ESTADÍSTICA

I

Material 1ra. Parte

Cátedra: Liliana Ghersi.
Autor: Facundo D. Tula.

INDICE

INTRODUCCIÓN: "REGLAS DE CONTEO".

0.1) EL NÚMERO FACTORIAL.

Se define como:

$$n\,! \Rightarrow \begin{cases} n \in E^{+} \Rightarrow n\,! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 \\ n = 0 \quad \Rightarrow 0\,! = 1 \end{cases}$$

0.2) PERMUTACIONES.

Son las distintas disposiciones ordenadas que se pueden formar con un conjunto de objetos diferentes.

Dados *n* elementos, permutar es ordenar de todas las maneras posibles.

Se define de la siguiente manera:

$$P_n = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 = n\,!$$

0.3) VARIACIONES.

Son las distintas formas en que podemos agrupar los elementos de un conjunto con *N* elementos tomados en grupos de *n* elementos. Es por esto que siempre se cumple que $N > n$. Estos grupos se diferencian por el orden de sus elementos.

Su cálculo es el siguiente:

$$V_N^n = \frac{N!}{(N-n)!}$$

En el caso que se de $N = n$, estamos frente a una simple permutación.

$$V_N^n = V_N^N = \frac{N\,!}{(N-N)!} = \frac{N\,!}{0!} = N\,!$$

0.4) COMBINACIONES.

Es la cantidad de grupos de *N* elementos tomados de a *n* elementos. En este caso los grupos son distintos únicamente cuando tienen un elemento diferente y no por el orden de los mismos.

Se define como:

$$C_N^n = \frac{N\,!}{(N-n)! \cdot n\,!}$$

UNIDAD N°1: PROBABILIDADES.

1.1) ALGUNAS DEFINICIONES INTRODUCTORIAS.

1.1.1. Probabilidad.

Es una medida que se aplica para cuantificar la incertidumbre de un evento o suceso aleatorio. Mide la expectativa de aparición de tal suceso.

Sueles notarse en porcentajes, pero de manera tradicional se expresa en términos relativos.

La notación matemática normalmente es $P(s)$, que se lee: Probabilidad de un suceso "s". El número expresado en términos relativos oscila entre 0 y 1.

1.1.2. Espacio Muestral.

Es el conjunto de todos los posibles resultados "elementales" de un experimento aleatorio.

1.1.3. Experimento Aleatorio.

Es la observación de un fenómeno del cual no sabemos qué resultado se va a obtener.

1.1.4. Suceso Aleatorio.

Dado un experimento aleatorio y su correspondiente espacio muestral se define como suceso aleatorio a un conjunto de resultados posibles, o sea, un subconjunto del espacio muestral.

1.1.5. Sucesos Incompatibles.

También conocidos como mutuamente excluyentes o disjuntos.

Son eventos definidos de modo que la ocurrencia de un elemento imposibilita la ocurrencia de cualquiera de los otros.

1.1.6. Sucesos Compatibles.

Son eventos que pueden presentarse simultáneamente.

1.1.7. Eventos dependientes o condicionados.

Son sucesos donde la presencia de uno condiciona la de otro.

1.1.8. Eventos independientes.

Cuando la ocurrencia o no de un suceso no afecta la posibilidad asignada a la ocurrencia de otro.

1.2) TEORÍAS OBJETIVAS DE PROBABILIDAD.

1.2.1. Definición Clásica (*Pierre Simon Laplace*).

La probabilidad en este caso se calcula *a priori*, o sea, antes de hacer cualquier prueba. Se calcula de la siguiente manera:

$$P(s) = \frac{Casos\ Faborables}{Casos\ Posbiles} = \frac{s}{EM}$$

1.2.2. Definición Frecuentalista (*Richard Von Mises*).

A diferencia de la definición Clásica, los frecuentalistas definen la posibilidad *a posteriori*, es decir, luego de experimentar.

Se calcula de esta forma:

$$P(s) = \lim_{n \to +\infty} \frac{Casos\ Favorables}{Casos\ Totales} = \frac{Casos\ Favorables}{Cantidad\ de\ veces\ que\ se\ experimentó}$$

1.2.3. Definición Axiomática.

Se constituye un modelo en el cual a cada uno de los resultados posibles de un experimento aleatorio se le asigna una probabilidad que cumple con las siguientes pautas:

$$\begin{cases} 0 \leq P(s) \leq 1 \\ P(EM) = 1 \end{cases}$$

Podemos agregar en base a estos límites una distinción entre las probabilidades asignadas a los Eventos Seguros y a los Imposibles.

La probabilidad de un evento seguro (que es lo mismo que decir la probabilidad del espacio muestral) es igual a 1. Pero que la probabilidad sea uno no implica que sea un evento seguro, es decir que no hay doble implicación en la afirmación.

$$Evento\ Seguro \Rightarrow P(Evento\ Seguro) = P(EM) = 1$$

En el caso de los Sucesos Imposibles (es decir, del conjunto vacío), la probabilidad asociada es 0. Pero que la probabilidad sea cero no implica que sea un suceso imposible.

$$Suceso\ Imposible \Rightarrow P(Suceso\ Imposible) = P(\phi) = 0$$

Una última aclaración en esto sería que si un suceso es seguro no implica que necesariamente vaya a aparecer, y que sea imposible no implica que no aparezca, ya que estamos hablando en términos probabilísticos, es decir de la expectativa sobre la aparición de esos sucesos.

1.3) CÁLCULO DE PROBABILIDADES.

1.3.1. Probabilidad Total.

Dados dos sucesos A y B, es la probabilidad de que ocurra A o B, o ambos. Según el *Teorema de la Adición de Probabilidades*:

$$Si\ son\ compatibles \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$Si\ son\ incompatibles \Rightarrow P(A \cup B) = P(A) + P(B)$$

Otras maneras de expresar la probabilidad total de dos sucesos, *A* y *B*, son las siguientes:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B) = P(A) + P(\overline{A} \cap B)$$
$$P(A \cup B) = P(A \cap \overline{B}) + P(A \cap B) + P(\overline{A} \cap B)$$
$$P(A \cup B) = P(B) + P(A \cap \overline{B})$$
$$P(A \cup B) = 1 - P(\overline{A} \cup \overline{B})$$
$$P(\overline{A \cup B}) = P(\overline{A} \cup \overline{B})$$

1.3.2. Probabilidad Conjunta.

Dados dos sucesos *A* y *B*, es la probabilidad de que ocurran, simultáneamente, ambos.

