A) = P(A \mid B)P(B) + P(A \mid \neg B)P(\neg B).$$

- Moreover, we have

$$P(A \mid \neg B) = 1 - P(\neg A \mid \neg B) = 1 - 0.95 = 0.05.$$

## Application: Positive Predictive Value iv

- Putting all this together, we get

$$\begin{aligned}P(A) &= P(A \mid B)P(B) + P(A \mid \neg B)P(\neg B) \\&= 0.9 \cdot 0.01 + 0.05 \cdot 0.99 \\&= 0.0585.\end{aligned}$$

## Application: Positive Predictive Value v

- We can finally use Bayes Theorem:

$$\begin{aligned}P(B \mid A) &= \frac{P(A \mid B)P(B)}{P(A)} \\&= \frac{0.9 \cdot 0.01}{0.0585} \\&\approx 0.1538.\end{aligned}$$

- In other words, the probability that a patient who tested positive actually has PACG is only about 15%.

- **Random variables** are variables that take (real) values according to a probability distribution
  - E.g.  $X = 1$  with probability  $p = 0.5$  and  $X = 0$  otherwise.
- RVs are characterized by their **cumulative distribution function** (CDF)

$$F(x) = P(X \leq x).$$

# Discrete RVs

- A **discrete** RV is one that takes only a finite, or countable, number of values.
- It can be characterized by its **probability mass function** (PMF):

$$f(x) = P(X = x).$$

- The main discrete distributions are Bernoulli, Binomial, Poisson, and Geometric.

# Continuous RVs

- A **continuous** RV is one for which  $P(X = x) = 0$  for all  $x$ .
- It can be characterized by its **probability density function** (PDF):

$$f(x) = \frac{d}{dx}F(x).$$

- The main continuous distributions are Gamma, Beta, Normal, and Student-t.

# Expectation

- For a discrete RV  $X$  with PMF  $f$ , we define its **expectation** as

$$E(X) = \sum_x x f(x).$$

- Analogously, for a continuous RV  $X$  with PDF  $f$ , we define its **expectation** as

$$E(X) = \int_x x f(x) dx.$$

# Variance

- If a RV  $X$  has expectation  $\mu$ , its **variance** is defined as

$$\text{Var}(X) = E\left((X - \mu)^2\right).$$

- **Very useful:** We have the following identity

$$\text{Var}(X) = E\left(X^2\right) - E(X)^2.$$

- **Exercise:** Prove this identity (*hint*: use the fact that the expectation is a linear operator).



# Statistics

---

# Descriptive Statistics

- **Descriptive statistics** are numerical summaries of a collection of observations.
- They give us an idea of the underlying distribution and relationship between variables.
  - E.g. mean, median, standard deviation, correlation.
- Often, a descriptive statistic matches a certain numerical property of a distribution.
  - E.g. The **sample** mean vs. the **population** mean.
- *Descriptive statistics alone shouldn't be used to draw general conclusions.*
  - We need more information/assumptions about the sampling mechanism, possible biases, etc.

# Measures of centrality

- The main ones are the **sample mean** and the **sample median**.
  - Mean:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
  - Median: Middle observation after ordering.
- Sample mean has better properties, but sample median is more robust to outliers and skewed distributions.

# Measures of variation

- The main ones are the **sample variance**, the spread, and the interquartile range (IQR).
  - Variance:  $S^2 = \frac{1}{n-1} \sum_{i=1} (X_i - \bar{X})^2$
  - Spread: Difference between largest and smallest observations.
  - IQR: Difference between 3rd and 1st quartile.
- The **sample standard deviation** is the square root of the sample variance.
- Sample variance has more better properties, but IQR is more robust to outliers.

# Chebyshev's inequality

- Chebyshev's inequality is a general statement about what proportion of the data should be a certain distance away from the population mean:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad k > 0.$$

- In other words, at least 75% of the data should be within 2 SDs from the mean, and 89% should be within 3 SDs.
- For **normal distributions**, we have tighter bounds:
  - 68% should be within 1 SD
  - 95% should be within 2 SDs
  - 99.7% should be within 3 SDs

## Application: Making educated guesses

- Assume we want to guess the distribution of heights in this class.
- We may guess that the distribution should be approximately normal.
  - We thus need the mean and variance.
- What is the average height? 170 cm?
- What is the range? 150 cm to 195 cm?
- The spread (i.e. 45 cm) should be approximately 6 times the standard deviation.
- **Final guess:**  $N(170, 7.5^2)$  should be a good approximation to the distribution of heights.

# Interpreting Variation

- **Variation** means that when we repeat measurements, we get a different answer.
- Data scientists often talk about the **signal-to-noise ratio**:
  - Is the treatment really improving patient outcomes (signal) or are differences in patient outcomes simply due to natural variation (noise)?
- For example: a great season by an athlete is sometimes followed by a more average season.
  - Was the athlete injured? Distracted?
  - Or was the great season an extreme value but normal variation?
- **Statistics** can help us disentangle these questions.