Teoria Homework 4

	Black	Brown	Blue	*total*
Low	5	10	5	20
Medium	15	10	5	30
High	10	0	40	50
total	30	20	50	

We can define this distribution as **bivariate** because it describes the joint distribution of two categorical (or qualitative) variables:

1. **Color** (with categories: Black, Brown, Blue)
2. **Level** (with categories: Low, Medium, High)

In a bivariate distribution, we are interested in the possible combinations between two variables and their joint frequencies—how often each combination of values occurs in the data. The Bivaried distribution shows how many elements belong to each combination of **Color** and **Level**.

> A bivariate distribution can show marginal distributions (univariate distributions) for each variable.

Marginal Distribution: The totals in the last row and column represent marginal distributions, summarizing the counts for each variable irrespective of the other.

The conditional distribution represents how the distribution of levels changes when conditioned on the color being Black (is like a subset because it is conditioned by the eye color). When we have this type of distribution we call it such as: **Conditional Distribution**

$$H \mid yC = \text{Black} \quad \text{One variable conditioned by the value of another variables}$$

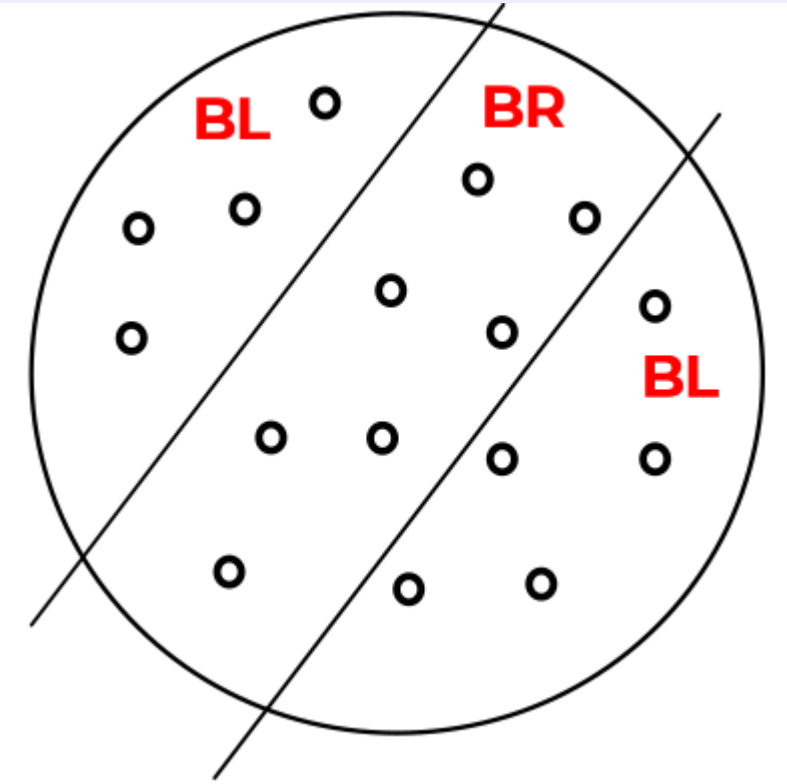
More generic:

$$X \mid Y_j = y_j \qquad Y \mid X_j = x_j$$

- Indicates the distribution of X given that Y takes on a specific value y_j.
- Indicates the distribution of Y given that X takes on a specific value x_j.

Example

Conditional distribution is like if we cut the population in the possible values that he can assume.



Statistical Independence

In statistics, independence refers to a situation where two or more variables do not influence each other. When analyzing data, particularly in the context of hypothesis testing or regression analysis, independence implies that the variability in one variable does not provide any information about the variability in another variable.

Example

For instance, if we have two variables X and Y, they are considered statistically independent if knowing the value of X does not change the distribution of Y. This is crucial when performing analyses that assume independence, such as linear regression or analysis of variance, as

