Tema 1 definiciones

1.1. Elementos básicos de la estadística

La estadística está ligada con los métodos científicos en la toma, organización, recopilación, presentación y análisis de datos, tanto para la deducción de conclusiones como para tomar decisiones razonables de acuerdo con tales análisis.

Población

En general, en estadística se denomina población a un conjunto de elementos (personas, objetos, etc.), que contienen una o más características observables que se pueden medir.

A cada elemento de la población se denomina unidad elemental o unidad estadística. El resultado de medir una característica observable en la unidad elemental se denomina dato estadístico.

Muestra

Se denomina muestra a una parte de la población seleccionada de acuerdo con un plan, con el fin de obtener información referente a la población de la que proviene. La muestra debe ser seleccionada de tal manera de que sea representativa y, por lo tanto, el método de selección de muestras debe garantizar la obtención de datos fidedignos.

Parámetro

Es una medida descriptiva que resume una característica de la población, por ejemplo, la media, la mediana, la varianza.

Estadístico

Llamado también estadígrafo, es una medida descriptiva que resume una característica de una muestra y es calculada a partir de datos observados de una muestra aleatoria.

Variables

Se llama variable estadística a una característica definida en la población por la investigación estadística y que puede tomar dos o más valores.

Las variables pueden ser cualitativas y cuantitativas. La primera describe cualidades mientras que la segunda describe cantidades. A su vez, la variable cuantitativa puede ser continua o discreta; es continua cuando puede tener valores en un intervalo, y es discreta cuando puede tomar exclusivamente valores exactos (enteros).

Los datos que vienen definidos por una variable discreta o continua se llaman datos discretos o datos continuos, respectivamente. En general, las medidas dan origen a datos continuos, mientras que las enumeraciones o conteos originan datos discretos.

1.2. Niveles o escalas de medición de las variables

Se denomina escala de medición a un instrumento de medida, con el que se asigna valores a las unidades estadísticas para una variable definida.

Las escalas de medición son de los siguientes tipos:
• **Variables nominales.** No tienen un orden o jerarquía determinados. **Ejemplos:** color, nacionalidad, religión, estado civil.

• **Variables ordinales.** Sí tienen un orden o jerarquía establecidos. **Ejemplos:** clase social (alta, media, baja), estado de conservación (excelente, muy bueno, bueno, regular, malo).

• **Variables de nivel de intervalo.** El cero es arbitrario y no representa la ausencia de la variable.
Ejemplo: temperatura: 0 grados centígrados (no representa la ausencia de variable, es decir, no significa que no haya temperatura).

• **Variables de nivel de razón o proporción.** El "0" sí representa la ausencia de variable.
Ejemplos: dinero (0 dólares representa que no hay dinero); n.° de personas (0 personas representa la ausencia de personas).

1.3. Importancia del muestreo

En estadística, una muestra estadística (llamada también muestra aleatoria o simplemente muestra) es un subconjunto de casos o individuos de una población estadística.

Las muestras se obtienen con la intención de inferir propiedades de la totalidad de la población, para lo cual deben ser representativas de la misma. Para cumplir esta característica la inclusión de sujetos en la muestra debe seguir una técnica de muestreo. En tales casos, puede obtenerse una información similar a la de un estudio exhaustivo con mayor rapidez y menor coste abajo.

Por otra parte, en ocasiones, el muestreo puede ser más exacto que el estudio de toda la población porque el manejo de un menor número de datos provoca también menos errores en su manipulación. En cualquier caso, el conjunto de individuos de la muestra son los sujetos realmente estudiados.

El muestreo es importante porque:

- Por lo general no se pueden estudiar a las poblaciones en su totalidad, entonces estaremos obligados a hacer el muestreo.

- Es más rápido y económico para conocer los parámetros (características) de interés de la población.
- Existe metodología clara y confiable para el muestreo (y tamaño de muestra).

Tema 2 Distribución de frecuencias y medidas

.1. Fuentes de información

En el proceso de cualquier investigación, la recolección de datos constituye el paso fundamental para asegurarse la obtención de resultados idóneos, ya sea dentro de una muestra o de una población.

La recolección de la información tiene que hacerse de acuerdo con una planificación y esquematización de la investigación, caso contrario vamos a tener datos que no son relevantes, repetitivos y excesivos, necesitamos organizar la información de tal manera que pueda contribuir favorablemente al logro de los objetivos propuestos.

Con el panorama claro de las necesidades de información del proyecto, se deben realizar las siguientes actividades:

- Instrumentos de medición o técnicas de recolección de información
- La aplicación de los instrumentos de medición
- La sistematización, codificación y estructura de la información

A continuación vamos a ver los diferentes instrumentos y técnicas que tenemos para recolectar información.

¿Qué entendemos por técnicas de recolección de datos?

Podemos decir que son las herramientas con las que se cuenta para la recolección de datos. Son los procedimientos especiales utilizados para obtener y evaluar las evidencias necesarias, suficientes y competentes que permiten formar un juicio profesional y objetivo; en resumen, podemos decir que son **cualquier recurso que recopile información referente a nuestro proyecto**.
Tenemos diferentes clases de formas para recolectar información: verbales, oculares, documentales, físicas, escritas y de auditoría.

a) Verbales. Consisten en obtener la información de manera oral mediante averiguaciones o indagaciones. Pueden ser entrevistas, encuestas y cuestionarios.

b) Oculares. Obtienen la información verificando visualmente en forma directa y paralela cómo los responsables de cada proceso desarrollan o documentan los procesos o variables evaluadas. Pueden ser: observación, comparación o confrontación, revisión selectiva y rastreo.

c) Documentales. Obtienen información escrita para soportar afirmaciones, análisis o estudios previos. Pueden ser: comprobación y revisión analítica.

d) Físicas. Reconocimiento real sobre hechos o situaciones dadas en tiempo y espacio determinados y se emplean como técnica de la inspección.

e) Escritas. Reflejan toda la información que se considera importante para sustentar los hallazgos del trabajo realizado.

f) De auditoría. Conducen a obtener información sobre el desarrollo de destrezas y habilidades dentro de un proceso específico.

Ahora veamos algunos de los instrumentos que nos ayudan a obtener la información.

- La observación
- La encuesta
- La entrevista

La observación

Es el registro visual de una situación real, estableciendo los acontecimientos de acuerdo con algún esquema ya planificado.

La entrevista

Es la comunicación interpersonal establecida entre el investigador y el sujeto de estudio, a fin de obtener respuestas verbales a las interrogantes planteadas sobre el problema propuesto.

La encuesta

Es una búsqueda sistemática de información en la que el investigador pregunta a los investigados sobre los datos que desea obtener y posteriormente reunir estos datos.

Dada la utilidad, complejidad y clasificación que tiene la encuesta, nos vamos a centrar en el estudio de la misma, además hay que mencionar que es una de las más utilizadas para la recolección de información.

Algunas de las características de una encuesta son:

- Es una observación no directa de los hechos, sino por medio de lo que manifiestan los interesados.
- Es un método preparado para la investigación.
- Permite una aplicación masiva que mediante un sistema de muestreo pueda extenderse universalmente.

La encuesta se puede clasificar en:

a) Exploratoria. Se usa cuando la información previa del fenómeno a estudiar es escasa o poco fiable o es la primera toma de contacto con un fenómeno no muy conocido.

Utilidad: - Desarrolla hipótesis de trabajo.
- Verificación factible de la investigación.

Estrategia: - Consulta a expertos o grupos de discusión.
- Revisión y análisis de datos disponibles en otras fuentes.

b) Descriptiva. Nos ayuda a definir la realidad, examinar un fenómeno para caracterizarlo y/o para diferenciarlo de otros. Es el paso previo en cualquier investigación mediante la encuesta (provoca los porqué de la investigación explicativa).

