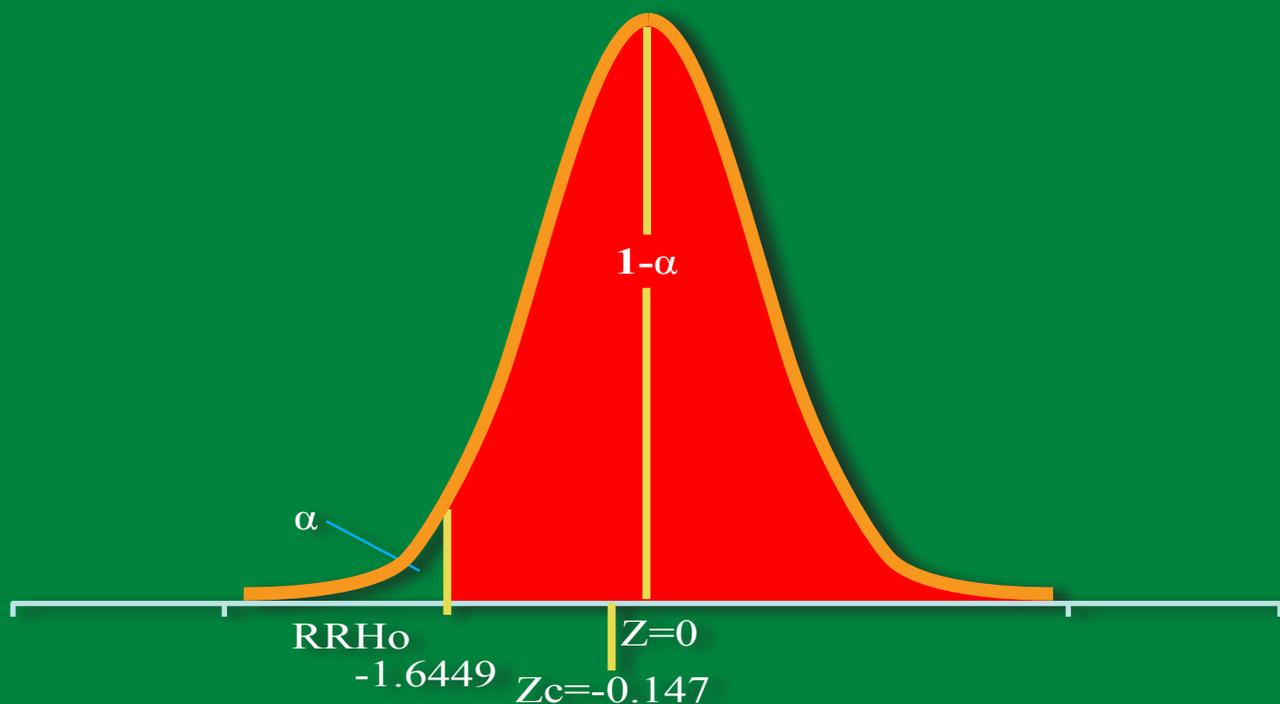


BIOESTADÍSTICA



M. en A. Teresa Guerra Dávila

PAPIME PE-104312

Universidad Nacional Autónoma de México
Facultad de Estudios Superiores Zaragoza



Datos para catalogación bibliográfica

Autora: Guerra Dávila, Teresa.

Bioestadística.

UNAM, FES Zaragoza, noviembre de 2014.

Primera edición.

PDF: 3,455 KB.

ISBN: 978-607-02-6101-5

Proyecto PAPIME PE-104312.

Diseño de portada y formación de interiores: Claudia Ahumada Ballesteros.

DERECHOS RESERVADOS

Queda prohibida la reproducción o transmisión total o parcial del texto o las ilustraciones de la presente obra bajo cualesquiera formas, electrónicas o mecánicas, incluyendo fotocopiado, almacenamiento en algún sistema de recuperación de información, dispositivo de memoria digital o grabado sin el consentimiento previo y por escrito del editor.

Bioestadística.

D.R. © Universidad Nacional Autónoma de México

Av. Universidad # 3000, Col. Universidad Nacional Autónoma de México, C.U.,
Delegación Coyoacán, C.P. 04510, México, D.F.

Facultad de Estudios Superiores Zaragoza

Av. Guelatao # 66, Col. Ejército de Oriente,
Delegación Iztapalapa, C.P. 09230, México, D.F.

Índice

	Página
Presentación	7
Introducción	9
Unidad 1. Elementos de probabilidad	11
1.1 Definiciones Básicas	11
1.2 Formas de Contar	12
1.3 Diagrama de Árbol	18
1.4 Definiciones de Probabilidad	19
1.5 Reglas para el cálculo de Probabilidades	20
1.6 Problemas adicionales de aplicación de las Leyes de Probabilidad	32
1.7 Teorema de Bayes	36
Unidad 2. Distribuciones de probabilidad	39
2.1 Conceptos Básicos	39
2.2 Modelos de Distribución Discreta	44
2.2.1 Distribución de Probabilidad de variable aleatoria Binomial	44
2.2.2 Distribución de Probabilidad de variable aleatoria de Poisson	48
2.2.2.1 Aproximación del proceso Binomial con la distribución Poisson	51
2.2.3 Distribución de Probabilidad de variable aleatoria Hipergeométrica	52
2.2.4 Distribución de Probabilidad de variable aleatoria de Pascal y Distribución Geométrica	55
2.2.5 Distribución de Probabilidad de variable aleatoria Multinomial	57
2.3 Modelos de Distribución Continua	59
2.3.1 Distribución de Probabilidad de una variable Continua	59
2.3.2 Parámetros de una Distribución Continua de Probabilidad	59
2.3.3 Distribución Normal	61

2.3.4	Distribución Normal Estándar	62
2.3.5	Aproximación Normal a la Distribución Binomial	66

Unidad 3. Estadística descriptiva **71**

3.1	Tipos de Datos	71
3.2	Tipos de Muestreo	72
3.3	Análisis Exploratorio de Datos	74
3.3.1	Diagrama de Tallo y Hoja	74
3.3.2	Diagrama de Caja con Bigotes	77
3.4	Medidas Descriptivas en la Muestra	82
3.4.1	Medidas de Tendencia Central	82
3.4.2	Medidas de Variabilidad	86
3.5	Estadística para Datos Agrupados	90
3.6	Representación Gráfica de los Datos	96
3.7	Medidas Descriptivas para Datos Agrupados	99
3.7.1	Medidas de Tendencia Central para datos agrupados	99
3.7.2	Medidas de Variabilidad para datos agrupados	102
3.7.3	Medidas Posicionales o Cuantiles	103

Unidad 4. Estadística inferencial **109**

4.1	Conceptos Básicos	109
4.1.1	Distribución Muestral	109
4.1.2	Teorema Central del Límite	109
4.2	Estimación de Parámetros por Intervalo	114
4.2.1	Ecuación General para la Estimación de Parámetros por Intervalo	115
4.2.2	Distribución t student	117
4.2.3	Distribución Ji Cuadrada (χ^2)	118
4.2.4	Aplicación de Estimación de Parámetros por Intervalo	119
4.3	Contrastes de Hipótesis para un Parámetro	122
4.3.1	Secuencia para realizar el Contraste	122
4.3.2	Aplicación del Proceso de Contraste	125
4.4	Inferencia en la Comparación de 2 Grupos Poblacionales	129
4.4.1	Estimación por Intervalo para la Comparación	129
4.4.2	Distribución de Fisher	130
4.4.3	Aplicación del Proceso de Estimación en la Comparación de 2 Grupos Poblacionales	132
4.5	Contraste de Hipótesis para la Comparación de 2 Grupos Poblacionales	141

4.5.1	Aplicación del Proceso de Contraste de Hipótesis en la Comparación de 2 Grupos Poblacionales	141
4.6	Pruebas con Datos Categóricos	148
4.6.1	Pruebas de Independencia	148
4.6.1.2	Pruebas de Independencia con Tablas 2×2	152
4.6.2	Pruebas de bondad de Ajuste	153
	Unidad 5. Diseño experimental y regresión	165
5.1	Relación entre Diseño de Experimentos y Análisis de Varianza	165
5.1.1	Análisis de Varianza de un Factor Completamente al Azar	166
5.1.1.1	Modelo de un Factor Completamente al Azar	167
5.1.1.2	Proceso de Contraste de Hipótesis en el Análisis de Varianza de un Factor Completamente al Azar	167
5.1.1.3	Definición Matemática de Sumas de Cuadrados para el modelo de un Factor Completamente al Azar	169
5.1.1.4	Prueba de la Diferencia Significativa Honesta de Tukey	171
5.1.2	Análisis de Varianza de un Factor con Bloques al Azar	173
5.1.2.1	Modelo de un Factor con Bloques al Azar	173
5.1.2.2	Definición Matemática de la Suma de Cuadrados para el Modelo de Bloques	174
5.1.3	Análisis de Varianza Factorial de 2 Factores, Completamente al Azar con Repetición	178
5.1.3.1	Modelo de 2 Factores, Completamente al Azar, con Repetición	179
5.1.3.2	Cálculo de las Sumas de Cuadrados para el Modelo	179
5.2	Análisis de Regresión	185
5.2.1	Análisis de Regresión Lineal Simple	185
5.2.1.1	Supuestos del Análisis de Regresión	185
5.2.1.2	Diagrama de Dispersión	186
5.2.1.3	Método de Mínimos Cuadrados para el cálculo de las Constantes de Regresión	187
5.2.1.4	Evaluación del Modelo Ajustado	188
5.2.2	Inferencia en el Análisis de Regresión	188
5.2.2.1	Estimación por Intervalo para los Parámetros de la Regresión	189
5.2.2.2	Contrastes de Hipótesis en Regresión Lineal	189
5.2.2.3	Aplicación de la Inferencia en el Análisis de Regresión	190
5.2.3	Análisis de Regresión No Lineal	202
5.2.3.1	Análisis Comparativo de los Modelos Exponencial y Potencial con el Lineal	202
5.2.3.2	Inferencia en la Regresión No Lineal	203
5.3	Análisis de Correlación Lineal	211
5.3.1	Definición Matemática del Coeficiente de Correlación Muestral	211

5.3.2	Contrastes de Hipótesis relacionados con la Correlación Poblacional	212
5.3.3	Estimación por Intervalo del Coeficiente de Correlación Poblacional	213
5.3.4	Aplicación de la Inferencia en el Análisis de Correlación	213
	Referencias	221



Presentación

Este material se ha desarrollado con la finalidad de que sirva como texto de apoyo al curso de Bioestadística, que forma parte del plan de estudios de la carrera de Ingeniería Química. Su contenido se ha limitado a cubrir las unidades temáticas de este curso, por lo que no pretende ser un texto para eruditos, es un texto elemental para ser usado por principiantes en el área de estadística y forma parte de los productos contemplados en el Proyecto PAPIME PE-104312.

Para elaborar este material se utilizó como referencia la carta descriptiva de Bioestadística con todos los temas establecidos en ella.

En la unidad I, Elementos de probabilidad, se definen conceptos importantes, se explica el uso de las leyes de probabilidad, se introduce el uso de las permutaciones y las combinaciones y algunos teoremas que permiten comprender los fundamentos del cálculo de probabilidades. Se presentan ejemplos resueltos totalmente, paso por paso, para que el estudiante conozca las técnicas de resolución de los problemas de aplicación.

En la unidad II, llamada Distribuciones de probabilidad, se identifican los tipos de variables aleatorias, se define el concepto de distribución de probabilidad y su clasificación. Se explica el uso de los diferentes modelos de distribución (discreta o continua) y se aplica cada función de probabilidad a la resolución de problemas. Se presentan también ejemplos totalmente resueltos.

La unidad III, Estadística descriptiva, se desarrolla a partir de los conceptos de población y muestra, tipos de muestreo y análisis exploratorio de datos. Se presentan y explican las medidas descriptivas de tendencia central y de variabilidad, su utilidad y la forma de calcularlas, tanto para datos sin agrupar como para datos agrupados. Se realiza e interpreta la representación gráfica de los datos.

Para la unidad IV, Estadística Inferencial, se inicia con el teorema central del límite y sus axiomas que dan cabida al concepto de error estándar, se establece la relación entre la estadística descriptiva y la estadística inferencial como fundamento para inferir probabilísticamente el comportamiento de las poblaciones objeto de estudio. Se trabajan los métodos de cálculo por intervalo y contraste de hipótesis para los parámetros o medidas que describen a la población, tanto para una sola población como para la comparación de 2 grupos poblacionales con la finalidad de tomar decisiones fundamentadas estadísticamente.

En la unidad V se tratan 3 tipos de análisis inferencial que por su utilidad son muy importantes: Análisis de Varianza, Análisis de Regresión y el Análisis de Correlación lineal. El análisis de varianza permite analizar más de 2 grupos por sus medias (sólo se incluyen los modelos de un factor completamente al azar, de un factor con bloques al azar y de 2 factores con repetición). El análisis de regresión permite definir el tipo de relación que guardan 2 o más variables de un experimento aleatorio mientras que el análisis de correlación permite analizar la asociación entre 2 variables no necesariamente dependientes una de otra. En este material se revisa sólo la regresión lineal simple y la regresión no lineal.

Consciente de la dificultad que representa este curso para los alumnos de la carrera, se redujo al mínimo necesario el tratamiento matemático. La simbología utilizada en cada fórmula o algoritmo está explicada para facilitar su comprensión y aplicación.

El texto incluye el desarrollo teórico de cada uno de los temas tratados, presentado de la forma más sencilla posible, para que el estudiante pueda apropiarse del conocimiento. Además, se incluyen ejemplos alusivos a cada tema, acompañados de su resolución completa, para hacer más accesible la comprensión y aplicación de las técnicas estadísticas que el alumno deberá demostrar haber aprendido al finalizar el curso.

Para que este material sea útil, se recomienda no saltarse la lectura de la teoría que acompaña a cada nuevo tema, con el fin de lograr una buena estructura del conocimiento pues esto permitirá ir incrementando la capacidad para comprender temas más avanzados de la estadística.

Agradecimientos

Quiero, en este espacio, agradecer y reconocer el esfuerzo de mis amigos y compañeros de trabajo, que hicieron el favor de revisar este material y hacerme las sugerencias y correcciones, que no fueron pocas, para mejorarlo.

Al Biólogo, **Luis Campos Lince** por sus sugerencias para aclarar las ideas y explicaciones que permitan el entendimiento de los temas, y así dar mayor apoyo al aprendizaje, pues su principal preocupación son los estudiantes y la formación de buenos profesionales.

A la Maestra en Ciencias, **María José Marques Dos Santos**, que con su experiencia y sus sugerencias permitió mejorar muchísimo este material, además del apoyo al proporcionarme algunos elementos adicionales para ejemplificar y aclarar ideas, en beneficio de los alumnos.

Al Biólogo, **Jorge Manuel López Reynoso**, por su paciencia al revisar tanto los temas como todos y cada uno de los ejemplos utilizados, señalando los errores e inconsistencias en los cálculos y haciendo sugerencias sobre el orden más adecuado para presentar los diferentes temas y así favorecer la comprensión de este material didáctico.



Introducción

Es importante aclarar que el uso de la estadística, es fundamental para el análisis de los resultados de una investigación y que todos los métodos de análisis se basan en la teoría de la probabilidad. Por esta razón es muy importante que se tengan nociones de esta teoría y de su aplicación para resolver problemas que involucran procesos al azar. Además, toda la estadística inferencial se basa en modelos probabilísticos llamados distribuciones de probabilidad.

Con objeto de situar adecuadamente los conceptos es necesario establecer algunas definiciones importantes.

0.1 Definiciones Básicas

0.1.1 Medición.- Es el proceso mediante el que se le asignan números que indican sus dimensiones, a los objetos o a los hechos.

0.1.2 Estadística.- Es una rama de la matemática aplicada, que proporciona los métodos para coleccionar, clasificar, resumir, organizar, analizar e interpretar datos numéricos como base para obtener conclusiones y tomar decisiones.

0.1.3 Población o Universo.- es el conjunto total de unidades elementales que al investigador le interesa conocer.

0.1.4 Muestra.- Subconjunto de unidades elementales extraído de la población objeto de estudio.

0.1.5 Relación entre la Probabilidad, la Estadística y la Investigación.- La estadística forma parte esencial de una investigación porque los datos obtenidos de un experimento (muestra) deben clasificarse, organizarse y analizarse para extraer toda la información posible y con base en ésta, generalizar el comportamiento observado. Entonces, será posible tomar decisiones, probar hipótesis etc. respecto al comportamiento de la población objeto de estudio. La probabilidad interviene en el proceso porque los modelos de medición que utiliza la estadística son fundamentalmente probabilísticos.

0.1.6 Diseño de Experimentos.- Es sumamente importante, al realizar investigación, diseñar y planificar la forma de hacerlo. Debe haber una idea clara de qué se va a medir, como se va a medir y en qué condiciones,

con objeto de que se cumplan los requisitos necesarios para utilizar un método estadístico que facilite el análisis y permita fundamentar la toma de decisiones. Si no se diseña el proceso desde el principio, el investigador podría llevarse la desagradable sorpresa de que todo su trabajo no es apto para ser manejado estadísticamente y entonces perder su valor ante la falta de un fundamento matemático sólido.

0.1.7 Relación entre la Estadística y la Probabilidad.- Los fenómenos en la naturaleza pueden clasificarse como determinísticos o como aleatorios. Sin embargo, la mayoría de los fenómenos que ocurren en el universo, son aleatorios, esto es, se producen de una forma o de otra, dependiendo de las circunstancias del momento, sin que el investigador pueda controlar todas las variables que influyen en el resultado. Por esta razón, los modelos matemáticos que podrían explicar un determinado fenómeno, se fundamentan en las leyes de la probabilidad. Así que es necesario que el investigador tenga una idea clara de cómo ocurren los procesos aleatorios y las leyes que rigen su comportamiento, para que sea capaz de obtener mejores resultados en su trabajo.

0.2 Etapas de una Investigación Estadística

0.2.1 Detección del problema.- Esta etapa ocurre cuando el investigador se enfrenta a una situación desconocida que puede influir en sus procesos de trabajo.

0.2.2 Delimitación del mismo.- El investigador debe delimitar perfectamente el problema para poder identificarlo, manejarlo y resolverlo de la mejor manera.

0.2.3 Planteamiento de la Hipótesis.- En esta etapa, el investigador establece algunas suposiciones respecto al comportamiento del fenómeno o situación observada, con la idea de comprobarlas.

0.2.4 Diseño del Experimento.- En esta fase del proceso de investigación, el investigador debe diseñar la forma como llevará a cabo el experimento para lograr los mayores beneficios: Definir la población, de qué tamaño tomará la muestra para que sea representativa de la población, cómo obtener la muestra, qué medir, cómo medir y cuándo medir, para adecuar los datos al método de análisis correcto, que permita obtener la mayor información posible.

0.2.5 Registro y Análisis de Resultados.- Realizado el trabajo de investigación, se registrarán los resultados obtenidos y se contrastan en función de los supuestos planteados.

0.2.6 Prueba de Hipótesis.- En esta etapa, el investigador, usará sus recursos para que, en forma experimental o teórica, recopile toda la información que pueda necesitar para sus pruebas.

0.2.7 Discusión e Interpretación de Resultados.- Los resultados del contraste nos permiten definir si los supuestos planteados son válidos o no dentro de cierto nivel de confiabilidad manejado por el investigador.

0.2.8 Conclusión y Toma de Decisiones.- En este punto, el investigador tomará decisiones respecto al problema planteado, con base en el proceso estadístico que eligió para hacer sus pruebas.

Elementos de probabilidad

Hablar de probabilidad, implica introducirse en el ámbito de los procesos aleatorios, aquellos que ocurren influenciados por las leyes del azar. Decimos que un proceso es aleatorio porque el resultado se ve influenciado por las situaciones del momento y no puede asegurarse, de manera anticipada, cuál será el resultado.

1.1 Definiciones Básicas

Para definir matemáticamente, la probabilidad de ocurrencia de un fenómeno aleatorio, tenemos que partir de algunas definiciones básicas:

1.1.1 Espacio Muestra o Espacio Muestral

Es el número de resultados totales, obtenidos al realizar un experimento al azar. Por ejemplo:

- Al tirar un dado, el número de resultados posibles es: 1, 2, 3, 4, 5 y 6 que, como espacio muestra se representan como el conjunto $S = \{1,2,3,4,5,6\}$.
- Si en un salón se encuentran 7 alumnos, 5 chicas y 2 varones, cuyos nombres son, Karen, Ana, Lourdes, Martha, Diana, Carlos y Vicente y se eligen al azar, 3 alumnos, los resultados posibles son todas las ternas ordenadas diferentes formadas por 3 nombres, identificados por sus iniciales:

$S = \{(K,A, L), (K,A, M), (K, A,D), (K,A, C), (K, A, V), (K, L, M), (K, L,D), (K, L, C), (K, L, V), (K, M,D), (K, M, C), (K, M, V), (K,D, C), (K,D,V), (K, C, V), (A,L, M), (A, L,D), (A, L, C), (A, L, V), (A, M,D), (A, M, C), (A, M, V), (A,D, C), (A,D, V), (A, C, V), (L, M,D), (L, M, C), (L, M, V), (L,D, C), (L,D, V), (L, C, V), (M,D, C), (M,D, V), (C, V, M), (C, V,D)\}$

Entonces, el Espacio Muestra es el conjunto formado por 35 triadas (ternas formadas por 3 elementos diferentes), que representan los nombres de los elementos a elegir. El cambio de orden y la repetición de sigla no son resultados apropiados, para formar parte de este conjunto.

Si el proceso aleatorio incluyera más elementos, sería cada vez más complicado definir los diferentes resultados posibles del experimento, por lo que se hace necesario usar formas adecuadas para contar los resultados totales.

1.1.2 Evento

Es cualquier subconjunto de resultados, definido dentro del Espacio Muestra.

Por ejemplo, si dentro del experimento aleatorio definido en el inciso **(a)**, especificamos que el evento E_1 es el número de puntos que son múltiplos de 3, al tirar el dado, tendremos que:

$$E_1 = \{3, 6\}$$

Esto es, hay 2 casos que favorecen la definición del evento E_1 .

Si tomamos el ejemplo del inciso **(b)** y definimos el evento E_2 , que Karen forme parte de los 3 alumnos elegidos, entonces:

$$E_2 = \{(K, A, L), (K, A, M), (K, A, D), (K, A, C), (K, A, V), (K, L, M), (K, L, D), (K, L, C), (K, L, V), (K, M, D), (K, M, C), (K, M, V), (K, D, C), (K, D, V), (K, C, V)\}$$

Como podemos observar, hay 15 formas en que Karen forme parte de los elegidos.

1.2 Formas de Contar

Cuando se desea saber el número de resultados totales de un experimento o el número de resultados favorables de un evento, es necesario utilizar formas de contar, de manera eficiente, las formas diferentes en que ocurre un proceso aleatorio.

Dado que los procesos aleatorios pueden ser ordenados o no ordenados, existen 2 formas de contar importantes:

- a) **Permutaciones.**- Se utilizan cuando el orden en que ocurren los resultados forma parte de las diferencias características del proceso aleatorio.
- b) **Combinaciones.**- Se utilizan cuando las diferencias en los resultados se refieren a cambios reales en los elementos, no a las diferentes ordenaciones o arreglos que se puedan lograr con ellos.

1.2.1 Permutaciones u ordenaciones

Existen 3 tipos de formas de contar con orden.

1.2.1.1 Ordenaciones con repetición

Si en la extracción aleatoria, hay reemplazo o reposición, de los elementos previamente extraídos, los resultados se pueden repetir, entonces, ocurren **ordenaciones con repetición** y el número de resultados diferentes se calcula como:

$$n^r$$

Donde:

n, es el número de resultados diferentes que pueden ocurrir en una extracción o ensayo.

r, es el número de ensayos o extracciones que se realizan sucesivamente.

EJEMPLO 1.1. Si en una caja tenemos 3 canicas: 1 roja, 1 verde y 1 azul y se extraen 4 canicas sucesivamente, reemplazando cada canica antes de la siguiente extracción, de tal manera que el número de canicas de cada color permanece constante durante todo el proceso, ¿cuál es el número total de ordenaciones, **por color**, que se pueden lograr en el Espacio muestra, si se permiten las repeticiones parciales o totales para cada color?

Si el resultado observado se refiere al color, **n** será 3 porque sólo disponemos de tres colores.

Puesto que se extraerán 4 canicas sucesivas, con reemplazo, **r** es 4.

Aplicando la definición de ordenaciones con repetición, tenemos que:

$$n^r = 3^4 = 81$$

Esto significa que habrá 81 conjuntos, de cuatro elementos, diferentes en donde se ordenan los 3 colores, incluyendo la repetición parcial o total de colores.

Es importante hacer notar que en este caso estamos definiendo los resultados desde el punto de vista cualitativo “color” y que no se han manejado todavía los datos cuantitativos. Esto significa que las canicas sólo se distinguen por el color.

Si cada canica fuera distinguible no sólo por el color sino por el número de canicas rojas, verdes o azules, el número de ordenaciones con repetición sería más grande que el que se logra al realizar el proceso sin repetición.

1.2.1.2 Permutaciones sin Reemplazo

Se utilizan para contar casos totales y casos favorables, cuando en el proceso de extracción aleatoria, no se permite el reemplazo, por lo que no puede ocurrir la repetición de resultados y entonces, el número de ordenaciones totales se calcula como:

Primer caso: Se extraen solamente parte de los n elementos para ser ordenados, esto es se extraen r elementos a la vez, de los n disponibles.

$$P_r^n = \frac{n!}{(n-r)!}$$

Donde:

P_r^n , es el número de permutaciones u ordenaciones que se pueden lograr con n los elementos disponibles.

n , es número total de elementos disponibles para ser ordenados.

$$n! = n(n-1)(n-2)(n-3)\dots(2)(1)$$

r es número de elementos que se ordenan dentro del total n ,* o el número de ensayos sucesivos, sin reemplazo

Segundo caso: Se ordenan los n elementos disponibles a la vez:

$$P_n^n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1} = n!$$

EJEMPLO 1.2. El profesor de teatro compró 12 boletos para los alumnos de su clase que irán a una representación teatral. Las 12 butacas se encuentran juntas en la misma fila del teatro. ¿De cuántas formas diferentes se pueden sentar los alumnos al ocupar las 12 butacas?

La pregunta implica considerar que los 12 alumnos son distinguibles entre sí y que se ordenarán en 12 lugares en forma aleatoria.

En este caso, n es 12 y r es también 12 porque se desea conocer el total de acomodos diferentes para 12 personas en una fila.

Por lo que:

$$P_{12}^{12} = \frac{12!}{(12-12)!} = \frac{12(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}{0!} = 479,001,600$$

Como puede verse, el número de ordenaciones es considerablemente grande.

EJEMPLO 1.3. En el laboratorio Ingeniería Química hay 7 manómetros de modelo diferente, identificados por los números del 1 al 7. Si en el laboratorio están trabajando 4 equipos y cada equipo requiere un manómetro ¿De cuántas formas diferentes se pueden asignar estos instrumentos si el inter-laboratorista los elige al azar?

Para resolver, utilizaremos la definición matemática de ordenaciones sin reemplazo porque necesitamos calcular el número de formas en que se pueden asociar los diferentes instrumentos a los 4 equipos. Esto es, formar ordenaciones de 4 en 4 con 7 elementos disponibles.

$$P_4^7 = \frac{7!}{(7-4)!} = \frac{7(6)(5)(4)(3!)}{3!} = \frac{(7)(6)(5)(4)}{1} = 840$$

También podemos esquematizar los lugares disponibles para ser ocupados en la ordenación mediante rayas o cajones en donde cada raya o cajón representa un lugar en la ordenación, por lo que, si vamos a ordenar 4 elementos de los 7 disponibles, tendremos 4 rayas o cajones vacíos que se llenarán en secuencia, de izquierda a derecha empezando con el valor n y descontando de uno en uno hasta incluir los r elementos que se desean ordenar:

$$\underline{7} \times \underline{6} \times \underline{5} \times \underline{4}$$

Esto quiere decir que hay 7 manómetros disponibles para elegir el del primer equipo, 6 para elegir el del segundo equipo, 5 para el tercero y cuatro para el último equipo, en este caso no se permite la repetición. Así, el producto de estos 4 dígitos será el número de formas diferentes de ordenar 4 elementos tomándolos de los 7 disponibles.

1.2.1.3 Permutaciones por subconjuntos

Estas formas de contar se utilizan cuando los elementos que se van a ordenar, se presentan en subconjuntos, de tal manera que habrá i elementos del tipo 1, j elementos del tipo 2, k elementos del tipo 3 etc. dentro de los n totales. Así, $n = i + j + k$. Esto es, cada subconjunto tiene un número determinado de elementos no distinguibles entre sí, pero cada subconjunto es diferente de los otros.

Entonces, el número de permutaciones disponibles, atendiendo al número de subconjuntos diferentes será:

$$P_{i,j,k,\dots}^n = \frac{n!}{i! j! k! \dots}$$

Donde:

n , es el total de elementos que se van a extraer aleatoriamente.

i , es el número de elementos que hay en el primer subconjunto.

j , es el número de elementos que hay en el segundo subconjunto.

k , es el número de elementos que hay en el tercer subconjunto, etc.

EJEMPLO 1.4. Si en una caja hay 5 canicas blancas, 3 canicas negras y 6 canicas anaranjadas y se desean ordenar, el número de ordenaciones posibles es, aplicando la definición de permutaciones por subconjuntos:

$$n = 14, i = 5 \quad j = 3 \quad k = 6$$

$$P_{5,3,6}^{14} = \frac{14!}{5!3!6!} = \frac{8.71782912 \times 10^{10}}{(20)(6)(720)} = 168168$$

NOTA. Observe que si las 14 canicas fuesen todas diferentes, entonces el resultado sería

$$P_{14}^{14} = 14! = 8.71782912 \times 10^{10}$$

Para utilizar cualquier tipo de permutación, debemos estar seguros de que el proceso aleatorio requiere orden.

EJEMPLO 1.5. ¿De cuántas formas se pueden arreglar las letras de la palabra *Paralelepípedo* tanto si las palabras resultantes tienen o no significado?

En este caso $n=14$ $p=3$ $a=2$ $r=1$ $l=2$ $i=1$ $e=3$ $d=1$ $o=1$

$$P_{2,2,1,2,1,3,1,1}^{14} = \frac{14!}{3!2!1!2!1!3!1!1!} = \frac{8.71782912 \times 10^{10}}{(3 \times 2 \times 1)(2 \times 1)(1)(2 \times 1)(1)(3 \times 2 \times 1)(1)(1)} = 605404800$$

Un proceso aleatorio incluye orden cuando, por la naturaleza de los elementos (números y letras) el orden en que se acomodan da origen a diferentes resultados o cuando la extracción se realiza tomando a los elementos uno por uno y el orden en que salgan es de interés particular.

Por ejemplo:

- a) Si en una carrera de caballos se toman apuestas y se premia de manera diferente a los tres primeros caballos en llegar a la meta, para el apostador es importante que el caballo de su preferencia llegue en primer lugar porque su ganancia aumentaría. Entonces, el número de arreglos en los que su caballo estaría en primer lugar, se calcularía usando una permutación u ordenación sin reemplazo.
- b) En el caso en que se asignan puestos de trabajo dependiendo del nivel de preparación de los aspirantes, también habría un número de ordenaciones diferentes para cubrir dichos puestos.