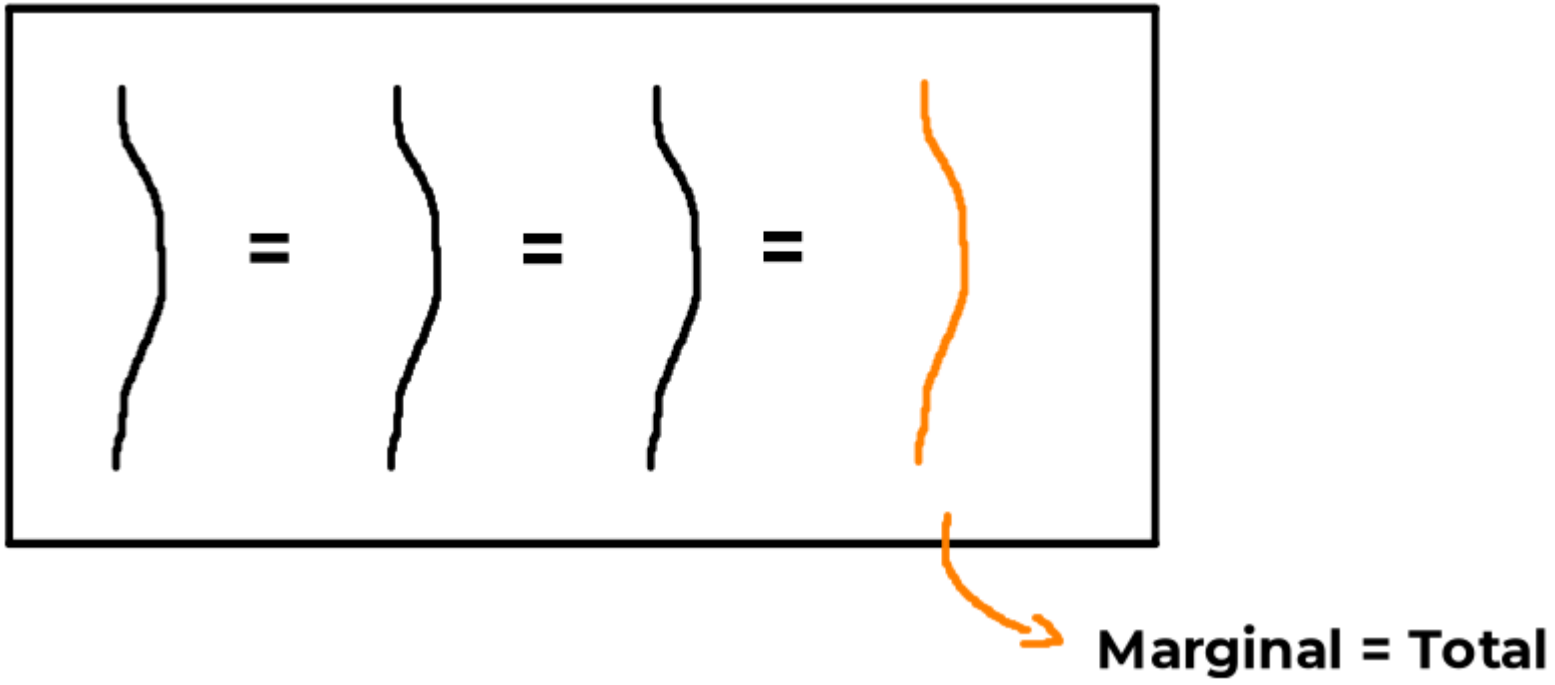
violations of this assumption can lead to incorrect conclusions.

> If two or more variables do not influence each other the distribution are the same. In the contrary the distribution will have different shape.

Matematically:
All condition distribution $(X \mid Y = y)$ are equal, in the sense of relative frequency. (histograms or shapes, same structure).

Definition of statistical independence

In statistic, dependence refers to a situation where two or more variables influence each other $\rightarrow$ coditional distributions are equal!



x/y	y_1	y_j	y_n	***total***
x_1	n_{11}	n_{1j}	n_{1n}	$n_{1.}$
x_i	n_{i1}	n_{ij}	n_{in}	$n_{i.}$
x_n	n_{n1}	n_{nj}	n_{nn}	$n_{n.}$
total	$n_{.1}$	$n_{.j}$	$n_{.n}$	n_{nn}

- i row
- j column
- $t_{.j}$ relative frequency of y_j
- $n_{i.}$ marginal frequency $\rightarrow \frac{n_{i,j}}{n_{.,j}} = \frac{n_{i,.}}{n}$ equal to $(n_{i,.} = \sum_j n_{i,j})$

$$\text{Conditional} = [f_x \mid Y = y_j] \text{ is equal to } [f(x = x_i)]$$

More specifically

$$\frac{n_{i,j}}{n} = \frac{n_{i,.}}{n} \cdot \frac{n_{.,j}}{m}$$

where:

- First fraction is the **relative frequency** of x_i : $f(x = x_i)$
- Second fraction is the **relative frequency** of y_i : $f(y = y_j)$

$$f(x = x_i) \cdot f(y = y_j) = f(x = x_i, y = y_j)$$

We say that the relative frequency of x_i and y_j are equal to this $f(x = x_i, y = y_j)$.
This relationship indicates that the likelihood of observing both x_i and y_j simultaneously can be captured through this joint probability expression. In the context of probability theory, this relationship can also be articulated as:

$$P(X \text{ and } Y) = P(X) \cdot P(Y)$$

$$P(X \mid Y) = P(X)$$

Donsker Distribution

Brownian Distribution

Brownian motion, also known as **Wiener process**, is a fundamental concept in stochastic processes and probability theory. Mathematically, Brownian motion is characterized by its continuous, nowhere differentiable paths, which exhibit stationary independent increments.

Donsker Distribution

The **Donsker Distribution** is named after Monroe Donsker, who formulated a key result in probability theory known as **Donsker's theorem**. This theorem establishes a profound connection between discrete random processes, such as random walks, and continuous stochastic processes, specifically Brownian motion.

Donsker’s Theorem

Donsker's theorem states that as the number of steps in a properly normalized random walk approaches infinity, the distribution of the scaled random walk converges to that of a standard Brownian motion.

In more technical terms, if we have a random walk defined by:

$$S_n = X_1 + X_2 + \cdots + X_n$$

where X_i are independent and identically distributed (i.i.d.) random variables, then, under certain conditions (such as appropriate normalization), the path of the scaled random walk:

$$\frac{S_n}{\sqrt{n}}$$

converges in distribution to a Brownian motion $B(t)$ as $n \to \infty$