Etapas

- Definición teórica del fenómeno a estudiar y selección/definición de las variables
- Definir la población
- Seleccionar muestras representativas

c) Explicativa. Determina las relaciones causa y efecto entre los fenómenos. Es imprescindible el control de las posibles explicaciones alternativas. Hay que considerar todas las variables del fenómeno.

Tipos de variables

- Variables independientes: causa de la explicación
- Variables dependientes: efecto producido por las anteriores.
- Variables extrañas: ajenas al objeto de investigación pero pueden afectar las variables explicativas.
- Variables controladas: bajo el control del investigador.
- Variables no controladas: aleatorias/perturbadoras.

d) Predictiva. Predice el funcionamiento de un fenómeno. Es necesario conocer la explicación de los fenómenos antes de tratar de establecer una predicción de estos.

Diseño de la encuesta

El diseño del cuestionario constituye el elemento principal de la encuesta, por lo tanto, su estructura es muy importante. Dado esto, tenemos que formular preguntas directas y concretas, que abarquen la necesidad de información que el proyecto requiere, de tal manera que en el momento de procesar toda esta información, esta no sea insuficiente ni inconsistente.

Las preguntas que se van a formular se pueden clasificar, entre otras, de la siguiente manera:

a) Pregunta cerrada. Se proporciona una serie de opciones donde se escoge una como respuesta, tiene la ventaja de ser fácil de procesar.

b) Pregunta abierta. No se proporciona opciones, se puede responder con libertad, tiene la ventaja de una mayor riqueza de respuestas pero es difícil de procesar.

c) Pregunta de profundización. Se utiliza para obtener una respuesta más amplia y completa a una pregunta abierta.

d) Pregunta parcialmente estructurada. Puede tener dos o más opciones o ser de tipo sí/no.
e) Pregunta de control. Es una pregunta que nos indica si el encuestado no está mintiendo.

La estadística está ligada con los métodos científicos en la toma, organización, recopilación, presentación y análisis de datos, tanto para la deducción de conclusiones como para tomar decisiones razonables de acuerdo con tales análisis.

Características del cuestionario

•Interesante.
• Sencillo de entender.
• Preciso y claro en las preguntas.
• Ordenado.
• Debe tener un vocabulario adecuado.
• Debe tener espacio suficiente para respuestas.

Una vez que ya tengamos listos los apartados anteriores, procedemos a realizar una prueba piloto o experimental, con el fin de identificar puntos críticos en los que exista controversia en las respuestas, debido a una mala formulación de alguna(s) pregunta(s), esta(s) se reformula(n) o en ciertos casos se elimina(n).

2.2. Distribuciones de frecuencia

Toda actividad que hagamos con un propósito investigativo, en el cual el resultado sean varias mediciones, tiene más que simples números o valores en una hoja. Este conjunto de datos puede ser organizado y tabulado. Se pueden realizar, entre otros procesos, gráficos que nos ayudan a captar tendencias y a establecer modelos de probabilidades. Lo que nos dice esto es que la organización de los datos es muy importante para los procesos y análisis estadísticos.

Métodos de organización de datos

Existen muchas herramientas para describir y resumir un gran conjunto de datos. Una de las más simples pero no menos importante es la ordenada, es decir ordenar los datos de forma ascendente o descendente. Para ciertos propósitos, este orden de los datos no es suficiente, y es cuando necesitamos

otras herramientas o métodos para organizarlos, de aquí tenemos los siguientes:

a) Distribuciones de frecuencias

Una distribución de frecuencias se puede tomar como una tabla donde los datos están categorizados por filas y su ocurrencia o frecuencia en columnas, la finalidad que tiene esta distribución es hacer más fácil la obtención de información de los datos.

El número de veces que aparece un valor o la frecuencia con que aparece un valor se llama **frecuencia absoluta** (fi), y la suma de estas frecuencias absolutas nos tiene que dar como resultado el número de datos.

Al ordenar nuestros datos tomando en cuenta estas consideraciones, podemos decir que hemos realizado una **distribución de frecuencias o tabla de frecuencias**.

En la tabla n.° 01 se muestra un ejemplo de distribución de frecuencias del peso de 100 estudiantes de una universidad de Perú.

Tabla n.° 01

Peso (lb)	Número de estudiantes (frecuencia absoluta)
140-149	8
150-159	15
160-169	46
170-179	22
180-189	9
	100

De acuerdo con esta tabla se pueden visualizar varias características que no podríamos ver si los datos no estuvieran organizados y tabulados. La primera clase comprende los pesos de los estudiantes que están entre 140 y 149 lb. En esta clase hay ocho estudiantes, es decir, la frecuencia de esta clase es 8.

Un dato importante que hay que mencionar es que cada clase tiene un **límite inferior** y un **límite superior** (por ejemplo la tercera clase 160-169 tiene como límite superior a 169 y como límite inferior a 160). Podemos ver que estos valores son exactos y más adelante notaremos su importancia, además es esencial que los valores no se solapen de una clase a otra, esto puede causar mucha confusión.

Existen también los denominados **límites reales** o **verdaderos de clase** que se obtienen sumando al límite superior de una clase 0,5 y restando al límite inferior del misma clase 0,5. Cuando los límites de la clase se expresan con un decimal, la regla nos indica que se debe sumar 0,05 al límite superior y restar el mismo valor al límite inferior. Por ejemplo los límites reales de la cuarta clase de la tabla n.° 01 serían: 170-0,5 = 169,5 y 179+0,5 = 179,5.

El **número de clases** (o número de intervalos) de una tabla de frecuencias de manera general debe tener entre 5 y 20 clases. Si tuviera pocas clases no se tendría ningún detalle sobre los datos, de la misma manera si hubiera muchas clases se crearía confusión.

El **intervalo de clase** son las divisiones o ancho de los valores que se encuentran dentro de una clase. Para determinar este valor se resta el límite superior menos el límite inferior. Lo mejor es hacer iguales los intervalos de clase de una distribución de frecuencias, de esta manera podemos facilitar la interpretación estadística.

El intervalo de clase se determina con la siguiente expresión:

$$Ancho\ de\ intervalo = \frac{Valor\ máximo - Valor\ mínimo}{Número\ de\ clase\ deseado}$$

Otro elemento necesario para una tabla de distribución de frecuencias es establecer un **punto medio o marca de clase**, que es el punto medio del intervalo de clase. Este se determina sumando el límite inferior con el límite superior y se divide por dos, por ejemplo, para la quinta clase de la tabla n.° 01, la marca de clase o punto medio sería:

$$X_{MC} = \frac{180 + 189}{2} = 184{,}5$$

b) Distribuciones de frecuencias acumuladas

Se conoce como frecuencia acumulada (Fi) a la suma de las frecuencias menores que el límite real superior de la clase de un intervalo de clase, también podemos decir que es la suma de todas las fi hasta el intervalo requerido. Por

ejemplo, la frecuencia acumulada hasta el intervalo de clase **170-179** de la tabla n.° 01 es: **8+15+46+22 = 91**.

Una tabla que represente estas frecuencias acumuladas se llama distribución de frecuencias acumuladas, tabla de frecuencias acumuladas o brevemente distribución acumulada.

Es importante mencionar que el proceso de acumulación puede basarse en “o más” o “menor que”, todo dependerá de lo requerido.

c) Distribuciones de frecuencias relativas

La frecuencia relativa (fr) se determina dividiendo cada frecuencia absoluta para el número de datos. Por ejemplo, en la tabla n.° 01, la frecuencia absoluta de la tercera clase es 46, este valor se lo divide para el número de datos que en este caso es 100, dándonos como resultado 0.46 que es la frecuencia relativa.