1.2.2 Combinaciones

Son formas de contar que se utilizan cuando el orden de extracción no significa diferencia de forma, porque los elementos presentes en el conjunto son lo importante y no su ordenación. La única diferencia reconocida es aquella que ocurre cuando se cambia a uno o más de los elementos que conforman el conjunto originalmente elegido. Esto equivale a tomar todos los elementos requeridos, juntos, en una sola extracción, por lo que la noción de orden desaparece.

Una combinación se define, matemáticamente de la siguiente manera:

$$C_r^n = \frac{n!}{r!(n-r)!}$$

Donde:

C , es el símbolo del número de combinaciones o de selecciones que se pueden lograr con los elementos disponibles.

n , es el número total de elementos disponibles.

r , es el número de elementos que se van a seleccionar.

EJEMPLO 1.6. El dueño de un restaurante va a elegir 6 meseros de los 10 que tiene contratados, para que se hagan cargo de atender un banquete en un salón de fiestas. Si todos son igualmente eficientes y el dueño hace la elección al azar, ¿cuántas formas tiene de seleccionar el conjunto de 6 meseros?

En este problema, un conjunto es diferente de otro si se cambia al menos una de las personas que constituyen el conjunto previamente elegido, por lo tanto el orden no forma parte de las diferencias observables entonces, es conveniente aplicar la definición de combinación para calcular las diferentes selecciones, de la siguiente forma:

n es el número total de personas disponibles para la selección (10)

r es el número de personas que se desean elegir(6), así que:

$$C_6^{10} = \frac{10!}{6!(10-6)!} = \frac{10!}{6!4!} = \frac{10(9)(8)(7)(6!)}{6!4!} = \frac{5040}{24} = 210$$

Este resultado es el número total de formas de seleccionar 6 elementos diferentes eligiéndolos de un grupo de 10.

Todas estas fórmulas, se aplican para definir casos totales y casos favorables o de interés, en el cálculo de probabilidades.

1.3 Diagrama de Árbol

Es una herramienta gráfica, útil para contar el número de formas en que ocurren los resultados de un proceso al azar. Consiste en un diagrama formado por ramas horizontales y divergentes, que parten de un origen y se van alargando conforme se agregan ensayos del proceso repetitivo hasta terminar el experimento aleatorio.

EJEMPLO 1.7. En una urna hay 10 tarjetas con la figura de un triángulo, 6 tarjetas con la figura de una esfera y 8 con la figura de un cuadrado. Si se extraen, una por una, 3 tarjetas, ¿Cuántos resultados diferentes de 3 tarjetas se pueden obtener tomando en cuenta sólo el resultado cualitativo (triángulo, esfera o cuadrado)?

Como la extracción es de una en una, existe cambio de resultado debido al orden y debido al tipo de figura. Si utilizamos un diagrama de árbol, tendremos:

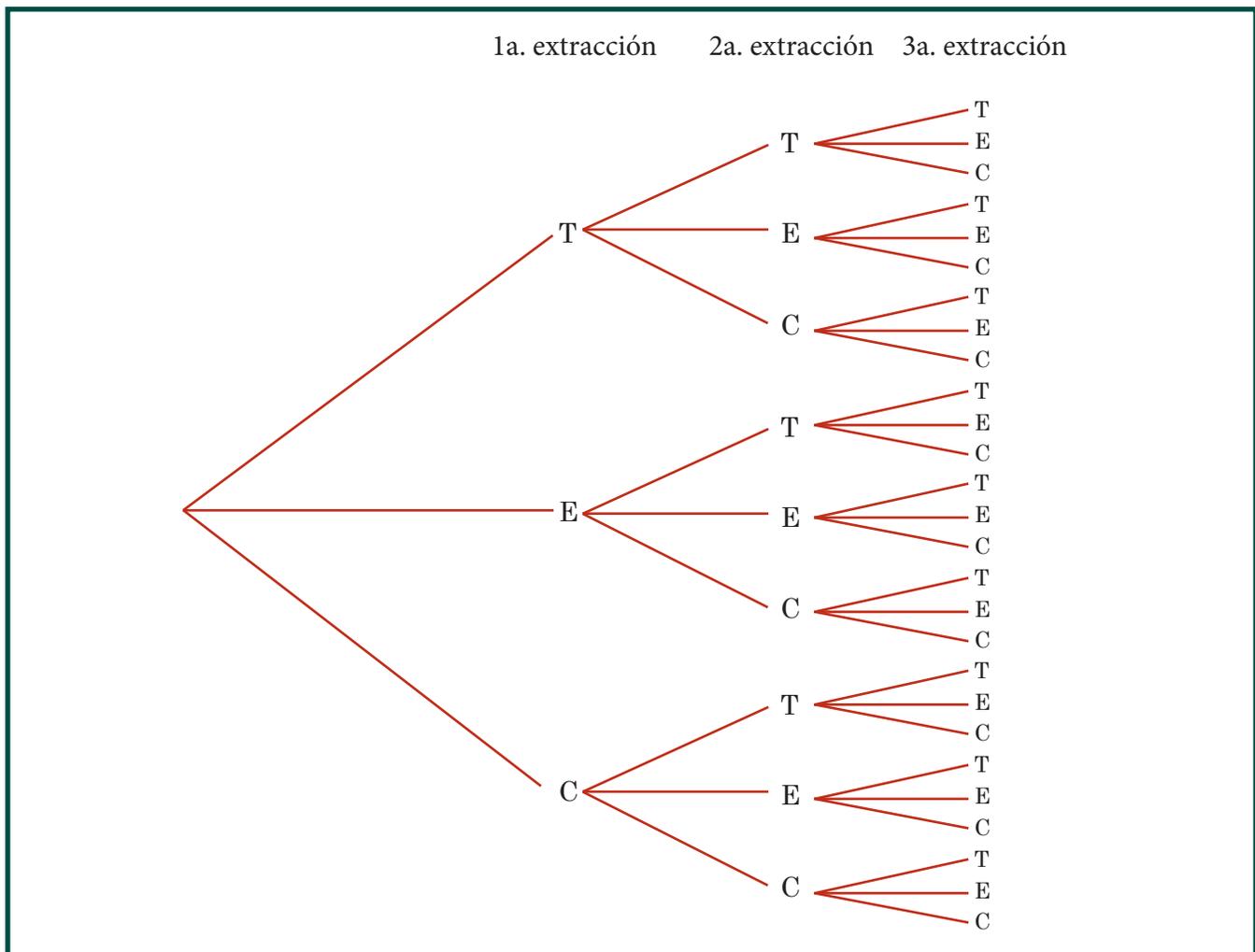


Figura 1.1. Diagrama de árbol para el proceso de extracción de tarjetas.

De acuerdo con este diagrama, hay 27 triadas que corresponden a los resultados totales del proceso aleatorio, antes mencionado.

$$S = \{TTT, TTE, TTC, TET, TEE, TEC, TCT, TCE, TCC, ETT, ETE, ETC, EET, EEE, EEC, ECT, ECE, ECC, CTT, CTE, CTC, CET, CEE, CEC, CCT, CCE, CCC\}$$

El diagrama de árbol es muy didáctico para entender cómo se ordenan los diferentes resultados. Sin embargo, a medida que se va ampliando el número de resultados por ensayo y el número de ensayos, se dificulta su uso, por lo que es mejor recurrir a las fórmulas de contar apropiadas, según sea el caso.

1.4 Definiciones de probabilidad

1.4.1 Definición Clásica de Laplace

Si se tiene un evento E , la probabilidad de que ocurra, se definirá como un cociente o relación entre, los casos que favorecen la ocurrencia de E y los casos totales definidos dentro del espacio muestra:

$$P(E) = \frac{n(E)}{n(S)}$$

Esto, siempre y cuando:

- a) El espacio muestra S sea finito (todos los resultados posibles de un experimento aleatorio).
- b) Los resultados del espacio muestra sean igualmente probables.

1.4.2 Definición Frecuencial de la Probabilidad

Cuando no conocemos a la población de resultados de un proceso aleatorio y trabajamos con una muestra de datos, podemos decir que la probabilidad de ocurrencia de un resultado específico es la frecuencia relativa que presenta este resultado, en el conjunto muestral:

$$f_r(E) = \frac{r}{n}$$

Sin embargo, si aumentamos el tamaño de la muestra o número de ensayos realizados, la frecuencia relativa del evento específico, tiende a regularizarse hasta llegar a una estabilidad de las frecuencias relativas, entonces podremos decir que es prácticamente cierto que:

$$P(E) = \lim_{n \rightarrow \infty} \frac{r}{n}$$

1.4.3 Definición Axiomática de la Probabilidad, de Kolmogorov

La probabilidad, por axioma, está definida matemáticamente dentro del intervalo $[0,1]$, esto quiere decir que si un evento está imposibilitado de ocurrir, en un momento dado, bajo ciertas condiciones, su probabilidad será cero. Mientras que si existen las condiciones para que ocurra, su probabilidad de ocurrencia será diferente de cero pero estará dentro del intervalo antes anotado.

Lo anterior significa que ningún evento tendrá una probabilidad de ocurrencia mayor de uno ni menor de 0.

Entonces la probabilidad es siempre un valor positivo entre 0 y 1.

Además, cuando se adicionan las probabilidades de los diferentes resultados de un espacio muestra, la suma debe ser 1.

1.4.3.1 Axiomas de la Probabilidad

a) La probabilidad, de un evento cualesquiera E :

$$0 \leq P(E) \leq 1$$

b) La probabilidad de todos los resultados de un espacio muestra debe totalizar 1.

$$P(S) = 1$$

c) Si $E_1, E_2, E_3, \dots, E_k$ es un conjunto finito de eventos, mutuamente excluyentes,

$$P(E_1 \cup E_2 \cup E_3 \dots \cup E_k) = \sum_{i=1}^k P(E_i)$$

d) Si E_1, E_2, E_3, \dots es un conjunto infinito de eventos, mutuamente excluyentes,

$$P(E_1 \cup E_2 \cup E_3 \dots) = \sum_{i=1}^{\infty} P(E_i)$$

Con la finalidad de poder realizar cálculos probabilísticos, de manera adecuada, debemos respetar los axiomas y utilizar las definiciones anteriores, dependiendo de las condiciones en que estemos, pero además, debemos aplicar las reglas o leyes que rigen al cálculo de probabilidades.

1.5 Reglas para el cálculo de Probabilidades

1.5.1 Probabilidad de un evento vacío

Si un evento o suceso, no tiene elementos, entonces, $P(\emptyset) = 0$.

1.5.2 Regla de Adición de eventos mutuamente excluyentes

Dos o más eventos son **mutuamente excluyentes cuando** no tienen elementos en común, entonces:

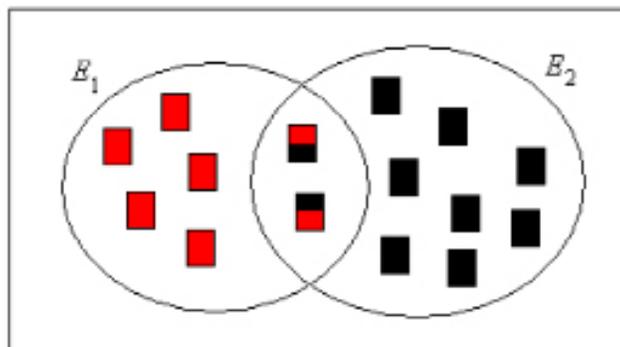
$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

1.5.3 Regla de Adición de eventos que no se excluyen mutuamente

Dos o más eventos no se excluyen mutuamente, cuando tienen elementos en común, por ejemplo, si en una urna hay tarjetas rojas por ambos lados, tarjetas negras por ambos lados y tarjetas con un lado negro y un lado rojo entonces es posible que al extraer una sola tarjeta ésta presente los dos colores, entonces los ensayos no son exclusivos.

De acuerdo con esta situación, si

$$E_1 = \{la\ tarjeta\ sea\ roja\}; E_2 = \{la\ tarjeta\ sea\ negra\} \text{ y } E_1 \cap E_2 = \{la\ tarjeta\ es\ roja\ y\ negra\}$$



Así, para evitar el conteo doble o triple de los elementos que forman intersecciones, se descuentan éstas como se observa en las siguientes definiciones matemáticas de la adición:

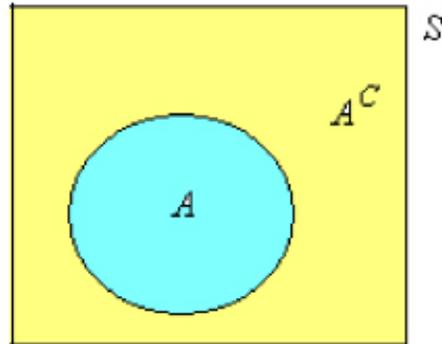
$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

1.5.4 Regla de complementación

Dos eventos son complementarios, si y sólo si se cumple que su intersección es el conjunto vacío y su unión es el Espacio Muestra.

$$P(A^c) = 1 - P(A)$$



Esto significa que, si probabilísticamente es más fácil definir la probabilidad de ocurrencia de A pero deseamos calcular la probabilidad de su complemento, a la unidad, que representa la probabilidad de ocurrencia de todo el espacio muestra se le resta la probabilidad de A y se obtiene la probabilidad del complemento, de manera semejante a la definición del complemento de un conjunto, donde, S es el universo, A es un subconjunto dentro del universo y A^c es el complemento de A .

$$A^c = S - A$$

1.5.5 Regla de la multiplicación de Eventos dependientes

Tomando en cuenta la situación de condicionalidad, si dos eventos ocurren de manera sucesiva y el segundo depende del primero, se define una intersección de los dos eventos que implica la multiplicación. Esto se denota como:

$$P(A \cap B) = P(A)P(B|A) \text{ cuando los eventos son dependientes}$$

El segundo factor de esta definición se denomina probabilidad condicional.

1.5.5.1 Probabilidad Condicional

Se dice que existe una probabilidad condicional cuando al realizar ensayos sucesivos de extracción dentro de una urna, los elementos previamente extraídos no se vuelven a introducir en la urna antes de realizar la extracción siguiente, de tal manera que las probabilidades de los ensayos sucesivos se van modificando porque dependen del resultado en los ensayos antecedentes.

EJEMPLO 1.8. Si de una caja que contiene 10 canicas rojas y 13 azules, se extraen sin reemplazo 3 canicas, ¿cuál es la probabilidad de que las 3 sean rojas?

De acuerdo con los datos: $n_{\text{rojas}}=10$ $n_{\text{azules}}=13$ y $n(S)=23$

Entonces,

$$P(3 \text{ rojas}) = P(\text{primera roja}) \times P(\text{segunda roja}) \times P(\text{tercera roja})$$

$$= P(3 \text{ rojas}) = \left(\frac{10}{23}\right) \left(\frac{9}{22}\right) \left(\frac{8}{21}\right) = \frac{720}{10626} = 0.06776$$

Podemos observar que al ir substituyendo en la fórmula de la definición de probabilidad, la cantidad de rojas disponibles va disminuyendo de una en una y ocurre lo mismo con la cantidad total de canicas en la urna, así la probabilidad de cada elemento sucesivo se va modificando y se deberá entender que la segunda extracción se ve afectada por la primera y que la tercera se ve afectada por las dos primeras, esto es el segundo resultado está condicionado al primero y el tercero estará condicionado a los dos primeros.

Entonces, cuando, el proceso aleatorio ocurre en **2 pasos sucesivos**, de tal manera que **B** sólo ocurrirá, si antes ha ocurrido **A**, se dice que el evento **B** está **condicionado** a la ocurrencia del evento **A**, y la probabilidad de que ocurra B, en estas condiciones se define matemáticamente como:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{ si } P(A) \neq 0$$

Donde:

$P(A)$ es la probabilidad del evento independiente (el que ocurre al iniciar la secuencia).

$P(A \cap B)$ es la probabilidad de la intersección entre los dos eventos.

$P(B|A)$ se lee como la probabilidad de que ocurra **B** si ha ocurrido **A**.

EJEMPLO 1.9. En un concurso realizado por una cadena comercial de alimentos, se colocan 100 latas idénticas de verduras variadas, sin etiquetar, en un contenedor. El juego consiste en que las amas de casa, seleccionadas al azar, tomen 4 latas que les serán regaladas. Se sabe que en el contenedor 25 latas contienen ejotes, 23 contienen granos de maíz, 40 contienen zanahorias y el resto contiene chícharos. Si la señora Suárez es elegida, ¿cuál es la probabilidad de que se lleve una lata de cada tipo de verdura?

De acuerdo con lo establecido en el juego, al tomar las latas en secuencia, el orden está implícito en el proceso, se genera una disminución de casos favorables o de interés y de los casos totales a cada paso del proceso, entonces, calculando la probabilidad pedida tenemos:

$$\begin{aligned} P(1e, 1m, 1z \text{ y } 1ch) &= P(e) \times P(m|e) \times P(z|eym) \times P(ch|e, m \text{ y } z) = \\ &= \left(\frac{25}{100}\right) \left(\frac{23}{99}\right) \left(\frac{40}{98}\right) \left(\frac{12}{95}\right) = \\ &= \frac{276000}{94109400} = 2.9327 \times 10^{-3} \end{aligned}$$

La explicación de este hecho tiene que ver con la definición de probabilidad condicional.

1.5.6 Regla de multiplicación para Eventos Independientes

Si dos o más eventos ocurren de manera independiente, ya sea porque se trabaja con reemplazo, en un espacio muestra único o se trabaja en 2 o más espacios muestra independientes, extrayendo sólo un elemento de cada espacio, la probabilidad de ocurrencia de los ensayos sucesivos no estará condicionada puesto que la cantidad de elementos disponibles permanece constante durante todo el proceso, para cada ensayo particular, entonces, el cálculo consiste en la multiplicación de las probabilidades específicas para cada ensayo.

$$P(A \cap B \cap C) = P(A)P(B)P(C), \text{ cuando los eventos son independientes}$$

EJEMPLO 1.10. Marcos, José y Daniel están compitiendo en tiro al blanco con dardos. Marcos acierta 3 de cada 4 tiros, José da en el blanco en 3 de cada 6 tiros y Daniel acierta 2 de cada 3 tiros. ¿Cuál es la probabilidad de que:

- a) Todos den en el blanco en la siguiente tirada?

Para calcular la probabilidad pedida, usaremos la regla multiplicación de eventos independientes como sigue:

$$P(M \cap J \cap D) = P(M) \cap P(J) \cap P(D) = \left(\frac{3}{4}\right)\left(\frac{3}{6}\right)\left(\frac{2}{3}\right) = \frac{27}{72} = 0.375$$

- b) Daniel y Marcos no acierten pero José sí?

Para resolver este inciso, usaremos primero la regla de complementación de probabilidades para obtener las probabilidades de no acertar de Daniel y Marcos y después la regla de multiplicación de eventos independientes.

$$P(D^c) = 1 - \frac{2}{3} = \frac{1}{3}, \quad P(M^c) = 1 - \frac{3}{4} = \frac{1}{4}, \quad P(J) = \frac{1}{2}$$

$$P(D^c \cap M^c \cap J) = \left(\frac{1}{3}\right)\left(\frac{1}{4}\right)\left(\frac{1}{2}\right) = \frac{1}{24} = 0.04166$$

Utilizando estas reglas de cálculo, las formas de contar apropiadas y la definición de probabilidad, podemos calcular probabilidades asociadas a diversos experimentos aleatorios.

EJEMPLO 1.11. En una caja hay papeletas con 10 diferentes nombres de personas, 6 de mujer y 4 de hombre. Si se extraen 4 papeletas, sin reemplazo, una por una, ¿Cuál es la probabilidad de que:

- a) Los distintos arreglos tengan solamente nombres de mujer.

Para resolver este inciso, podemos hacerlo usando 2 métodos:

- Si usamos el esquema de 4 lugares para ser llenados, obtendremos el total de arreglos para los casos favorables o de interés. Así tenemos, 6 nombres para llenar el primer lugar, 5 nombres para llenar el segundo, etc., por lo que el número de arreglos diferentes donde hay cuatro mujeres es:

$$n(E) = 6 \times 5 \times 4 \times 3 = 360$$

Por el principio de contar eventos que ocurren en secuencia, debemos multiplicar estos valores, lo que nos da un total de 360 casos favorables.

Los casos totales deben incluir todos los nombres, esto es, las 10 papeletas,

$$\text{Casos totales: } n(S) = 10 \times 9 \times 8 \times 7 = 5040$$

Con estos resultados parciales, sustituimos la definición de probabilidad como sigue:

$$P(E) = \frac{n(E)}{n(S)} = \frac{360}{5040} = 0.0714428$$

Nota: Si para resolver el problema, usamos formas de contar, debemos recordar que, las papeletas se extraen una por una y que no hay reemplazo. Esto nos lleva a tomar la decisión de usar permutaciones sin repetición, tanto para calcular casos favorables como casos totales:

Para casos favorables, como sólo 6 de las papeletas tienen nombres de mujer, de éstas se extraerán 4, la permutación será:

$$n(E) = P_4^6 = \frac{6!}{(6-4)!} = \frac{(6)(5)(4)(3)2!}{2!} = 360$$

Mientras que, para calcular casos totales, se tomarán al azar 4 papeletas de las 10 disponibles, que incluyen nombres de hombre y nombres de mujer, la permutación será:

$$n(S) = P_4^{10} = \frac{10!}{(10-4)!} = \frac{(10)(9)(8)(7)6!}{6!} = 5040$$

Resultados que al sustituirse en la definición de probabilidad, generan el mismo resultado, obtenido anteriormente.

$$P(E) = \frac{n(E)}{n(S)} = \frac{P_4^6}{P_4^{10}} = \frac{360}{5040} = 0.071428$$

- b) ¿Cuál es la probabilidad que se obtengan 2 papeletas con nombre de mujer y el resto, con nombre de hombre?

Para resolver este inciso, es necesario darse cuenta de que hay 2 grupos de resultados (H, M) y que las parejas de nombres, correspondientes a ambos géneros, se deben alternar en todos los posibles órdenes. Esto implica que, sólo para casos favorables, habrá que utilizar una permutación sin repetición para definir las diferentes formas en que se ordenan nombres de hombres, otra para definir las formas en que se ordenan nombres de mujeres y además una permutación para definir como se alternan los nombres de ambos géneros, entre sí, y después calcular los casos totales con la permutación adecuada.

Esquematizando la forma de contar el número de arreglos:

Queremos calcular las formas diferentes de ordenar 2 nombres de mujer, de los 6 disponibles:

$$\underline{6} \times \underline{5} = 30$$

Queremos, además, calcular las formas diferentes de ordenar 2 nombres de Hombres, de los 4 disponibles:

$$\underline{4} \times \underline{3} = 12$$

Entonces las diferentes formas de ordenar 2 nombres de mujer y 2 de hombre quedarían así:

$$\underline{6} \times \underline{5} \times \underline{4} \times \underline{3} = 360$$

Para calcular la forma como se alternan 2 nombres de mujer y 2 de hombre en cuatro lugares, tenemos:

Orden	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4	1 2 3 4
	H H M M	H M H M	M M H H	H M M H	M H M H	M H H M

Si contamos las cuartetos de orden diferente, vemos que son 6.

Así, los casos favorables serán el producto de las 2 ordenaciones, mujeres, hombres y de la alternancia entre los conjuntos, 2 femeninos y 2 masculinos, por lo tanto:

$$\text{Casos favorables o de interés} = 30(12)(6) = 2160$$

Y los casos totales, tomando en cuenta a las 10 personas disponibles para ordenar 4, tenemos:

$$\underline{10} \times \underline{9} \times \underline{8} \times \underline{7} = 5040$$

Que es el número total de formas diferentes de ordenar a 4 personas eligiéndolas de un conjunto donde hay 10.

Por lo que la probabilidad se obtiene haciendo el cociente entre casos favorables y casos totales, como sigue:

$$P\{2 \text{ de hombre y } 2 \text{ de mujer}\} = \frac{2160}{5040} = 0.42857$$

Si usamos las fórmulas de permutaciones tendremos:

Permutación para 2 nombres de mujer:

$$P_2^6 = \frac{6!}{(6-2)!} = \frac{(6)(5)4!}{4!} = 6(5) = 30$$

Permutación para 2 nombres de hombre:

$$P_2^4 = \frac{4!}{(4-2)!} = \frac{(4)(3)2!}{2!} = 4(3) = 12$$

Permutación por subconjuntos para definir la forma como se alternan 2 nombres de Mujer y 2 nombres de Hombre:

$$P_{2,2}^4 = \frac{4!}{2!2!} = \frac{4(3)2!}{2!2!} = \frac{4(3)}{2(1)} = \frac{12}{2} = 6$$

Y los casos favorables o de interés serán el producto de estas 3 cantidades:

$$30(12)(6) = 2160$$

Permutación para casos totales, ordenar 4 personas tomándolas de un conjunto donde hay 10:

$$P_4^{10} = \frac{10!}{(10-4)!} = \frac{10(9)(8)(7)6!}{6!} = 10(9)(8)(7) = 5040$$

Substituyendo sobre la definición de probabilidad:

$$P(2 \text{ de hombre y } 2 \text{ de mujer}) = \frac{P_{2,2}^4 P_2^6 P_2^4}{P_4^{10}} = \frac{6(30)(12)}{5040} = \frac{2160}{5040} = 0.42857$$

Comparando las formas de solución del inciso **a** y el inciso **b** podemos ver que en el primero, no se calculó alternancia. Esto se debe a que en el inciso **a**, sólo participa un subconjunto o resultado particular, mujeres, y por lo tanto no hay subconjuntos de resultados diferentes que se puedan alternar. En cambio, en el inciso **b** se piden 2 subconjuntos de resultados, mujeres y hombres, y como son de diferente especie, sí hay la posibilidad de que las ordenaciones de ambos subconjuntos se alternen en el proceso al azar.

Si hacemos un diagrama de árbol, podemos contar también casos favorables y casos totales:

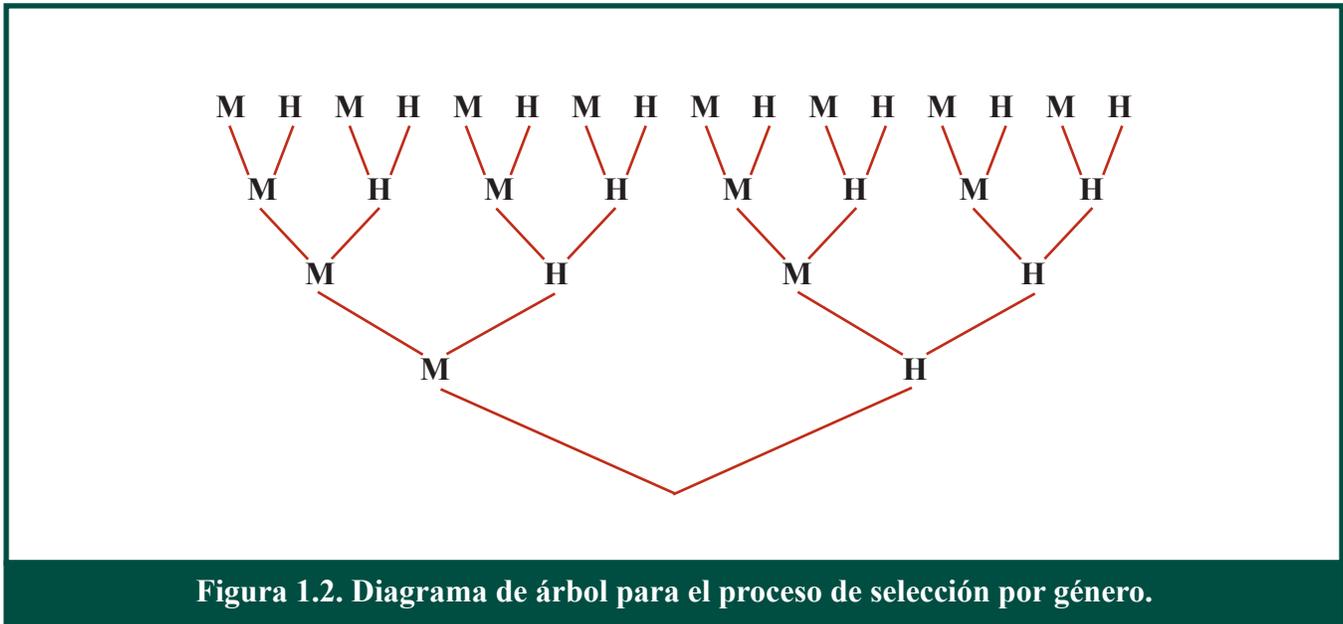


Figura 1.2. Diagrama de árbol para el proceso de selección por género.

Dibujando el diagrama de árbol, podemos obtener todas las ordenaciones de 4 elementos, H, M y definir los conjuntos de 4 en 4, que cumplen con los requisitos pedidos en el inciso **b**: 2 M y 2 H.

$$S = \{MMMM, MMMH, MMHM, MMHH, MHMM, MHMH, MHHM, MHHH, HMMM, HMMH, HMHM, HMHH, HHMM, HHMH, HHHM, HHHH\}$$

De éstos, sólo 6 subconjuntos cumplen con el requisito de 2M y 2H:

$$E_i = \{MMHH, MHMH, MHHM, HMMH, HMHM, HHMM\}$$

Si recordamos que el proceso se está realizando sin reemplazo y aplicando la definición de probabilidades para eventos sucesivos, tenemos:

$$MMHH \rightarrow \frac{6(5)(4)(3)}{10(9)(8)(7)} = \frac{360}{5040}$$

$$HMMH \rightarrow \frac{4(6)(5)(3)}{10(9)(8)(7)} = \frac{360}{5040}$$

$$MHMH \rightarrow \frac{6(4)(5)(3)}{10(9)(8)(7)} = \frac{360}{5040}$$

$$HMHM \rightarrow \frac{4(6)(3)(5)}{10(9)(8)(7)} = \frac{360}{5040}$$

$$MHHM \rightarrow \frac{6(4)(3)(5)}{10(9)(8)(7)} = \frac{360}{5040}$$

$$HHMM \rightarrow \frac{4(3)(6)(5)}{10(9)(8)(7)} = \frac{360}{5040}$$

Podemos ver que se trata de la misma fracción repetida 6 veces, porque representan las 6 diferentes alternativas de orden, entonces el resultado final para la probabilidad será:

$$P(2M \text{ y } 2H) = \left(\frac{360}{5040} \right) (6) = 0.42857$$

Que corresponde a la solución ya establecida en los cálculos anteriores.

EJEMPLO 1.12. El gerente de una pequeña tienda por departamentos, le pide al jefe del departamento de ropa, que elija, a 6 personas dentro de su departamento, para realizar el inventario de temporada. Si este departamento cuenta con 5 personas con carrera comercial, 8 personas con estudios de preparatoria y 10 sólo con estudios de secundaria y la selección se realiza al azar

a) ¿Cuál es la probabilidad de que en el grupo de trabajo queden 2 personas de cada nivel de estudios?

De acuerdo con la pregunta planteada en el problema, se debe hacer una selección y no una ordenación, por lo que se usarán combinaciones, tanto para calcular casos favorables como para calcular casos totales. Como no es un proceso ordenado, no tiene sentido contar los diferentes órdenes en que puede ocurrir.

