Multivariate Distributions

Goal

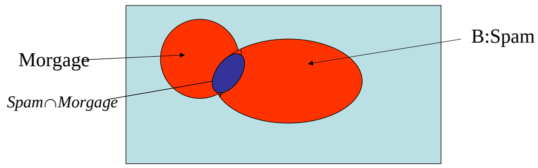
By the end of this class, you should be able to:

- compute joint distributions
- derive marginal distributions from joint distributions
- derive conditional distributions from joint distributions
- test whether two variables are statistically independent

Spam Example

- 1000 email, 100 spam, 900 ham
- 15 contains “mortgage”.
- 10 spam email has “mortgage”
- P(spam , ”mortgage”)=?

Venn Graph



The joint probability of events *A* and *B* is the probability of both events happen

Tabulated Data

	Spam	Ham	
Mortgage	10		15
~Mortgage			
	100	900	

Filled Table

	Spam	Ham	
Mortgage	10	5	15
~Mortgage	90	895	985
	100	900	1000

Relative frequency

Joint Prob

	Spam	Ham	
Mortgage	.01	.005	.015
~Mortgage	.09	.895	.985
	.1	.9	1

Joint probability distribution

Define the **joint probability distribution** $P_{XY}(x, y)$ to be a function that supplies the joint probability for each pair of values, *x* and *y*.

We know that $0 \le P_{XY}(x, y) \le 1$ and that $\sum\sum P_{XY}(x, y) = 1$ where the summation is over all values that *X* and *Y* take on.

Relative frequency

Joint Prob

	Spam	Ham	
Mortgage	.01	.005	.015
~Mortgage	.09	.895 $\sum$	.985
	.1	.9	1

Not to Scare You

1. $f(x, y) \geq 0$, for all (x, y)
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy = 1$
3. $P[(X, Y) \in A] = \int_A \int f(x,y)dxdy$

Marginal Probability Functions

The *marginal probability functions* of X and Y, denoted $p_X(x)$ and $p_Y(y)$ are given by

$$p_X(x) = \sum_y p(x, y) \qquad p_Y(y) = \sum_x p(x, y)$$

Filled Table

Joint Prob

	Spam	Ham	
Mortgage	.01	.005	.015
~Mortgage	.09	.895	.985
	.1	.9	1

marginal P(spam) P(ham)

Marginal Probability Density Functions

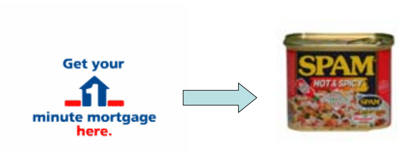
The *marginal probability density functions* of X and Y, denoted $f_X(x)$ and $f_Y(y)$, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \quad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx \quad \text{for } -\infty < y < \infty$$

Spam Example

- 1000 email, 100 spam, 900 ham
- 15 contains “mortgage”.
- 10 spam email has “mortgage”
- P(spam | ”mortgage”)=?



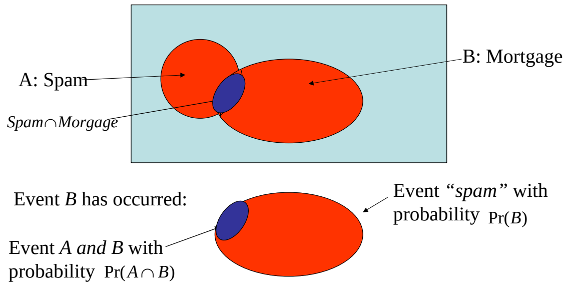
Computing Conditional Probabilities

- A conditional probability is the probability of one event, given that another event has occurred:

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

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

Where P(A and B) = joint probability of A and B
P(A) = marginal probability of A
P(B) = marginal probability of B



Example

- The probability that it is Friday and that a student is absent is 0.03. Since there are 5 school days in a week, the probability that it is Friday is 0.2. What is the probability that a student is absent given that today is Friday?