Es importante mencionar que la suma de estas frecuencias siempre dará 1, o también expresado en porcentaje la suma será del 100 %.

Si las frecuencias en la anterior tabla de frecuencias se sustituyen por las correspondientes frecuencias relativas, la tabla resultante se llama distribución de frecuencias relativas, distribución porcentual o tabla de frecuencias relativas.

d) Distribuciones de frecuencias relativas acumuladas

Una distribución de frecuencias relativas acumuladas (Hi) es la división de cada una de las frecuencias acumuladas para el total de datos. Por ejemplo, en la tabla n.° 01 la frecuencia acumulada de la cuarta clase es 91, esto dividido para el número de datos que es 100 nos da 0.0091, dato que nos indica frecuencia relativa acumulada.

El proceso de acumulación igual que en las distribuciones de frecuencia también se basa en un principio “o más” o “menor que”.

Ejemplo n.° 01

La siguiente tabla representa la distribución de frecuencias para los gastos semanales de 80 trabajadoras de una compañía de fabricación de pantalones en la ciudad de Ambato.

Gastos	Número de trabajadoras (frecuencia)
100-199	5
200-299	7
300-399	28
400-499	17
500-599	18
600-699	5

Elaborar una tabla de frecuencias acumuladas, frecuencias relativas y relativas acumuladas. Adicionalmente, incluir el valor de la marca de clase.

Desarrollo

Utilizando los conceptos revisados, calculamos lo solicitado (se puede hacer manualmente o utilizando una hoja de cálculo, en la cual, hay que tomar en cuenta la codificación de las fórmulas en cada celda).

Gastos	Marca de clase	Frecuencia absoluta	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
100-199	149,5	5	5	0,0625	0,0625
200-299	249,5	7	12	0,0875	0,1500
300-399	349,5	28	40	0,3500	0,5000
400-499	449,5	17	57	0,2125	0,7125
500-599	549,5	18	75	0,2250	0,9375
600-699	649,5	5	**80**	0,0625	**1,0000**
	Total	**80**			

Ejemplo n.° 02

La siguiente tabla representa la distribución de frecuencias para los valores pagados por horas extras de 70 trabajadores del área de Limpieza de la Universidad Internacional.

Salario	Número de trabajadores (frecuencia)
50-59,99	8
60-69,99	10
70-79,99	16
80-89,99	15
90-99,99	10
100-119,99	8
120-179,99	3

Elaborar una tabla de frecuencias acumuladas con criterio “menor que” y “o más”.

Desarrollo

Utilizando los contenidos estudiados, se sabe que se debe tomar en cuenta el límite inferior de cada intervalo de clase.

1. Criterio “menor que”

Salarios	Frecuencia acumulada	% frecuencia acumulada
Menos que 50	0	0,0
Menos que 60	8	11,43
Menos que 70	18	25,71
Menos que 80	34	48,57
Menos que 90	49	70,00
Menos que 100	59	84,29
Menos que 120	67	95,71
Menos que 179,99	70	100,00

2. Criterio “o más”

Salarios	Frecuencia acumulada	% frecuencia acumulada
50 o más	70	100,00
60 o más	62	88,57
70 o más	52	74.29
80 o más	36	51,43
90 o más	21	30,00
100 o más	11	15,71
120 o más	3	4,29
179,99 o más	0	0.0

e) Tablas de contingencias

La tabla de contingencia es una tabla que nos da el número de las observaciones en diferentes variables; es decir, tenemos varias variables con varias categorías a la vez. Con esto podemos determinar si dos características están relacionadas y de qué manera lo están.

La tabla n.° 02 muestra una tabla de contingencia en donde se evalúan dos variables: sexo y voto de 92 electores seleccionados de manera aleatoria.

Tabla n.° 02

	Candidato 1	Candidato 2	General
Hombres	25	20	45
Mujeres	31	16	47
General	56	36	92

La columna y fila general representan los totales por columna y por fila respectivamente.

2.3. Gráficos estadísticos

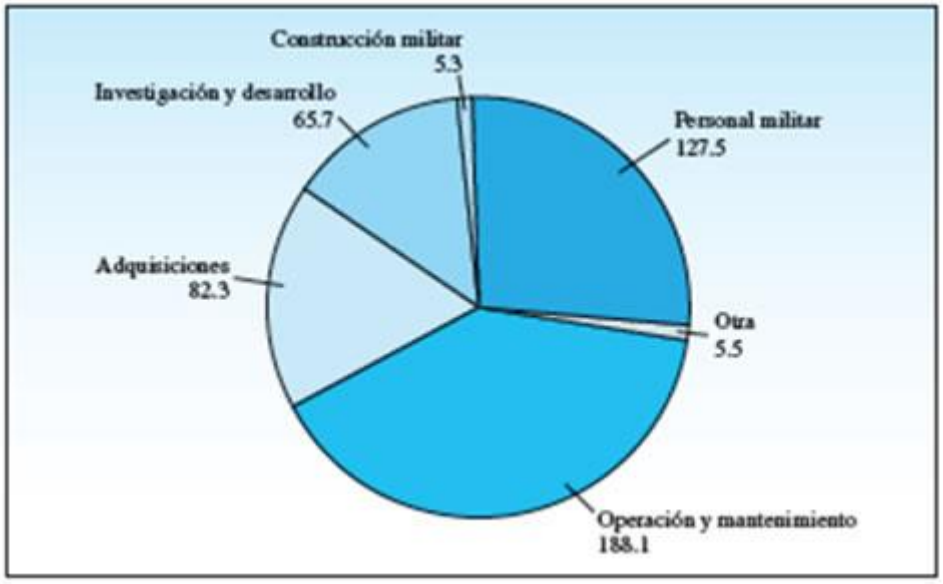
Las visualizaciones gráficas son una manera muy práctica de describir un conjunto de datos mediante la cual se puede adquirir enseguida una comprensión suficiente de los mismos. Entre estos podemos mencionar: gráfico circular, gráfico de barras, histograma, polígono de frecuencia y ojivas.

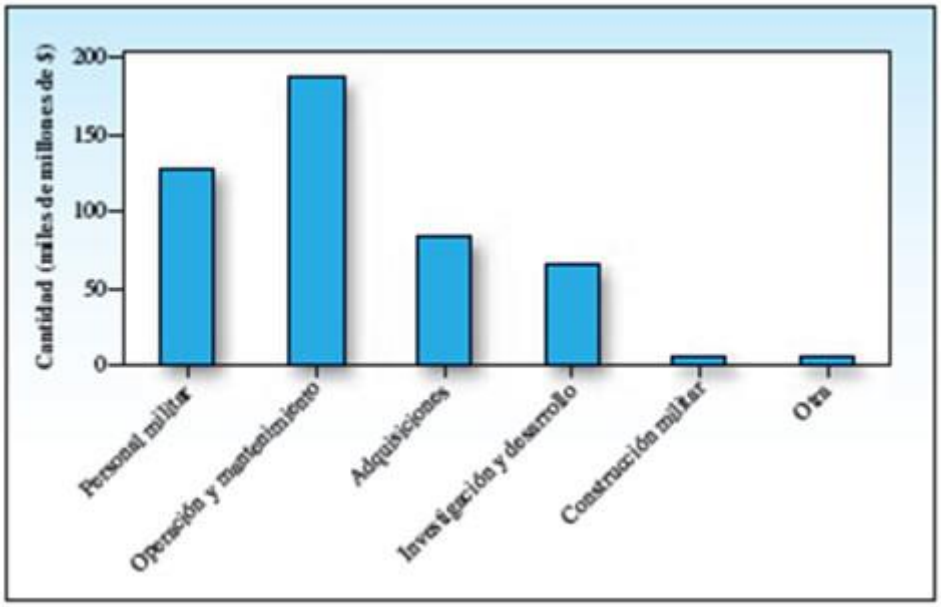
Gráfico circular

Presenta los datos en forma de círculo o tarta y de ahí que se lo llama también gráfico de pastel. El círculo que lo describe se encuentra dividido en segmentos en donde cada área de cada uno de los segmentos es proporcional al número de casos en esa categoría. De manera general, se usan porcentajes para cada categoría. Una muestra de cómo se ve el gráfico de pastel se presenta en el gráfico n.° 01.