Cálculo de casos favorables o de interés:

Combinación para seleccionar 2 personas con estudios comerciales:

$$C_2^5 = \frac{5!}{2!(5-2)!} = \frac{5!}{2!3!} = \frac{5(4)3!}{2!3!} = \frac{5(4)}{2(1)} = \frac{20}{2} = 10$$

Combinación para seleccionar 2 personas con estudios de preparatoria:

$$C_2^8 = \frac{8!}{2!(8-2)!} = \frac{8!}{2!6!} = \frac{8(7)6!}{2!6!} = \frac{8(7)}{2(1)} = \frac{56}{2} = 28$$

Combinación para seleccionar 2 personas con estudios de secundaria:

$$C_2^{10} = \frac{10!}{2!(10-2)!} = \frac{10!}{2!8!} = \frac{10(9)8!}{2!8!} = \frac{10(9)}{2(1)} = \frac{90}{2} = 45$$

Cálculo de casos totales:

Combinación para seleccionar 6 personas de un total de 23:

$$C_6^{23} = \frac{23!}{6!(23-6)!} = \frac{23!}{6!17!} = \frac{5(4)3!}{6!17!} = \frac{23(22)(21)(20)(19)(18)17!}{6(5)(4)...(1)17!} = \frac{23(22)...(18)}{6(5)(4)...(1)} = 100947$$

Como la tenemos todos los resultados parciales para casos favorables y para casos totales, sustituimos en la definición de probabilidad.

$$P(2 \text{ personas de cada nivel}) = \frac{n(E)}{n(S)} \Rightarrow \frac{C_2^5 C_2^8 C_2^{10}}{C_6^{23}} = \frac{(10)(28)(45)}{100947} = \frac{12600}{100947} = 0.124818$$

b) ¿Cuál es la probabilidad de que haya al menos 3 de carrera comercial?

La frase, **al menos 3**, significa como **mínimo 3** de esa categoría, lo que implica que dentro del grupo de 6 seleccionados **sea posible hallar, 3, 4 y 5** de carrera comercial. Por lo que, debemos sumar estos resultados parciales, para completar la respuesta al problema planteado.

Tenemos que formar conjuntos de 6 personas, tomando las necesarias de carrera comercial y completando con las personas de los otros niveles de estudio. Con objeto de facilitar la solución al problema, consideraremos al conjunto de elementos con nivel de preparatoria y al conjunto con nivel de secundaria como si fueran un gran conjunto llamado “personas con nivel de estudios no comercial, cuyo número total es de 18 personas, entonces.

Cálculo para casos favorables o de interés.

Tres personas con carrera comercial y 3 con estudios no comerciales:

$$C_3^5 C_3^{18} = \left(\frac{5!}{3!(5-3)!} \right) \left(\frac{18!}{3!(18-3)!} \right) = \left(\frac{5*4*3!}{3!2!} \right) \left(\frac{18*17*16*15!}{3!15!} \right) = \left(\frac{5*4}{2*1} \right) \left(\frac{18*17*16}{3*2*1} \right) = (10)(816) = 8160$$

Cuatro personas con carrera comercial y dos no comerciales:

$$C_4^5 C_2^{18} = \left(\frac{5!}{4!(5-4)!} \right) \left(\frac{18!}{2!(18-2)!} \right) = \left(\frac{5*4!}{4!*1!} \right) \left(\frac{18*17*16!}{2!*16!} \right) = 5* \left(\frac{18*17}{2} \right) = 5*(153) = 765$$

Cinco personas con carrera comercial y una no comercial.

$$C_5^5 C_1^{18} = \left(\frac{5!}{5!(5-5)!} \right) \left(\frac{18!}{1!(18-1)!} \right) = \left(\frac{5!}{5!0!} \right) \left(\frac{18*17!}{1!*17!} \right) = \left(\frac{5!}{5!(1)} \right) (18) = 1*18 = 18$$

Los casos totales los habíamos obtenido en el inciso anterior como:

$$C_6^{23} = \frac{23!}{6!(23-6)!} = \frac{23!}{6!17!} = \frac{(23)(22)(21)(20)(19)(18)(17!)}{(6!)(17!)} = \frac{(23)(22)...(18)}{(6)(5)(4)...(1)} = 100947$$

Entonces, la probabilidad pedida es:

$$P(\text{al menos 3 personas con EC}) = \frac{n(E)}{n(S)} = \frac{C_3^5 C_3^{18} + C_4^5 C_2^{18} + C_3^5}{C_6^{23}} = \frac{8160 + 765 + 18}{100947} = 0.08859$$

c) ¿Cuál es la probabilidad de que se hayan elegido cuando mucho 2 empleados con estudios máximos de secundaria?

Para resolver este inciso, hay que puntualizar que **“cuando mucho 2” significa máximo 2, o lo que es lo mismo, 0, 1 y 2 con secundaria.** Por lo que tendremos que calcular cada uno de estos resultados parciales y sumarlos después. (No debe olvidarse que el conjunto a elegir es de 6).

Casos Favorables o de interés:

Ninguno de secundaria:

$$C_0^{10} C_6^{13} = \left(\frac{10!}{0!(10-0)!} \right) \left(\frac{13!}{6!(13-6)!} \right) = \left(\frac{10!}{0!10!} \right) \left(\frac{13!}{6!7!} \right) = (1) \left(\frac{13*12*11*\dots*7!}{6!*7!} \right) = 1716$$

Uno de secundaria:

$$C_1^{10} C_5^{13} = \left(\frac{10!}{1!(10-1)!} \right) \left(\frac{13!}{5!(13-5)!} \right) = \left(\frac{10!}{1!9!} \right) \left(\frac{13!}{5!*8!} \right) = (10) \left(\frac{13*12*11*\dots*8!}{5!*8!} \right) = (10)(1287) = 12870$$

Dos de secundaria:

$$C_2^{10} C_4^{13} = \left(\frac{10!}{2!(10-2)!} \right) \left(\frac{13!}{4!(13-4)!} \right) = \left(\frac{10!}{2!8!} \right) \left(\frac{13!}{4!9!} \right) = (45) \left(\frac{13*12*11*\dots*9!}{4!9!} \right) = (45)(715) = 32175$$

y los casos totales C_6^{13} cuyo resultado ya teníamos, igual a 100,947.

Entonces haciendo la relación casos favorables entre casos totales tenemos.

$$P(\text{Máximo 2 de secundaria}) = \frac{1716 + 12870 + 32175}{100947} = 0.46322$$

1.6 Problemas adicionales de Aplicación de las Leyes de Probabilidad

EJEMPLO 1.13. Cuando un cliente llega a la caja de pago de una tienda por departamentos, puede usar 4 formas de pago, cheque, tarjeta de crédito, tarjeta de débito o efectivo, las probabilidades asociadas a cada forma de pago son, 0.05, 0.45, 0.15, 0.35, respectivamente. Si un día específico, un cliente llega a la caja, ¿cuál es la probabilidad de que:

- a) Pague con cheque o con tarjeta de crédito.

Para resolver este inciso, usaremos la ley de adición de eventos mutuamente exclusivos porque la suma de los resultados probables asignados a cada forma de pago es 1.

$$P(CH \text{ o } TC) = P(CH \cup TC) = 0.05 + 0.45$$

- b) No pague en efectivo.

Para resolver este inciso, podemos hacer uso de la ley de adición para eventos mutuamente exclusivos, sumando las probabilidades para las modalidades permitidas o usar la ley de complementación, restándole a la probabilidad total, la probabilidad de la forma no permitida.

Por adición:

$$P(\text{no efectivo}) = P(CH) + P(TC) + P(TD) = 0.05 + 0.45 + 0.15 = 0.65$$

Por complementación:

$$P(\text{no efectivo}) = 1 - P(\text{efectivo}) = 1 - 0.35 = 0.65$$

- c) Sólo use tarjetas.

Usando la ley de adición para eventos mutuamente excluyentes, tenemos:

$$P(\text{sólo Tarjetas}) = P(TC \cup TD) = 0.45 + 0.15 = 0.60$$

- d) Pague sus compras usando tarjeta de crédito y efectivo?

En este inciso se pide que use 2 formas de pago, lo que significa que deberán ocurrir ambos tipos de pago, entonces usamos la ley de multiplicación de eventos independientes:

$$P(TC \text{ y } Efec) = P(TC \cap Efec) = (0.45)(0.35) = 0.01575$$

EJEMPLO 1.14. Un juego consiste en sacar 3 pelotas, una detrás de otra, sin reemplazo, de una urna que tiene 10 pelotitas numeradas del cero al nueve. ¿Cuál es la probabilidad de que:

a) Todas tengan número par?

Primero hay que definir los casos favorables dentro del proceso aleatorio y después multiplicar las probabilidades sucesivas. Se usará la ley de eventos dependientes porque las extracciones sucesivas dependerán de los resultados anteriores.

En 10 dígitos hay 5 números pares y 5 impares, el cero se considera par, entonces la probabilidad pedida es:

$$P(3 \text{ con número par}) = \left(\frac{5}{10}\right)\left(\frac{4}{9}\right)\left(\frac{3}{8}\right) = \frac{60}{720} = 0.08333$$

También podemos resolver este inciso usando permutaciones sin repetición:

$$P(3 \text{ con número par}) = \frac{P_3^5}{P_3^{10}} = \frac{\frac{5!}{(5-3)!}}{\frac{10!}{(10-3)!}} = \frac{\frac{(5)(4)(3)(2!)}{2!}}{\frac{(10)(9)(8)(7!)}{7!}} = \frac{60}{720} = 0.08333$$

Aquí no se usó el cálculo de alternancia porque lo que se pidió fue un sólo tipo de resultado.

b) Se obtengan 2 impares y un par?

Este inciso pide 2 tipos de resultados al extraer 3 pelotitas, sin reemplazo, por lo que si se da alternancia de resultados pares e impares.

IPI, IIP, PII

En este diagrama de rayitas, podemos ver que existen tres posiciones diferentes para los pares al ordenarse en 3 lugares, entonces:

$$P(2 \text{ impares y un par}) = \left(\frac{5}{10}\right)\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)(3) = \frac{300}{720} = 0.4166$$

Si usamos formas de contar con orden, tenemos:

Para 2 impares:

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{(5)(4)(3!)}{3!} = 20$$

Para 1 par:

$$P_1^5 = \frac{5!}{(5-1)!} = \frac{(5)(4!)}{4!} = 5$$

Para la alternancia entre pares e impares:

$$P_{2,1}^3 = \frac{3!}{2!1!} = \frac{(3)(2!)}{2!} = 3$$

Para casos totales:

$$P_3^{10} = \frac{10!}{(10-3)!} = \frac{(10)(9)(8)(7!)}{7!} = 720$$

Entonces, substituyendo sobre la definición de probabilidad:

$$P(2 \text{ personas de cada nivel}) = \frac{P_2^5 P_1^5 P_{2,1}^3}{P_3^{10}} = \frac{(20)(5)(3)}{720} = \frac{300}{720} = 0.4166$$

EJEMPLO 1.15. En una escuela de nivel medio, el 55% de los alumnos son del sexo femenino. El 15% de las alumnas están interesadas en estudiar una carrera en ciencias, mientras que de los alumnos sólo el 8% manifestó su deseo de estudiar una carrera en ciencias. Si se elige al azar un alumno de esta escuela,

a) ¿cuál es la probabilidad de que sea uno de los que desean estudiar ciencias?

Para resolver este inciso, primero tenemos que definir nuestro espacio muestra, de acuerdo con los datos del problema. Podemos clasificar los datos por género y por preferencia en el tipo de estudios, usando una tabla de doble entrada. Después, para llenar la tabla, debemos tomar en cuenta que los datos, en las casillas interiores de la misma, son intersecciones entre las clasificaciones, género y carrera, y que las celdas exteriores son las probabilidades marginales o probabilidades definidas sin tomar en cuenta todos los niveles de clasificación. Como aquí sólo hay 2 niveles de clasificación, cada probabilidad marginal sólo puede referirse al género o al tipo de carrera.

Se ubican los porcentajes respectivos de alumnos y alumnas que van a ciencias en la parte interna de la tabla y después, por suma y resta se obtienen los valores faltantes, dado que la tabla debe totalizar una probabilidad de 1, tomando en cuenta totales de fila y totales de columna.

Si el total de mujeres es 55%, (0.55), al restarle el porcentaje que va a ciencias tenemos:

$0.55 - 0.15 = 0.40$, que corresponde a la fracción de mujeres que no va a ciencias.

Hacemos la misma operación para hombres y después completamos a 1.

	Hombres	Mujeres	Total
Ciencias	0.08	0.15	0.23
No Ciencias	0.37	0.40	0.77
Total	0.45	0.55	1

De la tabla podemos ver que $P(\text{ciencias}) = 0.23$.

b) ¿cuál es la probabilidad de que vaya a ciencias, si es mujer?

Esta pregunta se refiere a una probabilidad condicionada. La elección está condicionada al género y se escribe, de acuerdo con la definición de probabilidad condicional, como:

$$P(C|M) = \frac{P(M \cap C)}{P(M)} = \frac{0.15}{0.55} = 0.2727$$

c) ¿cuál es la probabilidad de que sea mujer dado que es de ciencias?

Esta probabilidad también es condicional por lo que se utiliza la misma regla que en el inciso **b**.

$$P(M|C) = \frac{P(C \cap M)}{P(C)} = \frac{0.15}{0.23} = 0.65217$$

d) ¿cuál es la probabilidad de que sea hombre y no desee ciencias?

Podemos ver que esta pregunta se refiere a la intersección entre 2 eventos, entonces, la probabilidad se puede seleccionar directamente de la tabla de doble entrada:

$$P(H \cap \text{no } C) = 0.37$$

e) ¿cuál es la probabilidad de que no desee estudiar ciencias?

Se lee directamente de la tabla, es una probabilidad marginal, complementaria de la respuesta al inciso **a**:

$$P(\text{no } C) = 0.77$$

1.7 Teorema de Bayes

Es una aplicación de la regla de multiplicación de eventos dependientes en donde lo que nos interesa es calcular probabilidades condicionales a partir de probabilidades marginales (son probabilidades que se definen usando sólo algunas de las características, pero no todas las que incluyen los elementos del problema planteado) y probabilidades condicionales cuya especificación esta invertida con respecto a las probabilidades condicionales pedidas y que son datos disponibles en el texto del problema. Con estos datos se desarrolla un diagrama de árbol que permita visualizar el espacio muestra.

Para poder aplicar este teorema, es necesario que los eventos posibles sean mutuamente exclusivos y exhaustivos.

La fórmula de cálculo utilizada se define matemáticamente como sigue:

$$P(A_i | B) = \frac{P(A_i) P(B | A_i)}{\sum_{i=1}^k P(A_i) P(B | A_i)}$$

EJEMPLO 1.16. Una tienda por departamentos, vende televisores de tres marcas. Debido al precio, vende 50% de TV de la marca 1, 30% de la marca 2 y 20% de la marca 3. Todas las marcas ofrecen un año de garantía en refacciones y mano de obra. Se sabe, por experiencia que, de la marca 1 el 25% requiere hacer uso de la garantía, mientras que de las otras dos marcas, los porcentajes de uso de esta garantía son, 20% y 10%, respectivamente.

- a) ¿Cuál es la probabilidad de que una TV, de cualquier marca requiera reparación dentro del tiempo de garantía?

El problema planteado incluye 2 características, la marca y el tener o no garantía, por lo que si nos referimos sólo a la marca, las probabilidades de elegir marca serán marginales y son las adecuadas para iniciar el diagrama que represente el espacio muestra.

Entonces, primero se plantea el problema usando un diagrama de árbol adecuado:

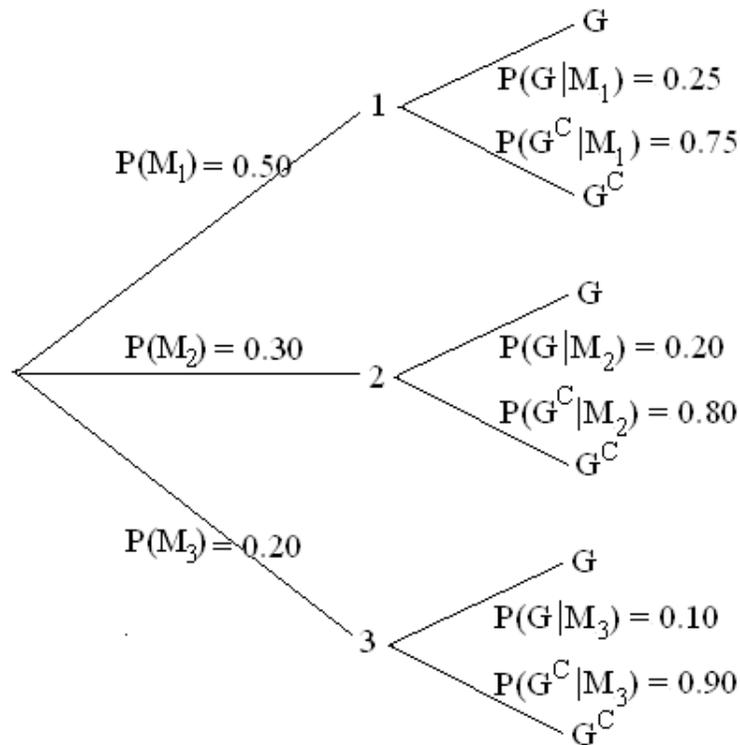


Figura 1.3.- Diagrama de árbol para el problema 1.16

Las tres ramas que inician el árbol, se refieren a las probabilidades marginales de adquirir cualquier marca. Las ramas secundarias, que parten de las primeras, se refieren a probabilidades condicionadas, con respecto a las primeras, para tener garantía o no. Al hacer el producto de cada rama marginal con sus respectivas condicionales obtenemos las intersecciones:

$$P(M_1 \cap G) = P(M_1)P(G | M_1) = (0.5)(0.25) = 0.125$$

$$P(M_1 \cap G^c) = P(M_1)P(G^c | M_1) = (0.5)(0.75) = 0.375$$

$$P(M_2 \cap G) = P(M_2)P(G | M_2) = (0.3)(0.2) = 0.06$$

$$P(M_2 \cap G^c) = P(M_2)P(G^c | M_2) = (0.3)(0.8) = 0.24$$

$$P(M_3 \cap G) = P(M_3)P(G | M_3) = (0.2)(0.1) = 0.02$$

$$P(M_3 \cap G^c) = P(M_3)P(G^c | M_3) = (0.2)(0.9) = 0.18$$

Como todos los resultados son igualmente probables, entonces puede ocurrir cualquiera de ellos. Así, la probabilidad de que se necesite reparación dentro del tiempo de garantía es la suma de todos aquellos que implican uso de garantía:

$$P(G) = P(M_1 \cap G) + P(M_2 \cap G) + P(M_3 \cap G) = 0.125 + 0.06 + 0.02 = 0.205$$

b) ¿Cuál es la probabilidad de que, si usó la garantía, haya comprado una TV de la marca 2.

Para resolver este inciso, se aplica el teorema de Bayes, como sigue:

$$P(M_2 | G) = \frac{P(G \cap M_2)}{P(G)} = \frac{0.06}{0.205} = 0.29268$$

Porque lo que se está calculando es la probabilidad de que si se usó la garantía, la TV haya sido de la marca 2.

Distribuciones de probabilidad

2.1 Conceptos Básicos

2.1.1 Variable Aleatoria

Es una entidad que toma valores al azar, dependiendo del tipo de experimento que se trabaje.

2.1.2 Variable Aleatoria Discreta

Es una entidad que toma valores, de unidad en unidad, porque surge del conteo de los resultados aleatorios que cumplen la característica especificada en el evento solicitado por lo que toma valores enteros pues responde a preguntas como: ¿ Cuántos alumnos en el grupo son de género femenino? ¿ Cuántas pulsaciones por minuto presenta Joel? ¿Cuántas pelotas son rojas?

Por ejemplo:

- El número de profesionales de las diferentes ramas de la ingeniería en una reunión.
- El número de votantes que prefieren al candidato A, por delegación.
- El número de personas que están de acuerdo con diferentes posturas políticas en época de elecciones, etc.

2.1.3 Variable Aleatoria Continua

Es una entidad que toma valores definidos al azar, dentro de un intervalo de la recta numérica, como resultado de la medición de los elementos aleatorios de un experimento por lo que la toma valores dentro del conjunto de los números reales (incluyen números enteros y fracciones).

Por ejemplo:

- La estatura de los alumnos del grupo 3304.
- El diámetro de los tubos de cobre, utilizados en un proyecto.
- El peso de los paquetes de café, llenados automáticamente, en un proceso de producción.

Es importante señalar que para cada valor de una variable aleatoria x siempre habrá un valor de probabilidad $f(x)$.

2.1.4 Distribución de Probabilidad de una variable discreta

Es una tabla o una gráfica o una función matemática, que asocia a cada valor de la variable aleatoria discreta, su probabilidad de ocurrencia.

EJEMPLO 2.1. Se tiran al azar 2 dados diferentes, de 6 caras y los resultados se anotan como la suma de los puntos que caen hacia arriba:

Si los 2 dados son diferentes en color, (negro y verde) se pueden identificar los resultados posibles, por el orden y el color, como sigue:

$$S = \{11, 12, 13, 14, 15, 16, 21, 22, 23, 24, 25, 26, 31, 32, 33, 34, 35, 36, 41, 42, 43, 44, 45, 46, 51, 52, 53, 54, 55, 56, 61, 62, 63, 64, 65, 66\}$$

Entonces los resultados probables se distribuyen de la siguiente manera:

Suma de puntos	2	3	4	5	6	7	8	9	10	11	12
Casos favorables	1	2	3	4	5	6	5	4	3	2	1
Probabilidad	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	4/36	2/36	1/36

Si eliminamos la fila de casos favorables, tendremos una distribución de probabilidad para una variable aleatoria discreta, en forma de tabla

$X_i =$ Suma de puntos	2	3	4	5	6	7	8	9	10	11	12
$P(X_i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	4/36	2/36	1/36

Podríamos graficar los resultados de la distribución discreta:

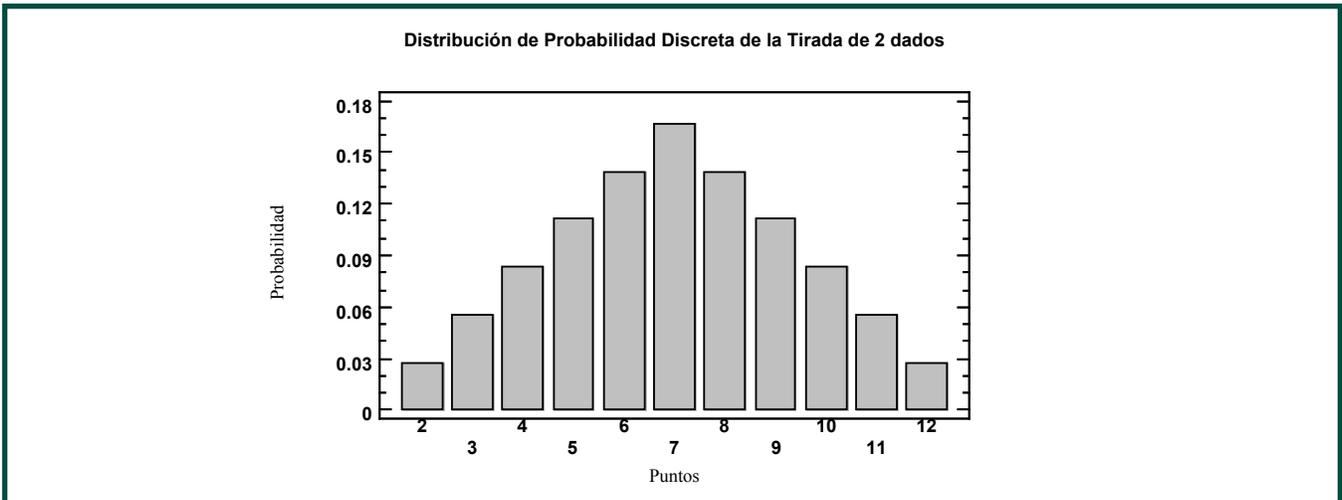


Figura 2.1.- Gráfica de la distribución de la tirada de dos dados.

Donde, los puntos observados marcan las probabilidades asociadas a cada valor de la variable X_i .

La función de probabilidad discreta que define como ocurren los resultados sería:

$$P(X_i) = \frac{\text{Casos favorables}}{\text{Casos totales}}$$

2.1.5 Función de distribución acumulada

Cuando los resultados parciales de un experimento aleatorio se van acumulando, desde el primero hasta el último, la probabilidad total acumulada será 1. Entonces, si deseamos encontrar el valor probabilístico de los resultados iguales o menores que X , tendremos que acumular las probabilidades hasta el límite marcado, entonces:

$$F(x) = P(X \leq x)$$

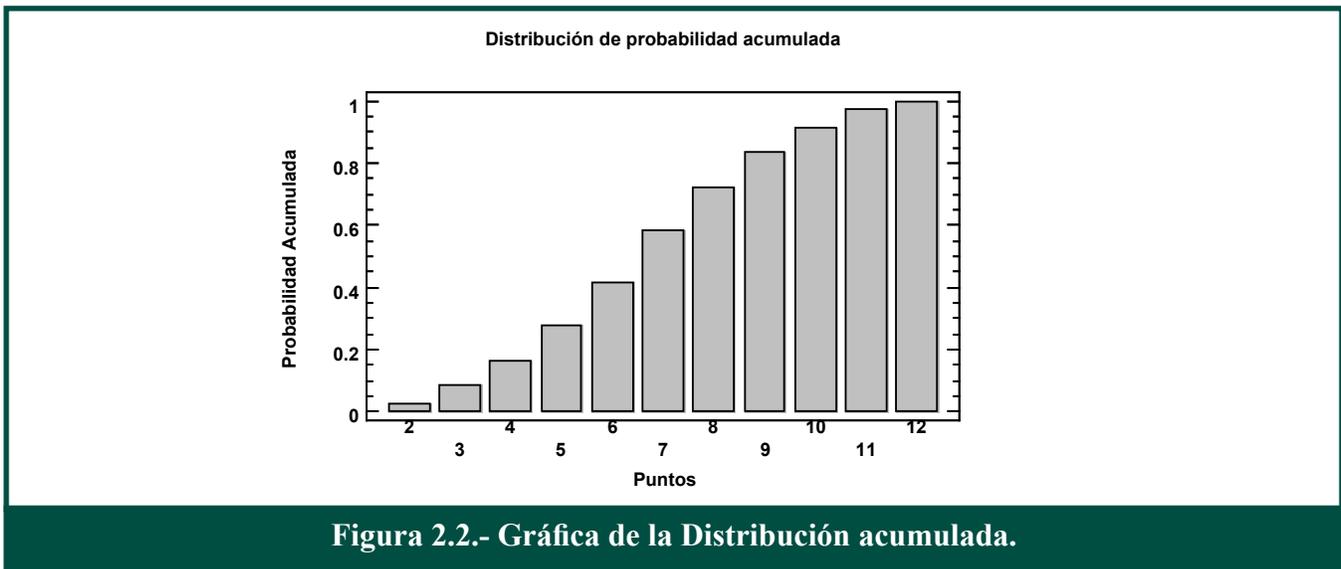
Para cualquier número x , $F(x)$ es la probabilidad de que el valor observado de X sea a lo sumo x .

Si tomamos como base el ejemplo de la tirada de un par de dados, anteriormente mencionada, tendremos:

Suma de puntos	2	3	4	5	6	7	8	9	10	11	12
$P(X)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
$F(X)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36

Al ir acumulando los resultados, desde el primero hasta el último, podemos observar que la última casilla muestra un total de una unidad. Esto se debe a que al área acumulada es unitaria porque representa el 100% de resultados del proceso aleatorio.

La gráfica de una distribución acumulada será una gráfica escalonada y creciente, cuando la distribución es discreta.



Los puntos representan el inicio del escalón correspondiente a cada probabilidad acumulada.

2.1.6 Parámetros de una distribución de probabilidad discreta

Toda distribución de probabilidad presenta características que la definen, esto es, parámetros o medidas importantes que definen el comportamiento de la distribución.

2.1.6.1 Media o Esperanza Matemática de una distribución discreta

Este parámetro nos define el comportamiento promedio o valor esperado de la función discreta que estemos analizando.

$$\mu_x = E(X) = \sum_{i=1}^n (X_i)P(X_i)$$

Tomando como base el ejemplo 3.1, tenemos que la esperanza matemática para la suma de los puntos, sería:

$$\mu_x = E(X) = 2 \left(\frac{1}{36} \right) + 3 \left(\frac{2}{36} \right) + 4 \left(\frac{3}{36} \right) + 5 \left(\frac{4}{36} \right) + 6 \left(\frac{5}{36} \right) + 7 \left(\frac{6}{36} \right) + \dots + 11 \left(\frac{2}{36} \right) + 12 \left(\frac{1}{36} \right) = 7$$

El valor obtenido para la esperanza matemática de la suma de las caras, nos está indicando que el valor más probable o esperado, al lanzar un par de dados, es 7.

2.1.6.2 Varianza de una distribución discreta

Es la variación o dispersión cuadrática que presenta una distribución de probabilidad discreta, se define, matemáticamente como sigue:

$$\sigma^2 = \sum_{i=1}^n (X_i - \mu)^2 P(X_i)$$

Con objeto de facilitar su cálculo, podemos usar la definición modificada, basada en la esperanza del valor cuadrático de la variable manejada: $E(X^2)$:

$$\sigma^2 = E(X^2) - \mu^2$$

Donde, $E(X^2) = \sum_{i=1}^n X_i^2 P(X_i)$ y μ es la esperanza de la variable X o Media.

Si aplicamos esta definición al ejemplo de la tirada de un par de dados, tendremos:

$$E(X^2) = (2^2) \left(\frac{1}{36} \right) + (3^2) \left(\frac{2}{36} \right) + (4^2) \left(\frac{3}{36} \right) + (5^2) \left(\frac{4}{36} \right) + \dots + (11^2) \left(\frac{2}{36} \right) + (12^2) \left(\frac{1}{36} \right) = 54.8333$$

Sustituyendo en la definición:

$$\sigma^2 = E(X^2) - \mu^2 = 54.8333 - 7^2 = 5.8333$$

Teniendo la varianza podemos obtener la desviación estándar de la distribución, que es una medida de la variación lineal:

$$\sigma = +\sqrt{\sigma^2} = \sqrt{5.8333} = 2.4152$$

Lo que significa, que las sumas más probables, al tirar 2 dados estarán entre:

$$\mu \pm \sigma = 7 \pm 2.4152 \Rightarrow \text{que las sumas más probables estarán entre 4 y 10}$$

Entonces, tratando de definir el comportamiento del proceso aleatorio, “suma obtenida en la tirada de un par de dados”, diremos que el valor más probable es el siete y que como la dispersión es de 2.4152 unidades, se abarcarían resultados de sumas, desde 4 hasta 110.

2.2 Modelos de Distribución Discreta

Dentro de las distribuciones discretas se encuentran algunas de gran importancia para el cálculo de probabilidades, debido a que existe una gran gama de experimentos aleatorios que se comportan de una manera característica, que se apega a un modelo específico. Por esta razón, se han generado los modelos matemáticos que explican el comportamiento probabilístico de la mayoría de los procesos al azar más comunes: Binomial, Poisson, etc.