Según el *Teorema del Producto de Probabilidades*:

$$\textit{Si son dependientes} \Rightarrow P(A \cap B) = P(A) \cdot P\left(B/A\right) \quad \Rightarrow P(B) \neq P\left(B/A\right)$$
$$\textit{Si son independientes} \Rightarrow P(A \cap B) = P(A) \cdot P(B) \quad \Rightarrow P(B) = P\left(B/A\right)$$

De esto se desprende el supuesto de que si $P(B) = P\left(B/A\right)$ se dice que ambos sucesos son *estocásticamente independientes*. Si $P(B) \neq P\left(B/A\right)$ se dice que son *estocásticamente dependientes*.

Para probar la independencia de con eventos, A y B, se parte, generalmente de la probabilidad condicional o directamente desde la conjunta.

Entonces, si $P(A \cap B) = P(A) \cdot P\left(B/A\right)$, y $P(A \cap B) = P(A) \cdot P(B)$, podemos igualar y decir que:

$$P(A \cap B) = P(A \cap B) \Rightarrow P(A) \cdot P\left(B/A\right) = P(A) \cdot P(B) \Rightarrow P\left(B/A\right) = P(B)$$

Finalmente llegamos a la expresión que buscábamos para probar la independencia de dos conjuntos, tenemos que probar que $P\left(B/A\right) = P(B)$.

Viéndolo desde un punto de vista más racional, y quizás un poco menos matemático, podemos ver para probar la dependencia o independencia, si el suceso A condiciona la aparición de B, o viceversa. Parecería obvio decir que $P(B) = P(B)$, pero es un buen punto de partida para probar lo que queremos, porque lo que hacemos desde esa igualdad para ver si ambos sucesos son independientes o no, es agregarle en uno de los lados la condición, en este caso A. Y así es como obtenemos, de forma inductiva, a la expresión que logramos más arriba $P(B) = P\left(B/A\right)$.

Entonces, si esta igualdad se cumple, decimos que A no condiciona a la aparición de B por lo que son sucesos estocásticamente independientes. En cambio, si no se cumple la igualdad, que aparezca A condiciona la aparición de B, por lo que se llaman sucesos estocásticamente dependientes.

1.3.3. Probabilidad Condicionada.

Dados dos sucesos *A* y *B*, es la probabilidad de que ocurra *A* sabiendo de ya ocurrió *B*, o viceversa.

$$P\left(A/B\right) = \frac{P(A \cap B)}{P(B)} \iff P(B) \neq 0$$

1.4) TEOREMA DE BAYES.

La regla de Bayes sirve para calcular probabilidades condicionales, pero su importancia tiene que ver con el uso de probabilidades subjetivas para tratar problemas de decisión bajo incertidumbre. El interés de Bayes fue buscar la probabilidad de una causa específica cuando se observa un fenómeno particular.

Entonces podemos decir que esta regla sirve para calcular la probabilidad de aparición de una causa (C_m) dado que ya se dio un efecto (E_k). La fórmula es la siguiente:

$$P\left(C_m/E_k\right) = \frac{P(C_m \cap E_k)}{P(E_k)} = \frac{P(C_m E_k)}{\sum_{i=1}^{n} P(C_i E_k)} = \frac{P(C_m)\cdot P\left(E_k/C_m\right)}{\sum_{i=1}^{n} P(C_i)\cdot P\left(E_k/C_i\right)}$$

UNIDAD N°2: VARIABLES ALEATORIAS.

2.1) DEFINICIÓN DE VARIABLE ALEATORIA.

Una Variable Aleatoria (VA) es una función cuya primera coordenada es un elemento de la colección exhaustiva definida para el espacio muestral (EM) y la segunda coordenada es la "*Probabilidad del Evento*" o la "*Ley de Densidad de ese evento*".

La primera coordenada va a ser un número que identifica al evento en cuestión por lo tanto una VA es un conjunto de pares ordenados numéricos donde el valor de la primera depende para su presentación de la aparición del evento de la colección exhaustiva.

Una VA puede estar dentro del campo real ($VA \in \mathbb{R}$) como el complejo ($VA \in \mathbb{C}$).

2.2) TIPOS DE VARIABLES ALEATORIAS. PROBABILIDAD ACUMULADA. LEY DE CIERRE.

Cuando el conjunto de los elementos de la colección exhaustiva, sea finita o infinita, es numerable se llama "*Ley de Probabilidad*" la que determina el valor de la segunda coordenada y se la denomina *Variable Aleatoria Discreta* (VAD). En este tipo de VA si tomamos dos valores de variable existen infinitos puntos interiores que no son tomados en cuenta, es decir que la gráfica de distribución va a ser un conjunto de puntos aislados unos de otros.

Cuando el conjunto es no numerable (es decir, ilimitado) se usa para definir la segunda coordenada la "*Ley o Función de Densidad*". Este tipo de VA se llama *Variable Aleatoria Continua* (VAC). A diferencia de una VA Discreta el gráfico va a ser una curva completa ya que toma todos los valores posibles dentro del campo que se determine, ya sea el real ($VAC \in \mathbb{R}$) o el complejo ($VAC \in \mathbb{C}$).

Para calcular, en ambos casos de variables, la probabilidad acumulada a un punto "*a*" se hace aplica el siguiente razonamiento:

$$VAD \Rightarrow P(X \leq a) = \sum_{i=1}^{X_i=a} P(X_i)$$

$$VAC \Rightarrow P(X \leq a) = \int_{L_i;-\infty}^{a} f(x)\,dx$$

Para los dos casos se cumple la Ley de Cierre de la siguiente forma:

$$VAD \Rightarrow P(X \leq L_s) = \sum_{i=1}^{X_i=a} P(X_i) = 1$$

$$VAC \Rightarrow P(X \leq L_s) = \int_{L_i;-\infty}^{L_s;+\infty} f(x)\,dx = 1$$

2.3) PARÁMETROS POBLACIONALES DE UNA VARIABLE ALEATORIA.

