Probabilities: Joint, Marginal, and Conditional

BIOS 6611

CU Anschutz

Week 2

Probability for Bivariate Distributions

Probability

Probability is the likelihood that a given event, or combination of events, will occur. There are two important rules:

- A probability can range from 0 to 1
- The sum of all probabilities in a sample space must sum to 1

Most often in class we will be working with probability distributions (e.g., binomial, normal, t, etc.).

Bivariate Distributions

A **bivariate distribution** is a distribution that is defined by *two* random variables.

For our motivating example, let's consider a respiratory disease (Y) and smoking (X):

	Smoker (X=1)	Non-Smoker (X=0)	Total
Respiratory Disease (Y=1)	0.15	0.05	0.20
No Respiratory Disease (Y=0)	0.30	0.50	0.80
Total	0.45	0.55	1.00

Joint Distribution

Based on our bivariate distribution, we can calculate the **joint distribution** for any two observed values of X (smoking) and Y (respiratory disease):

$$P(X = x \text{ and } Y = y) = P(X = x \cap Y = y)$$

What is the probability of randomly sampling a non-smoker with a respiratory disease?

	X=1	X=0	Total
Y=1	0.15	0.05	0.20
Y=0	0.30	0.50	0.80
Total	0.45	0.55	1.00

Marginal Distribution

The **marginal distribution** is the individual probability distribution for each random variable (i.e., $P(X = x), P(Y = y)$). For our bivariate example we can add the cells together that represent a given random variable value (i.e., the "Total" column/row).

What is the probability of randomly sampling a non-smoker?

	X=1	X=0	Total
Y=1	0.15	0.05	0.20
Y=0	0.30	0.50	0.80
Total	0.45	0.55	1.00

What is the probability of randomly sampling someone without respiratory disease?

Conditional Distribution

The **conditional distribution** is the probability distribution of a random variable when another random variable is known to be a particular value:

$$P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

For a non-smoker, what is the probability of having a respiratory disease?

	X=1	X=0	Total
Y=1	0.15	0.05	0.20
Y=0	0.30	0.50	0.80
Total	0.45	0.55	1.00

For a smoker, what is the probability of having a respiratory disease?

Independence of Two Random Variables

Two random variables X and Y are independent iff (if and only if)

$$P(X = x \cap Y = y) = P(X = x) \times P(Y = y), \text{ or}$$

$$P(Y = y | X = x) = P(Y = y)$$

Are X and Y independent in our example?

	X=1	X=0	Total
Y=1	0.15	0.05	0.20
Y=0	0.30	0.50	0.80
Total	0.45	0.55	1.00

Bayes' Theorem

Bayes' Theorem (or Rule or Law)

Bayes' Theorem calculates the posterior probability of an event based on some prior probability *by utilizing conditional probabilities*.

The theorem shows how to take prior probabilities (e.g., assumed prevalence of disease), incorporate new information (e.g., diagnostic test results), and obtain revised (posterior) probabilities (e.g., predictive values).

In BIOS 6611 we will encounter this most heavily when discussing diagnostic testing performance (e.g., sensitivity and specificity). However, we can note that this nifty theorem serves as the foundation for the entire *Bayesian paradigm* of statistics!

Bayes' Theorem Defined - Part 1

Before defining Bayes' Theorem, it is helpful to note an extension from our conditional probability formula from earlier:

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \implies P(X \cap Y) = P(X|Y)P(Y)$$

By moving our probabilities around, we see that $P(X \cap Y) = P(X)P(X|Y) = P(Y)P(Y|X)$.

Bayes' Theorem Defined - Part 2

Before defining Bayes' Theorem, it is helpful to recall the **Total Law of Probability** (also found in Lecture SA3, Section D in the "Packet for Enrolled Students"):

$$P(X) = \sum_{i=1}^{k} P(X \cap Y_i) = \sum_{i=1}^{k} P(X|Y_i)P(Y_i)$$

This is how we calculated the marginal probability earlier, with $k = 2$ for our bivariate distribution.

Bayes' Theorem in General

Without further ado, Bayes' theorem states:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

(Yes, I know, a bit anticlimactic. . .)

Bayes' Theorem Future Motivation for BIOS 6611

To motivate our future diagnostic testing work, let's define two new random variables: D_i represents i mutually exclusive and exhaustive disease states $(i = 1, ..., k)$ and T represents a positive test or presence of a symptom. Then Bayes' theorem states:

$$P(D_i|T) = \frac{P(T \cap D_i)}{P(T)} = \frac{P(T|D_i)P(D_i)}{\sum_{i=1}^{k} P(T|D_i)P(D_i)}$$

Where the first equality is by our definition of conditional probability, and the second is by the definition of joint probability and total probability.

Bayes' Theorem as Applied in Bayesian Statistics

While beyond the scope of BIOS 6611, I think it is important to at least introduce the roots of Bayesian analysis in practice. Let H represent a hypothesis to be tested and D be the data which may given evidence for (or against) H. Then Bayes' theorem is

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Bayes' Theorem as Applied in Bayesian Statistics

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Each of these terms plays a different role:

- $P(H)$ is the **prior (probability)** that H is true *before* the data is considered
- $P(D|H)$ is the **likelihood** and represents the evidence for H provided by the observed data D
- $P(D)$ is the **total probability** of the data which takes into account all possible hypotheses
- $P(H|D)$ is the **posterior (probability)** that H is true after the data has been considered