2.2.1 Distribución de probabilidad de variable aleatoria Binomial

Dentro de las distribuciones más utilizadas está la Distribución Binomial. Se llama así porque sus resultados se distribuyen de acuerdo con el desarrollo de un binomio a la potencia n (binomio de Newton)

Un proceso aleatorio es **Binomial** si cuenta con las siguientes características:

- Hay n ensayos finitos en el proceso.
- Cada ensayo es independiente.
- La variable manejada es discreta, con x número de éxitos en los n ensayos.
- Para cada ensayo sólo hay 2 resultados probables: Éxito y Fracaso
- Los ensayos se realizan con reemplazo.
- El orden en que ocurren los resultados es importante.
- La probabilidad de éxito p , en un ensayo, es constante a lo largo del proceso y es dato.
- La probabilidad de fracaso q , es complementaria a la probabilidad de éxito, así:

$$p + q = 1 \Leftrightarrow q = 1 - p$$

Por lo anterior, siempre que un proceso aleatorio tenga estas características se le llamará proceso Binomial y se resolverá de acuerdo con la función matemática que lo define:

$$P(b, x; n, p) = C_x^n p^x q^{n-x}$$

Donde:

b, indica que es un proceso binomial.

n, es el total de ensayos que se realizarán.

p, es la probabilidad asignada al éxito, en el proceso.

q , es la probabilidad de fracaso en el proceso.

x , es el número de éxitos que se desea ocurran en el proceso.

$n-x$ es el número de ensayos no exitosos en el proceso.

$$C_x^n = \frac{n!}{x!(n-x)!}$$

EJEMPLO 2.2. En una población determinada, la probabilidad de encontrar personas con cabello rubio, es de 45%. Se hace un muestreo en esta población, eligiendo al azar a 10 personas. Se considera éxito encontrar una persona con cabello rubio. ¿Cuál es la probabilidad de que:

a) Al menos 5 tengan cabello rubio?

Al menos 5, significa que 5 o más tengan cabello rubio, por lo tanto:

$$P(\text{al menos } 5) = P(x \geq 5) = P(x = 5) + P(x = 6) + P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10)$$

Entonces, aplicando la función binomial tenemos:

$$P(x = 5) = C_5^{10}(0.45)^5 (0.55)^5 = 0.2340$$

$$P(x = 6) = C_6^{10}(0.45)^6 (0.55)^4 = 0.15956$$

$$P(x = 7) = C_7^{10}(0.45)^7 (0.55)^3 = 0.0746$$

$$P(x = 8) = C_8^{10}(0.45)^8 (0.55)^2 = 0.02289$$

$$P(x = 9) = C_9^{10}(0.45)^9 (0.55)^1 = 0.004162$$

$$P(x = 10) = C_{10}^{10} (0.45)^{10} (0.55)^0 = 0.00034$$

Sumando las probabilidades obtenidas tenemos el resultado deseado:

$$P(\text{al menos } 5 \text{ con cabello rubio}) = 0.4955$$

b) a lo más 3 tengan cabello rubio?

A lo más 3 significa máximo 3, por lo tanto:

$$P(\text{a lo más } 3) = P(0)+P(1)+P(2)+P(3)$$

Aplicando la función, tenemos:

$$P(x = 0) = C_0^{10} (0.45)^0 (0.55)^{10} = 0.002533$$

$$P(x = 1) = C_1^{10} (0.45)^1 (0.55)^9 = 0.0207$$

$$P(x = 2) = C_2^{10} (0.45)^2 (0.55)^8 = 0.0763$$

$$P(x = 3) = C_3^{10} (0.45)^3 (0.55)^7 = 0.16648$$

Sumando las probabilidades respectivas, tenemos:

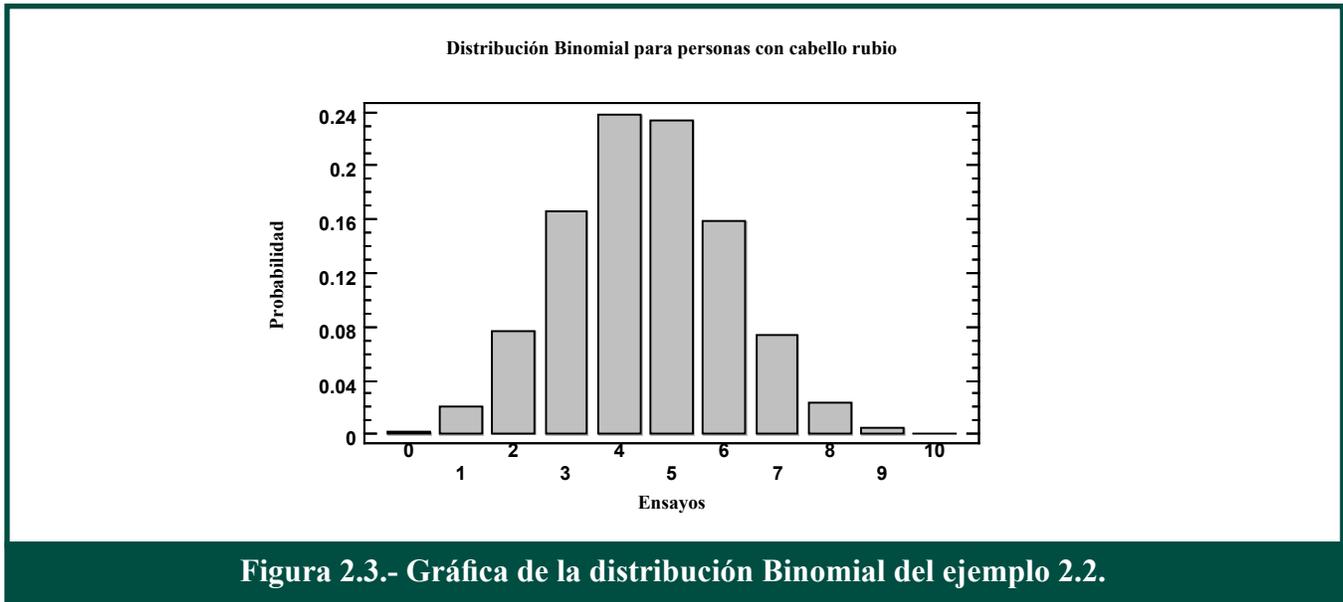
$$P(\text{a lo más } 3) = 0.266013$$

Existen tablas para la distribución binomial, en donde ya se encuentran los resultados acumulados de las probabilidades de, x éxitos en n ensayos, para determinados valores de la probabilidad de éxito y del número de ensayos realizados, por lo que cuando los datos estén disponibles en tablas, pueden tomarse los valores de ellos sin tener que sustituir la función binomial, sobre todo cuando el intervalo de cálculo tiene varios términos. Ver Tabla T2, páginas 207 a 215 del Cuaderno de Problemas de Probabilidad y Estadística, de Guerra Dávila T. Marques Dos Santos M. J. y López Reynoso Jorge M., UNAM, FES Zaragoza, 2009.

c) sólo 4 tengan cabello rubio?

$$P(\text{sólo } 4) = C_4^{10} (0.45)^4 (0.55)^6 = 0.2384$$

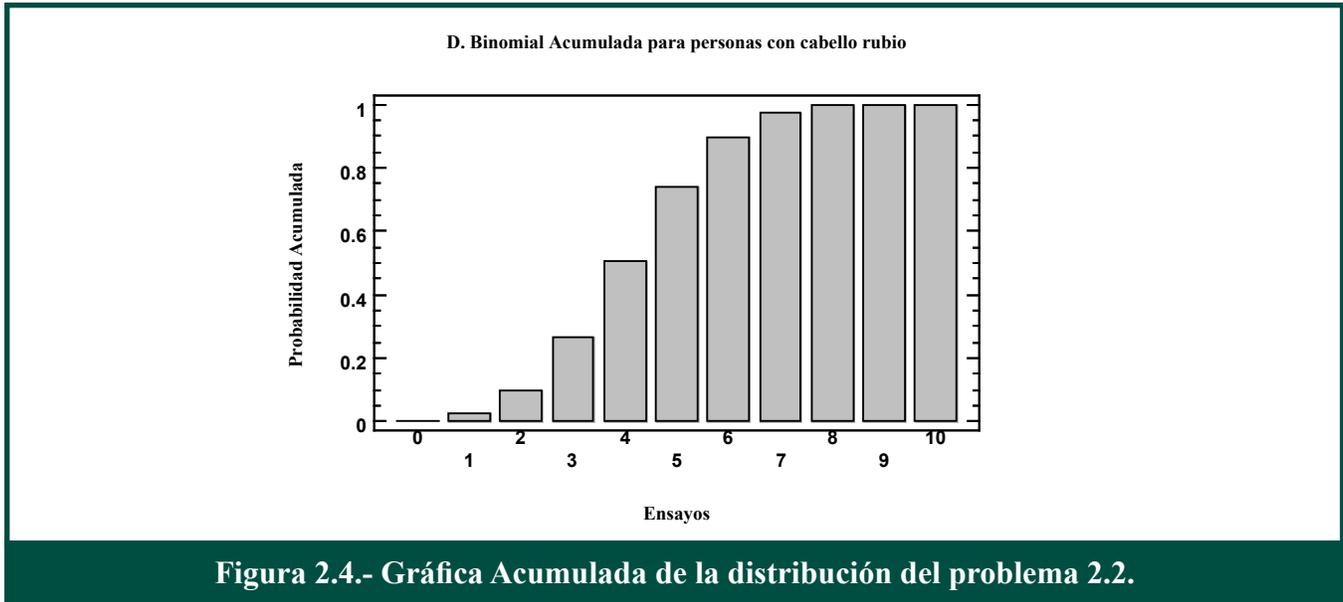
Si quisiéramos representar gráficamente el proceso, tendríamos que tener todas las probabilidades de ocurrencia de personas con cabello rubio y graficarlas contra el valor de la variable X .



También podríamos tener la función acumulada del proceso, $F(X)$:

X	$P(X)$	$F(X)$
0	0.00253	0.00253
1	0.02070	0.02323
2	0.07630	0.09953
3	0.16640	0.26600
4	0.23840	0.50430
5	0.23400	0.73830
6	0.15960	0.89790
7	0.07460	0.97250
8	0.02289	0.99540
9	0.00416	0.999580
10	0.00034	0.999920

Cabe hacer notar, que si se usaran todas las cifras decimales, el resultado final sería exactamente 1.



Cuando ya se sabe que la distribución es binomial, podemos calcular los parámetros de la distribución como sigue:

$$\mu = np = (10)(0.45) = 4.5$$

$$\sigma^2 = npq = (10)(0.45)(0.55) = 2.475$$

$$\sigma = \sqrt{npq} = \sqrt{2.475} = 1.5732$$

2.2.2 Distribución de probabilidad de variable aleatoria de Poisson

La distribución de Poisson, es muy utilizada para resolver procesos probabilísticos que tienen como característica principal, definir la ocurrencia de hechos poco comunes, en donde la probabilidad de ocurrencia en un intervalo dado es sumamente pequeña y el número de ensayos es muy grande.

Un proceso aleatorio es **Poisson** si cumple con las siguientes características:

- **Hay un número de ensayos que tiende a infinito en el proceso, esto es **n** es muy grande.
- Cada ensayo es independiente.
- La variable manejada es discreta.
- Para cada ensayo sólo hay 2 resultados posibles, éxito y fracaso pero sólo nos interesan los éxitos.
- Los ensayos se realizan con reemplazo.
- El orden en que ocurren los resultados es importante.

- Dentro del intervalo en el que sucede del experimento, ocurre un promedio de éxitos λ , que es proporcional al número de intervalos o a la longitud del intervalo si se trata de tiempo, longitud, área, etc.

Nota: Las características marcadas con asteriscos son las que hacen la diferencia respecto a la distribución Binomial.

$$P(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

Donde:

λ , es la media de la distribución Poisson.

x , es el número de éxitos deseados.

e , es la base de los logaritmos naturales.

EJEMPLO 2.3. Suponga que el número X de tornados observados en una región particular, durante un período de un año, tiene una distribución de Poisson con $\lambda = 8$. Calcule la probabilidad de que el número de tornados sea:

- a) como máximo 5 en un año dado.
- b) entre 6 y 9 inclusive en un año cualquiera.
- c) de al menos 7 en el próximo año.
- d) Si se observara el fenómeno durante 2 años, cuál sería la media y la varianza del mismo?
- e) Si se observara el fenómeno durante 2 años, ¿cuál sería la probabilidad de que hubiera exactamente 15 tornados?

De acuerdo con los datos del problema, podemos darnos cuenta de que se cuenta con un valor medio de tornados por año, y esto nos lleva directamente a las características de un proceso de Poisson. (Los tornados son independientes, la variable es discreta, la probabilidad de ocurrencia de tornados es muy pequeña, en general, se considera un éxito la ocurrencia de tornado, etc.), entonces el problema se resuelve utilizando un modelo de Poisson.

$P\{X \text{ sea como máximo } 5\}$ se resolvería así:

$$P(X \text{ sea como máximo } 5) = P(X \leq 5) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5)$$

Entonces:

$$P(0) = \frac{e^{-\lambda} \lambda^0}{0!} = \frac{e^{-8} 8^0}{0!} = \frac{(3.3546 \times 10^{-4})(1)}{1} = 3.3546 \times 10^{-4}$$

$$P(1) = \frac{e^{-\lambda} \lambda^1}{1!} = \frac{e^{-8} 8^1}{1!} = \frac{(3.3546 \times 10^{-4})(8)}{1} = 2.6837 \times 10^{-3}$$

$$P(2) = \frac{e^{-\lambda} \lambda^2}{2!} = \frac{e^{-8} 8^2}{2!} = \frac{(3.3546 \times 10^{-4})(64)}{2} = 0.0107348$$

$$P(3) = \frac{e^{-\lambda} \lambda^3}{3!} = \frac{e^{-8} 8^3}{3!} = \frac{(3.3546 \times 10^{-4})(512)}{6} = 0.02863$$

$$P(4) = \frac{e^{-\lambda} \lambda^4}{4!} = \frac{e^{-8} 8^4}{4!} = \frac{(3.3546 \times 10^{-4})(4096)}{24} = 0.05725$$

$$P(5) = \frac{e^{-\lambda} \lambda^5}{5!} = \frac{e^{-8} 8^5}{5!} = \frac{(3.3546 \times 10^{-4})(32768)}{120} = 0.09160$$

La probabilidad para este inciso es la suma de las probabilidades anteriores:

$$P(X \leq 5) = 0.191234$$

Para resolver el inciso anterior, también podemos hacer uso de las tablas de la distribución Poisson, acumulando las probabilidades de 0 a 5 anotadas en la columna con $\lambda = 8$. Ver Tabla T3, páginas 216 a 218 del Cuaderno de Problemas de Probabilidad y Estadística, de Guerra Dávila T., Marques Dos Santos M. J. y López Reynoso J. M., UNAM, FES Zaragoza, 2009.

a) $P(\text{entre } 6 \text{ y } 9 \text{ inclusive})$ se resuelve de la manera siguiente:

$$P(\text{entre } 6 \text{ y } 9 \text{ inclusive}) = P(6 \leq X \leq 9) = P(6) + P(7) + P(8) + P(9)$$

$$P(6) = \frac{e^{-\lambda} \lambda^6}{6!} = \frac{e^{-8} 8^6}{6!} = \frac{(3.3546 \times 10^{-4})(262144)}{720} = 0.122138$$

$$P(7) = \frac{e^{-\lambda} \lambda^7}{7!} = \frac{e^{-8} 8^7}{7!} = \frac{(3.3546 \times 10^{-4})(2097152)}{5040} = 0.139587$$

$$P(8) = \frac{e^{-\lambda} \lambda^8}{8!} = \frac{e^{-8} 8^8}{8!} = \frac{(3.3546 \times 10^{-4})(16777216)}{40320} = 0.139587$$

$$P(9) = \frac{e^{-\lambda} \lambda^9}{9!} = \frac{e^{-8} 8^9}{9!} = \frac{(3.3546 \times 10^{-4})(134217728)}{362880} = 0.124077$$

Sumamos las probabilidades parciales y obtenemos el resultado deseado:

$$P(6 \leq X \leq 9) = 0.525388$$

b) $P(\text{al menos } 7)$ se resuelve de acuerdo con el planteamiento siguiente.

$$P(\text{al menos } 7) = P(X \geq 7) = P(7) + P(8) + P(9) + \dots + P(\infty)$$

Como no se conoce el valor de ∞ (es un número indeterminado), no es conveniente calcular la probabilidad pedida, por adición directa. La forma más fácil es trabajar con la regla de complementación de probabilidades, entonces:

$$P(\text{al menos } 7) = P(X \geq 7) = 1 - [P(6) + P(5) + P(4) + P(3) + P(2) + P(1) + P(0)]$$

Por lo tanto:

$$P(\text{al menos } 7) = P(X \geq 7) = 1 - (0.122138 + 0.0916 + 0.05725 + \dots + 3.3546 \times 10^4)$$

$$P(\text{al menos } 7) = P(X \geq 7) = 1 - 0.313372 = 0.686628$$

c) Media y Varianza cuando se realiza la observación en 2 años:

Para resolver este inciso, de acuerdo a la forma en que se definen los parámetros de esta distribución, tenemos que tomar en cuenta que, al modificar el lapso de observación del fenómeno, también se modifica, proporcionalmente la media y la varianza del experimento Poisson, por lo que al duplicar el lapso de observación, también se duplica la media y por lo tanto la varianza.

Así, si $\lambda = 8$ en un año, en dos años $\lambda = (2)(8) = 16$ y como $\sigma^2 = \lambda$, entonces $\sigma^2 = 16$.

d) $P(X=15 \text{ en dos años})$ se resolverá tomando en cuenta la siguiente que se ha modificado el lapso de observación y por lo tanto, se debe trabajar con la media modificada:

$$P(X=15) = \frac{e^{-16} 16^{15}}{15!} = \frac{(1.125351747 \times 10^{-7})(1.15292155 \times 10^{18})}{1.307674368 \times 10^{12}} = 0.099217$$

2.2.2.1 Aproximación del proceso Binomial con la distribución de Poisson

Cuando un proceso aleatorio se comporta como binomial pero el tamaño de la muestra n es muy grande y p pequeña, se dificulta el cálculo de la combinación C_r^n por falta de capacidad de la calculadora, entonces es conveniente hacer uso de la distribución de Poisson. Para que el cálculo sea adecuado y lo más próximo al del modelo binomial, es necesario que $n \geq 20$ y $p \leq 0.05$ o si $n \geq 100$ la aproximación es muy buena siempre y cuando la media $\mu = n \times p \leq 10$.

EJEMPLO 2.4. En un instituto de educación media superior la probabilidad de obtener una beca de estudios es de 0.025 debido a la limitación de los recursos. Si 170 de ellos solicitan una beca, ¿Cuál es la probabilidad de que:

- ¿Como máximo 10 consigan la beca?
- ¿No más de 5 reciban la beca?
- ¿Exactamente 100?

Solución:

- a) Usaremos la aproximación de la Poisson a la Binomial

$$n = 170 \quad p = 0.025 \quad \mu = n \times p = 170 \times 0.025 = 4.25 \quad x \leq 10$$

Para resolver este inciso es necesario sumar todos los términos sustituidos de la distribución de Poisson desde $x=0$ hasta $x=10$ como sigue:

$$P(x \leq 10, \lambda = 4.25) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{4.25^0 e^{-4.25}}{0!} + \frac{4.25^1 e^{-4.25}}{1!} + \dots + \frac{4.25^{10} e^{-4.25}}{10!} = 0.995566$$

- b) Para resolver este inciso es necesario sumar los términos de la distribución Poisson desde $x=0$ hasta $x=5$ como sigue:

$$P(x \leq 5, \lambda = 4.25) = \frac{4.25^0 e^{-4.25}}{0!} + \frac{4.25^1 e^{-4.25}}{1!} + \frac{4.25^2 e^{-4.25}}{2!} + \frac{4.25^3 e^{-4.25}}{3!} + \frac{4.25^4 e^{-4.25}}{4!} + \frac{4.25^5 e^{-4.25}}{5!} = 0.74493$$

- c) Exactamente 70 obtengan beca:

$$P(x = 70) = \frac{4.25^{70} e^{-4.25}}{70!} = 1.156 e^{-58}$$

2.2.3 Distribución de probabilidad de variable aleatoria Hipergeométrica

La distribución Hipergeométrica se utiliza para calcular probabilidades cuando el proceso aleatorio consiste en una selección de elementos sin reemplazo.

Un proceso aleatorio es Hipergeométrico si cumple con las siguientes características:

- El número de elementos que participan en el proceso aleatorio es finito.
- Sólo hay 2 resultados, esto es, se dicotomizan los resultados: los elementos con la característica deseada (favorables), y lo que no presentan tal característica.
- El proceso termina cuando se han seleccionado todos los elementos deseados.
- **El proceso no incluye el orden como cambio de resultado.
- **No hay reemplazo.
- **No existen probabilidades asociadas a los 2 resultados, definidas desde el principio, sólo hay cantidades de elementos que cumplen o no una característica específica.

Nota: Las características marcadas con asteriscos son las que hacen la diferencia respecto a las distribuciones antecedentes

La función que define a esta distribución es la siguiente:

$$P(N, K; n, x) = \frac{C_x^K C_{n-x}^{N-K}}{C_n^N}$$

Donde:

x = número de elementos favorables en la muestra.

K = número de elementos favorables en la población

N = número de elementos totales en la población

$N - K$ = número de elementos no favorables en la población

N = tamaño de la muestra

EJEMPLO 2.5. En un grupo de alumnos de la carrera de Ingeniería Química hay 60 personas de las cuales, 22 llevaron un curso propedéutico de matemáticas. ¿Cuál es la probabilidad de que:

- En una muestra de 20 alumnos de este grupo, ocho hayan llevado el curso propedéutico?
- Al menos 4 de los 20 elegidos no hayan llevado el curso propedéutico.

Para resolver este problema se hace uso de la distribución Hipergeométrica.

- En este inciso, los casos favorables se refieren al hecho de haber llevado el curso propedéutico, entonces, sustituyendo la función:

$$P(60, 22; 20, 8) = \frac{C_8^{22} C_{20-8}^{60-22}}{C_{20}^{60}} = \frac{C_8^{22} C_{12}^{38}}{C_{20}^{60}} = 0.206537$$

b) Para este inciso la característica favorable es no haber cursado propedéutico, entonces:

$$\begin{aligned} &P(\text{al menos 4 no hayan llevado propedéutico}) = \\ &1 - [P(0, 1, 2, 3 \text{ no hayan llevado propedéutico})] = \\ &= 1 - \left[\frac{C_0^{38} C_{20}^{22}}{C_{20}^{60}} + \frac{C_1^{38} C_{19}^{22}}{C_{20}^{60}} + \frac{C_2^{38} C_{18}^{22}}{C_{20}^{60}} + \frac{C_3^{38} C_{17}^{22}}{C_{20}^{60}} \right] = 1 - 5.4237 \times 10^{-8} = 0.9999 \end{aligned}$$

Los parámetros del proceso Hipergeométrico se calculan de la siguiente manera:

Media o Esperanza Matemática:

$$\mu_H = n \left(\frac{K}{N} \right)$$

Varianza:

$$\sigma_H^2 = n \left(\frac{K}{N} \right) \left(\frac{N-K}{N} \right) \left[\frac{N-n}{N-1} \right]$$

Desviación Estándar:

$$\sigma_H = \sqrt{\left(n \left(\frac{K}{N} \right) \left(\frac{N-K}{N} \right) \left(\frac{N-n}{N-1} \right) \right)}$$

Los parámetros para este ejemplo son:

$$\mu_H = n \left(\frac{K}{N} \right) = 20 \left(\frac{22}{60} \right) = 7.333$$

$$\sigma_H^2 = n \left(\frac{K}{N} \right) \left(\frac{N-K}{N} \right) \left[\frac{N-n}{N-1} \right] = 20 \left(\frac{22}{60} \right) \left(\frac{38}{60} \right) \left(\frac{60-20}{59} \right) = 3.149$$

$$\sigma_H = \sqrt{n \left(\frac{K}{N} \right) \left(\frac{N-K}{N} \right) \left(\frac{N-n}{N-1} \right)} = \sqrt{(20) \left(\frac{22}{60} \right) \left(\frac{38}{60} \right) \left(\frac{60-20}{59} \right)} = 1.774$$

2.2.4 Distribución de probabilidad de variable aleatoria de Pascal y Distribución Geométrica

La distribución de Pascal, también llamada binomial negativa, se utiliza para calcular la probabilidad de que el último resultado exitoso, en un proceso aleatorio, ocurra en un ensayo X , determinado.

Un proceso aleatorio es de Pascal cuando cumple las características siguientes:

- **El número de ensayos realizados X , es variable.
- La variable manejada es discreta
- Hay orden en el proceso
- Hay reemplazo
- Hay 2 resultados por ensayo: Éxito y Fracaso
- La probabilidad p , de éxito en un ensayo es constante y es un dato conocido
- La probabilidad de fracaso q , es complementaria a la de éxito
- Los ensayos son independientes
- **El proceso termina cuando se ha logrado el último éxito pedido.

Nota: Las características marcadas con asteriscos son las que hacen la diferencia respecto a las distribuciones antecedentes

La función de Pascal se denota como:

$$P(x, k, p) = C_{k-1}^{x-1} p^k q^{x-k}$$

Donde:

x = el número de ensayos necesarios para alcanzar el último éxito.

k = Número total de éxitos deseados en el experimento

p = probabilidad de éxito

q = probabilidad de fracaso

EJEMPLO 2.6. Entre los profesionales de las áreas de ingeniería, el 8% corresponde a los ingenieros químicos. Con base en lo anterior, suponiendo selecciones aleatorias:

- a) ¿Cuál es la probabilidad de que sea necesario elegir 20 profesionales de estas áreas para encontrar el segundo ingeniero químico?
 b) ¿Cuál es la probabilidad de que sea necesario seleccionar a 14 ingenieros para encontrar un ingeniero químico?

a) Para resolver este inciso se hace uso de la distribución de Pascal donde los valores de las variables son:

$$x = 20; k = 2; p = 0.08$$

Entonces:

$$P(20, 2, 0.08) = C_{2-1}^{20-1} (0.408)^2 (0.92)^{20-2} = C_1^{19} (0.408)^2 (0.92)^{20-2} = 0.027109$$

b) Para este inciso se utiliza la distribución Geométrica, que es la un caso particular de la distribución de Pascal cuando sólo se requiere un éxito en el proceso aleatorio:

La función Geométrica es:

$$P(x, 1, p) = C_{1-1}^{x-1} p^1 q^{x-1} = C_0^x p q^{x-1} = p q^{x-1}$$

La combinación cero de cualquier número siempre es 1, por eso desaparece el término $C_0^x = 1$.

Entonces:

$$x = 14; k = 1; p = 0.08$$

$$P(14, 1, 0.08) = (0.08)(0.92)^{14-1} = 0.02706$$

Los parámetros de la distribución de Pascal son:

Media o Esperanza Matemática: $\mu_p = \frac{k}{p}$

Varianza: $\sigma_p^2 = \frac{kq}{p^2}$

Desviación Estándar: $\sigma_p = \sqrt{\frac{kq}{p^2}}$

Para este ejemplo, los parámetros serían:

$$\mu_p = \frac{2}{0.08} = 25$$

$$\sigma_p^2 = \frac{2(0.92)}{0.08^2} = 287.5$$

$$\sigma_p = \sqrt{\frac{2(0.92)}{0.08^2}} = 16.9558$$

Los parámetros de la distribución geométrica son:

$$\text{Media: } \mu_G = \frac{1}{p}$$

$$\text{Varianza: } \sigma_G^2 = \frac{q}{p^2}$$

$$\text{Desviación Estándar: } \sigma_G = \sqrt{\frac{q}{p^2}}$$

Para este ejemplo, los parámetros serían:

$$\mu_G = \frac{1}{0.08} = 12.5$$

$$\sigma_G^2 = \frac{0.92}{0.08^2} = 143.75$$

$$\sigma_G = \sqrt{\frac{0.92}{0.08^2}} = 11.9896$$

2.2.5 Distribución de probabilidad Multinomial

Las características que definen un proceso aleatorio multinomial son las siguientes

- Hay n ensayos finitos en el proceso.
- Cada ensayo es independiente.
- Las variables manejadas son discretas.
- **Existen más de 2 resultados por ensayo
- Los ensayos se realizan con reemplazo.
- El orden en que ocurren los resultados es importante.
- **Las probabilidades de ocurrencia de cada diferente resultado son conocidas y constantes
- **La suma de las probabilidades de los diferentes resultados debe ser 1
- **La suma de los diferentes resultados en el proceso, debe ser igual a n

Nota: Las características marcadas con asteriscos son las que hacen la diferencia respecto a las distribuciones antecedentes.

La función multinomial se escribe como:

$$P(x_1, x_2, \dots, x_k, p_1, p_2, \dots, p_k, n) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

Donde:

n = Total de ensayos realizados.

x_i = Número de resultados específicos deseados.

p_i = Probabilidades específicas para cada resultado deseado.

EJEMPLO 2.7 En el proceso de fabricación de recipientes de aluminio, se pueden presentar 3 tipos de defectos: coloración opaca 0.45%, paredes rayadas 0.33% y laminado delgado 0.28%. Si en el departamento de control de calidad se analiza una muestra aleatoria de 20 recipientes, ¿cuál es la probabilidad de que se encuentren:

- 2 recipientes con coloración opaca, uno con paredes rayadas y 3 con laminado delgado?
- Dos con cada tipo de defecto?

Para resolver problemas con la distribución multinomial debe recordarse que las p_i deben sumar uno y que la suma de las x_i debe ser igual a n .

a) Para este inciso, de la muestra de 20 se piden 2 con coloración opaca, uno con paredes rayadas y 3 con laminado delgado, que en total suman 6 lo que implica que 14 recipientes no presentan ningún tipo de defecto. Por otro lado, la suma de las probabilidades de defecto es de 0.0106 por lo que la probabilidad de no tener defecto es el complemento para 1, esto es:

$$P(\text{no defecto}) = 0.9894$$

Entonces, sustituyendo en la función multinomial tenemos:

$$\begin{aligned} P(2, 1, 3, 14, 0.0045, 0.0033, 0.0028, 0.9894, 20) &= \\ &= \frac{20!}{2!1!3!14!} (0.0045)^2(0.0033)^1(0.0028)^3(0.9894)^{14} = 2.9387 \times 10^{-9} \end{aligned}$$

b) Ahora, se piden dos de cada tipo de defecto entonces:

$$\begin{aligned} P\{2, 2, 2, 14, 0.0045, 0.0033, 0.0028, 0.9894, 20\} &= \frac{20!}{2!2!2!14!} (0.0045)^2(0.0033)^2(0.0028)^2(0.9894)^{14} = \\ &= 5.1952 \times 10^{-9} \end{aligned}$$

Como el modelo multinomial incluye más de 2 resultados por ensayo, los parámetros de la distribución deben obtenerse para cada resultado particular como en el caso de la binomial. Se considerará $P(\text{éxito})$ la del resultado específico de interés y $P(\text{fracaso})$ el complemento con respecto a 1.

2.3 Modelos de Distribución Continua

2.3.1 Distribución de Probabilidad de una Variable Continua

Cuando se define una función de densidad de probabilidad, hay que integrarla entre límites específicos para obtener la función acumulada $F(X)$ que define un área bajo la curva igual a uno. Esto es:

$$F(X) = \int_a^b f(x)dx = 1, \text{ si } a < x < b$$

Entonces, calcular la probabilidad de ocurrencia de un fenómeno aleatorio de variable continua implica integrar entre límites establecidos para la función densidad de probabilidad.

Las distribuciones continuas, son descritas en su comportamiento, por los parámetros.