2.3.1. *Parámetros de Tendencia Central.*

2.3.1.A. Valor Esperado o Esperanza.

El Valor Esperado o Esperanza Matemática, es una construcción a partir de las cosas que suceden y a la fuerza con que estas pasan, que denota una expectativa. Es, junto con la Dispersión, un parámetro fundamental dentro del cálculo estadístico.

Su resultado es un valor de variable, del cual tenemos expectativa de que resulte, en promedio, al experimentar.

Su cálculo es el siguiente:

$$\text{Situación Discreta} \Rightarrow E(X) = \mu = \sum_{i=1}^{n;+\infty} X_i \cdot P(X_i)$$

$$\text{Situación Continua} \Rightarrow E(X) = \mu = \int_{L_i;-\infty}^{L_s;+\infty} X \cdot f(X)\, dx$$

La Esperanza [$E(X)$] presenta las siguientes propiedades:

$$E(k) = k \qquad \forall k \in \mathbb{R}$$

$$E(k + X) = k + E(X)$$

$$E(k \cdot X) = k \cdot E(X)$$

$$E(X \pm Y) = E(X) \pm E(Y)$$

$$E(X \cdot Y) = E(X) \cdot E(Y) \quad \Leftrightarrow \quad X,Y \text{ Independientes}$$

$$mín(X) \leq E(X) \leq máx(X)$$

2.3.1.B. Moda.

Es el valor de variable que posee la mayor probabilidad asociada. Es decir, es el valor con mayor expectativa de aparición.

2.3.1.C. Mediana.

Es el valor del Espacio Muestral que divide simétricamente los valores poblacionales de la variable.

2.3.2. *Parámetros de Variabilidad*.

2.3.2.A. Varianza.

Se conoce a la Varianza como la sumatoria de los desvíos cuadráticos de la variable (X_i) con respecto de su valor esperado [$E(X)$] ponderados por sus probabilidades asociadas [$P(X_i)$].

Es por esto que cálculo es el siguiente:

$$\text{Situación Discreta} \Rightarrow Var(X) = \sigma^2 = \sum_{i=1}^{n} \left[X_i - E(X)\right]^2 \cdot P(X_i)$$

$$\text{Situación Continua} \Rightarrow Var(X) = \sigma^2 = \int_{L_i;-\infty}^{L_s;+\infty} \left[X - E(X)\right]^2 \cdot f(X)\, dx$$

La Varianza [$Var(X)$] posee, entre otras, estas propiedades fundamentales:

$$Var(k) = 0 \qquad \forall k \in \mathbb{R}$$
$$Var(k + X) = Var(X)$$
$$Var(k \cdot X) = k^2 \cdot Var(X)$$

2.3.2.B. Desvío Esperado o Dispersión.

Como habíamos anticipado, conjuntamente con el valor esperado, éste es un parámetro fundamental para el estudio de las poblaciones como también para el de las muestras.

Representa la diferencia entre el Valor esperado y lo que en realidad esperamos. De alguna manera es un tipo de valor esperado ya que nos determina cuánto esperamos desviarnos de lo que en realidad deseamos alcanzar.

Por esto es que su cálculo es, de cierta forma, la simplificación del resultado de la $Var(X)$, ya que ésta muestra los desvíos cuadráticos ponderados por las probabilidades asociadas, y la $Disp(X)$ los desvíos con respecto del valor esperado [$E(X)$]. Se calcula de la siguiente manera:

$$Disp(X) = +\sqrt{Var(X)} = +\sqrt{\sigma^2} = \sigma$$

Al igual que el valor esperado y la varianza, la dispersión posee propiedades aplicables al cálculo. Las más comunes son las siguientes:

$$Disp(k) = 0 \qquad \forall k \in \mathbb{R}$$
$$Disp(k + X) = Disp(X)$$
$$Disp(k \cdot X) = |k| \cdot Disp(X)$$

En torno a este punto podemos hablar de una clasificación de las Variables Aleatorias con respecto de su $E(X)$. Podemos dividir a las variables en: *Homogéneas* y *Heterogéneas*.

Una VA es *Homogénea* cuando los valores de la variable está próximos a su $E(X)$. En cambio decimos que es *Heterogénea* cuando los valores de la variable están lejos de su $E(X)$.

Desde el punto de vista de la Dispersión, se puede decir que: cuando ésta se acerca a 0 (cero) la variable tiende a ser Homogénea ya que los desvíos son más pequeños, en cambio cuando la Dispersión tiende a ser un número más alto, la variable va a ser Heterogénea ya que los desvíos son mayores.

2.3.3. *Coeficiente de Variabilidad*.

Este coeficiente nos muestra cuán representativa es la $E(X)$ con respecto de los valores de la variable. Por lo general se expresa como un porcentaje. Es por esto que su cálculo es el siguiente:

$$C.V. = \frac{Disp(X)}{E(X)} \cdot 100 = \frac{\sigma}{\mu} \cdot 100 = \frac{+\sqrt{\sum_{i=1}^{n} \left[X_i - E(X)\right]^2 \cdot P(X_i)}}{\sum_{i=1}^{n} X_i \cdot P(X_i)} \cdot 100$$

En el caso de no querer expresar en valores porcentuales y verlo en valores relativos, basta con obviar la multiplicación por el 100.

Los resultados de este coeficiente pueden clasificarse en varios grupos de representación. La manera más común es la siguiente.

C.V.	*C.V. (%)*	*E(X) con respecto de los valores de variable*
0 – 0,10	0% – 10%	Muy Representativa
0,10 – 0,30	10% – 30%	Representativa
0,30 – 0,50	30% – 50%	Poco Representativa
0,50 y más	50% y más	Nada Representativa

2.4) DESIGUALDAD DE CHÉBISHEV.