Gráfico de barras

Otra manera de representar los datos es mediante el llamando gráfico de barras que consiste en exhibir los datos mediante un número de rectángulos, del mismo ancho, en donde cada uno de ellos representa una categoría particular. La longitud (y por lo tanto el área) de cada rectángulo es proporcional al número de casos en la categoría que representa. Se usa de manera general para datos cualitativos. Una muestra de cómo se ve el gráfico de barras se presenta en el gráfico n.° 02.





2.4. Medidas descriptivas

Es importante saber que los datos recolectados para un estudio estadístico no son generalmente constantes, es necesario ver una medida que nos indique la variabilidad de estos datos y nos dé una referencia sobre alrededor de qué valor fluctúan. Por otro lado, también es necesario conocer la simetría y la forma en la que los datos tienden a agruparse.

Las medidas que permiten esto son las llamadas medidas descriptivas y usualmente se encuadran en los siguientes cuatro tipos:

- Medidas de posición (o de tendencia central)
- Medidas de dispersión
- Medidas de simetría (sesgo)
- Medidas de forma (curtosis)

Medidas de tendencia central

Los estadísticos de ubicación o de tendencia central (también llamados promedios) proporcionan una estimación de la puntuación típica, común o normal encontrada en una distribución de puntuaciones en bruto. Es muy importante que a más de saber calcular las medidas de tendencia central, se pueda dar una interpretación correcta de la información que estas proporcionan.

Una primera medida es la **media poblacional**, que es la suma de todos los valores observados en la población divididos por el número de datos en la población. La **media muestral** es la suma de todos los valores de la muestra divididos por el número de datos en la muestra.

Si analizamos las propiedades de la media aritmética, se destaca que es única y que su cálculo incluye todos los datos de la muestra. Por esto, es la medida

de tendencia central más utilizada; sin embargo, el valor de la media aritmética se ve afectado por la presencia de uno o más valores sumamente grandes o pequeños (valores extremos). En tales casos, la medida de tendencia central más representativa es la **mediana**. La **media ponderada** es un caso especial de la media aritmética.

Otra medida de tendencia central que es utilizada es la **media geométrica**, que resulta útil para determinar el cambio promedio de porcentajes, razones, índices o tasas de crecimiento. La media geométrica es la raíz enésima del producto de n datos.

La **mediana** es el punto medio de los valores una vez que se han ordenado de menor a mayor. Si el número de datos es par, la mediana es la media aritmética de los dos valores centrales. Si el número de datos es impar, la mediana es el único dato central. Las principales propiedades de la mediana son que no es influida por la presencia de valores extremos y que es calculable en el caso da datos de nivel ordinal o más altos.

La **moda** es el dato que aparece con mayor frecuencia. En una distribución puede haber una o más modas o no haber ninguna. La moda puede determinarse para todos los niveles de datos y tiene la ventaja de que no influyen en ella los valores extremos. Sin embargo, se usa menos que la media o la mediana, ya que en muchos casos no hay moda o hay más de una. Si Media = Mediana = Moda, la distribución es **simétrica**. Si Media > Mediana > Moda, la distribución **no es simétrica** y **tiene sesgo positivo**. Si Moda > Mediana >Media, la distribución **no es simétrica y tiene sesgo negativo**.

Ejemplo
Con los siguientes datos: 8, 2, 3, 5, 4, 2, 6, 3, 1, 3, 13, 4, calcular la media aritmética, la media geométrica, la mediana y la moda. Indicar además si hay un valor extremo y cuál es el tipo de sesgo de la distribución.

1) Media aritmética:

$$\overline{X} = \frac{\Sigma X}{n} = \frac{54}{12} = 4.5$$

2) Media geométrica:

3) Mediana: Para determinar la mediana ordenar los datos: 1,2,2,3,3,3,4,4,5,6,8,13

Como n= 12 es par, la mediana es la media de las dos puntuaciones centrales es decir mediana = (3+4)/2 = 3.5

4) Moda = 3 (el valor con la frecuencia mayor)

5) Valor extremo: 13 (claramente separado de los demás valores)

6) Tipo de sesgo: Media > Mediana > Moda sesgo positivo o a la derecha

Para datos agrupados en una distribución de frecuencias, en el cálculo de la media aritmética intervienen el producto de la frecuencia y el punto medio de cada intervalo de clase.

Las ecuaciones que se utilizarán para este tipo de datos serán;

Media aritmética

Donde:

X = punto medio o marca de clase
f = frecuencia

Mediana

Donde:

= límite inferior de la clase de la mediana

= frecuencia acumulada de la clase anterior a la clase de la mediana.

= frecuencia de la clase que contiene a la mediana

Moda

Se la puede aproximar por el punto medio de la clase modal.

Un valor más preciso se obtiene aplicando la siguiente fórmula:

Donde:

= límite inferior de la clase modal

= (frecuencia de la clase modal) – (frecuencia de la clase que le antecede)

= (frecuencia de la clase modal) - (frecuencia de la clase que le sigue)

= es el ancho del intervalo de clase.

Medidas de dispersión

Una medida de ubicación, como la media o la mediana, solo describe el centro de los datos pero no dice nada sobre la dispersión de los datos. Por eso son necesarias las medidas de dispersión.

Una medida de dispersión pequeña indica que los datos se acumulan con proximidad alrededor de la media aritmética mientras que una medida de dispersión grande indica que hay uno o varios datos alejados de la media aritmética.

El **rango** es la medida de dispersión más simple. Es la diferencia entre el valor máximo y el valor mínimo de un conjunto de datos. Es muy fácil de calcular y entender, sin embargo, es una medida de dispersión que da una información limitada ya que solo toma en cuenta dos valores (el máximo y el mínimo) de la distribución.

La **varianza** es la media aritmética de las desviaciones de la media elevadas al cuadrado. La desviación estándar es la raíz cuadrada de la varianza. Podemos hablar de una varianza poblacional y de una varianza muestral.

La diferencia principal es que en la varianza poblacional, el numerador se divide para N (tamaño de poblacional) y en la varianza muestral para n-1 (donde n es el tamaño muestral) ya que se debe compensar el hecho de que la distribución muestral tiene menor dispersión que la distribución poblacional.

Ejemplo

Calcular la desviación estándar de los siguientes datos considerando que son datos a) de un población y b) de una muestra.

4 5 8 7 9 6

a) $\mu = 39/6 = 6.5$

X	X-μ	(X-μ)2
4	-2.5	6.25
5	-1.5	2.25
8	1.5	2.25
7	0.5	0.25
9	2.5	6.25
6	0.5	0.25

_____

b) Para la desviación estándar muestral, el cálculo de la sumatoria es el mismo solo que en lugar de μ se usa $\bar{X}$.

La desviación estándar es una medida de dispersión más adecuada que el rango ya que en su cálculo entran todos los datos. Esta medida se utiliza normalmente para comparar la dispersión de dos o más conjuntos de datos.

Para datos agrupados en una distribución de frecuencias, la desviación estándar toma en cuenta también la frecuencia de cada clase, como se muestra en las ecuaciones a continuación.