2.3.2 Parámetros de una distribución continua de probabilidad

2.3.2.1 Media o Esperanza Matemática.

$$\mu = E(x) = \int_a^b (x)f(x)dx, \text{ cuando } a < x < b$$

2.3.2.2 Varianza.

$$\sigma^2 = \int_a^b (x-\mu)^2 f(x) = E(x^2) - \mu^2 = \int_a^b x^2 f(x)dx - \mu^2$$

EJEMPLO 2.8. Un maestro universitario, nunca termina su clase antes de que suene la campana y siempre termina su clase a lo más 2 minutos después de que suena la campana. Sea X el tiempo que transcurre entre el toque de la campana y el término de la clase. Suponga que la función de densidad de probabilidad de la variable X es:

$$f(x) = \begin{cases} kx^2, & 0 \leq x \leq 2 \\ 0, & \text{de otra manera} \end{cases}$$

- Encuentre el valor de k .
- ¿Cuál es la probabilidad de que la clase termine como máximo un minuto después de que suene la campana.

- c) ¿Cuál es la probabilidad de que la clase continúe entre 60 y 90 segundos después de que suene la campana?
- d) ¿Cuál es el tiempo medio para que termine la clase después de que suena la campana?
- e) ¿Cuál es la varianza del tiempo para terminar la clase después de que suena la campana?

Solución

- a) Para encontrar el valor de k , debemos recordar que la integral de la función debe dar 1, entonces, integramos en todo el dominio de la función e igualamos a 1.

$$f(x) = \begin{cases} kx^2, & 0 \leq x \leq 2 \\ 0, & \text{de otra manera} \end{cases} \Rightarrow F(x) = k \int_0^2 x^2 dx = 1 \Rightarrow F(x) = k \left(\frac{x^3}{3} \Big|_0^2 \right) = 1$$

$$\Rightarrow k \left(\frac{8}{3} - 0 \right) = 1 \Rightarrow k \left(\frac{8}{3} \right) = 1, \text{ despejando } k, \text{ se tiene } k = \frac{3}{8}$$

Entonces la función densidad de probabilidad es:

$$f(x) = \begin{cases} \frac{3}{8} x^2, & 0 \leq x \leq 2 \\ 0, & \text{de otra manera} \end{cases}$$

- b) $P(\text{la clase termine como máximo un minuto después de que suene la campana}).$

Para resolver este inciso, tenemos que integrar la función entre 0 y 1 inclusive:

$$\frac{3}{8} \int_0^1 x^2 dx = \frac{3}{8} \left(\frac{x^3}{3} \Big|_0^1 \right) = \frac{3}{24} = \frac{1}{8}$$

Así, la probabilidad de que la clase termine en como máximo un minuto es $\frac{1}{8} = 0.125$

- c) $P(\text{clase continúe entre 60 y 90 seg. después del toque}).$

Para resolver este inciso, debemos integrar la función entre 1 y 1.5, entonces:

$$P(60 \leq X \leq 90) = \frac{3}{8} \int_1^{1.5} x^2 dx = \frac{3}{8} \left(\frac{x^3}{3} \Big|_1^{1.5} \right) = (0.421875 - 0.125) = 0.296875$$

d) Tiempo medio después de que suena la campana:

Para resolver este inciso, tenemos que utilizar la definición de esperanza matemática de una función continua:

$$\mu = E(x) = \int_0^2 x f(x) dx = \frac{3}{8} \int_0^2 x(x^2) dx = \frac{3}{8} \int_0^2 x^3 dx = \frac{3}{8} \left(\frac{x^4}{4} \Big|_0^2 \right) = (1.5 - 0) = 1.5$$

e) Para calcular la varianza en el tiempo después de que suena la campana, utilizamos la definición de varianza como sigue.

$$\sigma^2 = E(x^2) - \mu^2$$

Primero calculamos la esperanza de las x^2 :

$$E(x^2) = \frac{3}{8} \int_0^2 x^2(x^2) dx = \frac{3}{8} \int_0^2 x^4 dx = \frac{3}{8} \left(\frac{x^5}{5} \Big|_0^2 \right) = (2.4 - 0) = 2.4$$

Sustituyendo en la fórmula de la varianza, se tiene:

$$\sigma^2 = 2.4 - 1.5^2 = 2.4 - 2.25 = 0.15$$

2.3.3 Distribución Normal

La distribución Normal es una de las distribuciones más importantes en estadística, ya que muchas poblaciones numéricas tienen distribuciones normales o se pueden ajustar con mucha aproximación mediante una curva normal. Aún cuando la distribución fundamental sea discreta, la curva normal proporciona, con frecuencia una excelente aproximación. Además, cuando las variables individuales no están normalmente distribuidas, en condiciones apropiadas, las sumas y promedios de las variables, tendrán aproximadamente una distribución normal.

Una variable aleatoria continua presenta una distribución normal si la función de densidad de probabilidad está definida como sigue:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}, \quad -\infty < x < \infty$$

Para calcular las probabilidades se tendrían que calcular las áreas bajo esta curva, sin embargo esta función no es integrable por los métodos de integración usuales sino por métodos numéricos.

Las curvas generadas por la función para cada par (μ, σ) tendrán forma de campana y serán simétricas con respecto a la media que estará en el punto central de la campana, al igual que la mediana. La desviación estándar, σ , es la distancia desde la media a los puntos de inflexión de la curva (los puntos donde hay cambio de concavidad).

2.3.4 Distribución Normal Estándar

Como ninguna de las técnicas de integración usuales se puede emplear para evaluar la expresión que define a la curva normal, es conveniente definir la curva normal estándar mediante la introducción de un eje relativo Z , en donde se considera que los valores de μ, σ con valores específicos sobre el eje X , corresponderán al 0 y 1 respectivamente, medidos sobre este eje relativo Z . El modelo estandarizado de la normal tampoco es integrable por los métodos usuales. Sin embargo, se utiliza porque permite leer las áreas bajo la curva usando valores relativos Z aun cuando los pares de valores μ, σ cambien. La función matemática para la normal estándar quedaría así:

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad -\infty < z < \infty$$

Donde z queda entonces definida como:

$$z = \frac{x - \mu}{\sigma}$$

que se conoce como fórmula de estandarización de los valores x .

Para calcular las áreas bajo la curva normal, es necesario localizar los valores estandarizados z , en las tablas probabilísticas de la curva normal estándar, Tabla T-4, páginas 219 a 226 del Cuaderno de Problemas de Probabilidad y Estadística, de Guerra Dávila T. Marques Dos Santos, M.J. y López Reynoso J. M., UNAM, FES Zaragoza, 2009.

Nota: Existen calculadoras y paquetes estadísticos que dan las áreas bajo la curva normal para cualquier par. (μ, σ)

EJEMPLO 2.9. La presión de aire de un neumático, seleccionado al azar, instalado en un automóvil nuevo, está distribuida normalmente con valor de 31 lb/in² y desviación estándar de 0.2 lb/in².

- a) ¿Cuál es la probabilidad de que la presión de un neumático, seleccionado al azar, exceda 30.5 lb/in²?

- b) ¿Cuál es la probabilidad de que la presión de un neumático, seleccionado al azar, se encuentre entre 30.5 y 31.5 lb/in²?
- c) ¿Cuál es la probabilidad de que la presión de un neumático, seleccionado al azar, esté entre 31.2 y 31.4 lb/in²?
- d) Suponga que un neumático se considera con presión baja si está debajo de 30.4 lb/in². ¿Cuál es la probabilidad de que al menos uno de los 4 neumáticos de un automóvil se encuentre bajo?

Solución:

a) $P(x > 30.5)$

Para resolver este inciso, primero se estandariza el valor x substituyendo la fórmula:

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{30.5 - 31}{0.2} = -2.5$$

Se localiza el valor 2.5 en las tablas de la normal, y construimos el modelo adecuado. Note que el valor de z es negativo. Esto implica que el límite inferior del área pedida se encuentra localizado a la izquierda de la media, en -2.5 del eje Z y que corresponde al límite mínimo de presión 30.5. El área solicitada es la que aparece sombreada.

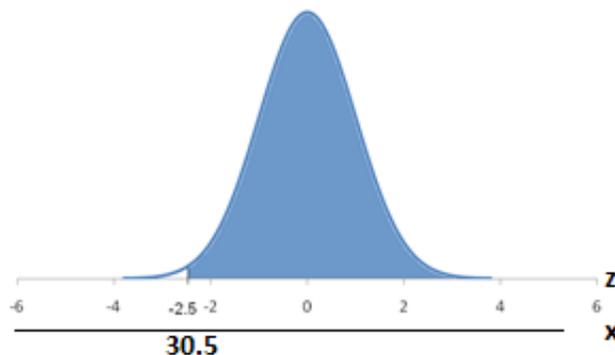


Figura 2.5. Área bajo la curva para presión mayor a 30.5.

El modelo nos muestra el área sombreada que corresponde a la probabilidad solicitada. Usando las tablas de la normal, la columna E (para áreas mayores al 50%) se obtiene la probabilidad deseada:

$$P(x > 30.5) = 0.9938$$

b) Para resolver este inciso, es necesario estandarizar los 2 valores límites del área pedida, entonces:

$$P(30.5 < x < 31.5) = P\left(\frac{30.5 - 31}{0.2}\right) < z < \left(\frac{31.5 - 31}{0.2}\right) = P(-2.5 < z < 2.5)$$

Esto implica que el área solicitada es la que se encuentra entre -2.5 y +2.5 del eje Z y que corresponde a los límites mínimo 30.5 y máximo 31.5 de la presión del neumático.

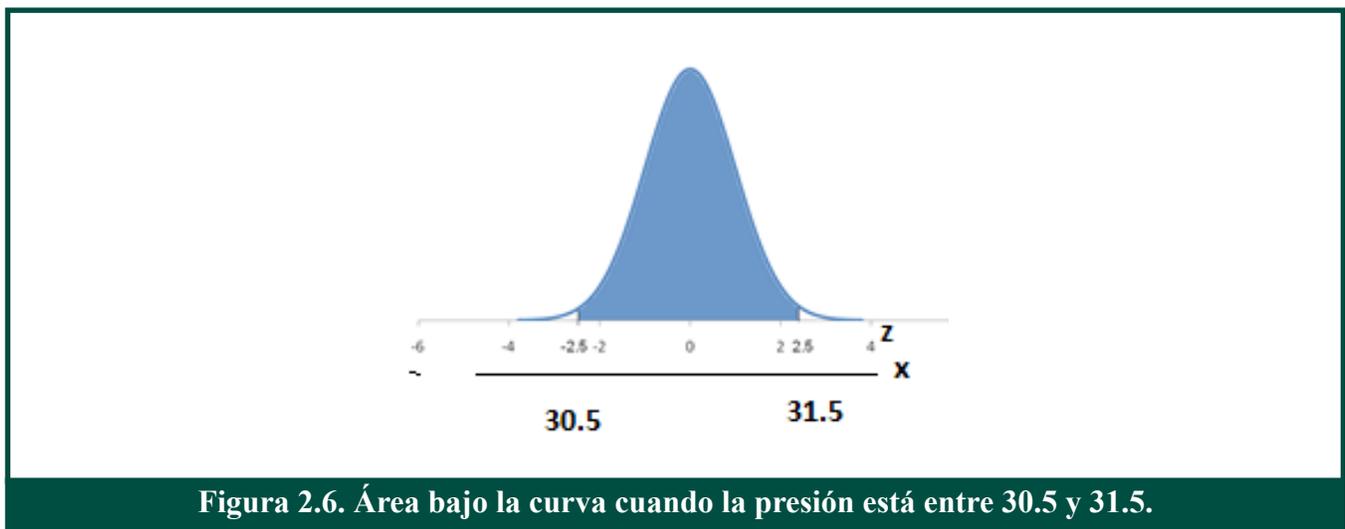


Figura 2.6. Área bajo la curva cuando la presión está entre 30.5 y 31.5.

Así, el área pedida se lee en la columna C de las tablas, y su valor es 0.9872.

c) En este inciso se pide que el área seleccionada esté entre 31.2 y 31.4 lb/in² entonces:

$$P(32 < x < 34) = \left(\frac{31.2 - 31}{0.2}\right) < z < \left(\frac{31.4 - 31}{0.2}\right) = P(1 < z < 2)$$

Como ambos valores Z son positivos, se encuentran a la derecha de la media, es decir, el límite 31.2 corresponde al valor 1 de Z y el límite 31.4 corresponde al valor 2 de Z, como se muestra en el modelo.

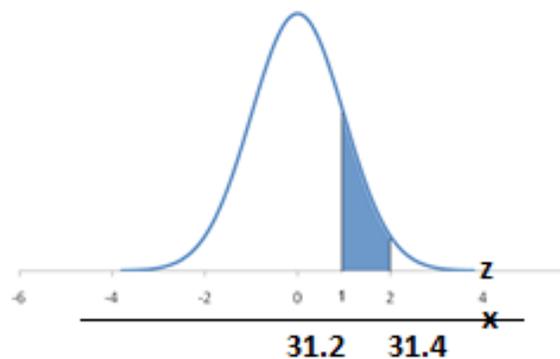


Figura 2.7 Área bajo la curva cuando la presión está entre 31.2 y 31.4.

Entonces, es necesario leer dos veces la tabla de áreas bajo la curva normal, en la columna A (para 2 y para 1) y restar las áreas respectivas:

$$A_{requerida} = A_{mayor} - A_{menor}$$

$$\text{Área}_{z=2} - \text{Área}_{z=1} = 0.4772 - 0.3413 = 0.1359$$

- d) P (al menos 1 con presión baja), para resolver este inciso, primero calculamos la probabilidad normal de tener un neumático con presión baja.

$$P(\text{neumático con presión baja}) = P(\text{presión menor que } 30.4) = P(x < 30.4) =$$

$$= P\left(z < \frac{30.4 - 31}{0.2}\right) = P(z < -3)$$

Recuerde que la presión baja corresponde a un área menor o igual a 30.4, entonces, el área bajo la curva normal, corresponde a la cola izquierda a partir de -3 y queda como sigue:

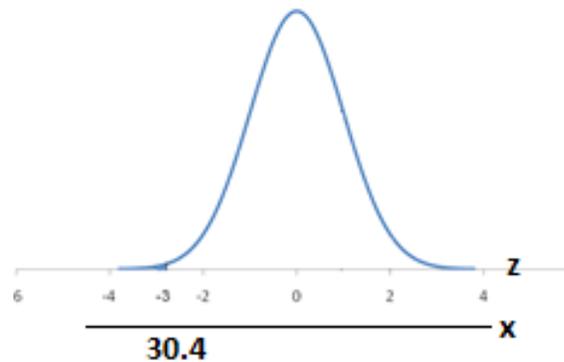


Figura 2.8.- Área bajo la curva cuando la presión es menor a 30.4.

El área es el espacio pequeño a la izquierda de la marca colocada en -3 y corresponde a la probabilidad de encontrar un neumático con presión baja, leyendo el área en la columna B, con $z = 3$, tenemos:

$$P(\text{neumático con presión baja}) = 0.0013$$

Ahora que ya tenemos la probabilidad de tener un neumático con presión baja, utilizamos este valor como probabilidad de éxito para sustituir en la distribución Binomial con $n=4$ neumáticos y calcular la probabilidad solicitada.

$$P(\text{al menos 1 con presión baja}) = P(1) + P(2) + P(3) + P(4) = 1 - P(0)$$

De acuerdo con el planteamiento, nos conviene calcular la probabilidad de ninguno con presión baja y restarlo de 1.

$$P(x = 0) = C_0^4 (0.0013)^0 (0.99879)^4 = 0.995169$$

$$P(\text{al menos 1 con presión baja}) = 1 - P(0) = 1 - 0.995169 = 0.0048312$$

Como puede verse, la probabilidad de al menos un neumático con presión baja es muy pequeña.

2.3.5 Aproximación Normal a la Distribución Binomial

Cuando un proceso aleatorio, de variable discreta, presenta todas las características de un proceso Binomial, pero el número de ensayos es muy grande y el intervalo de cálculo es también muy grande, es conveniente utilizar los parámetros de la distribución Binomial y sustituirlos en la fórmula de estandarización de la normal, con objeto de estimar el área bajo la curva o probabilidad para todo el intervalo de valores de la variable discreta. Al aceptar este método de cálculo aproximado de la probabilidad real, es necesario

recordar que la variable normal es continua, por lo que es necesario “convertir” los valores discretos a valores continuos midiéndolos como intervalos y no como puntos. Por ejemplo, si la variable discreta tiene valor 1, viéndolo como área, sería el espacio comprendido entre 0.5 y 1.5; si se desea calcular la probabilidad entre 20 y 35 inclusive, tendríamos que calcular el área comprendida entre 19.5 y 35.5, como valor continuo. Esto es, darle continuidad a un valor discreto es sumar o restar 0.5 al valor discreto, según sea el caso y estandarizar estos límites continuos.

Estandarizar una variable discreta implica, utilizar la fórmula de z modificada incluyendo los parámetros de la Binomial:

$$z = \frac{(x \pm 0.5) - np}{\sqrt{npq}}$$

El valor obtenido de z , se lee en tablas de la Normal, de acuerdo con el área sombreada en el modelo construido.

EJEMPLO 2.10. La probabilidad de que un infante empiece a mudar los dientes a los 6 años de edad es de 80 %. Con base en esta información, ¿cuál es la probabilidad de que:

- en una muestra de 30 infantes que ya han cumplido 6 años, encontremos 15 que ya están mudando los dientes?
- si la muestra es de 300 infantes que ya han cumplido 6 años, encontremos al menos 220 que ya están mudando los dientes?
- si la muestra es de 250, encontremos entre 180 y 215 infantes inclusive, que ya están mudando los dientes?

Solución

- Para resolver este inciso, no necesitamos aproximar con la normal porque el proceso es Binomial con un número de ensayos finito y podemos resolverlo directamente como sigue:

$$b(15; 30, 0.8) = C_{15}^{30}(0.8)^{15}(0.2)^{15} = 1.7884 \times 10^{-4}$$

$$P(\text{al menos } 220, \text{ de } 300) = P(x \geq 220; 300, 0.8)$$

En este caso, es necesario aproximar la solución usando la distribución normal, calculando la media y la desviación estándar como sigue:

$$\mu = np = (300)(0.8) = 240 \quad \text{y} \quad \sigma = \sqrt{npq} = \sqrt{(300)(0.8)(0.2)} = 6.9282$$

Este planteamiento nos está indicando que debemos calcular desde 220 hasta 300 como valores discretos. Pero si usamos la aproximación normal, tendremos que buscar el área cuyo límite inferior comienza en 219.5, entonces:

$$P(x \geq 220) \approx P\left(z > \frac{219.5 - 240}{6.9282}\right) = P\left(z > \frac{-217.1}{6.9282}\right) = P(z > -2.96) = 0.9985$$

Entonces el modelo Normal queda así:

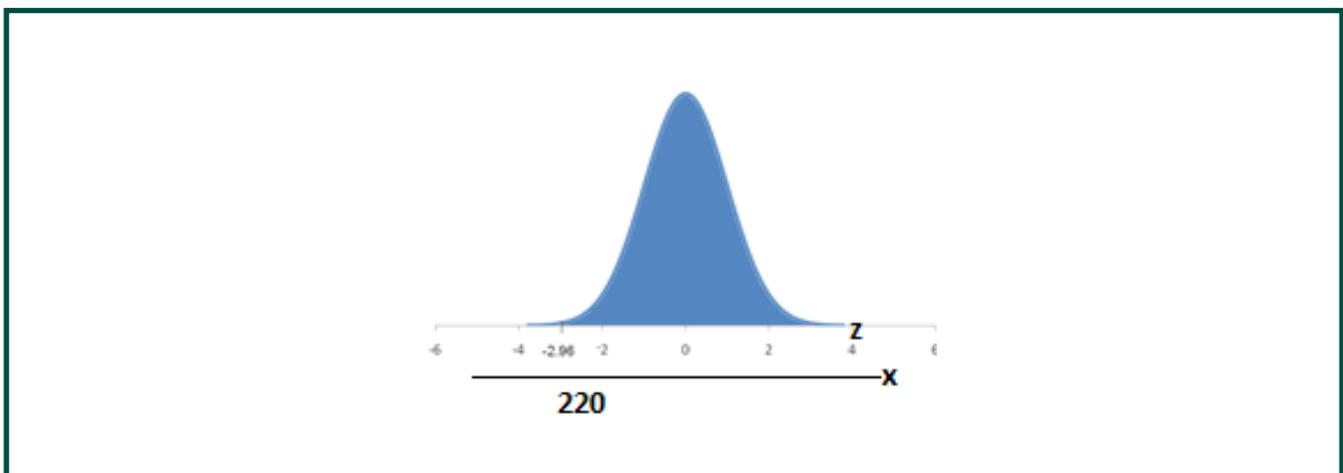


Figura 2.9.- Área bajo la curva correspondiente a la probabilidad de encontrar un número de niños igual o mayor que 220, que están mudando los dientes.

En donde está la línea colocada en $z > -2.96$, estará el valor de la variable discreta, X , correspondiente a 220. Entonces el área pedida es la de la derecha, la que está sombreada. Por lo tanto, buscaremos el área correspondiente en la columna E, con z de 2.96 y tenemos que:

$$P(x \geq 220) \approx 0.9985$$

- b) Para este inciso, también aproximamos pero debemos sustituir para 2 valores Z y calcular una nueva media y desviación estándar:

$$\mu = np = (250)(0.8) = 200 \quad \text{y} \quad \sigma = \sqrt{npq} = (250)(0.8)(0.2) = 6.32455$$

$$P(180 < x < 215) \approx P\left(\frac{179.5-200}{6.32455} < z < \frac{215.5-200}{6.32455}\right) = P(-3.24 < z < 2.45)$$

De acuerdo con los valores z , el área estará en la parte central del modelo, como sigue:

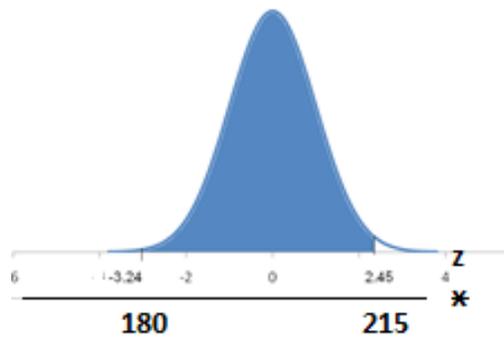


Figura 2.10.- Área bajo la curva correspondiente a la probabilidad de encontrar entre 180 y 215 niños que estén mudando los dientes.

Leyendo las áreas correspondientes a los valores 3.24 y 2.45 de z , en la columna A, se obtiene el área total requerida, sumando las dos probabilidades obtenidas:

$$A_{\text{Requerida}} = A_{z=3.24} + A_{z=2.45} = 0.4938 + 0.4929 \approx 0.99228$$

Por lo tanto:

$$P(180 < x < 215) \approx 0.99225$$

Estadística Descriptiva

La estadística descriptiva proporciona los métodos para recabar información acerca de una determinada población que se desea conocer o investigar con fines específicos, entonces es importante obtener muestras adecuadas que permitan inferir el comportamiento de dicha población.

3.1 Tipos de Datos

Los datos a los que se tiene acceso durante el muestreo que se realiza para obtener la información, se clasifican como sigue:

3.1.1 *Cualitativos*, que se dividen en 2 grandes grupos

- **Nominales o Categóricos.**- Son datos que califican a los elementos estudiados, como por ejemplo: el género de los elementos, masculino o femenino; el color de cabello, negro, café o rubio; la constitución de los elementos, delgado, grueso, regular; las características de un objeto grueso, liso, etc.
- **Ordinales.**- Son datos que indican una jerarquía entre los elementos estudiados, como por ej. Primero y segundo; presidente, vicepresidente y vocal; excelente, bueno, regular, etc.

3.1.2 *Cuantitativos*.-Estos datos también se dividen en 2 grandes grupos

- **Discretos.**- Son datos que surgen del conteo, de los elementos estudiados, generalmente son valores enteros, como por ejemplo, número de miembros en una familia, número de alumnos zurdos, número de alumnos evaluados con calificación de 10, etc.
- **Continuos.**- Son datos que se generan al medir a los elementos estudiados, incluyen valores enteros y fraccionarios, como por ejemplo. peso, estatura, largo del pie, circunferencia craneal, espesor de una viga, etc.

3.2 Tipos de Muestreo

Para obtener una muestra, dependiendo de las necesidades y recursos del investigador, el muestreo puede ser:

3.2.1 No aleatorio.- Consiste en seleccionar una muestra sin permitir la aleatoriedad en el proceso, esto es, los elementos de la población no tienen la misma probabilidad de ser elegidos para participar en el experimento, por lo que la selección es parcializada.

3.2.2 Aleatorio.- Consiste en generar un proceso que permita que todos los miembros de la población participen en el momento de seleccionar la muestra de tal manera que todos y cada uno de ellos, tengan la misma probabilidad de ser elegidos para formar parte de la muestra.

Por lo general, cuando se hace una investigación estadística, se usa muestreo aleatorio, porque favorece la imparcialidad de la selección y evita las tendencias en la información que se obtiene al finalizar el estudio. Un buen proceso de muestreo aleatorio evita que se obtengan conclusiones inadecuadas y que la toma de decisiones sea errónea.

3.2.2.1 Muestreo Aleatorio Simple.- Este tipo de muestreo es el que se usa básicamente para los sorteos, como por ejemplo, la Lotería Nacional, donde, actualmente se utilizan urnas llenas de pelotitas numeradas, que se seleccionan al azar. Las pelotitas en la urna se someten a movimiento continuo para asegurar que todos los elementos dentro de ella tengan la misma probabilidad de ser elegidos y mediante corrientes de aire, las pelotitas vuelan y ocupan un lugar dentro de un aditamento, que permite formar la cifras que componen el número ganador. Para llevarlo a cabo es necesario que:

- a) La población esté codificada o sea fácil de codificar en el momento de realizar el muestreo.
- b) Todos los elementos en la población, tengan la misma probabilidad de ser elegidos para formar parte de la muestra.

En los sorteos, el código es el número del billete o boleto de participación comprado.

3.2.2.2 Muestreo Aleatorio Sistemático.- Este tipo de muestreo se utiliza preferentemente cuando, la población a muestrear presenta un orden, pues esto facilita que se genere el sistema de muestreo. Por ejemplo, en las fábricas de refrescos o productos envasados, es común que el muestreo de producto terminado se realice de esta forma, dado que los elementos producidos son transportados por bandas que permiten la ordenación, en la salida de los productos. Se realiza estableciendo un intervalo de toma de muestra, que se define como un cociente entre el tamaño de la población y el tamaño de la muestra deseada:

$$k = \frac{N}{n}$$

El inicio del muestreo es el que le confiere la aleatoriedad al proceso, ya que se elige cualquiera de los elementos al azar, a , (arranque de muestreo) y a partir de este se cuenta el intervalo k , de tal manera que los elementos que forman parte de la muestra son: $1^{\circ} = a$, $2^{\circ} = a+k$, $3^{\circ} = a+2k$, $4^{\circ} = a+3k$..., y así sucesivamente, hasta terminar la toma de muestra. El que se defina el intervalo de toma de muestra, favorece que se recorra toda la población durante el proceso.

3.2.2.3 Muestreo Aleatorio por Conglomerados o Grupos.- Este tipo de muestreo se usa principalmente para estudios de mercado porque es muy barato. Para aplicarlo, se elige un lugar o región densamente poblada que permita fácil acceso a personas o elementos con características muy diversas, como por ejemplo, ingresos diferentes, creencias diferentes, nivel educativo diferente, nivel socioeconómico diferente, etc. De tal manera que no sea necesario tomar muestras muy grandes, para obtener una gran diversidad de opinión o un consenso respecto a cualquier situación que nos interese analizar.

Los requisitos que se deben de cumplir al hacer este muestreo son:

- a) Los elementos que conforman un mismo conglomerado deben presentar la mayor diversidad en las características que los definen.
- b) Entre un conglomerado y otro, deberá haber la mayor similitud en su conformación, de tal manera que la información obtenida sea semejante sin importar de que conglomerado provenga la información.

Se considera que las tiendas por departamentos, como por ejemplo, Wal-Mart, Comercial Mexicana, Chedraui, etc., son conglomerados natos, porque cumplen las características de alta densidad y fácil acceso a la población.

3.2.2.4 Muestreo Aleatorio Estratificado.- Este tipo de muestreo es muy caro y sólo se utiliza cuando la decisión que se va a tomar, con base en el análisis, afecte de manera muy diferenciada a los diversos sectores de la población. Por lo que es muy importante, permitir que la muestra nos deje ver estas diferencias de opinión, que se generan por la misma diversidad en la conformación de la población, favoreciendo así que la toma de decisión sea la más adecuada, evitando afectar de manera negativa a los sectores más desprotegidos.

El muestreo estratificado, consiste en definir niveles o estratos específicos de clasificación de los elementos de la población. De tal manera que un elemento no pueda pertenecer a 2 o más estratos a la vez. Se puede estratificar por ingresos, nivel socioeconómico, cultural, escolar, profesional, etc. Este muestreo puede hacerse en forma proporcional, respetando la proporción de participación de cada sector en la población, o desproporcionado, dependiendo de las necesidades propias del investigador.

En el muestreo estratificado deben cumplirse los requisitos siguientes:

- a) Los elementos pertenecientes a un mismo estrato o nivel, deberán ser lo más semejantes entre sí, en cuanto a la característica de estratificación.
- b) Entre estrato y estrato, deberá haber la mayor diferencia posible.

Por ejemplo, cuando se muestrea un cuerpo de agua, es necesario usar muestreo estratificado definiendo como estratos las diferentes profundidades en donde los niveles de oxígeno y de luz cambian y por lo tanto la flora y la fauna pueden diferir.

En un estudio estadístico, puede ser necesario hacer más de una etapa de muestreo y utilizar más de un tipo de muestreo. Por ejemplo, estratificar y después, en cada estrato aplicar muestreo aleatorio para obtener la muestra requerida al final del proceso.

3.3 Análisis Exploratorio de Datos

Cuando se está en las etapas iniciales de un estudio estadístico es importante recurrir a herramientas gráficas que nos permitan analizar los datos crudos, obtenidos por muestreo, y definir la forma de su distribución, si la dispersión es poca o mucha, si hay simetría, si hay huecos o datos fuera de contexto, etc. Básicamente se utilizan 2 herramientas gráficas para el análisis exploratorio: El **diagrama de tallo y hoja** ayuda a ordenar los datos crudos de una muestra aleatoria y a definir la forma de la distribución y sus tendencias mientras que el **diagrama de caja con bigotes** permite definir datos atípicos, aparentemente fuera de contexto que podrían significar errores de registro, de toma de muestra o representar comportamiento real de la muestra.

3.3.1 Diagrama de Tallo y Hoja

Consiste en descomponer la información numérica en 2 partes, una llamada tallo y otra llamada hoja, que se grafican siguiendo reglas que facilitan la representación de los mismos.

Para generar un diagrama de tallo y hoja, se traza una línea vertical donde, del lado izquierdo se colocan los tallos y del lado derecho las hojas tratando de completar el número original. Se pueden hacer diagramas de tallo único, de doble o de 5 tallos, dependiendo del tipo de datos y de la cantidad. Para definir los valores del tallo y de las hojas, deben observarse los valores máximos del grupo de datos. Un elemento importante en este gráfico se denomina profundidad y sirve para ubicar medidas posicionales importantes para definir el comportamiento de la muestra.

Para tener una idea de cómo se hace, veremos 2 ejemplos.

EJEMPLO 3.1. Una muestra consta de los siguientes valores que corresponden al número de pizzas de peperoni entregadas por un repartidor en 40 días consecutivos:

12	15	14	9	7	11	13	6	10	13
7	4	7	9	11	12	14	15	23	12
8	7	5	3	2	0	2	3	15	23
0	3	6	7	5	11	13	14	10	9

- Trece el diagrama de tallo y hoja para este ejemplo.
- Trace el diagrama de caja y bigotes respectivo.
- ¿Existen datos extraordinarios o atípicos en la muestra?