Para Variables Aleatorias Discretas o Continuas dado un intervalo simétrico respecto del valor esperado y sea $E(X)=\mu$, es decir que $E(X)$ existe y es finito, y la $Disp(X)=\sigma$ y es finita:

$$P\left[\left|X-E(X)\right|\leq t\cdot\sigma\right]\geq 1-t^{-2}$$

Se define $P(X)$ de que la variable tome valores equis-distantes en, a lo sumo, *t* veces el desvío ($t\in\mathbb{R}^{+}$). Cuanto mayor es *t*, la $P(X)$ se acerca más a 1, y cuando $t=1$ la $P(X)=0$.

Desarmando el valor absoluto la relación se puede expresar de la siguiente manera:

$$P\left[E(X)-t\cdot\sigma\leq X\leq E(X)+t\cdot\sigma\right]\geq 1-\frac{1}{t^{2}}$$

En general no se conoce la distribución de la Variable Aleatoria, ni su $E(X)$, ni su $Disp(X)$. Este teorema sirva para hallar la $P(X)$ en un rango dado.

2.5) COVARIANZA.

Dadas dos Variables Aleatorias *X* e *Y*, la $Cov(X;Y)$ es una medida de la forma en que varían conjuntamente las variables y nos dice cómo se relacionan entre ellas.

Se puede decir que es el valor esperado de los desvíos de las variable *X* e *Y*, de manera conjunta.

Su cálculo se expresa de estas dos maneras equivalentes:

$$Cov(X;Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}\left[X_i-E(X)\right]\cdot\left[Y_j-E(Y)\right]\cdot P(X_i;Y_j)$$

$$Cov(X;Y)=\left[\sum_{i=1}^{n}\sum_{j=1}^{m}X_i\cdot Y_j\cdot P(X_i;Y_j)\right]-\left[E(X)\cdot E(Y)\right]$$

Si la $Cov(X;Y)$ es *positiva*, la relación entre las variables es *Directa*, lo que implica que ambas variables crecen y decrecen de manera conjunta.

Si la $Cov(X;Y)$ es *negativa*, se dice que la relación es *Indirecta*, es decir que cuando una variable crece la otra decrece y viceversa.

Cuando X e Y son independientes $Cov(X;Y)=0$, pero que la $Cov(X;Y)=0$ no implica independencia entre ambas variables.

UNIDAD N°3: DISTRIBUCIONES DISCRETAS DE PROBABILIDADES.

3.1) EXPERIMENTO O DISTRIBUCIÓN DE BERNOULLI.

La Variable Aleatoria toma los valores 0 ó 1, es decir $X \in \{0,1\}$.

Es por esto que hablamos de una prueba dicotómica, ya que solamente existen dos posibles resultados y son excluyentes uno del otro.

Si la variable *X* toma el valor 1 su probabilidad asociada va a ser la del suceso favorable *p*. Si, en cambio, toma el valor 0, la probabilidad asociada será la del no favorable definido como $1-p$, usualmente denotado como *q*.

Analíticamente:

$$\left.\begin{array}{l} X = 1 \Rightarrow P(SF) = p \\ X = 0 \Rightarrow P(\overline{SF}) = 1 - p = q \end{array}\right\} \Rightarrow X : \{(0;q),(1;p)\}$$

Es por esto que los parámetros son los siguientes:

$$E(X) = 0 \cdot (1-p) + 1 \cdot p = p$$

$$Var(X) = (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p = p \cdot (1-p) = p \cdot q$$

$$Disp(X) = +\sqrt{(0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p} = +\sqrt{p \cdot (1-p)} = +\sqrt{p \cdot q}$$

3.2) DISTRIBUCIÓN BINOMIAL.

Se dice que una Variable Aleatoria Discreta sigue una Distribución Binomial cuando se conforma de *n* repeticiones independientes del experimento de Bernoulli.

Se busca, generalmente, la probabilidad de que determinado suceso ocurra un finito número (X_n) de veces en *n* repeticiones independientes, para cada una de las cuales la probabilidad de que ocurra el suceso favorable es *p*.

Analíticamente, la notación para este tipo de distribución es la siguiente:

$$\left.\begin{array}{l} P(SF) = p \\ P(\overline{SF}) = 1 - p = q \end{array}\right\} \Rightarrow X \approx Bi(n;p)$$

La probabilidad va a estar dada por la siguiente estructura:

$$P(X = X_n) = C_n^{X_n} \cdot p^{X_n} \cdot q^{(n-X_n)}$$

Ya que este tipo de distribución discreta surge de repetir *n* veces el experimento de Bernoulli, los parámetros van a estar dados de la siguiente forma:

$$E(X) = n \cdot p$$

$$Var(X) = n \cdot p \cdot q$$

$$Disp(X) = +\sqrt{n \cdot p \cdot q}$$

$$m_o(X) = z_o \Leftrightarrow n \cdot p - q \leq z_o \leq n \cdot p \quad \forall z_o \in E^+ + \{0\}$$

3.3) DISTRIBUCIÓN HIPERGEOMÉTRICA.

Al igual que en la distribución Binomial, se basa en la existencia de un universo dicotómico, pero a diferencia de la anterior las repeticiones son dependientes unas de otras.

Es por esto que la probabilidad del suceso favorable va a estar dada por:

$$P(SF) = p = \frac{k}{N}$$

Es decir, luego de cada repetición la probabilidad del suceso favorable se va a ver alterada, a diferencia del caso de la Binomial donde se mantiene constante a lo largo del experimento.

Podemos decir que si tenemos un número N total de elementos en una población finita, de manera tal que k de estos elementos presenta una cierta característica en la modalidad A (suceso "favorable"), y $N-k$ presenta una cierta característica en la modalidad $\overline{A}$ (suceso "no favorable"); si se toma una *muestra aleatoria* de la población construida por n elementos, la probabilidad de que X_n elementos presenten la característica de A y $n-X_n$ la presenten de la modalidad $\overline{A}$ es la siguiente:

$$P(X = X_n) = \frac{C_k^{X_n} \cdot C_{N-k}^{n-X_n}}{C_N^n}$$

En este tipo de distribución los parámetros están dados de la siguiente forma:

$$E(X) = n \cdot p = n \cdot \frac{k}{N}$$

$$Var(X) = n \cdot p \cdot q \cdot (FC) = n \cdot \frac{k}{N} \cdot \frac{N-k}{N} \cdot \frac{N-n}{N-1}$$

$$Disp(X) = +\sqrt{Var(X)} = +\sqrt{n \cdot p \cdot q \cdot (FC)} = +\sqrt{n \cdot \frac{k}{N} \cdot \frac{N-k}{N} \cdot \frac{N-n}{N-1}}$$

"*FC*" representa el *Factor de Corrección* que sirve para corregir los resultados por el hecho de que las pruebas no son independientes sino que unas dependen de otras.