Amplitud de variación o rango

AV = límite superior de la clase más alta – límite inferior de la clase más baja

Desviación estándar

Donde:

= punto medio o marca de clase

= frecuencia

= media aritmética

Dispersión relativa

Coeficiente de variación

De la población:

De la muestra:

Coeficiente de asimetría o sesgo, denominado coeficiente de Pearson.

Cuartiles, deciles y centiles

Para calcular la posición de un cuartil, decil o percentil se usa la fórmula:

Una vez calculada la posición del percentil, proceda a calcular el percentil conforme se explica en los ejemplos 1 y 2. (No confunda la posición del centil o percentil con su valor).

Una vez calculada la posición del percentil, proceda a calcular el percentil conforme se explica en los ejemplos 1 y 2. (No confunda la posición del centil o percentil con su valor).

Ejemplo

Para la posición del primer cuartil Q1 use C = 25, para el tercer cuartil Q3 use C= 75.
(Q1 = C25; Q3 = C75), en algunos textos en vez de C se usa P, así P25

Para calcular la posición de un decil, por ejemplo, D3 use C = 30; para el decil 7 D7 use.
C = 70

Si Lc es entero el centil es el dato de la posición Lc

Si Lc no es entero, por ejemplo, si L25 = 7.62, el centil o percentil 25 se encontrará a 0.62 de la distancia entre el séptimo y el octavo dato. Su valor se calcula del siguiente modo:

C25 = Q1 = Dato7 + 0.62(Dato8 – Dato7)

En el cálculo de los cuartiles, recuerde, por ejemplo, que el primer cuartil Q1 es aquel valor que es mayor o igual que el 25 % de los datos y menor o igual que el 75 % de ellos.

Ejemplo

Calcular el primer y tercer cuartiles de los siguientes datos:

8.4 8.8 9.2 10 11.3 12.5 12.9 13.6 14 15

Solución
En este caso: n = 10, para Q1 C = 25 y para Q3 C = 75

Es la posición de Q1, mientras que su valor es:

Q1 = Dato 2 + 0.75 (Dato 3 – Dato 2)

De igual manera:

Es la posición de Q3, mientras que su valor es:

Q3 = Dato 8 + 0.25 (Dato 9 – Dato 8)

La mediana es se calcula del mismo modo que los otros cuartiles.

Medidas de sesgo y curtosis

Aparte de las medidas de tendencia central y de dispersión, otra característica de un conjunto de datos es la forma. Hay cuatro formas: simétrica, con sesgo positivo, con sesgo negativo y bimodal. En un conjunto simétrico media, mediana y moda son iguales y los valores de los datos se dispersan uniformemente en torno a estos valores. Un conjunto de valores se encuentra sesgado a la derecha o positivamente sesgado si existe un solo pico y los valores se extienden mucho más allá a la derecha del pico que a la izquierda de este.

En una distribución sesgada a la izquierda o negativamente sesgada existe un solo pico pero las observaciones se extienden más a la izquierda, en dirección negativa (gráfico n.° 01).

Gráfico n.° 01

La medida más sencilla para calcular el sesgo es el coeficiente de sesgo de Pearson, que se puede calcular con dos fórmulas distintas. La primera basada en la media, mediana y distribución estándar de una distribución, la segunda se puede obtener mediante programas estadísticos.

La **curtosis** mide cuán puntiaguda es una distribución en general en referencia a la distribución normal. Si tiene un pico alto, se dice leptocúrtica mientras que si es aplastada se dice platicúrtica. La distribución normal, que no es ni muy puntiaguda ni muy aplastada, se llama mesocúrtica.

El coeficiente de curtosis se calcula con la fórmula siguiente:

Si el coeficiente es mayor a 3, la forma es leptocúrtica. Si es igual a 3, la forma es mesocúrtica y si es menor a 3, la forma es platicúrtica (gráfico n.° 02).

Gráfico n.° 02

Ejemplo
Considerando los siguientes datos, calcular el coeficiente de curtosis e indicar el tipo de curtosis. 2 3 4 4 5 6

La curva es platicúrtica.

Otra manera de visualizar la simetría es utilizando el llamado diagrama de caja.

El **diagrama de caja** permite visualizar la simetría o la asimetría de una distribución de datos.

Para construir un diagrama de caja se requieren cinco valores: la media, la mediana, el dato menor o mínimo, el dato mayor o máximo y el primero y tercer cuartiles.

Rango intercuartílico

Es la diferencia entre el tercer y el primer cuartil:

Ejemplo

Suponga que en el servicio de entrega a domicilio de cierta pizza, el tiempo mínimo de entrega es de 15 minutos, que el tiempo máximo es de 40 minutos,

que la mediana es 25 minutos y que los cuartiles son:
minutos.

a) Calcular el rango intercuartílico.

b) Trazar el diagrama de caja y sobre la base de este indique si la distribución de los datos es o no simétrica.

Solución

a) Rango intercuartílico = 32.5 – 20 = 12.5

b) El diagrama de caja es el que se muestra a continuación:

El diagrama muestra que:

1. El bigote izquierdo es más corto que el derecho.

2. Que Q1 está más cerca de la mediana que Q3.

Comentario

Se observa que la cola o el bigote de la derecha es más largo que el de la izquierda, y también la distancia entre la mediana y es mayor que la distancia entre y la mediana, lo que indica que la distribución de los datos es asimétrica, con sesgo positivo.

Las líneas que van desde el mínimo a Q1 y desde Q3 al máximo se denominan bigotes.

Tema 3

3.1. Teoría de probabilidades

En la naturaleza y en la vida cotidiana, se presentan fenómenos cuyo resultado se lo da anticipadamente a través de la aplicación de leyes o fórmulas; sin embargo, existen otros cuyo resultado no puede ser anticipado con certeza, sino que existe una probabilidad de que un cierto resultado se dé. Para dar una explicación matemática a aquellos resultados que podrían aparecer, se desarrolló lo que se llama Teoría de la probabilidad.

En general, la probabilidad es la posibilidad de que algo pase; es decir, una probabilidad provee una descripción cuantitativa de la posibilidad de ocurrencia de un evento particular y se puede pensar que es su frecuencia relativa en una serie larga de repeticiones de una prueba, en la que uno de los resultados es el evento de interés.

Terminología

Para una mejor comprensión, se utilizan ciertas definiciones generales.

a) Experimento

Es un proceso que genera un conjunto de datos, ya sean estos cualitativos o cuantitativos; en su mayoría, los resultados dependen del azar, siendo imposible pronosticar con exactitud.

Ejemplos

- Registrar el tiempo de los competidores en una carrera.
- Medir los cambios en la bolsa de valores.
- Lanzar un dado.

b) Evento

Son todos los resultados posibles de un experimento u otra situación que genere incertidumbre. Podemos clasificar los eventos en dos tipos:

Los **elementales,** aquellos que constan de un solo resultado.

Los **compuestos,** que consisten en dos o más resultados.

Ejemplo

Al lanzar un dado, el evento “sale uno” es un evento elemental porque es un único evento posible; mientras que el evento “sale impar” es un evento compuesto porque está formado de los eventos elementales “sale uno”, “sale tres” y “sale cinco”.

Debemos indicar que dos eventos son **mutuamente excluyentes** si cuando ocurre un evento los otros no pueden ocurrir y viceversa. Por ejemplo, al lanzar una moneda al aire, si cae y “sale cruz” ya no puede darse el evento “sale cara”.

c) Espacio muestral

Es el conjunto de todos los resultados posibles de un experimento. Se lo identifica con el simbolo Ω.

Ejemplos

En el experimento: lanzar la moneda, el espacio muestral sería: evento “cae cruz” y evento “cae cara”.