Solución:

a) Se puede ver que el valor máximo de los datos corresponde a las decenas, así que para construir el diagrama de tallo y hoja, usaremos las decenas como tallo y las unidades como hoja. Con objeto de que el diagrama no quede amontonado y se disperse adecuadamente usaremos un diagrama de doble tallo, en este caso, las reglas son:

- Hojas de 0 al 4 se colocan en el tallo con asterisco (*).
- Hojas de 5 a 9 se colocan en el tallo con punto (•).

Profundidades	tallos	hojas
8	0*	00223334
(13)	0•	556677778999
19	1*	00111222333444
5	1•	555
2	2*	333

Tallos: decenas 10.0
 Hojas: unidades 1.0
 Ejm. 1* | 0 representa 10

Figura 3.1 Diagrama de Tallo y Hoja (Doble tallo).

Las unidades, que forman las hojas, se van colocando en orden creciente, en los tallos correspondientes, hasta terminar. Podemos ver que la forma como se distribuyen los datos crudos es asimétrica positiva. Esto se debe a que el valor 23 es muy grande con respecto a la mayoría de los datos dispersos.

Para sacarle mayor provecho a este diagrama y poder construir el diagrama de caja, es conveniente calcular las Profundidades y colocarlas al lado izquierdo de la columna de tallos. El cálculo de la profundidad consiste en ir acumulando los elementos (hojas) hasta encontrar el tallo que contiene a la mediana, en este tallo no se acumulan los valores sino que, se especifica la cantidad de elementos en el mismo, encerrando esta cantidad en un paréntesis. A partir de este momento se empiezan a acumular los elementos contenidos desde el último tallo, ascendiendo hasta encontrar el tallo que contiene a la mediana. (ver diagrama arriba).

La mediana es el valor de la variable que divide al conjunto en 2 partes, dejando el 50% de valores más bajos a la izquierda y el 50% de valores más altos a la derecha. Para obtenerla, calculamos la posición de la mediana como sigue:

$$P_{Md} = \frac{n + 1}{2} = \frac{40 + 1}{2} = 20.5$$

Este valor nos indica que la mediana es el promedio entre el dato 20 y el dato 21, ordenados de menor a mayor. Por lo tanto, cuando se arreglan los datos en el diagrama de doble tallo, la mediana está en el segundo tallo, y su valor es:

$$Md = \frac{D_{20} + D_{21}}{2} = \frac{9 + 9}{2} = 9$$

Si con estos mismos valores muestrales hacemos el diagrama con quintuple tallo, usamos las letras de los números en inglés, de la siguiente forma:

- En asterisco se colocan hojas 0 y 1
- En “t” se colocan hojas 2 y 3
- En “f” se colocan hojas 4 y 5
- En “s” se colocan hojas 6 y 7
- En (•) se colocan hojas 8 y 9

Diagrama de Tallo y Hoja para los datos: unidades = 1.0 1|2 representa 12.0.

Profundidades	Tallo	Hojas
2	0*	0
7	0t	22333
10	0f	455
17	0s	6677777
(4)	0•	8999
19	1*	111
14	1t	222333
8	1f	444555
2	1s	
2	1•	
2	2*	
2	2t	33

Figura 3.2 Diagrama de Tallo y Hoja de quintuple tallo.

Nota: El diagrama comienza desde el registro de los datos (hojas) mínimos y termina en el registro de los datos máximos, es por ello que en este ejemplo sólo se llega al nivel 2^t. La ausencia de datos en categorías intermedias (v.g. 1^s y 1^t) no hace que estas se eliminen.

3.3.2 Diagrama de Caja con Bigotes

Para hacer este diagrama se toma como base el diagrama de tallo y hoja, porque es necesario que los datos estén ordenados. Este diagrama consiste en un rectángulo cuyos límites son el cuarto inferior y el cuarto superior de la distribución ordenada de datos y unas extensiones llamadas bigotes

- b) Para construir el diagrama de caja, primero se calculan el cuarto inferior y el cuarto superior (C_I y C_S), definiendo la posición de los mismos a partir de la posición de la mediana truncada (sin decimales) a la que se le suma 1 y al resultado se le divide entre 2.

Posición de los cuartos tomando para el ejemplo 3.1:

$$P_{\text{cuartos}} = \frac{P_{MdT} + 1}{2} = \frac{20 + 1}{2} = 10.5$$

Este resultado nos indica que los cuartos corresponden al promedio entre el dato 10 y el dato 11. Para localizar C_I se cuenta de arriba hacia abajo y de izquierda a derecha en el diagrama de tallo y hoja y para C_S se cuenta de abajo hacia arriba y de derecha a izquierda, en el mismo diagrama.

Por lo tanto:

$$C_i = \frac{5 + 6}{2} = 5.5 \quad \text{y} \quad C_s = \frac{13 + 13}{2} = 13$$

En seguida se calcula la dispersión de los cuartos, que nos sirve para definir las cotas inferior y superior internas y las cotas inferior y superior externas. Estas cotas se calculan para delimitar aquellos valores que por su dispersión se conocen como casos extraordinarios o atípicos. Todos los valores que se localizan entre las cotas inferiores internas y externas se consideran casos extraordinarios moderados o leves y los que se localizan después de las cotas superior e inferior externas se consideran casos extraordinarios graves o severos.

Dispersión de los cuartos DC:

$$DC = C_s - C_i = 13 - 5.5 = 7.5$$

Para calcular la cota inferior interna, (C.I.I) se tiene la fórmula siguiente:

$$C.I.I. = C_i - (1.5)DC = 5.5 - (1.5)7.5 = 5.5 - 11.25 = -5.75$$

Como el valor es negativo, se concluye que ningún valor, en la distribución se encuentra fuera de la cota, inferior interna. Si calculamos la cota inferior externa, el valor será todavía más negativo, como podemos observar:

$$C.I.E. = C_s - (3)DC = 5.5 - (3)7.5 = 5.5 - 22.5 = -17$$

De acuerdo con los valores de ambas cotas inferiores podemos concluir que no hay casos extraordinarios o atípicos por la izquierda.

Ahora calculamos la cota superior interna como sigue:

$$C.S.I. = C_s + (1.5)DC$$

$$C.S.I. = 13 + (1.5)7.5 = 13 + 11.25 = 24.25$$

Este valor nos indica que no hay ningún caso extraordinario o atípico por la derecha que rebase la cota superior interna pues nuestro valor máximo, en la distribución de datos, es 23. Por lo anterior, tampoco es necesario calcular la cota superior externa, cuya fórmula es:

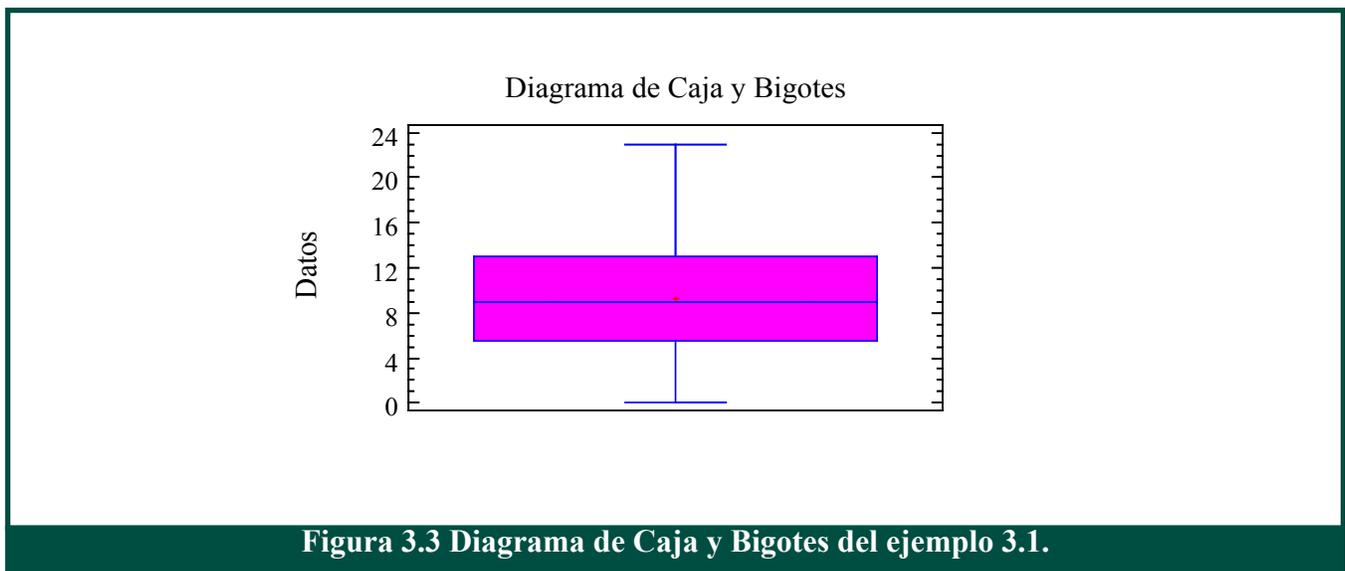
$$C.S.E. = C_s + (3)DC$$

¿Hasta dónde llegan los bigotes?

El bigote izquierdo parte de cuarto inferior o borde inferior de la caja hasta el máximo de los valores $\{X_1, C.I.I.\}$, es decir, el mayor de estos dos valores.

El bigote derecho parte del cuarto superior o borde superior de la caja hasta el mínimo de los valores $\{X_n, C.S.I.\}$, es decir, el menor de los estos dos valores.

Al no haber casos atípicos en nuestro ejemplo, los bigotes se pintan a partir del primer y último dato de la distribución y terminan en la caja. En la caja debe marcarse también la mediana o segundo cuarto, C_2 .



- c) Este diagrama nos muestra que la caja no es simétrica, pues la parte de arriba de la mediana es más grande que la de abajo, lo que nos indica una mayor dispersión hacia los datos mayores, pero también nos muestra que no hay valores fuera de contexto. (casos extraordinarios o atípicos).

Nota: Este tipo de diagramas puede hacerse tanto de forma vertical como horizontal.

EJEMPLO 3.2. Se registró el número de meses de servicio de refrigeradores industriales vendidos por una empresa, antes de necesitar el primer servicio de reparación, que también presta la misma empresa, como se muestra en la tabla siguiente:

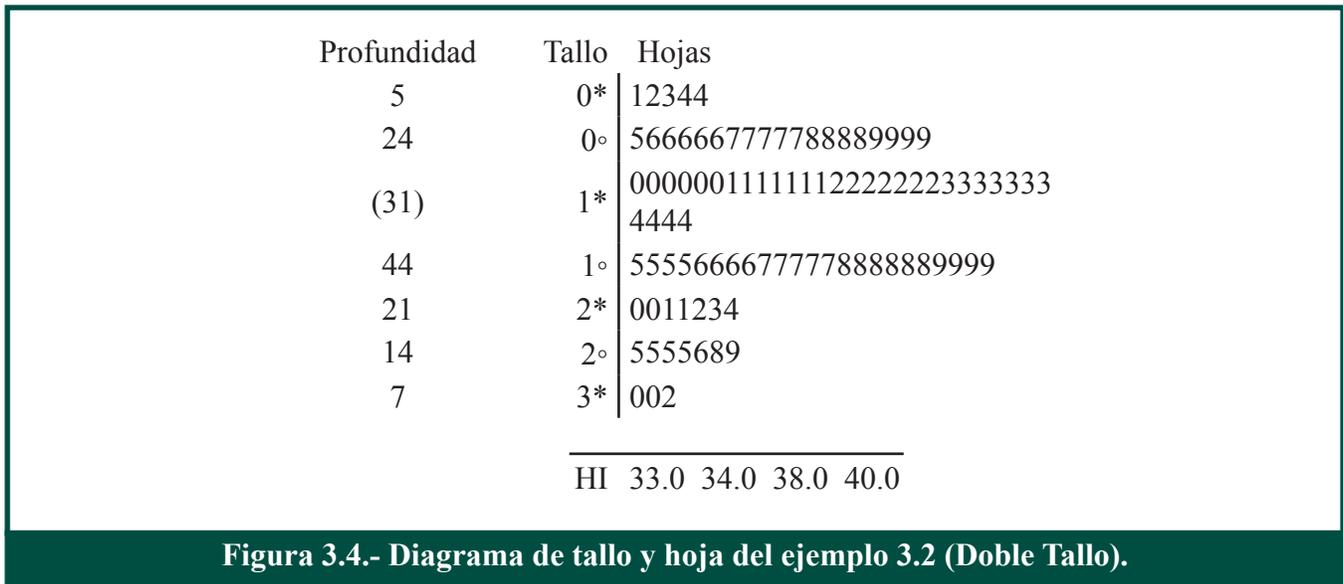
1	2	7	23	17	32	15	6	40	6
6	7	9	30	24	17	6	34	20	8
9	9	3	4	19	25	8	10	18	38

10	11	12	7	9	14	33	25	25	17
12	12	14	9	11	19	25	19	13	15
13	13	11	14	14	10	11	13	6	20
16	16	22	18	7	13	13	15	10	13
18	21	29	11	4	5	19	8	8	10
21	28	18	16	30	7	10	11	12	12
26	18	16	12	12	11	17	17	15	18

- a) Con base en estos datos, construya un diagrama de tallo y hoja.
 b) Construya el diagrama de caja correspondiente.
 c) De acuerdo con el diagrama anterior, ¿se puede considerar que existen datos atípicos?

Solución:

- a) Diagrama de Tallo y hoja.



En este diagrama de tallo y hoja (doble tallo) puede observarse que en la parte inferior existe una línea HI, que identifica los datos atípicos, como se podrá comprobar al realizar el diagrama de caja.

- b) Ahora, realizaremos el diagrama de caja y bigotes del ejemplo 3.2 con base en los datos del diagrama de tallo y hoja, respectivo, siguiendo las instrucciones del ejemplo 3.1.

$$P_{Md} = \frac{n+1}{2} = \frac{100+1}{2} = 50.5$$

$$M_d = \frac{D_{50} + D_{51}}{2} = \frac{13 + 13}{2} = 13$$

Entonces la mediana o cuarto 2 es 13.

Para obtener los cuartos, tenemos:

$$P_{\text{cuartos}} = \frac{P_{MdT} + 1}{2} = \frac{50 + 1}{2} = 25.5$$

$$\text{Cuartos} = \frac{D_{25} + D_{26}}{2} \Rightarrow C_I = \frac{10 + 10}{2} = 10 \quad \text{y} \quad C_S = \frac{19 + 18}{2} = 18.5$$

Así, la dispersión de los cuartos es la diferencia entre ellos:

$$DC = C_S - C_I = 18.5 - 10 = 8.5$$

Calculando las cotas internas y externas:

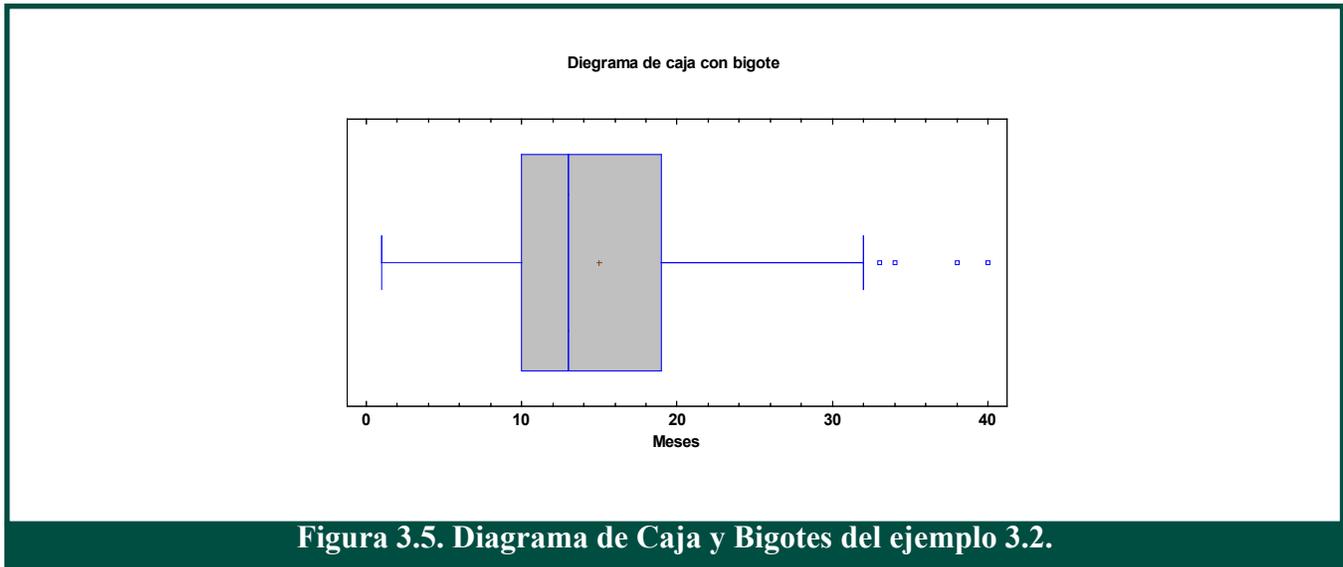
$$C.I.I. = 10 - (1.5)8.5 = -2.75$$

Como se puede observar, la cota inferior interna es negativa, como la muestra no contiene datos negativos, concluimos que no hay datos atípicos por la izquierda y tampoco tiene caso calcular la cota inferior externa.

$$C.S.I. = 18.5 + (1.5)8.5 = 31.25$$

$$C.S.E. = 18.5 + (3)8.5 = 44$$

Así, el diagrama de caja para el ejemplo 3.2 es:



- c) Puede verse que la cota superior interna es 31.25, esto indica que todos los datos cuyo valor sea mayor a 31.23 corresponden a datos atípicos por la derecha. Como la cota superior externa es 44 y el dato máximo en la muestra es 40, se concluye que los datos 33, 34, 38 y 40 son datos atípicos leves que en el diagrama se ven como puntos al lado derecho del bigote superior.

3.4 Medidas Descriptivas en la Muestra

Para estudiar el comportamiento de una muestra, obtenida por muestreo aleatorio, dentro de una población, se utilizan medidas descriptivas o estimadores que se dividen en 2 grandes grupos:

- Medidas de Tendencia Central (Media Aritmética, Mediana y Moda).
- Medidas de variabilidad o de dispersión (Recorrido, Varianza, Desviación Estándar y Coeficiente de Variación).

3.4.1 Medidas de Tendencia Central

3.4.1.1 Media Aritmética

Esta medida es la más usada para centralizar los datos. Se utiliza ampliamente, debido a que está definida algebraicamente y es fácil, entonces, introducirla en procesos de análisis más complejos, aprovechando sus propiedades algebraicas. Tiene como desventaja, el hecho de moverse hacia los valores extremos cuando en la muestra hay valores atípicos o dispersos con respecto a la generalidad. Lo anterior, favorece que en un momento dado se subestime o se sobrestime el valor medio real. Este defecto se controla si en un análisis, la media aritmética va acompañada de una medida adecuada de variabilidad.

La media aritmética de una muestra se representa mediante el símbolo \bar{X} y se define así:

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Donde x_i , representa cada valor diferente, adquirido por la variable y n es el total de datos en la muestra. Así que, la definición nos dice, básicamente, que hay que sumar todos los datos, desde el primero al n -ésimo y dividir por n . Esta fórmula es adecuada cuando no hay repetición de datos y la cantidad de ellos es pequeña.

EJEMPLO 3.3. Un vendedor de licuados de fruta, consigue fresas, de la misma clase, a diferente precio al comprar con 5 diferentes distribuidores:

Distribuidor	1	2	3	4	5
Precio(\$/Kg)	28	26	27.30	25.75	27.50

¿Cuál es el precio promedio, pagado por kilogramo de fresas?

En este caso, se suman todos los precios y se divide esta suma entre el total de elementos (precios) que contribuyen al promedio:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{28+26+27.30+25.75+27.50}{5} = \frac{134.55}{5} = 26.91$$

Este resultado nos dice que el cliente pagó 26.91, en promedio, por kilo de fresas.

Sin embargo, cuando se tiene una muestra grande y hay datos repetidos, es conveniente hacer una tabla de distribución con los datos y aplicar una fórmula, de la media aritmética, que introduce el término “frecuencia” en la definición, porque facilita el cálculo:

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n}$$

Donde, f_i es la frecuencia o número de veces que se repite un dato específico x_i dentro de la muestra, y k es el número de categorías diferentes que presenta la variable.

EJEMPLO 3.4. El profesor de estadística, le pidió a Rosy, que hiciera una encuesta entre los alumnos de la facultad, respecto al número de llamadas que estos reciben por día en su celular, el tamaño de la muestra fue de 110. También le pidió que organizara la información de manera que pudiera obtener fácilmente

el promedio de llamadas por día. Rosy, organizó una tabla de frecuencias con los datos, de la siguiente manera:

N° de llamadas (x_i)	3	4	5	6	7	8	9	10 o más
Frecuencia (f_i)	10	13	16	23	18	15	10	5

Con base en esta tabla, calculó el promedio de llamadas recibidas por celular, en un día.

$$x = \frac{\sum_{i=1}^k f_i x_i}{n} = \frac{10(3)+13(4)+16(5)+23(6)+18(7)+15(8)+10(9)+5(10)}{110} = 6.2363$$

Por la definición algebraica de la media aritmética, el resultado tiende a dar valores dentro de un intervalo y por esta razón, el valor promedio presenta decimales. Sin embargo, por el hecho de que los datos originales son discretos (no definidos dentro de una escala continua), el resultado se interpreta diciendo que el número de llamadas promedio por día, recibidas por alumnos de esta facultad es de 6.

3.4.1.2 Mediana

Como se indicó en apartados anteriores, esta una medida de tendencia central que se calcula ubicando su posición en el grupo de datos. Para calcular esta medida, es obligatorio ordenar los datos, de menor a mayor y localizar el dato o datos, que dividen a la mitad, a la distribución ordenada. Esto es, el 50% de los datos queda a la izquierda de ese valor y el otro 50% queda a la derecha del mismo.

Esta medida no se ve afectada por valores extremos, como la media aritmética, porque su definición es posicional y por lo tanto, es más justa para valorar el promedio de una distribución. Su desventaja es que no puede definirse algebraicamente por lo que no se utiliza mucho para análisis más complejos.

Tomando los datos del Ejemplo 3.3: 28, 26, 27.30, 25.75, 27.50.

Observamos que nuestra muestra tiene 5 datos, esto es, el tamaño de la muestra n es de 5, entonces calculamos primero la posición de la mediana como sigue:

$$P_{Md} = \frac{n+1}{2} = \frac{5+1}{2} = 3$$

El resultado nos está indicando que la mediana es el tercero de los datos, ordenados de menor a mayor.

Ordenando los datos: 25.75, 26, 27.30, 27.50, 28, podemos apreciar que el dato que divide a la mitad es 27.30 por lo tanto, la mediana es 27.30.

Cuando hay muchos datos y aparecen más de una vez, es necesario acumular los datos dentro de la distribución, para localizar la mediana.

Tomando los 110 datos del Ejemplo 3.4 tenemos que:

$$P_{Md} = \frac{n + 1}{2} = \frac{110 + 1}{2} = \frac{111}{2} = 55.5$$

El valor decimal, en la posición de la mediana, nos está indicando que la mediana corresponde a la media aritmética de los datos ordenados, que está entre el dato que ocupa la posición 55 y el que ocupa la posición 56.

Para ordenar los datos es conveniente hacer una distribución de frecuencias que contenga frecuencias acumuladas. Acumular las frecuencias consiste en adicionar las frecuencias, categoría por categoría, hasta terminar con un número de valores acumulados de 110 en la última categoría:

N° de llamadas (X_i)	Frecuencia (f_i)	Frecuencia Acumulada (F_i)
3	10	10
4	13	23
5	16	39
6	23	62
7	18	80
8	15	95
9	10	105
10	5	110

Observamos que el dato 55 y el dato 56 están incluidos en la cuarta categoría. (Note que la tercera categoría incluye hasta el dato 39 pero la cuarta incluye desde el dato 40 hasta el dato 62) Entonces podemos concluir que la mediana del número de llamadas por celular, para estudiantes de la facultad, corresponde a 6.

3.4.1.3 Moda

Esta medida de tendencia central se define como el dato que aparece con mayor frecuencia, esto es, el dato que más se repite.

Aunque la moda está considerada como una medida de tendencia central, no siempre está colocada en el centro de la distribución. Es más, podría no haber moda (porque todos los datos son únicos) o inclusive haber más de una moda. Debido a estas características, no es factible utilizar a la moda para hacer análisis más complejos.

Si tomamos los datos del ejemplo 3.3: 28, 26, 27.30, 25.75, 27.50, vemos que son datos únicos, no hay repeticiones, por lo tanto, la moda no está definida.

Tomando el ejemplo 3.4:

N° de llamadas (X_i)	3	4	5	6	7	8	9	10
Frecuencia (f_i)	10	13	16	23	18	15	10	5

Vemos que el dato que se repite más, el de mayor frecuencia, es el 6 (23 personas contestaron que reciben 6 llamadas), por lo que podemos asegurar que el número de llamadas, más usual es 6.

Con objeto de ejemplificar el caso de distribuciones de datos multimodales, se trabajará el siguiente problema.

EJEMPLO 3.5. Catalina es empleada en una mercería y el dueño de la misma, la envía a obtener el inventario del número de botones blancos. Estos botones se venden en 9 tamaños identificados por un número. Ella registra la siguiente información:

N°	5	7	9	10	13	14	15	17	20
Cantidad	45	28	30	29	30	18	16	16	45

En este caso hay dos modas porque los botones de tamaño 5 y 20 presentan la misma frecuencia (45) y los botones de tamaño 9 y 13 también presentan frecuencia semejante (30). Por esta razón, identificamos a la distribución de datos como bimodal.

3.4.2 Medidas de Variabilidad

3.4.2.1 Recorrido. - Es una medida burda de la variabilidad porque representa sólo la distancia entre el extremo superior o valor más alto y el y el inferior o valor más bajo de la distribución de datos.

$$R = \text{Valor máximo} - \text{Valor mínimo} = X_n - X_1$$

Tomando como referencia los Ejemplos 3.2 y 3.3, presentados anteriormente tenemos:

Recorrido de los precios de fresas: $R = 28 - 26 = 2$.

Recorrido en el número de llamadas por celular: $R = 10 - 3 = 7$.

3.4.2.2 Varianza

Es la variación o dispersión cuadrática de una distribución de datos. Se considera una medida absoluta de la variación, porque sólo tiene significado cuando va acompañada de las dimensiones, al cuadrado, de la variable que se analiza.

Se define como el promedio corregido de las distancias cuadradas de cada valor de la variable, en la distribución, con respecto a su media aritmética. Matemáticamente se define, en su forma más sencilla, como:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Donde el término, $n - 1$ se denomina grados de libertad, que representa el número de observaciones realmente aleatorias, que intervienen en el cálculo de esta medida.

Si en la distribución hay valores repetidos, es conveniente introducir el término de frecuencia en el cálculo de la varianza:

$$s^2 = \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{n - 1}$$

f_i representa el número de veces que cada distancia cuadrática se repite y k es el número de categorías diferentes que se presentan en la distribución.

Cuando el número de datos y categorías que se manejan es grande, es conveniente utilizar una modificación algebraica de esta fórmula, para agilizar el cálculo. (Cabe hacer notar que esta segunda fórmula sale de la anterior desarrollando el cuadrado del binomio y utilizando propiedades de las sumatorias):

$$s^2 = \frac{\sum_{i=1}^k f_i x_i^2 - n\bar{x}^2}{n - 1}$$

Tomando como base los datos del Ejemplo 3.3:

Distribuidor	1	2	3	4	5
Precio(\$/Kg)	28	26	27.30	25.75	27.50

y la media aritmética calculada previamente $\bar{x} = 26.91$ podemos obtener la varianza del proceso como sigue:

$$s^2 = \frac{(28-26.91)^2 + (26-26.91)^2 + (27.30-26.91)^2 + (25.75-26.91)^2 + (27.50-26.91)^2}{5-1}$$

$$s^2 = 0.9655 (\$/ Kg)^2$$

Para utilizar la fórmula modificada, se calculan los cuadrados de los valores de la variable:

Distribuidor	1	2	3	4	5
Precio(\$/Kg)	28	26	27.30	25.75	27.50
Precio ² (\$ ² /Kg ²)	784	676	745.29	663.0625	756.25

La suma de los valores cuadrados es: 3624.6025 entonces, substituyendo en la fórmula modificada tenemos:

$$s^2 = \frac{3624.6025 + 5(26.91)^2}{5-1} = \frac{3.862}{4} = 0.9655 (\$/Kg)^2$$

De los cálculos anteriores, concluimos que la variación cuadrática en el precio por kg de fresa es de 0.9655 (pesos/kilogramo)².

Tomando los datos del Ejemplo 3.4 se tiene:

N° de llamadas (X_i)	3	4	5	6	7	8	9	10
Frecuencia (f_i)	10	13	16	23	18	15	10	5

Usando la fórmula original, con el término frecuencia, tenemos:

$$s^2 = \frac{10(3-6.2363)^2 + 13(4-6.2363)^2 + 16(5-6.2363)^2 + \dots + 5(10-6.2363)^2}{110-1} = 3.6684 \text{ llamadas}^2$$

Usando la fórmula modificada es más sencillo el cálculo:

N° de llamadas	3	4	5	6	7	8	9	10 o más
Frecuencia	10	13	16	23	18	15	10	5
N° de llamadas ² (x_i) ²	9	16	25	36	49	64	81	100
$F_i (X_i)^2$	90	208	400	828	882	960	810	500

Al sumar los valores de la última fila, obtenemos 4678, entonces, substituyendo.

$$s^2 = \frac{4678 - 110(6.2363)^2}{110 - 1} = \frac{399.8545}{109} = 3.6684 \text{ llamadas}^2$$

Vemos que, sin importar que fórmula que se utilice, el resultado de la variación cuadrática será el mismo, siempre y cuando se usen correctamente.

3.4.2.3 Desviación Estándar

Es la medida real de la dispersión que presentan los datos. Cuando se grafican los datos, lo que realmente se presenta como dispersión es la desviación estándar, cuyo símbolo es s . Es el promedio corregido, de las variaciones que presenta la muestra. En este estimador, las dimensiones ya no son cuadráticas.

Así, la desviación estándar es la raíz cuadrada positiva de la varianza.

Entonces, si los datos no se repiten.

$$s = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{n - 1}}$$

Si hay repeticiones en los datos:

$$s = \sqrt{\frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{n - 1}} \quad \text{o} \quad s = \sqrt{\frac{\sum_{i=1}^k f_i x_i^2 - n(\bar{x})^2}{n - 1}}$$

Para el Ejemplo 3.3, cuya varianza es $s^2 = 0.9655 (\$/Kg)^2$ tendremos que

$$s = \sqrt{0.9655 (\$/Kg)^2} = 0.9826 \$/Kg$$

Para el Ejemplo 3.4, cuya varianza es $s^2 = 3.6684 \text{ llamadas}^2$ tendremos que

$$s = \sqrt{3.6684 \text{ llamadas}^2} = 1.9153 \text{ llamadas}$$

3.4.2.4 Coeficiente de Variación.- Es una medida relativa de la variabilidad que presenta un conjunto de datos. Es una medida adimensional que se reporta como un porcentaje de la variación observada en la muestra, con respecto a la media aritmética. Se define matemáticamente como:

$$C.V. = \left(\frac{s}{\bar{x}} \right) 100$$

Si aplicamos esta definición al Ejemplo 3.3, obtendremos:

$$C.V. = \left(\frac{0.9826\$/Kg}{26.91\$/Kg} \right) 100 = 3.65145\%$$

Aplicándola el Ejemplo 3.4, tenemos:

$$C.V. = \left(\frac{1.9153 \text{ llamadas}}{6.2363 \text{ llamadas}} \right) 100 = 30.71\%$$

Cuando se comparan 2 grupos de datos, por ejemplo el peso y la longitud de una muestra de peces de una especie determinada, el coeficiente de variación nos permite identificar cual de las características presenta mayor variación.