3.4) DISTRIBUCIÓN DE POISSON.

En este tipo de distribución, si bien los sucesos son independientes y definidos de manera discreta, las probabilidades están en términos continuos.

Se utiliza cuando un experimento se repite de forma independiente una cantidad lo suficientemente grande de veces, la probabilidad asociada al suceso favorable es lo adecuadamente pequeña; y la cantidad de veces de que se presente el suceso también es pequeña.

La probabilidad en un punto X_0 está definida de la siguiente manera:

$$P(X = X_0) = \frac{\lambda^{X_0} \cdot e^{-\lambda}}{X_0!}$$

Los parámetros están definidos de la siguiente forma:

$$E(X) = n \cdot p = \lambda$$
$$Var(X) = E(X) = n \cdot p = \lambda$$
$$Disp(X) = +\sqrt{Var(X)} = +\sqrt{E(X)} = +\sqrt{\lambda}$$

3.5) DISTRIBUCIÓN MULTINOMIAL.

El concepto y el razonamiento de los sucesos es el mismo que en la Binomial, la diferencia es que en la Binomial se trabaja sobre un universo dicotómico donde las posibilidades son que en las repeticiones se de el suceso favorable o no, en cambio en este tipo de distribución hablamos de múltiples sucesos simultáneos en las repeticiones. Además podríamos comprobar a partir de un ejemplo muy sencillamente la distribución Binomial es el caso de la Multinomial con dos posibilidades.

La forma de calcular las probabilidades es la siguiente:

$$P(X_1 = x_1; X_2 = x_2; \ldots; X_n = x_n) = \frac{n!}{x_1! \cdot x_2! \cdot \ldots \cdot x_n!} \cdot P_1^{x_1} \cdot P_2^{x_2} \cdot \ldots \cdot P_n^{x_n}$$

Los parámetros definidos para cada variable en particular tienen la misma forma que en el caso de la Binomial.

$$\mu_{X_i} = E(X_i) = n \cdot p_i$$
$$\sigma^2_{X_i} = Var(X_i) = n \cdot p_i \cdot q_i$$
$$\sigma_{X_i} = Disp(X_i) = +\sqrt{n \cdot p_i \cdot q_i}$$

Si se desea analizar la situación de manera conjunta, los parámetros se definen de la siguiente manera:

$$\mu = E(X_1; X_2; X_3; \ldots; X_k) = \sum_{i=1}^{k} n \cdot p_i = n$$
$$\sigma^2 = Var(X_1; X_2; X_3; \ldots; X_k) = \sum_{i=1}^{k} n \cdot p_i \cdot q_i = \sum_{i=1}^{k} Var(X_i)$$
$$\sigma = Disp(X_1; X_2; X_3; \ldots; X_k) = \sum_{i=1}^{k} + \sqrt{n \cdot p_i \cdot q_i} = \sum_{i=1}^{k} + \sqrt{Var(X_i)}$$

La Covarianza para dos variables cualesquiera distintas, por ejemplo X_1 y X_2, es la siguiente:

$$\lambda_{1,2} = E\{[X_1 - E(X_1)] \cdot [X_2 - E(X_2)]\} = E[(X_1 - n \cdot p_1) \cdot (X_2 - n \cdot p_2)]$$

Sintéticamente, la podemos expresar mediante la siguiente fórmula:

$$\lambda_{1,2} = -n \cdot p_1 \cdot p_j$$

UNIDAD N°4: DISTRIBUCIONES CONTINUAS DE PROBABILIDADES.

4.1) DISTRIBUCIÓN NORMAL.

4.1.1. Función de Densidad Normal.

Cuando una VA toma ilimitados valores dentro del campo de los números reales, se dice que es una VA Continua. La distribución Normal es un tipo de distribución continua, y se denota de la siguiente manera:

$$\left.\begin{array}{l}\mu = E(X)\\ \sigma = Disp(X)\end{array}\right\} \Rightarrow X \approx N(\mu;\sigma)$$

En las VA continuas las probabilidades vienen dadas por la función de densidad. En el caso de la Normal, la función de densidad es la siguiente:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$$

El dominio de esta función, como la VA se mueve dentro del campo de los números reales va a ser ese mismo conjunto. Y su ámbito el intervalo real comprendido entre 0 y el máximo admitido de la función, ya que estos son los límites entre los cuales puede tomar valor la función de densidad. Analíticamente:

$$Dom_f = \{x / x \in \mathbb{R}\}$$

$$Ambito_f = \left\{ f(x) \in \mathbb{R} \ / \ 0 < f(x) \leq \frac{1}{\sigma \cdot \sqrt{2\pi}} \right\}$$

Por otra parte la función es simétrica con respecto a la esperanza (μ). Es decir $f(x+\mu) = f(-x+\mu)$.

