

# Essentials of Applied Data Analysis

## IPSA-USP Summer School 2017

### Joint Distributions

Leonardo Sangali Barone  
leonardo.barone@usp.br

jan/17

### Joint Distribution

Causal theories of politics necessarily involve expected relationships among concepts or variables. As such, we want to study joint distributions; marginal distributions are a natural extension as we will see in a moment.

For discrete variables, we can use contingency tables to represent the joint frequency distribution for two random variables.

### Joint Distribution - legislators

Example: two variables, gender ( $Y$ ) and political party ( $X$ )  
If we take the relative frequency of the cells we get:

Y/X	Party A ( $A$ )	Party B ( $B$ )	Party C ( $C$ )
Women ( $W$ )	$P(W \cap A)$	$P(W \cap B)$	$P(W \cap C)$
Man ( $M$ )	$P(M \cap A)$	$P(M \cap B)$	$P(M \cap C)$

Note: joint distributions are represented, guess what, by joint probabilities!

## Joint Distribution

We can write the joint probabilities as

$$P(\text{Women} \cap \text{Party B}) \text{ or } P(W \cap B)$$

or, if we have named the variables

$$P(\text{Gender} = \text{Women}, \text{Party} = \text{Party B})$$

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

These notations are equivalent if everything is well named.

## Joint Distribution - dice

Joint distributions can be build from the process that generate the data (dice) or from a sample.

Example: roll a dice. Prime *vs* not prime (Y); and even *vs* odd (X). If we take the relative frequencies we get:

Y/X	Even ( $E$ )	Odd ( $O$ )
Prime ( $I$ )	$P(I \cap E) = 1/6$	$P(I \cap O) = 2/6$
Not Prime ( $N$ )	$P(N \cap E) = 2/6$	$P(N \cap O) = 1/6$

## Joint Distribution - sex and political affiliation in Fakeland

We could do the same using our Fakeland example (let's not use numbers here).

Example:

Sex/Party	Conservative ( $C$ )	Independent ( $I$ )	Socialist ( $S$ )
Women ( $W$ )	$P(W \cap C)$	$P(W \cap I)$	$P(W \cap S)$
Man ( $M$ )	$P(M \cap C)$	$P(M \cap I)$	$P(M \cap S)$

Now, use the fake dataset to build the table above. Remember that what you will build is the joint *sample* distribution (not the “True” Fakeland distribution).

## Joint Distribution - conditional probability notation

All of the notations below are equivalent.

$$P(X = x_i|Y = 1) = P(X|Y = 1) = \frac{P(X = x, Y = 1)}{P(Y = 1)} = \frac{P(X \cap (Y = 1))}{P(Y = 1)}$$

## Joint Distribution - Marginal Probabilities

The marginal probability of an event A is the probability that A will occur unconditional on all the other events on which A may depend. It is very easy to comprehend that in our example. If we take the relative frequencies we get:

Example:

Sex/Party	Conservative ( <i>C</i> )	Independent ( <i>I</i> )	Socialist ( <i>S</i> )	Marginal
Women ( <i>W</i> )	$P(W \cap C)$	$P(W \cap I)$	$P(W \cap S)$	P(W)
Man ( <i>M</i> )	$P(M \cap C)$	$P(M \cap I)$	$P(M \cap S)$	P(M)
Marginal	$P(C)$	$P(I)$	$P(S)$	1

## Joint Distribution - Marginal Probabilities

We can calculate the Marginal Probability by simply summing the probability of *A* happening conditional on all other events on which *A* depend (partitions of *B*):

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n) = \sum_{i=1}^n P(A \cap B_i)$$

or

$$\begin{aligned} P(A) &= P(B_1) * P(A|B_1) + P(B_2) * P(A|B_2) + \dots + P(B_n) * P(A|B_n) = \\ &= \sum_{i=1}^n P(B_i) * P(A|B_i) \end{aligned}$$

This means that one averages over other events and focuses on the one event, *A*, of interest.

## Joint Distribution - Independence - Cards

What does happen with the joint distribution of two random variables if they are independent of each other?

Example: choose a card from a deck. Calculate the relative frequencies:

Y/X	Hearts	Spades	Clubs	Diamonds	Marginal
King	1/52	1/52	1/52	1/52	4/52
Queen	1/52	1/52	1/52	1/52	4/52
Other	11/52	11/52	11/52	11/52	44/52
Marginal	13/52	13/52	13/52	13/52	52/52

We got a king. What is the probability that it is the king of hearts?

$$P(H|K) = 1/4$$

We got a queen. What is the probability that it is the queen of hearts?

$$P(H|Q) = 1/4$$

We got any other card. What is the probability that it is a card of hearts?

$$P(H|O) = 1/4$$

If the marginals probabilities are equal to the conditional probabilities, than the two variables are independent from each other.

$$P(H) = P(H|K) = P(H|Q) = P(H|O) = 1/4$$

Under independence:

$$P(H \cap K) = P(H) * P(K) = 1/4 * 1/13 = 1/52$$

or

$$P(Y = Hearts, X = King) = P(Y = Hearts) * P(X = King)$$

## Joint Distribution - sex and political affiliation in Fakeland

Let's go back to our empirical example. If you completed the sample joint distribution, you got the table below:

Count:

Sex/Party	Conservative ( $C$ )	Independent ( $I$ )	Socialist ( $S$ )	Marginal
Women ( $W$ )	3	8	4	15
Man ( $M$ )	3	7	5	15
Marginal	6	15	9	30

Proportion relative to Total:

Sex/Party	Conservative ( $C$ )	Independent ( $I$ )	Socialist ( $S$ )	Marginal
Women ( $W$ )	0.10	0.27	0.13	0.50
Man ( $M$ )	0.10	0.23	0.17	0.50
Marginal	0.20	0.50	0.30	1.00

Proportion relative to Rows:

Sex/Party	Conservative ( $C$ )	Independent ( $I$ )	Socialist ( $S$ )	Marginal
Women ( $W$ )	0.20	0.54	0.26	1.00
Man ( $M$ )	0.20	0.47	0.33	1.00
Marginal	0.20	0.50	0.30	1.00

Proportion relative to Columns:

Sex/Party	Conservative ( $C$ )	Independent ( $I$ )	Socialist ( $S$ )	Marginal
Women ( $W$ )	0.50	0.53	0.13	0.45
Man ( $M$ )	0.50	0.47	0.17	0.55
Marginal	1.00	1.00	1.00	1.00

Can you tell if the two variables are independent of each other just by looking at the tables? We are going to train this a lot with real data using Stata.