Conditional probability distribution

Just as we defined the conditional probability of B given A as: $\Pr(B \mid A) = \dfrac{\Pr(A \cap B)}{\Pr(A)}$, we can define the **conditional distribution** of Y given $X = x$ as:

$P_{Y|X}(y \mid x)$ **Conditional distribution**

Thus, for any value of Y,

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x \cap Y = y)}{\Pr(X = x)} = \frac{P_{XY}(x, y)}{P_X(x)}$$

Relative frequency

Conditional Distribution

	Spam	Ham	
P(*am | mortgage)	.01/.015	.005/.015	.015
P(*am | ~mortgage)	.09/.985	.895/.985	.985
	.1	.9	1

We have a distribution of x for each value of y {mortgage, ~mortgage}

Independence of Events

Two events *E* and *F* are **independent** if the occurrence of event *E* in a probability experiment does not affect the probability of event *F*.

Two events are **dependent** if the occurrence of event *E* in a probability experiment affects the probability of event *F*.

Statistical Independence

- Two events are independent if and only if:

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

Independence of random variables

we can define the **independence of random variables** as:

$$P_{Y|X}(y \mid x) = P_Y(y) \quad \text{for all } x, y$$
$$P_{XY}(x, y) = P_X(x) \cdot P_Y(y) \quad \text{for all } x, y$$

Independent or Dependent

- X: Weekend or Not, Y: Sales Volume
- X: Temperature, Y: Your shoe size
- X: Changes in Bond Yield Tomorrow, Y: Changes in Dow Jones Index tomorrow
- X: GOOGLE's Stock return tomorrow
- Y: Yahoo's stock return tomorrow
- Mother's blood type, Father's blood type

		y				
		0	**1**	**2**	**3**	$P_X(x)$
x	**0**	2/84	3/84	4/84	5/84	14/84
	1	4/84	6/84	8/84	10/84	28/84
	2	6/84	9/84	12/84	15/84	42/84
	$P_Y(y)$	12/84	18/84	24/84	30/84	

In this case, *X* and *Y* are independent.

$$\frac{14}{84} \cdot \frac{12}{84} = \frac{2}{84}, \quad \frac{28}{84} \cdot \frac{12}{84} = \frac{4}{84}, \quad \frac{42}{84} \cdot \frac{12}{84} = \frac{6}{84}, \text{ etc.}$$

We can also see it by examining the distribution probability of Y given X.

		y			
		$P_{Y\|X}(0\|x)$	$P_{Y\|X}(1\|x)$	$P_{Y\|X}(2\|x)$	$P_{Y\|X}(3\|x)$
x	0	2/14	3/14	4/14	5/14
	1	2/14	3/14	4/14	5/14
	2	2/14	3/14	4/14	5/14
	$P_Y(y)$	2/14	3/14	4/14	5/14

We can also see it by examining the distribution probability of X given Y.

		y				
		0	1	2	3	$P_X(x)$
x	$P_{X\|Y}(0\|y)$	1/6	1/6	1/6	1/6	1/6
	$P_{X\|Y}(1\|y)$	2/6	2/6	2/6	2/6	2/6
	$P_{X\|Y}(2\|y)$	3/6	3/6	3/6	3/6	3/6

Table A

	1	2	3
1	.25	.15	.1
2	.15	.09	.06
3	.1	.06	.04

Table B

	1	2	3
1	.35	.1	.05
2	.15	.1	.05
3	0	.1	.1

Table C

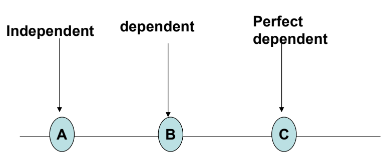
	1	2	3
1	.5	0	0
2	0	.3	0
3	0	0	.2

Anything Special About this?

It's diagonal!

This is, in some sense, a maximum amount of interaction: if you know one, you know the other.

Level of Dependency



How to Quantify? Next Week.....

After Class

- Think about Monte Hall Problem
- True or False Problem at the Back
- Install R and Data Analysis Add In in Excel and Let me know any problem during the Demo.

Monty Hall Problem

- You are participating in a TV game show. You are shown three doors and are told that there is a car behind one door and a goat behind each of the other two doors.
- You are allowed to select one door (without opening it). The game show host opens one of the other doors, showing you a goat.
- Should you now pick the other closed door, or stick with your original pick?

True or False

If we know that $P(A|B_1) = P(A|B_2)$, then

$$\frac{P(B_1|A)}{P(B_2|A)} = \frac{P(B_1)}{P(B_2)}$$