En el experimento: Registre el tipo de sangre de una persona, el espacio muestral está formado por cuatro eventos, que son mutuamente excluyentes: “sangre tipo A”, “sangre tipo B”, “sangre tipo AB”, “sangre tipo O”.

Clasificación de la probabilidad

Existen tres maneras básicas de estudiar a la probabilidad; sin embargo, representan planteamientos conceptuales para el estudio de la teoría de probabilidades generando que los expertos no logren ponerse de acuerdo en el más apropiado. Estas son:

a) El planteamiento clásico

Supóngase un suceso E, que de un total de n casos posibles, **todos igualmente posibles,** puede presentarse en h de los casos. Entonces la probabilidad de aparición del suceso, es decir la probabilidad de ocurrencia, viene dada por:

La probabilidad de **no** aparición del suceso, llamada no ocurrencia del suceso, viene dada por:

Esta relación se da partiendo de que la probabilidad de un evento A es una función que cumple con:

b) <u>El planteamiento como frecuencia relativa</u>

La definición anterior de probabilidad tiene el inconveniente de que las palabras "igualmente posibles" son poco concretas generando el efecto circular porque se define a la probabilidad en términos de ella misma. Se ha generado, entonces, una definición con mucho más rigor, en donde la **probabilidad empírica de un suceso se toma como la frecuencia relativa de la aparición del suceso,** cuando el número de observaciones es muy grande. La probabilidad por sí misma es el límite de la frecuencia relativa cuando el número de observaciones crece indefinidamente.

A pesar de ser práctica, esta definición tiene problemas desde el punto de vista matemático, ya que no puede existir un número límite generándose una moderna teoría de probabilidades en donde la probabilidad es un concepto **no** definido como ocurre con el punto y la línea en geometría.

c) <u>El planteamiento subjetivo</u>

Está basado en las creencias de las personas que efectúan la estimación de la probabilidad definiéndose como la probabilidad asignada a un evento por parte de un individuo, basada en la evidencia que tenga disponible.

Las asignaciones de probabilidades subjetivas se dan con más frecuencia cuando los eventos se presentan una sola vez o un número muy reducido de veces.

Ejemplo de probabilidad clásica

Si se lanza al aire una moneda equilibrada, cuál será la probabilidad de que se obtenga cruz o cara.

a) Cruz es: P(cruz) = 1/2 porque una de las dos alternativas.

b) Cara es: P(cara) = 1/2 porque una de las dos alternativas.

Ejemplo de probabilidad empírica

Suponga que en un experimento se realizan 1000 ensayos y se produjo un evento E en 200 ocasiones. ¿Cuál es la probabilidad de que en un ensayo cualquiera se produzca el evento E?

R: P(E) =200/1000 = 1/5 = 0.20

Relación entre la probabilidad y la teoría de conjuntos

El estudio de las reglas de probabilidad está estrechamente relacionado con la teoría de conjuntos, para ello se asimila un evento con un conjunto.

Axiomas de Kolmogorov

1. $P(E) \geq 0$ La probabilidad de un evento es un número comprendido entre 0 y 1.
2. $P(S) = 1$ La probabilidad del espacio muestral es 1.
3. $P(E_1 \text{ o } E_2 \text{ o } \ldots \text{o } E_n) = P(E_1) + P(E_2) + \ldots + P(E_n)$ donde E_1, E_2,,… son eventos mutuamente excluyentes.

Propiedades de las probabilidades

Diagramas de Venn

Conjunto Universo

Conjunto A

Complemento de A es la parte del universo que no es A.

Conjuntos A y B disjuntos

Conjuntos A y B secantes o solapados

Reglas de probabilidad

- **Regla especial de adición**

Se aplica cuando los eventos son **mutuamente excluyentes o disjuntos.**

Para un par de eventos A, B: P(A o B) = P(A) + P(B)

Para tres eventos A, B, C: P(A o B o C) = P(A) + P(B) + P(C)

En el ejemplo de las tres bolas rojas, dos blancas y cinco azules, calcular la probabilidad de que al sacar una bola de la urna esta sea:

a) Roja o blanca: P(roja o blanca) = P(roja) + P(blanca) = 3/10 + 2/10 = ½

b) Blanca o azul: P(blanca o azul) = P(blanca) + P(azul) = 2/10 + 5/10 = 7/10

- **Regla general de la adición**

Se aplica para calcular la probabilidad de ocurrencia de uno u otro evento que **no sean mutuamente excluyentes**. (La fórmula es válida también para eventos mutuamente excluyentes dado que P(A y B) = 0).

Para los eventos A, B: P(A o B) = P(A) + P(B) – P(A y B)

Ejemplo

Un estudiante está tomando Álgebra y Castellano. Si la probabilidad de que apruebe Álgebra es 0.75, la de que apruebe Castellano es 0.90 y la probabilidad de que apruebe Álgebra y Castellano es 0.70. Se pregunta cuál es la probabilidad de que apruebe Álgebra o Castellano.

P(A o C) = P(A) + P(C) – P(A y C)

= 0.75 + 0.90 - 0.70 = 0.95

Para resolver estos problemas debe realizar un diagrama de Venn como el de la figura.

- **Regla especial de la multiplicación**

Se aplica para calcular la **probabilidad conjunta de ocurrencia de eventos independientes.**

Para dos eventos A y B: $P(A \text{ y } B) = P(A)\,P(B)$

Para tres eventos A, B y C: $P(A \text{ y } B \text{ y } C) = P(A)\,P(B)\,P(C)$

Ejemplo

Se lanza un dado por dos ocasiones. ¿Cuál es la probabilidad de que en los dos lanzamientos caiga en 3?

$P(3, 3) = P(3)\,P(3) = (1/6)\,(1/6) = 1/36$

Observe que el resultado del segundo lanzamiento es independiente del primero.

- **Probabilidad condicional**

Es la probabilidad de que ocurra un evento B, dado que ya ocurrió un evento A. O también la probabilidad de que ocurra un evento A dado que ya ocurrió el evento B. Esto se escribe:

Si se cumple que los eventos o sucesos A y B son estadísticamente independientes.

- **Regla general de la multiplicación**

Se aplica para calcular la **probabilidad conjunta de eventos dependientes**, es decir, cuando la ocurrencia de uno de ellos está condicionada a la ocurrencia del otro.

$$P(A \text{ y } B) = P(A)\, P(B/A) \quad \text{o también} \quad P(A \text{ y } B) = P(B)\, P(A/B)$$

Estas fórmulas y las de la probabilidad condicional están relacionadas, ya que las unas se obtienen de las otras mediante despejes.

Tomemos el ejemplo de las tres bolas rojas, dos blancas y cinco azules y supongamos que se desea calcular la probabilidad de que al sacar una bola y luego otra, la primera sea roja y la segunda blanca:

Observe que la probabilidad de que la primera vez salga roja es 3/10, pero al haber sacado una roja ahora nos quedan en total nueve bolas, de las cuales dos son blancas.

Calculemos ahora la probabilidad de sacar una bola roja y una azul. Como no se indica el orden tendremos que:

Tabla de contingencia o matriz de probabilidad

Los problemas de probabilidades se resuelven fácilmente usando una tabla de contingencia o matriz de probabilidad, en ella se pueden leer las probabilidades a priori y las probabilidades conjuntas o de intersección. Además, permite calcular fácilmente las probabilidades de la unión de eventos y las condicionales, tal como se ilustra a continuación.