En el mismo punto de la esperanza (μ) es donde la función alcanza su valor máximo:

$$máx_f = \frac{1}{\sigma \cdot \sqrt{2\pi}} \qquad en\ x = \mu$$

Con esto podemos definir los intervalos de crecimiento y decrecimiento que van a estar dados por:

$$\begin{cases} f(x)\ es\ creciente\ en\ (-\infty;\mu) \\ f(x)\ es\ decreciente\ en\ (\mu;+\infty) \end{cases}$$

Esta función también posee un límite asintótico inferior en $f(x) = 0$, ya que se cumple que:

$$\lim_{x \to -\infty} f(x) = \lim_{x \to +\infty} f(x) = 0 \Rightarrow Asíntota\ Horizontal\ en\ f(x) = 0$$

Además, posee dos puntos de inflexión situados en:

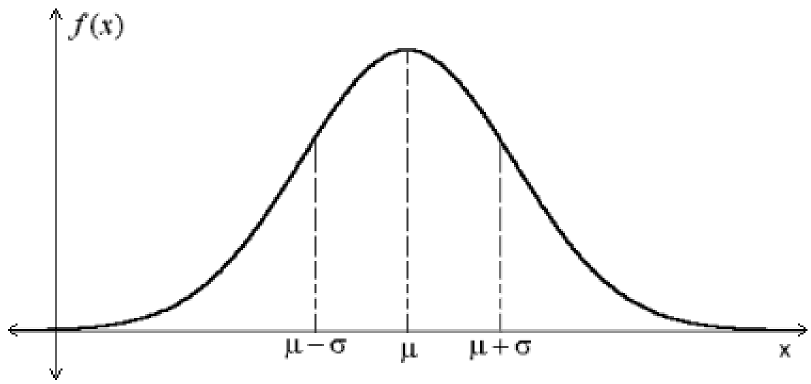
$$x_1 = \mu - \sigma$$

$$x_2 = \mu + \sigma$$

Es por esto que los intervalos de concavidad van a ser los siguientes:

$$\begin{cases} f(x) \text{ es concava o con concavidad hacia abajo en } (\mu - \sigma; \mu + \sigma) \\ f(x) \text{ es convexa o con concavidad hacia arriba en } (-\infty; \mu - \sigma) \cup (\mu + \sigma; +\infty) \end{cases}$$

Conjugando todo esto, el gráfico típico de una función de densidad simétrica (la característica de una Distribución de tipo Normal) es el siguiente:



Este gráfico es comúnmente conocido como la *Campana de Gauss*.

4.1.2. Probabilidades Acumuladas.

Para calcular la Probabilidad Acumulada a un punto *a*, basándonos en la función de densidad normal dada, esta Probabilidad va a ser igual al área generada entre la curva y el eje de las abscisas con límite superior en ese punto *a*. Analíticamente:

$$P(X \leq a) = P(X < a) = \int_{-\infty}^{a} f(x)\, dx = \int_{-\infty}^{a} \left[\frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \right] dx$$

Como la gráfica es simétrica y posee su máximo en su valor esperado (μ), la probabilidad acumulada hasta ese valor es igual a 0,50 como consecuencia de la Ley de Cierre.

$$P(X < +\infty) = \int_{-\infty}^{+\infty} f(x)\, dx = 1 \Rightarrow \begin{cases} P(X \geq \mu) = P(X > \mu) = \int_{\mu}^{+\infty} f(x)\, dx = 0{,}50 \\ P(X \leq \mu) = P(X < \mu) = \int_{-\infty}^{\mu} f(x)\, dx = 0{,}50 \end{cases}$$

Es decir que el valor esperado divide la gráfica en dos regiones iguales con probabilidades acumuladas de 0,50.

Además del gráfico de la Función de Densidad, existe otro que representa directamente las Probabilidades Acumuladas. Se la puede definir como:

$$f_a(x) = \int_{-\infty}^{x} \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du$$

Esta función de Probabilidades Acumuladas tiene como dominio el campo de los números reales y como ámbito el intervalo abierto entre 0 y 1.

$$Dom_{f_a} = \{x / x \in \mathbb{R}\}$$

$$Ambito_{f_a} = \{f(x) \in \mathbb{R} \ / \ 0 < f_a(x) < 1\}$$

No posee valores máximos ni mínimos.

La función es estrictamente creciente en todo su dominio y no decrece, esto es porque en su avance va acumulando probabilidades y éstas son no negativas.

$$\begin{cases} f_a(x)\ es\ creciente\ en\ Dom_{f_a} \Rightarrow Crece\ en\ \mathbb{R} \\ f_a(x)\ no\ decrece \end{cases}$$

Posee dos asíntotas horizontales, una en 0 y otra en 1. Esto se da ya que:

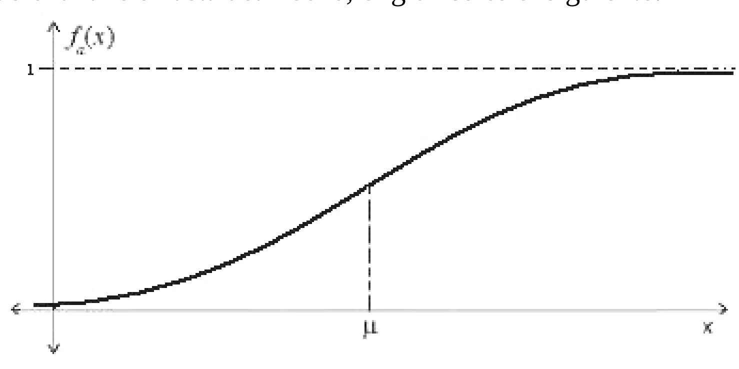
$$\begin{cases} \lim\limits_{x \to -\infty} f_a(x) = 0 \Rightarrow \text{Asíntota Horizontal en } f(x) = 0 \\ \lim\limits_{x \to +\infty} f_a(x) = 1 \Rightarrow \text{Asíntota Horizontal en } f(x) = 1 \end{cases}$$

Tiene un solo punto de inflexión y está en la esperanza, es decir en $x = \mu$. En este valor, el punto que se define es el $(\mu; f_a(\mu))$, y como hasta μ se acumula el 50% de las probabilidades, el punto de inflexión es el $(\mu; 0{,}50)$.

Como consecuencia de este punto de inflexión pueden definirse los intervalos de concavidad y convexidad, que son los siguientes:

$$\begin{cases} f(x)\ es\ convexa\ o\ con\ concavidad\ hacia\ arriba\ en\ (-\infty; \mu) \\ f(x)\ es\ concava\ o\ con\ concavidad\ hacia\ abajo\ en\ (\mu; +\infty) \end{cases}$$

Conjugando el análisis hasta acá hecho, el gráfico es el siguiente:



UNIDAD N°5: ESTADÍSTICA DESCRIPTIVA.

5.1) DEFINICIONES BÁSICAS.