Si los datos se refieren a grupos diferentes como por ejemplo niños y niñas es posible identificar cuál de estos grupos presenta menor variación en sus características.

Suponga que se están comparando dos grupos de alumnos por su rendimiento. Para que la información sea más completa, se toman, el promedio de calificación y la desviación estándar de cada grupo y se calculan los coeficientes de variación respectivos. Al comparar los resultados es posible definir cual grupo fue mejor, más consistente en su rendimiento porque su variación relativa es menor.

3.5 Estadística para datos Agrupados

Los datos se agrupan, básicamente para construir gráficas representativas de la distribución que guardan los datos. Antiguamente, cuando no existía la facilidad de cálculo que dan las computadoras, se agrupaban los datos para facilitar el cálculo de las medidas descriptivas.

Para agrupar los datos, existen algunas reglas empíricas que fueron formuladas con la intención de lograr distribuciones lo más parecidas a una normal. Es importante mencionar que los estimadores obtenidos por agrupación son aproximados a los reales, esto es, tienen discrepancias pequeñas y serán lo más parecidos a los reales cuando se logre una muy buena agrupación.

En general, no es bueno forzar los datos hacia un modelo normal, cuando no son normales porque, se pierden características importantes del comportamiento muestral.

3.5.1 Reglas Empíricas para agrupar datos

3.5.1.1 En 1926, Sturges, estableció un algoritmo para definir el número adecuado de intervalos, categorías o grupos que debería tener la agrupación para lograr resultados óptimos.

$$K = 1 + (3.322)\log_{10}n$$

El problema con este algoritmo es que el número de intervalos no cambia mucho a medida que aumenta el tamaño de la muestra y al final se tienen distribuciones con pocas categorías y frecuencias muy altas. Por ejemplo: si $n=60$, $K=7$; si $n=200$, $K=9$; si $n=760$, $K=11$.

3.5.1.2 En 1965, Dixon y Kronmal, propusieron una nueva forma de calcular el número adecuado de intervalos para agrupar datos, siempre y cuando el tamaño de la muestra fuese mayor que 50:

$$K = (10)\log_{10}n$$

Este algoritmo, al contrario del anterior, hace crecer suficiente el número de intervalos, a medida que crece n .

Por ejemplo, si $n=60$, $K=18$; si $n=300$, $K=24$; si $n=760$, $K=28$.

Sin embargo, un número muy grande de intervalos puede generar distribuciones con intervalos intermedios con frecuencia 0.

3.5.1.3 En 1976, Velleman, estableció que si el tamaño de la muestra era como máximo 50, una manera adecuada de calcular el número de intervalos era:

$$k = 2\sqrt{n}$$

Así, si $n=20$, $K=9$; si $n=40$, $K=13$ y si $n=50$, $K=14$.

Vemos que esta fórmula tiende a sobrestimar el número de intervalos, dada la cantidad de datos que se manejan.

3.5.1.4¹ Existe otro método para agrupar los datos, basado en la fórmula de Scott, que en 1979 derivó una fórmula para calcular la amplitud óptima asintótica resultante en un error cuadrado integrado medio mínimo (ECIM) para histogramas. Es necesario conocer previamente la verdadera función de densidad, pero como ésta raramente se conoce, se parte del supuesto de que la densidad es normal (Gaussiana) y entonces propone la ecuación:

$$\hat{h} = 3.5 \hat{\sigma}(n)^{-1/3}$$

Donde:

\hat{h} , es la amplitud de banda estimada

$\hat{\sigma}$, es una estimación de la desviación estándar de los datos

El parámetro obtenido, suaviza a una distribución log-normal pero cuando el índice del sesgo es tan grande como 1, la diferencia con la amplitud de intervalo óptima verdadera, es menor que 30%, es insensible a curtosis moderada y sobreesuaviza datos bimodales, cuando la distancia entre las modas es mayor que 2.

3.5.1.5¹ Freedman y Diaconis, en 1981, propusieron una regla más robusta en la que se utiliza un múltiplo del rango intercuartílico (RIC) o de la dispersión de los cuartos en lugar de la estimación de la desviación estándar, como sigue:

$$\hat{h} = 2(RIC)(n)^{-1/3} \text{ o } \hat{h} = 2(DC)(n)^{-1/3}$$

3.5.1.6 Se ha observado que una buena opción, basada en la experiencia práctica para calcular el número adecuado de intervalos es:

$$K = \sqrt{n}$$

Porque no tiende a subestimar ni a sobrestimar el número adecuado de intervalos y además es sencilla.

Por ejemplo: si $n=40$, $K=6$; si $n=100$, $K=10$; si $n=300$, $K=17$; y si $n=760$, $K=28$.

De todas estas reglas empíricas, las más convenientes son las 2 últimas aunque en particular se prefiere la de Freedman y Diaconis. Para ejemplificar el método de agrupación se usarán estas dos últimas reglas usando los datos del ejemplo 3.4.

Nota: *Es conveniente agrupar con intervalos del mismo tamaño, con objeto de facilitar la representación de los datos.*

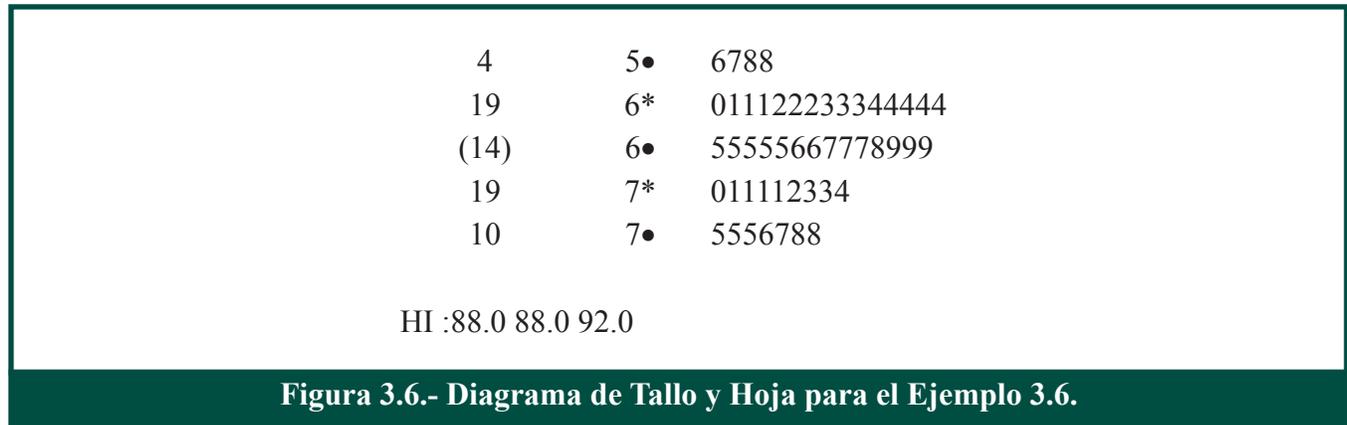
EJEMPLO 3.6. Los siguientes datos corresponden al número de litros de leche vendidos en un mini súper, en 52 sábados consecutivos:

67	75	63	71	65	73	71	88	61
65	56	62	58	72	66	76	77	75
61	70	64	71	63	61	63	64	62
69	60	66	78	92	64	64	69	64
65	75	72	67	88	74	65	73	
78	62	68	69	67	57	65	58	

Agrupe los datos en una distribución de frecuencias, que contenga Límites reales de clase, centros de clase, frecuencias absolutas, frecuencias relativas, frecuencias acumuladas y frecuencias acumuladas porcentuales.

Realizar el conteo de las frecuencias correspondiente a cada grupo o intervalo de clase, se facilita si previamente hemos realizado un diagrama de Tallo y Hoja, que automáticamente nos permite ordenar los datos.

Diagrama de Tallo y Hoja para Litros: unidad = 1.0 Ejm. 1|2 representa 12.0.



3.5.2 Cálculos para una buena agrupación usando el método de Freedman y Diaconis y método basado en la experiencia práctica

Para el método de Freedman y Diaconis, usando como base el diagrama de tallo y hoja se calculan los cuartos inferior y superior a partir de la posición de la mediana truncada:

$$P_{Md} = \frac{n+1}{2} = \frac{54+1}{2} = 27.5; \quad P_{c_i} = \frac{P_{Md \text{ truncada}} + 1}{2} = \frac{27+1}{2} = 14$$

Entonces:

$$C_l = \text{Dato } 14 \text{ contado de arriba hacia abajo y de izquierda a derecha} = 63$$

$$C_s = \text{Dato } 14 \text{ contado de abajo hacia arriba y de derecha a izquierda} = 72$$

Por lo que la dispersión de los cuartos se calcula como:

$$DC = C_s - C_l = 72 - 63 = 9$$

Así que la amplitud del intervalo para realizar la agrupación, de acuerdo con la fórmula de Freedman será:

$$\hat{h} = 2(DC)(n)^{-\frac{1}{3}} = 2(9)(54)^{-\frac{1}{3}} = 4.76 \approx 5$$

De acuerdo con el resultado, se puede elegir una amplitud, \hat{h} de 4 o de 5 unidades para el intervalo, se elige amplitud de 5.

Si se utiliza el método basado en la experiencia práctica donde $K = \sqrt{n}$

$$K = \sqrt{52} = 7.21 \approx 7$$

El recorrido o rango es:

$$R = \text{Valor máximo} - \text{Valor mínimo}$$

$$R = 92 - 56 = 36$$

Y por último calculamos la amplitud de intervalo:

$$a = \frac{R}{K}$$

$$a = \frac{36}{7} = 4.15 \approx 5$$

Entonces, la amplitud o tamaño del intervalo a utilizar para agrupar es 5, al igual que con la regla de Freedman y Diaconis.

Se utiliza nomenclatura de intervalo abierto por la izquierda y cerrado por la derecha, para definir correctamente los límites. Con objeto de que todos los datos de la muestra estén incluidos en la distribución de frecuencias, se acostumbra bajar una décima, una centésima o una unidad al dato inicial, según corresponda.

Como nuestro dato menor es 56, bajaremos una unidad, para empezar el conteo en 55 y la distribución de frecuencias queda como sigue:

Tabla 3.1.-Distribución de Frecuencias con Intervalos, Límites Reales y Centros de Clas para los Litros de Leche vendidos.

Intervalo de Clase	Frecuencia f_i	Frecuencias Acumuladas F_i	Límites Reales de Clase LRC	Centros de clase (m_i)
(55,60]	5	5	55-60	57.5
(60,65]	19	24	60-65	62.5
(65,70]	10	34	65-70	67.5
(70,75]	11	45	70-75	72.5
(75,80]	4	49	75-80	77.5
(80,85]	0	49	80-85	82.5
(85,90]	2	51	85-90	87.5
(90,95]	1	52	90-95	92.5

3.5.2.1 Marcas o Centros de Clase o puntos medios (m_i), se consideran los valores representativos de cada clase o intervalo. De tal manera que si la agrupación se realiza para facilitar los cálculos de los estimadores, las (m_i) son la base de los cálculos.

$$\text{Centro de clase} = \frac{\text{Límite Inferior} + \text{Límite Superior}}{2}$$

Como los intervalos tienen la misma amplitud, basta con calcular el centro del primer intervalo y agregar consecutivamente la amplitud para ir obteniendo los centros de clase restantes:

$$\text{Centro de clase} = \frac{55 + 60}{2} = 57.5$$

A partir de este centro de clase, completaremos los demás sumando cada vez 5 al punto medio antecedente (ver tabla 3.1).

3.5.2.2 Frecuencias relativas f_r , representan la proporción que guarda cada clase con el total. Se calculan dividiendo la frecuencia absoluta f_i entre el total de datos, n .

$$f_r = \frac{f_i}{n}$$

3.5.2.3 Frecuencias Acumuladas.- Se calculan para obtener las frecuencias acumuladas porcentuales (llamado también, porcentaje acumulado (% acum)), que a la vez nos permitirán graficar un Polígono de Frecuencias Acumuladas u Ojiva y obtener medidas posicionales llamadas cuantiles.

Tabla 3.2.- Frecuencias relativas, acumuladas y porcentuales.

Frecuencia f_i	Frecuencias acumuladas F_i	Frecuencias relativas f_i	Frecuencias acumuladas Porcentuales (% acum)
5	5	$5/52 = 0.096$	$(5/52)100 = 9.6$
19	24	$19/52 = 0.3654$	$(24/52)100 = 46.14$
10	34	$10/52 = 19.23$	$(34/52)100 = 65.38$
11	45	$11/52 = 21.15$	$(45/52)100 = 86.54$
4	49	$4/52 = 0.077$	$(49/52)100 = 94.23$
0	49	$0/52 = 0$	$(49/52)100 = 94.23$
2	51	$1/52 = 0.01923$	$(51/52)100 = 98.08$
1	52	$1/52 = 0.01923$	$(52/52)100 = 100$

3.6 Representación Gráfica de los Datos

Como ya se había mencionado, antes, la agrupación se lleva a cabo para poder hacer representaciones gráficas de la distribución de datos. Las gráficas más usadas son:

3.6.1 Histograma

Es uno de los gráficos más útiles en el análisis estadístico, porque nos permite visualizar la forma de la distribución y la tendencia de los datos.

Es un gráfico de barras continuas, que se construye trazando, sobre el eje de las abscisas, los límites reales o fronteras de cada clase, y sobre el eje de las ordenadas, las frecuencias absolutas respectivas, siempre y cuando los intervalos sean del mismo tamaño.

Si tomamos como base, la distribución de datos del **Ejemplo 3.6**, tendremos que graficar los límites reales de clase contra las frecuencias absolutas y entonces, el histograma queda de la siguiente forma:

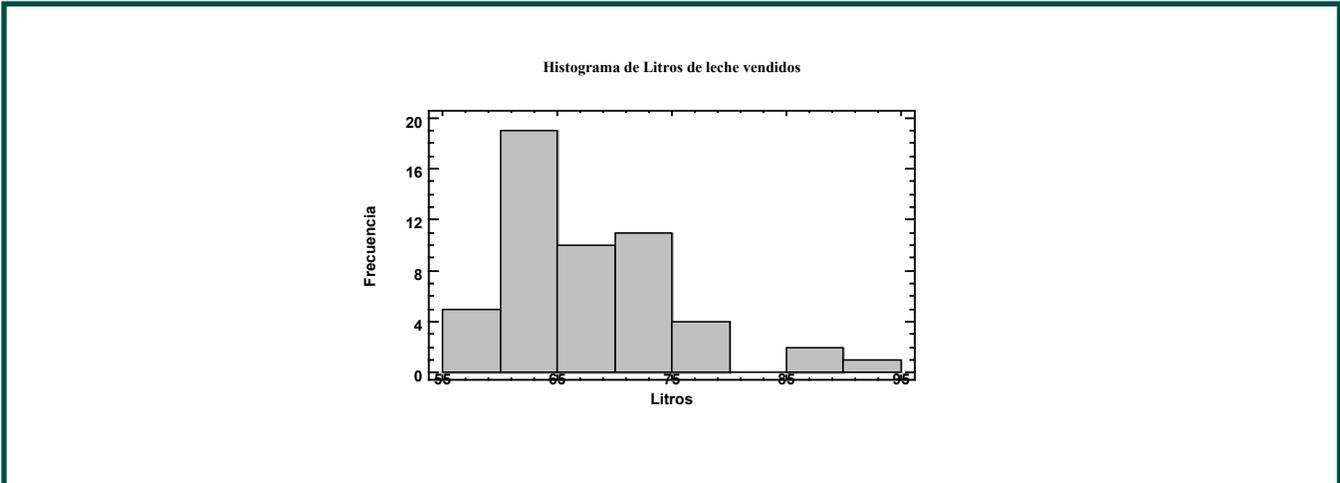


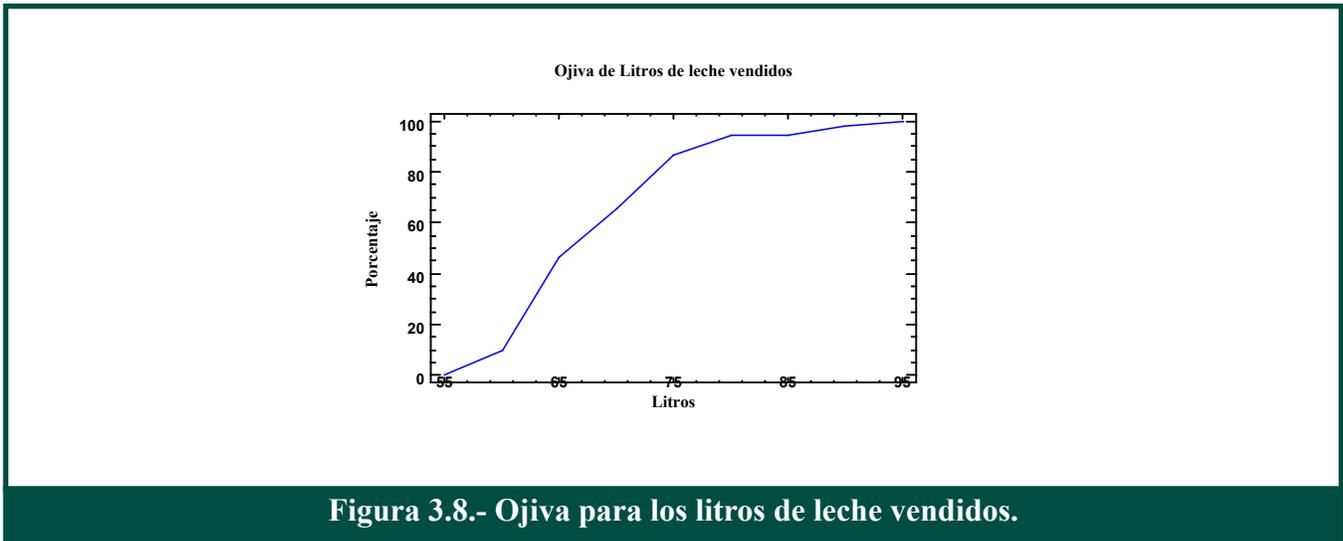
Figura3.7- Histograma de frecuencias para la cantidad de litros de leche vendidos.

Podemos observar en la gráfica, que los últimos 3 datos, están muy alejados del resto. Aparentemente son casos extraordinariamente altos. Vemos que la distribución de datos tiende a alejarse hacia el lado derecho, con respecto al centro. Concluimos así, que la distribución es asimétrica positiva.

3.6.2 Polígono de frecuencias acumuladas u Ojiva.

Es un gráfico de línea ascendente, que se construye trazando sobre el eje de las abscisas, los **límites reales superiores** de cada clase, y sobre el eje de las ordenadas, las frecuencias acumuladas o las frecuencias acumuladas porcentuales. Trabajaremos la Ojiva con frecuencias acumuladas porcentuales, para el ejemplo 3.4, usando las columnas adecuadas de la tabla que construimos:

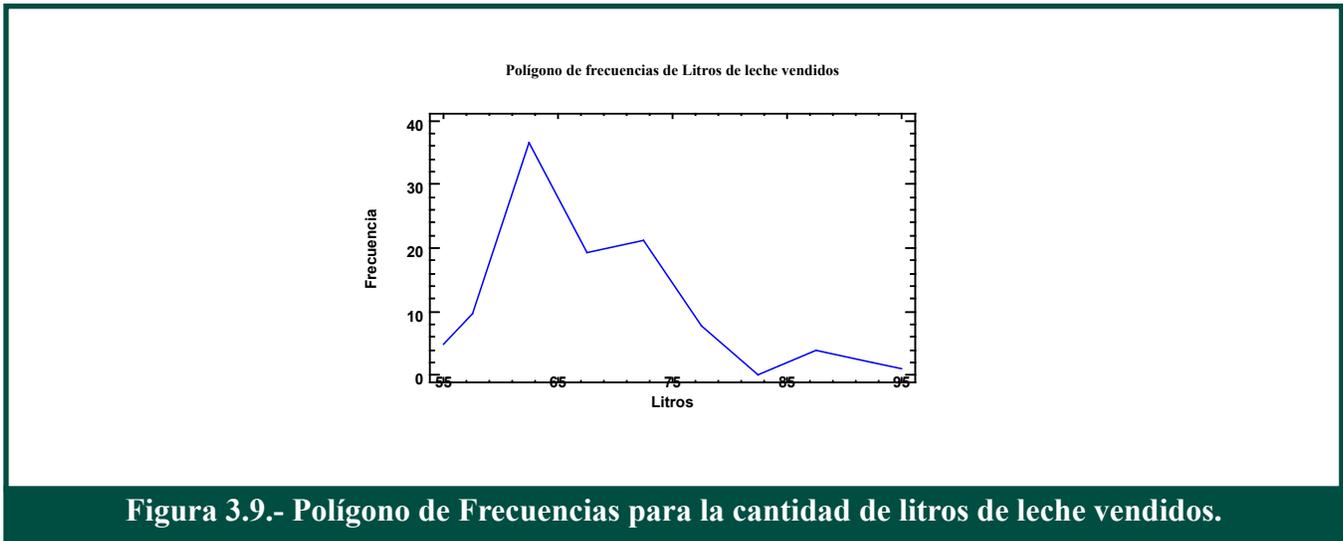
Límites reales de clase LRC	Frecuencia f_i	Frecuencias acumuladas F_i	Frecuencias acumuladas porcentuales
55-60	5	5	$5/52 = 0.096$
60-65	19	24	$19/52 = 46.14$
65-70	10	34	$34/52 = 65.38$
70-75	11	45	$45/52 = 86.54$
75-80	4	49	$49/52 = 94.23$
80-85	0	49	$49/52 = 94.23$
85-90	2	51	$51/52 = 98.08$
90-95	1	52	$52/52 = 100$



3.6.3 Polígono de Frecuencias

Es un gráfico de línea quebrada, que se construye trazando, sobre el eje de las abscisas, los centros de clase o marcas y sobre el eje de las ordenadas, las frecuencias absolutas. El gráfico no debe quedar volando, por lo que se prolongan sus extremos, hasta los centros de clase anterior y posterior a las de nuestra distribución, con objeto de que quede asentado sobre el eje X.

Límites Reales de Clase LRC	Frecuencia f_i	Centros de Clase (m_i)
55-60	5	57.5
60-65	19	62.5
65-70	10	67.5
70-75	11	72.5
75-80	4	77.5
80-85	0	82.5
85-90	2	87.5
90-95	1	92.5



Si deseáramos utilizar los datos agrupados, para realizar los cálculos de las medidas de tendencia central, tendríamos que revisar las fórmulas modificadas.

3.7 Medidas Descriptivas para Datos Agrupados

3.7.1 Medidas de Tendencia Central, para datos Agrupados

3.7.1.1 Media aritmética

$$\bar{x} = \frac{\sum_{i=1}^k f_i(m_i)}{n}$$

Substituyendo los datos de la tabla de frecuencias para datos agrupados, tenemos:

$$\bar{x} = \frac{5(57.5) + 19(62.5) + 10(67.5) + 11(72.5) + 4(77.5) + 2(87.5) + 1(92.5)}{n} = 67.788$$

Este resultado nos está mostrando que el número promedio de litros de leche vendidos en sábado es entre 67 y 68, el valor puntual es 67.788.

3.7.1.2 Mediana

Cuando los datos están agrupados, el valor de la mediana se localiza, dentro del intervalo que contiene al 50% acumulado de los datos ordenados, por lo que deberá interpolarse el valor real a partir del límite inferior del intervalo correspondiente.

Desde luego que, es más fácil localizar el intervalo mediano usando la definición de Posición de la mediana:

$$P_{Md} = \frac{n + 1}{2} = \frac{52 + 1}{2} = 26.5$$

El valor de la posición nos está indicando que la mediana está entre el dato 26 y el dato 27, por lo que habremos de interpolar en el intervalo que contenga estos 2 datos.

Observando nuestra distribución vemos que los datos 26 y 27 están incluidos en el tercer intervalo, cuyo límite inferior es 65.

La fórmula que nos permite interpolar a la mediana es:

$$M_d = L_{infMd} + \left(\frac{\frac{n}{2} - F_{Md-1}}{f_{Md}} \right) a$$

Donde L_{infMd} es el límite inferior del intervalo que contiene a la mediana

$$L_{infMd} = 65$$

F_{Md-1} , es la frecuencia acumulada en el intervalo antecedente al intervalo mediano

$$F_{Md-1} = 24$$

f_{Md} , es la frecuencia absoluta correspondiente al intervalo mediano

$$f_{Md} = 10$$

a , es la amplitud utilizada para agrupar los datos

$$a = \hat{h} = 5$$

n , es el tamaño de la muestra agrupada

$$n = 52$$

Sustituyendo, tenemos:

$$M_d = 65 \left(\frac{\frac{52}{2} - 24}{10} \right) 5 = 66$$

Este resultado nos está indicando que si ordenamos de menor a mayor, los registros correspondientes al número de litros de leche vendidos, la medida que limita el 50% acumulado es 66 litros.

3.7.1.3 Moda

La moda para datos agrupados, también se calcula por interpolación en el intervalo que presenta mayor frecuencia, en la tabla de datos agrupados.

Para el ejemplo 3.4, la moda se encuentra en el segundo intervalo, cuya frecuencia absoluta es 19.

La fórmula para interpolar a la moda es:

$$M_o = L_{infMo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) a$$

Donde L_{infMo} , es el límite inferior del intervalo con mayor frecuencia absoluta.

$$L_{infMo} = 60$$

Δ_1 , es la diferencia entre la frecuencia absoluta del intervalo modal y la frecuencia absoluta del intervalo antecedente,

$$\Delta_1 = 14$$

Δ_2 , es la diferencia entre la frecuencia absoluta del intervalo modal y la frecuencia absoluta del intervalo posterior al modal, tomada con valor absoluto

$$\Delta_2 = 9$$

a o \hat{h} , es la amplitud utilizada para agrupar los datos.

Substituyendo en la fórmula, tenemos:

$$M_o = 60 + \left(\frac{14}{14 + 9} \right) (5) = 63.043$$

Este resultado nos permite interpretar que el número de litros de leche que se compran más frecuentemente es 63.

Cuando en una distribución hay dos intervalos consecutivos con la mayor frecuencia la moda es el valor correspondiente al límite superior del primer intervalo puesto que la diferencia entre las frecuencias de dichos intervalos es cero.

Si hay más de un intervalo con la frecuencia más alta y no son consecutivos, se tendrá una distribución multimodal.

3.7.2 Medidas de Variabilidad, para datos Agrupados

3.7.2.1 Varianza

Como ya se comentó anteriormente, al agrupar los datos, cada intervalo queda representado por su centro de clase o marca (m_i) y entonces, este elemento forma parte de los cálculos de los estimadores. La forma de cálculo de la varianza es básicamente la misma pero en lugar de datos únicos, introducimos la marca de clase en la definición algebraica.

$$s^2 = \frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{n - 1}$$

$$s^2 = \frac{5(57.5-67.788)^2 + 19(62.5-67.788)^2 + \dots + 1(92.59-67.788)^2}{52 - 1} = 60.209$$

Cuando el número de datos y categorías que se manejan es grande, es conveniente utilizar una modificación algebraica de esta fórmula, para agilizar el cálculo:

$$s^2 = \frac{\sum_{i=1}^k f_i m_i^2 - n\bar{x}^2}{n - 1}$$

$$s^2 = \frac{5(57.5^2) + 19(62.5^2) + 10(67.5^2) + 4(72.5^2) + \dots + 1(92.5^2) - 52(67.788^2)}{52 - 1} = 60.209$$

3.7.2.2 Desviación Estándar

La fórmula de cálculo, quedaría así:

$$s = \sqrt{\frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{n - 1}} = \sqrt{60.209} = 7.759$$

Con la fórmula modificada algebraicamente, el resultado debe de ser el mismo:

$$s = \sqrt{\frac{\sum_{i=1}^k f_i m_i^2 - n(\bar{x}^2)}{n - 1}} = \sqrt{60.209} = 7.759$$

3.7.2.3 Coeficiente de variación

En cuanto al coeficiente de variación, su definición algebraica no cambia, con respecto a la fórmula utilizada para datos sin agrupar. Sin embargo, tanto la desviación estándar como la media, habrán sido calculadas para datos agrupados. Así, el coeficiente de variación para los datos del ejemplo 3.6 será:

$$C.V. = \left(\frac{s}{\bar{x}} \right) 100 = \left(\frac{7.759}{67.7888} \right) 100 = 11.446\%$$

De acuerdo con este resultado, podemos decir que la variación relativa en la cantidad de litros de leche vendidos es de 11.5%, con respecto a la media.

3.7.3 Medidas Posicionales o Cuantiles

Los cuantiles, son medidas posicionales que nos permiten definir medidas máximas, debajo de las cuales, se encuentran acumuladas ciertas proporciones de datos. Por ejemplo, la Mediana es un cuantil que define el límite máximo abajo del cual, se encuentra acumulado el 50% de los datos. Es importante enfatizar que, todas las medidas posicionales se obtienen cuando los datos en la muestra están ordenados de menor a mayor y estén agrupados para que haya cálculos de porcentajes acumulados.

3.7.3.1 Cuantiles, Q_i

Son medidas posicionales que dividen a la muestra en cuatro partes. Así, definir el límite para cualquiera de las cuartas partes de una muestra, implica localizar el número de datos ordenados y acumulados para delimitar esta porción, entonces hablamos de la frecuencia acumulada. Genéricamente, la vamos a definir así:

$$F_{Q_i} = \frac{n(i)}{4}$$

Donde:

F_{Q_i} es la frecuencia acumulada requerida para localizar el cuartil buscado.

n , es el tamaño de la muestra o número total de datos manejados.

i , es el número de cuarta parte requerida, $i=1,2$ y 3 .

Una vez que se calcula la frecuencia requerida, se utiliza una fórmula de interpolación, semejante a la que utilizamos para obtener la mediana agrupada. Para localizar el Q_i buscado, debemos ubicarnos en el intervalo adecuado, de acuerdo con el cálculo de F_{Q_i} .

$$Q_i = L_{inf Q_i} + \left(\frac{F_{Q_i} - F_{Q_{i-1}}}{f_{Q_i}} \right) a$$

Donde:

Q_i , es el cuartil buscado.

$L_{inf Q_i}$, es el Límite inferior del intervalo que contiene al cuartil buscado.

F_{Q_i} , es la frecuencia acumulada necesaria para localizar el cuartil buscado.

$F_{Q_{i-1}}$, es la frecuencia acumulada, hasta el intervalo anterior al que contiene el cuartil buscado.

f_{Q_i} , es la frecuencia absoluta correspondiente al intervalo que contiene al cuartil buscado.

a , es la amplitud utilizada para hacer la agrupación.

Si deseamos obtener el cuartil 3 de la distribución de datos, en el ejemplo 3, tenemos:

$$F_{Q_i} = \frac{n(i)}{4} = \frac{52(3)}{4} = 39$$

Entonces, el intervalo que contiene a Q_3 será aquel en donde la frecuencia acumulada incluya al dato 39 (**).