Ejemplo

El personal que labora en una empresa está formado por hombres y mujeres que trabajan en las siguientes secciones: Gerencia, Profesional y Técnica, cuyos datos se resumen en esta tabla:

Complete esta tabla de contingencia y luego suponiendo que se elige al azar un empleado, calcule las siguientes probabilidades.

a) La probabilidad de que sea mujer.

b) La probabilidad de que sea hombre y trabaje en la Sección Técnica.

c) La probabilidad de de que trabaje en Gerencia o en la Sección Profesional.

d) La probabilidad de que trabaje en Gerencia, dado que sea mujer.

e) La probabilidad de que sea hombre dado que trabaje en la Sección Técnica.

Solución

A la tabla de los datos le añadimos una fila y una columna para los totales parciales de las filas y de las columnas. En la celda del extremo inferior derecho se coloca el total horizontal y vertical.

a) P(Mujer) = 54/136

b) P(Hombre y Técnica) = 50/136

c) P(Gerencia o Profesional) = P(Gerencia) + P(Profesional) = 11/136 + 40/136 = 51/136

d) P(Gerencia/Mujer) = 3/54. En la columna de MUJER vemos que tres de las 54 trabajan en Gerencia. También se puede aplicar la fórmula de la probabilidad condicional.

e) P(Hombre/Técnica) = 50/85. En la fila TÉCNICA se ve que 50 de los 85 técnicos son hombres.

Aplicando la fórmula:

3.2. Teorema de Bayes y diagramas de árbol

En la siguiente gráfica, sea S el espacio muestral y sean los eventos mutuamente excluyentes y colectivamente exhaustivos, de modo que:

Y sea el evento B tal que:

Entonces la probabilidad de que ocurra B viene dada por: ; luego:

Esta es la **probabilidad total** de que ocurra B.

De la probabilidad condicional sabemos que:

De aquí se tiene que:

Es decir:

Si ahora suponemos que $P(A_1)$ es una probabilidad a priori, $P(B/A_1)$ es la probabilidad condicional de que ocurra B dado que ocurrió A_1; y pensemos que

se quiere calcular la probabilidad a posteriori de que ocurra A_1 dado que ocurrió B, simplemente despejemos $P(A_1/B)$; según la fórmula anterior.

Donde P(B) es la probabilidad total de B.
sustituyendo P(B) en el denominador se obtiene la fórmula del Teorema de Bayes.

Diagrama de árbol

Un **diagrama de árbol** es una herramienta que se utiliza para determinar todos los posibles resultados de un experimento aleatorio (su uso es más característico en el Teorema de Bayes).

El diagrama de árbol es una representación gráfica de los posibles resultados del experimento, el cual consta de una serie de pasos, donde cada uno de estos tiene un número finito de maneras de ser llevado a cabo.

Para la construcción de un diagrama en árbol se partirá poniendo una rama para cada una de las posibilidades, acompañada de su probabilidad. Cada una de estas ramas se conoce como rama de primera generación.

En el final de cada rama de primera generación, se constituye a su vez, un nudo del cual parten nuevas ramas conocidas como ramas de segunda generación, según las posibilidades del siguiente paso, salvo si el nudo representa un posible final del experimento (nudo final).

Hay que tener en cuenta que la construcción de un árbol no depende de tener el mismo número de ramas de segunda generación, estas salen de cada rama de primera generación y la suma de probabilidades de las ramas de cada nudo debe ser 1.

Existe un principio sencillo de los diagramas de árbol que hace que estos sean mucho más útiles para los cálculos rápidos de probabilidad: multiplicamos las probabilidades si se trata de ramas adyacentes (contiguas).

Ejemplos

Una universidad está formada por tres facultades:

- La 1.ª con el 50 % de estudiantes.
- La 2.ª con el 25 % de estudiantes.
- La 3.ª con el 25 % de estudiantes.

Las mujeres están repartidas uniformemente, siendo un 60 % del total en cada facultad.

¿Probabilidad de encontrar una alumna de la primera facultad?

¿Probabilidad de encontrar un alumno varón?

Pero también podría ser lo contrario.

3.4. Distribuciones de probabilidad discreta y continua

El objetivo fundamental de la Estadística es inferir las propiedades de una población de la observación de una muestra o subconjunto de esta. La construcción y estudio de los modelos estadísticos, con variables discretas o continuas, están, entonces, íntimamente ligados al cálculo de probabilidades y, por consiguiente, a las distribuciones de probabilidad que se generan.

Distribuciones de probabilidad

En Estadística Descriptiva se observó a las distribuciones de frecuencias como una forma útil de resumir variaciones en los datos observados. Las distribuciones de probabilidad están relacionadas con las distribuciones de frecuencias. De hecho, podemos pensar que una distribución de probabilidad es una distribución de frecuencia teórica.

Una distribución de frecuencia teórica es una distribución de probabilidades que describe la forma en que se espera varíen los resultados. Como estas distribuciones representan las expectativas de que algo suceda, resultan modelos útiles para hacer inferencias y tomar decisiones en condiciones de incertidumbre.

Como existen variables discretas y continuas se puede hablar entonces de distribuciones de probabilidad discreta y distribuciones de probabilidad continua.

Distribuciones de probabilidad discreta

La **distribución de probabilidad binomial** es una distribución de probabilidad discreta, en la cual solo hay dos resultados posibles en cada ensayo: éxito o fracaso, y los ensayos son independientes entre sí. La variable aleatoria cuenta el número de éxitos (x) en n ensayos. La probabilidad de éxito en un ensayo se representa por Π y permanece constante en todos los ensayos.

Ejemplo

Se estima que 20 % de las personas de un país están afectadas por una enfermedad. Se selecciona al azar a nueve personas. Determinar la probabilidad de que seis de ellas estén afectadas por la enfermedad.

En este ejemplo: $p = 0.2$, $n = 9$ y $x = 6$

Otra distribución de variable discreta que se va a examinar es la **distribución hipergeométrica.** En la distribución binomial, la probabilidad de éxito permanece constante en cada ensayo mientras que en la distribución hipergeométrica esto no ocurre, ya que el muestreo se realiza sin reemplazo. Si se selecciona una muestra de una población finita sin reemplazo y si el tamaño de muestra n es mayor al 5 % del tamaño de la población N, se aplica la distribución hipergeométrica. S es el número de éxitos en la población y x el número de éxitos en la muestra.

Ejemplo

Una empresa tiene 50 empleados en la sección de ensamblado, 40 de ellos pertenecen al sindicato. Se elige al azar cinco empleados para formar un comité. ¿Cuál es la probabilidad de que cuatro de ellos formen parte del sindicato?

En este ejemplo: N = 50 S = 40 n = 5 X = 4

Distribuciones de probabilidad continua

Las distribuciones de probabilidad continua resultan de medir una variable continua. En este tipo de distribuciones generalmente se desea conocer el porcentaje de observaciones que se encuentra dentro de cierto margen. Es importante señalar que una variable aleatoria continua tiene un número infinito de valores dentro de cierto intervalo en particular.

La **distribución de probabilidad uniforme** es la más simple de una variable aleatoria. Tiene forma rectangular y queda definida por el valor máximo b y el valor mínimo a. La altura de la distribución es constante e igual a 1/(b-a).

Ejemplo

Una distribución uniforme se define en el intervalo de 6 a 10. Hallar la media y calcular la probabilidad de un valor mayor a 7.

En este ejemplo: a = 6 y b = 10, entonces $\mu = (a+b)/2 = 8$

P (X>7) = (10-7)/(10-6) = 3/4 = 0.55 (basta con dividir las longitudes de los segmentos ya que la altura es la misma para ambas áreas).

La **distribución normal** se la identifica como la piedra angular de la Estadística moderna. Esto se debe, en parte, al papel que desempeña en el desarrollo de la teoría estadística y, en parte, al hecho de que las distribuciones de datos observados frecuentemente tienen el mismo patrón general que las distribuciones normales.