5.1.1. Dato.

Es el resultado de una evaluación cuantitativa o cualitativa de un suceso. Por ejemplo el de compra de una PC –medición cuantitativa– o el grado de satisfacción de un usuario respecto de un servicio médico asistencial –medición cualitativa–.

5.1.2. Información.

Representa el conjunto de datos interrelacionados. La interrelación puede aplicarse a datos que correspondan a la misma variable presentes en distintos elementos, o bien a datos que correspondan a distintas variables presentes en un mismo elemento. En ambos casos, el fin de la información es alimentar el proceso de la toma de decisiones.

De acuerdo a John Burch y Gary Grudnitski en *Diseño de Sistemas de Información* se tiene que: "*La información la componen datos que se han colocado en un contexto significativo y útil y se ha comunicado a un receptor, quien la utiliza para tomar decisiones. (…) Especialmente en los negocios, la información debe dar señales oportunas de aviso y anticipar el futuro*".

Por ejemplo, la gerencia de compras de una empresa analiza los precios de un determinado insumo entre los diversos proveedores existentes en el mercado –lo que implica varios valores para una misma variable–; o bien un analista financiero analiza el monto de las inversiones de una empresa en las áreas: productiva, capacitación y perfeccionamiento del personal, mantenimiento y modernización de la infraestructura, y administración comercial –lo cual lleva a obtener varios valores que surgen de distintas variables–.

5.1.3. Población.

Es el conjunto de todos los valores que toma una característica predeterminada. Por ejemplo: lo valores del salario básico correspondiente a los empleados de la administración pública central.

5.1.4. Universo.

Es el conjunto del total de elementos que serán los portadores de la información sobre la característica en estudio. En nuestro ejemplo anterior, sería el conjunto de todos los empleados de la administración pública central.

5.1.5. Muestra.

Es un subconjunto del universo en estudio y, consecuentemente, subconjunto de la población respectiva. En nuestro ejemplo podríamos pensar en todos los empleados de la administración pública central que fueron seleccionados por alguna técnica muestral específica.

5.2) ANÁLISIS DE DATOS Y GENERACIÓN DE RESULTADOS.

5.2.1. Medidas de Tendencia Central.

5.2.1.A. Media.

El concepto es el mismo que el de la esperanza matemática, la diferencia radica en que el valor esperado se calcula a priori –antes de experimentar– ya que muestra cuál es el valor esperable de aparecer post-experimentación, en cambio la media se calcula a posteriori, es decir sobre una muestra.

Es por esto que su cálculo es el siguiente:

$$M(X) = \overline{X} = \sum_{i=1}^{n} X_i \cdot P(X_i) = \frac{\sum_{i=1}^{n} X_i \cdot F}{N} = \sum_{i=1}^{n} X_i \cdot f_r$$

Como mencionamos que mantiene el mismo concepto que la esperanza, las propiedades de su cálculo son iguales.

5.2.1.B. Mediana.

La mediana es un valor que separa la serie ordenada en dos grupos que contiene la misma cantidad de datos, de tal manera que en el primer grupo se encuentran todos aquellos valores menores o iguales a ella y en el segundo grupo el resto de los valores.

Se calcula sobre el intervalo en donde se logra acumular el 50% de las frecuencias, y su cálculo es el siguiente:

$$M_e(X) = L_{\inf} + \frac{\frac{N}{2} - f_{aia}}{f_i} \cdot w_i$$

5.2.1.C. Modo/a.

Es el valor de variable que mayor cantidad de observaciones presenta, es decir es el valor de mayor frecuencia relativa o absoluta asociada. Es por esto que en general, cuando la serie de datos está conformada por pocos valores, el modo no es para nada relevante.

En el caso que se tenga valores agrupados con amplitudes de intervalos constantes o no, la modalidad de cálculo viene dada por la fórmula:

$$M_o(X) = L_{\inf} + \frac{d_1}{d_1 + d_2} \cdot w_i = \left[\frac{\left(\frac{f_i}{w_i} - \frac{f_{ia}}{w_{ia}} \right)}{\left(\frac{f_i}{w_i} - \frac{f_{ia}}{w_{ia}} \right) + \left(\frac{f_i}{w_i} - \frac{f_{ip}}{w_{ip}} \right)} \right] \cdot w_i$$

El valor modal se calcula sobre el intervalo que posee la mayor altura en la muestra. Para conocer las alturas se recurre a la siguiente fórmula: $h_i = \frac{f_i}{w_i}$.

Entonces la moda se calcula sobre el intervalo que arroje mayor resultado.

5.2.2. Medidas de Tendencia No Central.

La mediana permite dividir la muestra en dos grupos iguales, pero se podría pensar en dividir la serie en cuatro, diez o cien grupos, que contengan cada uno de ellos la misma cantidad de elementos. De esta idea surgen los Percentiles, Deciles y Cuartiles.

5.2.2.A. Percentiles.

Dividen a la serie ordenada en cien grupos de igual cantidad de elementos. Su cálculo es el siguiente:

$$P_k = L_{\inf} + \frac{k \cdot \frac{N}{100} - F_{aia}}{F_i} \cdot w_i$$

Si lo relacionamos con la mediana, vamos a decir que el Percentil cincuenta (P_{50}) va a ser igual a la mediana.

5.2.2.B. Deciles.

Dividen a la serie ordenada en diez grupos de igual cantidad de elementos. Su cálculo es el siguiente:

$$D_k = L_{\inf} + \frac{k \cdot \frac{N}{10} - F_{aia}}{F_i} \cdot w_i$$

Si lo relacionamos con la mediana, vamos a decir que el Decil cinco (D_5) va a ser igual a la mediana.

5.2.2.C. Cuartiles.

Dividen a la serie ordenada en cuatro grupos de igual cantidad de elementos. Su cálculo es el siguiente:

$$Q_k = L_{\inf} + \frac{k \cdot \frac{N}{4} - F_{aia}}{F_i} \cdot w_i$$

Si lo relacionamos con la mediana, vamos a decir que el Cuartil dos (Q_2) va a ser igual a la mediana.

5.2.3. Medidas de Dispersión.

Estas medidas nos proporcionan información adicional a los efectos de juzgar sobre la confiabilidad de nuestras medidas de tendencia central. Es decir para los casos en que los datos se encuentran altamente dispersos, estas medidas nos informan que la media no es representativa de la muestra, y para los casos en que los datos estén altamente concentrados nos informarán que la media es altamente representativa de la serie ordenada.

5.2.3.A. Rango.

Es la distancia entre el mayor y el menor de los valores observados. Nos permite analizar la extensión de las variaciones observadas. Es muy sensible a la presentación de valores atípicos.

5.2.3.B. Varianza.

Es una medida de variabilidad basada en las desviaciones respecto de la media de la variable. Nos da una distancia promedio –en unidades cuadráticas– entre cualquier observación de datos y la media del mismo.

$$Var(X) = s^2 = \sum_{i=1}^{n} \left[X_i - \overline{X}\right]^2 \cdot f_r(X_i) = \frac{\sum_{i=1}^{n} \left[X_i - \overline{X}\right]^2 \cdot F(X_i)}{N}$$

5.2.3.C. Coeficiente de Dispersión o Desviación Estándar.

Es la simplificación del valor –presente en valores cuadráticos– de la Varianza. Muestra cuánto los valores de la variable se desvían, en promedio, respecto de la media.

Su cálculo es el siguiente:

$$Disp(X) = s = +\sqrt{Var(X)} = +\sqrt{s^2}$$

5.3) ANÁLISIS DE SIMETRÍA.

5.3.1. Comparación de la Media con la Mediana y la Moda.

Surge de hacer una comparación entre los tres valores de tendencia central, con lo que surge:

$$M(X) < M_e(X) < M_o(X) \Rightarrow \textit{Distribución Asimétrica Izquierda}$$
$$M(X) = M_e(X) = M_o(X) \Rightarrow \textit{Distribución Simétrica}$$
$$M(X) > M_e(X) > M_o(X) \Rightarrow \textit{Distribución Asimétrica Derecha}$$

Se entiende que basta con comparar la $M(X)$ con la $M_e(X)$, ya que una distribución puede tener más de una $M_o(X)$ y podría distorsionar el resultado de esta comparación.

5.3.2. *Comparación relativa de Media y Mediana con respecto a la Desviación.*

Este tipo de relación para conocerla Asimetría permite sacar el sesgo que presenta la serie –en caso de ser asimétrica–. Compara relativamente la $M(X)$ con la $M_e(X)$ de la siguiente manera:

$$A_s = \frac{M(X) - M_e(X)}{s(x)}$$

El análisis de los resultados es el siguiente:

$$A_s \Rightarrow \begin{cases} A_s > 0 \Rightarrow \textit{Distribución Asimétrica Derecha} \\ A_s < 0 \Rightarrow \textit{Distribución Asimétrica Izquierda} \end{cases}$$

5.3.3. *Coeficiente de Asimetría.*

Viene dado por el cociente entre el momento centrado tercero de la variable y el cubo del coeficiente de dispersión.

$$CA_s = \frac{\dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{N}}{\left[\sqrt{\dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{N}}\right]^3} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3 \cdot f_r}{\left[\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \cdot f_r}\right]^3} = \frac{E(X - \overline{X})^3}{Disp(X)^3}$$

El análisis de los resultados es el siguiente:

$$CA_s \Rightarrow \begin{cases} CA_s > 0 \Rightarrow \textit{Distribución Asimétrica Derecha} \\ CA_s = 0 \Rightarrow \textit{Distribución Simétrica} \\ CA_s < 0 \Rightarrow \textit{Distribución Asimétrica Izquierda} \end{cases}$$

5.3.4. *Coeficiente de Curtosis.*

Este coeficiente nos permite ver qué tan campanular es la distribución de la variable. Su cálculo surge del cociente entre el momento centrado cuarto de la variable y el coeficiente de dispersión elevado a la cuarta.

$$K = \frac{\dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{N}}{\left[\sqrt{\dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{N}}\right]^4} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4 \cdot f_r}{\left[\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \cdot f_r}\right]^4} = \frac{E(X - \overline{X})^4}{Disp(X)^4}$$

Sus resultados se interpretan de la siguiente manera:

$$K \Rightarrow \begin{cases} K > 3 \Rightarrow \textit{Distribución Leptocúrtica} \\ K = 3 \Rightarrow \textit{Distribución Mesocúrtica} \\ K < 3 \Rightarrow \textit{Distribución Platicúrtica} \end{cases}$$

5.4) ANÁLISIS DE CAJA Y BIGOTES.

Para comenzar el análisis se debe calcular el Rango Intercuartílico, que surge de la diferencia entre el tercer y el primer Cuartil:

$$\textit{Rango Intercuartílico} = RIC = q_3 - q_1$$

Una vez conocido esto podemos ubicar la "caja" sobre el intervalo delimitado por los límites inferior y superior de la distribución. Esto nos va a mostrar qué tan simétrica o asimétrica es la distribución, ya que cuanto más al centro la caja esté ubicada más simétrica va a ser, en cambio cuando se ubique más hacia los extremos va a mostrar mayor asimetría.

La segunda parte, la que corresponde a los "bigotes", surge del cálculo y ubicación en el diagrama de cuatro barreras o bisagras que van a determinar la tipicidad o atipicidad de los valores de la muestra. Estas barreras se calculan así:

$$\textit{Barreras} \Rightarrow \begin{cases} \textit{Barrera Exterior Inferior} = BEI = q_1 - 3 \cdot RIC \\ \textit{Barrera Interior Inferior} = BII = q_1 - 1{,}5 \cdot RIC \\ \textit{Barrera Interior Superior} = BIS = q_3 + 1{,}5 \cdot RIC \\ \textit{Barrera Exterior Superior} = BES = q_3 + 3 \cdot RIC \end{cases}$$

Se dice que los valores de la distribución que quedan comprendidos entre la "caja" y las Barreras Interior Inferior y Superior son "*Valores Típicos*". Los que queden entre las Barreras Interiores y Exteriores, tanto Inferiores como Superiores, son "*Valores Atípicos*". Y los valores que menores a la Barrera Exterior Inferior superan la Barrera Exterior Superior, son "*Valores Muy Atípicos*".