Las aplicaciones de la distribución normal son muy numerosas en diferentes disciplinas académicas, en la industria y en las empresas. Observe que en cada caso se debe realizar un gráfico de la curva normal y ubicar la media y el o los valores de X. Luego, se debe calcular el valor normal estándar Z. Con el valor de Z, se busca el valor del área, gracias a la tabla de áreas bajo la curva normal. Por simetría, el valor del área correspondiente para un valor negativo de Z es el mismo que para su correspondiente valor positivo de Z. Es indispensable que sepa usar correctamente dicha tabla.

El rango percentilar es el porcentaje de datos que se encuentra por debajo de un valor determinado (el área a la izquierda de un valor dado en la curva normal).

Ejemplo

Una distribución normal tiene media 10 y desviación estándar 2. Determinar la proporción de datos que se encuentra:

a) entre 10 y 12

b) por encima de 12

c) por debajo de 12

En este ejemplo, el rango percentilar es 84.13.

d) entre 7 y 11 (un valor menor y otro mayor a la media, se suman las áreas)

e) entre 11.5 y 13.5 (dos valores mayores -o menores- a la media, se restan las áreas)

P = 0.4599 – 0.2734 = 0.1865 (**el resultado de p siempre debe ser positivo**).

- Cuando se debe calcular porcentajes, multiplicar la proporción por 100: (p)x100.
- Cuando se debe calcular cantidades, realizar p x n, donde n = tamaño de la muestra.

Resumen

Distribuciones de probabilidad discreta

Variable aleatoria: discreta y continua

Media, varianza y desviación estándar de una distribución de probabilidad

Media o valor esperado:

Varianza:

Puede usar la fórmula alterna:

Desviación estándar:

Distribución de probabilidad binomial

Características

a. Solo tiene dos resultados posibles en cada ensayo de un experimento: éxito o fracaso.
b. La variable aleatoria es el resultado del conteo del número de éxitos en n ensayos.
c. La probabilidad de éxito en cada ensayo es siempre igual en cada ensayo.
d. Los ensayos son independientes.

Fórmula

Tablas de probabilidad binomial

Se las usa para calcular las probabilidades binomiales para **n** desde 1 a 15.

Media de una distribución binomial:

Varianza de una distribución binomial:

Desviación estándar:

Probabilidad acumulada

Se calcula sumando las probabilidades de cada uno de los eventos involucrados.

Distribución de probabilidad de Poisson

Características

a. La variable aleatoria es el número de veces que ocurre un evento en un intervalo dado.
b. La probabilidad de que ocurra un evento es proporcional al tamaño del intervalo.
c. Los intervalos no se superponen y son independientes.

Fórmula

Nota: En los ejercicios de probabilidad acumulada, se suman las probabilidades de cada uno de los eventos involucrados.

Distribución hipergeométrica

Esta distribución se refiere a los experimentos estadísticos que consisten en tomar una muestra sin reemplazo, de un conjunto finito el cual contiene algunos elementos considerados “éxitos” y los restantes son considerados “fracasos”.

Tomar una muestra sin reemplazo significa que los elementos son tomados uno a uno, sin devolverlos. Podemos concluir, entonces, que los ensayos ya no pueden ser considerados independientes porque la probabilidad de “éxito“ al tomar cada nuevo elemento es afectada por el resultado de los ensayos anteriores, debido a que la cantidad de elementos de la población está cambiando.

Sean:

Ejemplo

Una caja contiene nueve pelotas de tenis, de las cuales cuatro están en buen estado y las restantes defectuosas. Se toma una muestra eligiendo al azar tres pelotas. Calcule la probabilidad de que en la muestra se obtengan:

1. Ninguna pelota en buen estado.
2. Al menos una pelota en buen estado.
3. No más de dos pelotas en buen estado.

Este es un experimento de muestreo sin reemplazo, por lo tanto, es un experimento hipergeométrico con:

N = 9 ; K = 4 ; n = 3 ; x: Cantidad de pelotas en buen estado en la muestra (variable aleatoria discreta)

Familia de distribuciones de probabilidad normal

Tenga presente que la distribución de probabilidad normal estándar N(0,1) es aquella en la que $\mu = 0$ y $\sigma = 1$.

Una distribución normal cualquiera $N(\mu, \sigma)$ se convierte en una distribución normal estándar N(0,1) mediante el cálculo del valor normal z, el cual expresa el número de desviaciones estándar de la distribución normal dada.

Valor normal estándar

Cálculo de áreas bajo la curva normal

Una vez calculado el valor z para un X, μ y σ dados, se usa la tabla.

Para encontrar el área bajo la curva normal estándar, comprendida entre la media μ y el valor z, tenga presente que a z le corresponde la primera columna y la primera fila, mientras que el área bajo la curva o probabilidad se lee en la intersección de la fila con la columna. Por ejemplo, para z = 1.84 nos situamos

en la fila **1.8** y en la columna **0.04**, allí leemos **0.467,** este es el valor del área bajo la curva o la probabilidad.

Problema inverso

El otro uso de la tabla de distribución normal estándar es aquel en el que dada una probabilidad (área bajo la curva normal estándar) hay que buscar el valor z correspondiente y, a partir de este, se puede a su vez calcular el X de la distribución normal dada. Para ello recuerde el signo de z, el cual será negativo si se halla a la izquierda de la media y positivo en caso contrario.

Ejemplo

Un estudio del INEC determinó que la media del gasto mensual en alimentación de una familia integrada por cuatro personas es de $480 y que este rubro sigue una distribución normal con una desviación estándar de $100.

Si se elige al azar una familia:

1. ¿Cuál es la probabilidad de que esta gaste entre $480 y $580?
2. ¿Cuál es la probabilidad de que gaste entre $300 y $480?
3. ¿Cuál es la probabilidad de que gaste entre $300 y $580?
4. ¿Cuál es la probabilidad de que gaste un mínimo de $650?
5. ¿Cuál es el gasto mínimo del 12 % de las familias que mas gastan?

Solución

En este caso: $\mu = 480 \quad \sigma = 100$

Aproximación de la distribución normal a la binomial

Una aplicación importante de la distribución normal es **la distribución normal como una aproximación de la distribución binomial.** Se aplica cuando **n** sea grande, entendiéndose como grande cuando n>25 y además se

cumpla con la condición de que

Como la distribución normal es continua y la binomial discreta, se debe aplicar el **factor de corrección por continuidad.**

Regla práctica

Una idea sencilla que permite una correcta aplicación de la corrección por continuidad es la de pensar que la gráfica de la distribución normal está formada por un conjunto de rectángulos, cada uno de los cuales se forma al fundir varillas (las varillas están separadas unas de otras), las cuales al fundirse formarían una sola lámina continua, de manera que, por ejemplo, la varilla 6 al fundirse formará el rectángulo que va desde 5.5 hasta 6.5.

Entonces es fácil comprender que para $x \geq 6$ se incluye el resultado de la fundición de la varilla 6, por lo que irá desde 5.5; para $X > 6$, no se incluye el resultado de la fundición de la varilla 6, por lo que irá desde 6.5; para $x < 15$, el resultado de la fundición de la varilla 15 no se incluye, por lo que se tomará hasta 14.5; etc.

Lo dicho es equivalente a las siguientes **reglas para aplicar el factor de corrección por continuidad.**

Reglas

Ejemplo

Suponga una distribución binomial con $n = 40$, $p = 0.55$. Calcular:

a) La media y la desviación estándar de la variable aleatoria

b) La probabilidad de que $x \geq 25$

c) La probabilidad de que $X < 15$

d) La probabilidad de que $15 \leq x \leq 25$

Solución