Límites Reales de Clase LRC	Frecuencia f_i	Frecuencias Acumuladas F_i
55-60	5	5
60-65	19	24
65-70	10	34
(**) 70-75	11	45
75-80	4	49
80-85	0	49
85-90	2	51
90-95	1	52

En la tabla vemos que el dato 39 está incluido en el cuarto intervalo, lo que significa que:

$$L_{inf Q_i} = 70; \quad F_{Q_{i-1}} = 34, \quad f_{Q_i} = 11 \quad \text{y} \quad a = 5$$

Por lo que, sustituyendo en la ecuación de cálculo correspondiente tenemos:

$$Q_i = L_{inf Q_i} + \left(\frac{F_{Q_i} - F_{Q_{i-1}}}{f_{Q_i}} \right) a = 70 + \left(\frac{39 - 34}{11} \right) 5 = 72.27$$

Este valor nos indica que si partimos la distribución de datos ordenados, al 75%, el número máximo de litros de leche comprados, sería de 27.

3.7.3.2 Percentiles P_i

Cuando llamamos percentil a la partición buscada, es porque, la muestra se está dividiendo en 100 partes. Eso quiere decir que podemos calcular percentiles desde el 1 al 99 ($i=1,2,3,\dots,98, 99$).

La fórmula para interpolar sería la misma que la de Q_i pero el cálculo de las frecuencias acumuladas, para localizar el intervalo de interpolación se haría dividiendo entre 100.

Si deseamos calcular el percentil 60, correspondiente a los datos del ejemplo 3, tendríamos:

$$F_{P_i} = \frac{n(i)}{100} = \frac{52(60)}{100} = 31.2$$

De acuerdo con este cálculo, el percentil 60 estaría localizado en el tercer intervalo.

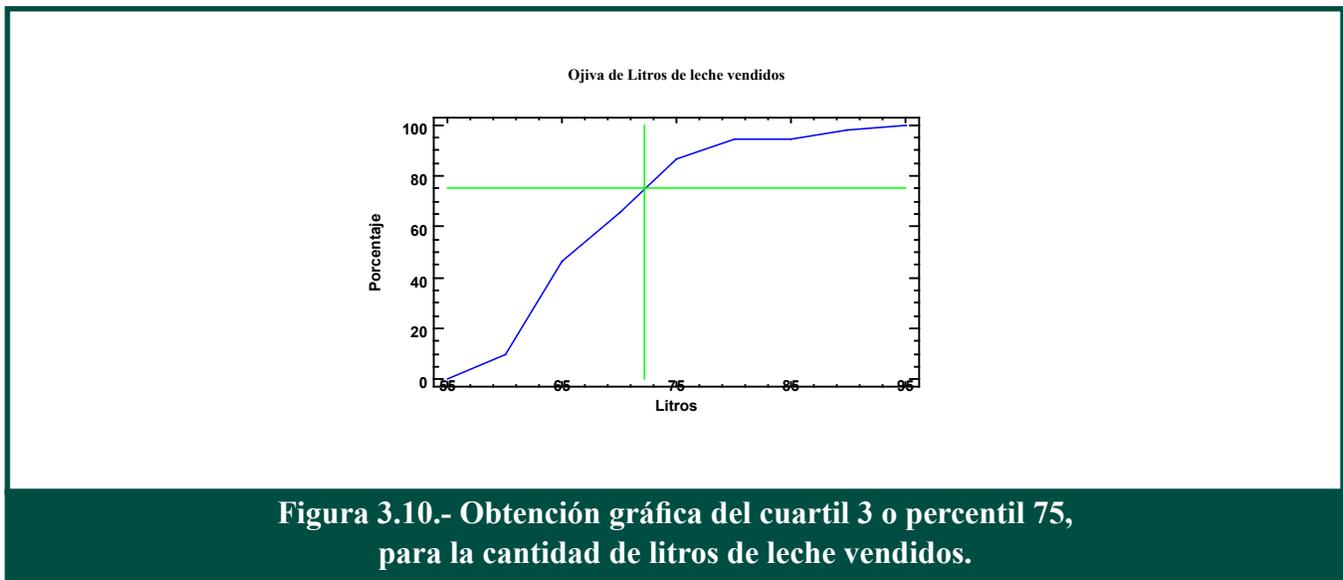
Entonces:

$$P_i = L_{inf P_i} + \left(\frac{F_{P_i} - F_{P_{i-1}}}{f_{P_i}} \right) a = 65 + \left(\frac{31.2 - 24}{10} \right) 5 = 68.6$$

Este resultado, nos está indicando que si ordenamos de menor a mayor, los datos de venta de leche y separamos el 60% acumulado, encontraremos que el número máximo de litros sería 68.6.

Hemos calculado, analíticamente, el cuartil 3 y el percentil 60, pero también podemos obtener estas medidas gráficamente. Para lograrlo, tendremos que hacer una Ojiva y sobre ella localizar los cuantiles anteriores:

Primero trazamos una línea horizontal que parte del 75, en el eje de las ordenadas hasta chocar con la línea que describe a la ojiva, en el punto de corte con la ojiva, trazamos otra línea vertical hasta cruzar el eje de las abscisas y en el punto de cruce, se localiza el valor de la variable, correspondiente a la partición cuartil 3. Desde luego que la gráfica debe estar bien acotada para que sea fácil definir el valor de la variable.



Para calcular gráficamente el percentil 60, hacemos la misma operación anterior pero trazando la horizontal sobre el valor 60 en el eje de las ordenadas. Chocamos con la ojiva y bajamos hasta cruzar el eje de las abscisas.



Figura 3.11.-Obtención gráfica del Percentil 60 para la cantidad de litros de leche vendidos.

El valor de la partición correspondiente al 60% se ubica en el eje de las abscisas 68.55 y será más preciso mientras mejor acotada esté la ojiva.

El

Nota: Quizá los Cuartiles y los Percentiles sean las medidas posicionales más relevantes por su uso, pero los cuantiles incluyen otras medidas de esa índole como los Quintiles (que dividen en cinco grupos nuestros datos totales), Deciles (en diez partes), dodeciles (en doce), etc.

Estadística Inferencial

4.1 Conceptos Básicos

Cuando hablamos de Inferencia Estadística, nos referimos a los procesos de análisis mediante los cuales, estudiamos el comportamiento de una o más poblaciones, basándonos en las distribuciones de probabilidad y los estimadores o medidas que describen el comportamiento de la(s) muestra(s). Es decir, a partir del muestreo calculamos probabilísticamente o contrastamos, los parámetros que describen el comportamiento poblacional, con objeto de predecir y tomar decisiones respecto al comportamiento poblacional.

Para inferir, debemos basarnos en la distribución muestral del estadístico que se estudia.

4.1.1 Distribución Muestral

La distribución muestral de un estadístico, es la distribución de probabilidad de todos los valores posibles que puede tomar dicho estadístico, calculados a partir de muestras del mismo tamaño, extraídas aleatoriamente de una población, con reemplazo o sin reemplazo.

Para generar una distribución muestral, se extraen todas las muestras posibles, de tamaño n , de una población discreta, de tamaño N y se calcula el estadístico de interés para cada muestra. Se grafica la distribución de valores del estadístico con sus probabilidades respectivas y se calculan los **parámetros puntuales** μ y σ .

4.1.2 Teorema Central del Límite

La distribución muestral de las medias, calculada a partir de todas y cada una de las muestras del mismo tamaño que sea posible obtener, de una población cualquiera con media μ y varianza σ^2 , será aproximadamente normal, con Media de las medias muestrales, Varianza de las medias muestrales definidas como sigue:

$$\mu_x = \mu$$

$$\sigma_x^2 = \frac{\sigma^2}{n},$$

entonces, el error estándar de las medias será: $\sigma_x = \frac{\sigma}{\sqrt{n}}$

Estos axiomas nos indican que es posible calcular un parámetro o medida poblacional en forma puntual, si tomamos todas y cada una de las muestras del mismo tamaño que sea posible obtener dentro de una población, siempre y cuando el muestreo se realice con reemplazo.

Mientras que la Media de las medias muestrales y el error estándar de las medias muestrales cuando el muestreo es sin reemplazo se definen como:

$$\mu_x = \mu$$

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \left(\frac{N-n}{N-1} \right)$$

Nótese que el error estándar, en este caso está multiplicado por un factor de corrección, esto se debe a que al no permitir el reemplazo, el número de muestras es tan pequeño que no permite valorar adecuadamente la varianza.

La distribución de medias muestrales tendrá distribución más aproximada a la normal, mientras mayor sea **n**.

EJEMPLO 4.1. De la población cuyos elementos son **4, 5, 6, 8**, extraiga todas las muestras de tamaño 3, con reemplazo y obtenga las distribuciones muestrales de la media y de la varianza.

Solución:

En la tabla 4.1 se desglosan las muestras con sus estimadores de media y varianza respectivos.

Tabla 4.1. Distribución de las muestras de tamaño 3 obtenidas de la población N.

Muestras	Frecuencia*	\bar{x}_i	s^2
4,4,4	1	4	0
4,4,5	3	4.333	0.333
4,4,6	3	4.666	1.333
4,4,8	3	5.333	5.333
5,5,5	1	5	0
5,5,4	3	4.666	0.333
5,5,6	3	5.333	0.333
5,5,8	3	6	3
6,6,6	1	6	0
6,6,4	3	5.333	1.333
6,6,5	3	5.666	0.333
6,6,8	3	6.666	1.333
8,8,8	1	8	0
8,8,4	3	6.666	5.333
8,8,5	3	7	3
8,8,6	3	7.333	1.333
4,5,6	6	5	1
4,5,8	6	5.666	4.333
5,6,8	6	6.333	2.333
4,6,8	6	6	4
Total	64		

*Los datos de frecuencia que aparecen en la tabla representan el número de veces que ocurre la misma media, debido al cambio de orden de los dígitos en las muestras y al muestreo con reemplazo.

Usando las definiciones matemáticas de la media y la varianza **de la población** obtenemos sus valores:

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{4 + 5 + 6 + 8}{4} = 5.75$$

$$\sigma^2 = \frac{\sum_{i=1}^N (\bar{x}_i - \mu)^2}{N} = \frac{(4 - 5.75)^2 + (5 - 5.75)^2 + (6 - 5.75)^2 + (8 - 5.75)^2}{4} = \frac{8.75}{4} = 2.1875$$

Las distribuciones muestrales, Media de las medias muestrales y Varianza de las medias muestrales se definen respectivamente como:

$$\mu_{\bar{x}} = \sum_{i=1}^K \bar{x}_i P(\bar{x}), \quad \sigma_{\bar{x}}^2 = \sum_{i=1}^K (\bar{x}_i)^2 P(\bar{x}_i) - \mu_{\bar{x}}^2$$

Para realizar los cálculos, se hacen unas tablas de resumen de los datos.

\bar{x}_i	$P(\bar{x}_i)$	$(\bar{x}_i)P(\bar{x}_i)$	$(\bar{x}_i)^2P(\bar{x}_i)$
4	1/64	4/64	0.25
4.333	3/64	13/64	0.880073
4.666	6/64	28/64	2.04108
5	7/64	35/64	2.7344
5.333	9/64	48/64	4
5.666	9/64	51/64	4.515
6	10/64	60/64	5.625
6.333	6/64	38/64	3.76
6.666	6/64	40/64	4.166
7	3/64	21/64	2.297
7.333	3/64	22/64	2.5206
8	1/64	8/64	1

$$\Sigma = \frac{368}{64}$$

$$\Sigma = 33.7892$$

Entonces, la media de las medias muestrales es: $\mu_{\bar{x}} = \frac{368}{64} = 5.75$

Note que este valor es idéntico al de la media poblacional, por lo tanto se cumple el axioma $\mu_{\bar{x}} = \mu$.

La varianza de las medias muestrales sería:

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^K (\bar{x}_i)^2 P(\bar{x}_i) - \mu_{\bar{x}}^2 = 33.7892 - 5.75^2 = 0.7267$$

Por axioma, en el teorema central del límite,

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \Rightarrow \sigma^2 = n(\sigma_{\bar{x}}^2) = 3(0.7267) = 2.1801$$

Como se observa, podemos despejar σ^2 a partir de la varianza de las medias muestrales, que representa el valor puntual de la varianza poblacional.

Es claro que $2.1875 \neq 2.1801$, pero esta pequeña diferencia se debe a errores de aproximación en los cálculos.

De acuerdo con lo anterior, el teorema central del límite nos establece la relación que guarda la muestra con respecto a la población y cómo, a partir de la muestra podemos obtener valores puntuales para los parámetros de la población.

Por otro lado, Media de las varianzas muestrales se define como $\mu_{s^2} = \sum_{i=1}^K s_{\bar{x}}^2 P(s_{\bar{x}}^2)$ y se sabe que $\mu_{s_{\bar{x}}^2} = \sigma^2$, por lo que, basándonos en el muestreo anterior, realizaremos la distribución de las varianzas muestrales como sigue:

Tabla 4.2. Distribución de las Varianzas muestrales.

$s_{\bar{x}}^2$	f_i	$s_{\bar{x}}^2 P(\bar{x})$
0	4	0/64
0.333	12	4/64
1	6	6/64
1.333	12	16/64
2.333	6	14/64
3	6	18/64
4	6	24/64
4.333	6	26/64
5.333	6	32/64
	$\sum f_i = 64$	$\sum = 140/64$

Entonces, la media de las varianzas muestrales sería:

$$\mu_{s^2} = \sum_{i=1}^K (s_{\bar{x}}^2) P(s_{\bar{x}}^2) = \frac{140}{64} = 2.1875$$

Por lo anterior confirmamos que:

$$\mu_{s_{\bar{x}}^2} = \sigma^2$$

Como se puede ver, es impráctico calcular los parámetros poblacionales, en forma puntual, porque el trabajo de muestreo y aritmético es arduo. El ejemplo trabajado se basó en una población muy pequeña, de sólo 4 elementos y la cantidad de cálculos para obtener los parámetros puntuales fue grande, entonces, muestrear una población grande, como sucede en la realidad, implicaría un trabajo todavía más oneroso.

Sin embargo, algo muy importante, y que hay que destacar, es un valioso concepto definido en el teorema central del límite. Nos referimos al concepto de error estándar de la media que es la raíz cuadrada de la varianza de las medias muestrales:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \Rightarrow \sigma_{\bar{x}} \sqrt{\sigma_{\bar{x}}^2} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

La definición del error estándar de la media, es importante porque sirve de base para definir los errores estándar correspondientes a las demás distribuciones muestrales, pues, conociendo el error estándar de una distribución, podemos realizar cálculos probabilísticos, por intervalo, para los diferentes parámetros de una población.

Hay que puntualizar que el error estándar de un parámetro puede minimizarse si se aumenta el tamaño de la muestra tomada para el análisis.

4.2 Estimación de Parámetros por Intervalo

Es importante, para iniciar este tema, recordar que los parámetros son medidas que describen el comportamiento de una población, y que para calcular dichos parámetros, debemos trabajar por muestreo dentro de la misma población, ante lo impráctico de censarla, por el tiempo y el costo que representa.

Como ya se demostró anteriormente, el cálculo de parámetros puntuales como los realizados en la discusión del teorema central del límite, no es una forma eficiente de trabajar y por esta razón se decide hacer uso de los modelos probabilísticos para evaluar los parámetros por intervalo.

Un intervalo de estimación se define como el segmento, sobre la recta numérica real, donde es posible localizar una medida poblacional (parámetro) buscada, con una confiabilidad establecida por el investigador.

Por ejemplo, para evaluar el parámetro Media poblacional (μ), con una confiabilidad del 95%, podríamos usar una distribución normal y desde el punto de vista gráfico se vería así:

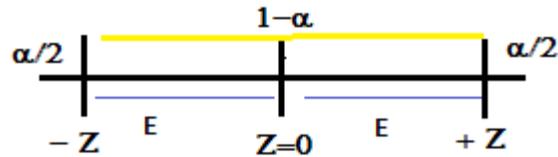


Figura 4.1. Intervalo de confianza con distribución Normal.

Donde $1-\alpha$ representa el nivel de confiabilidad con el que se realizará el cálculo de dicho parámetro.

Este símbolo se interpreta, generalmente, como la fracción central, donde se encuentra el parámetro buscado. Esto es, si $1-\alpha=0.95$, decimos que entre los límites $-Z$ y $+Z$ de la distribución normal, hay un 95% de confianza de encontrar este parámetro.

El símbolo, α , es el error probabilístico que el investigador está dispuesto a aceptar en la estimación del parámetro, también se le llama nivel de significación. Este error se divide entre dos para que se defina un área equidistante del centro, que representará el intervalo más probable para localizar al parámetro en cuestión.

La distancia E , marcada en el intervalo, se conoce como **error máximo de estimación** y está formado, generalmente por el producto del valor relativo de la distribución utilizada, (para el ejemplo, como se utilizó la normal, el valor es Z) obtenido de las tablas probabilísticas, páginas 233 a 240 del Cuaderno de problemas resueltos y propuestos de probabilidad y estadística, de Guerra D. T; Marques D. S. M. J. y López R. J. M., UNAM, FES Zaragoza y el error estándar de la misma distribución.

La línea central, en la gráfica, corresponde a un estimador cualesquiera, como por ejemplo la media muestral, \bar{X} . De lo anterior, podemos darnos cuenta que el intervalo de estimación se construye, sumando y restando el error máximo de estimación, al estimador correspondiente para el parámetro buscado, y que el error máximo queda definido por la confiabilidad, el error estándar y la distribución del estimador utilizado.

4.2.1 Ecuación General para la Estimación de Parámetros por Intervalo:

$$\hat{\theta} - E < \theta < \hat{\theta} + E \quad \text{con nivel de confianza de } 1 - \alpha$$

Donde:

$\hat{\theta}$, representa cualquier estimador o medida muestral.

E , es el error máximo de estimación.

θ , representa cualquier parámetro o medida poblacional.

Para el cálculo de parámetros por intervalo se tienen diferentes casos según los datos con que se cuente y el parámetro que se vaya a estimar.

Caso 1.- Estimación de la Media poblacional, μ , cuando la desviación estándar poblacional es dato (se conoce σ).

Este caso se utiliza para calcular el parámetro Media, por intervalo, cuando los datos con que se cuenta son el tamaño de la muestra n y la desviación estándar de la población σ . Esto implica que la distribución normal es el modelo apropiado para realizar el cálculo, por lo que valiéndonos de las tablas probabilísticas de esta distribución obtenemos el valor relativo Z y sustituimos la ecuación que define al error máximo de estimación:

$$E = Z_{1-\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

En seguida, sabiendo el valor del estimador media muestral, sustituimos la ecuación general para estimación por intervalo, anotada arriba, quedando como sigue:

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{con una confianza de } 1 - \alpha$$

Caso 2.- Estimación de la Proporción o Fracción poblacional, π .

Este caso se utiliza cuando la variable manejada es discreta, proveniente de un experimento binomial, cuya probabilidad de éxito es p y con probabilidad de fracaso q . Entonces, para hacer inferencia sobre el comportamiento poblacional es necesario hacer una aproximación mediante la distribución normal.

El error estándar de la fracción o proporción es:

$$\sqrt{\frac{p q}{n}}$$

y Y el error máximo de estimación queda así:

$$E = Z_{1-\alpha/2} \sqrt{\frac{p q}{n}},$$

Donde:

$$p = \frac{X_i}{n}, \quad q = 1 - p$$

X_i es el número de casos favorables y n el número de casos totales.

La ecuación para este intervalo es:

$$p - E < \mu < p + E \quad \text{con una confianza de } 1 - \alpha$$

Caso 3.-Estimación de la Media poblacional, μ , cuando la desviación estándar poblacional no es dato.

En este caso, no se conoce σ por lo que se utiliza la desviación estándar muestral s , y la distribución utilizada se conoce como t de student. Esta distribución es considerada una “normal” para muestras pequeñas.

El error máximo en este caso se calcula como:

$$E = t_{1-\alpha/2, n-1} \left[\frac{s}{\sqrt{n}} \right]$$

Así, la ecuación para el intervalo de la media queda como sigue:

$$\bar{x} - E < \mu < \bar{x} + E \quad \text{con una confianza de } 1 - \alpha$$

4.2.2 Distribución t de Student

Es una curva semejante a la normal, simétrica, pero más achatada en el centro y ancha en las colas, que se considera típica de las muestras pequeñas. Sin embargo, se asocia a aquellas muestras obtenidas de poblaciones cuya desviación estándar σ no se conoce, independientemente del tamaño.

La curva t toma su forma dependiendo del número de grados libres. **El número de grados libres es la cantidad de observaciones en la muestra, que son estadísticamente independientes.**

Cuando se analiza una sola variable, los **grados libres de la “ t ”** se calculan como **$n-1$** .

Es importante comentar que, a medida que el tamaño de la muestra crece, más se aproxima la forma de la t a la curva normal z .

Las tablas probabilísticas de esta distribución, tabla T-5, se encuentran en las páginas 241 y 242 del Cuaderno de Problemas resueltos y propuestos de Probabilidad y Estadística, de Guerra D. T.; Marques D. S. M. J. y López R. J. M., UNAM, FES Zaragoza, 2009. Para leer estas tablas, se calculan los grados libres **$n-1$** y se localizan en la columna izquierda de la tabla, después, en el cintillo horizontal, localizado en la parte superior, se ubica el percentil buscado, $1 - \alpha/2$, y donde se intersecten la fila y columna respectivas se tendrá el valor límite t , que se usará para definir el error máximo.

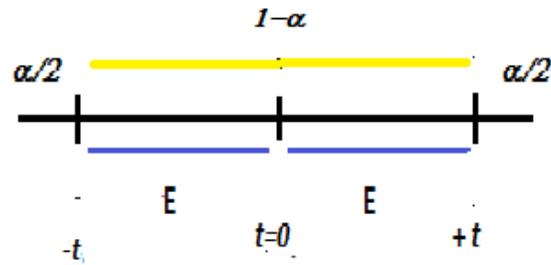


Figura 4.2. Intervalo de confianza con la distribución t de student.

Caso 4.- Estimación de la Varianza poblacional σ^2 .

4.2.3 Distribución Ji Cuadrada χ^2

Es una distribución, con área unitaria que mide la variabilidad cuadrática de un proceso aleatorio. Es asimétrica positiva, esto es, no tiene valores negativos, por lo que para definir los límites del intervalo, es necesario leer dos veces la tabla de la distribución, primero para los percentiles menores al 50% con $\alpha/2$, y después para los mayores o iguales al 50% con $1 - \alpha/2$ pues ambos valores son mayores que cero, dependiendo del tamaño de la muestra. La distribución χ^2 , toma su forma dependiendo del número de grados libres, por lo que se genera una gran familia de curvas χ^2 .

Para obtener los valores límites de la distribución Ji cuadrada, se utiliza la tabla T-6, páginas 243 a 246 del cuaderno de problemas ya mencionado. Esta tabla, se maneja en forma similar a la de la distribución t con los grados libres a la izquierda y los percentiles en el cintillo superior. El valor específico que se busca, es aquel donde se intersecta la fila de grados libres con la columna del percentil deseado. Note que hay cuatro páginas en esta tabla, las primeras dos corresponden a los percentiles menores que 50%, que pertenecen al límite izquierdo cercano al eje Y, las dos últimas páginas corresponden a los valores límites a la derecha de la distribución.

La gráfica de una distribución χ^2 , se vería así:

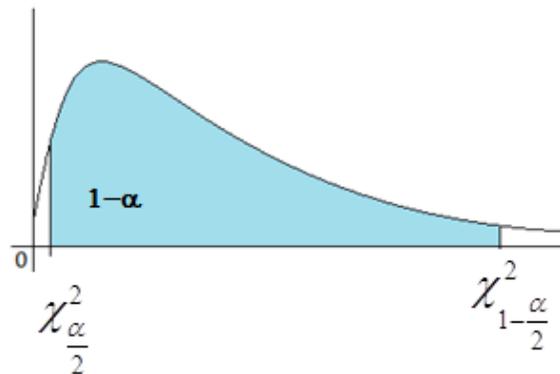


Figura 4.3.-Gráfica de la distribución Ji Cuadrada bilateral.

Cuando se desea obtener a la varianza poblacional por intervalo, se debe tomar en cuenta que el estimador apropiado es la desviación estándar al cuadrado (s^2), pero este valor forma parte a la vez del error máximo, por lo que no es posible calcular este error por separado como en los casos anteriores, por lo tanto, se sustituirá directamente la ecuación del intervalo para la varianza como sigue:

$$\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2, n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{(\alpha/2, n-1)}} = 1 - \alpha$$

4.2.4 Aplicación de Estimación de Parámetros por Intervalo

EJEMPLO 4.2. Los datos siguientes, corresponden al número de libras por hectárea, en miles, cosechadas de lúpulo y obtenidas por muestreo aleatorio en una región productora.

Cosecha de lúpulo (Lb/Ha)			
3.4	5.0	4.8	6.2
5.8	4.6	5.1	4.7
4.4	3.6	4.0	5.0
3.1	3.7	4.6	5.4
4.8	3.5	3.6	6.8
2.7	6.0	5.5	2.2
4.5	5.3	5.0	6.0

Con base en esta información:

- Obtenga la media y varianza de la muestra.
- Estime por intervalo de 95% de confianza, la cosecha media de la población.

- c) Estime por intervalo de 90% de confianza la varianza poblacional de lúpulo cosechado.
 d) Estime por intervalo de 99% de confianza, la proporción de la población con cosechas menores a 5 Lb/Ha.

Solución:

- a) Cálculo de la media y la varianza.

Para calcular estos estimadores, es conveniente usar la calculadora en formato estadístico para una variable e introducir los datos en la memoria para después pedir \bar{x} y s^2 .

$$\bar{x} = 4.6178 \quad \text{y} \quad s^2 = 1.222$$

- b) Para estimar a la media poblacional, hay que hacer notar que no tenemos el dato de la varianza poblacional y trabajaremos con la varianza muestral y esto, nos obliga a usar la distribución t de student.

Datos con que contamos

Estimadores: <i>Media muestral:</i> $\bar{x} = 4.6178$ <i>Varianza muestral:</i> $s^2 = 1.222$	Con base en el nivel de confianza calculamos: <i>Nivel de confianza:</i> $1 - \alpha = 0.95$ $\Rightarrow \alpha = 0.05$ $\alpha/2 = 0.05/2 = 0.025$ $1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$	Por lo tanto, buscaremos: $t_{(0.975, 27)} = 2.0518$
---	--	---

Calculamos el Error máximo:

$$E = t_{(1-\alpha/2, n-1)} \frac{s}{\sqrt{n}} = 2.0518 \left(\frac{\sqrt{1.222}}{\sqrt{28}} \right) = 0.4287$$

Sustituyendo en la ecuación que define el intervalo tenemos:

$$\bar{x} - E < \mu < \bar{x} + E$$

$$4.6178 - 0.4287 < \mu < 4.6178 + 0.4287 \text{ con } 95\% \text{ de confianza}$$

$$(4.1892, 5.0465) \text{ con } 95\% \text{ de confianza}$$

Interpretación.- El resultado obtenido se interpreta diciendo que: de cada 100 intervalos que se calculen, en las mismas condiciones, en 95 de ellos, la cantidad media verdadera de lúpulo cosechado estará entre 4.1892 y 5.0465 Lb/Ha, aproximadamente.

- c) Para estimar a la varianza poblacional, usamos el valor muestral de la varianza y la distribución χ^2 , leída en tablas, con $n-1$ grados de libertad.

$$1 - \alpha = 0.90 \Rightarrow \alpha = 1 - 0.90 = 0.10 \Rightarrow \alpha/2 = 0.10/2 = 0.05$$

Buscamos los 2 valores de la distribución como sigue:

$$\chi^2_{(\alpha/2, n-1)} = \chi^2_{(0.05, 27)} = 16.1514 \quad \text{y} \quad \chi^2_{(0.95, 27)} = 40.1133$$

Sustituyendo en la fórmula del intervalo para la varianza tenemos:

$$\frac{(n-1)s}{\chi^2_{(1-\alpha/2, n-1)}} < \sigma^2 < \frac{(n-1)s}{\chi^2_{(\alpha/2, n-1)}} = 1 - \alpha$$

$$\Leftrightarrow \frac{27(1.222)}{40.1133} < \sigma^2 < \frac{27(1.222)}{16.1514}$$

$$\Leftrightarrow 0.822696 < \sigma^2 < 2.043232 \text{ con } 90\% \text{ de confianza}$$

Interpretación.- Este resultado se interpreta en forma semejante a la anterior: de cada 100 intervalos que se calculen, en las mismas condiciones, en 90 de ellos se observará que la varianza verdadera, en la producción de lúpulo estará entre 0.8227 y 2.04323 (Lb/Ha)², aproximadamente.

- d) Para estimar a la proporción verdadera de cosechas menores a 5 Lb/Ha, se cuenta el número de elementos menores que 5, en la tabla de datos.

Encontramos 16 datos que cumplen con esta condición, por lo que:

$$p = \frac{X_i}{n} = \frac{16}{28} = 0.5714 \Rightarrow q = 1 - 0.5714 = 0.4286$$

Con una confianza de 99%, buscamos los valores Z de la distribución normal:

$$1 - \alpha = 0.99 \Rightarrow \alpha = 1 - 0.99 = 0.01, \text{ entonces } \alpha/2 = 0.01/2 = 0.005$$

$$1 - \alpha/2 = 1 - 0.005 = 0.995$$

Por lo tanto

$$Z_{1-\alpha/2} = Z_{0.995} = 2.5758$$

Ahora, calculamos el error máximo, como sigue:

$$E = Z_{1-\alpha/2} \sqrt{\frac{pq}{n}} = 2.5758 \sqrt{\frac{(0.5714)(0.4286)}{28}} = 0.2409$$

Sustituyendo en el intervalo correspondiente:

$$p - E < \pi < p + E$$

$$0.5714 - 0.2409 < \pi < 0.5714 + 0.2409$$

$$0.3305 < \pi < 0.8123$$

Interpretación.- De cada 100 intervalos calculados, en las mismas condiciones, en 99 de ellos, la proporción real de cosechas menores que 5 Lb/Ha estará entre 33% y 81% aproximadamente.

4.3 Contrastes de Hipótesis para un Parámetro

El proceso de contraste de hipótesis consiste en establecer un supuesto estadístico para el valor de un parámetro y seguir una secuencia de pasos para probar la validez del mismo.

4.3.1 Secuencia para Realizar el Contraste

1. **Identificar y anotar los datos con que se cuenta y la pregunta que origina el proceso.** Es muy importante que antes de empezar a hacer el análisis se tenga bien claro, que tipo de datos están disponibles, porque dependiendo de ellos, se elegirá el tipo de planteamiento, el parámetro, el estadístico de contraste, etc.
2. **Planteamiento.-** Consiste en establecer un par de hipótesis llamadas **Hipótesis Nula** e **Hipótesis Alterna** que defienden posiciones contrarias y complementarias entre sí, acerca del parámetro de interés.
 - **Hipótesis Nula: H_0 .**- Se plantea siempre usando el signo de igualdad (=) o (\leq) o (\geq) solo o acompañado de un signo de desigualdad. A esta hipótesis se le concede la mayor área bajo la curva de la distribución que mide el proceso porque corresponde a lo más usual o conocido. Ocupa un área de $1 - \alpha$.

- **Hipótesis alterna H_a .**- Se plantea en forma complementaria a la nula y sólo puede usar los signos de desigualdad diferente de (\neq), mayor que ($>$) y menor que ($<$), respectivamente.

Dependiendo del signo de desigualdad que presente H_a , se generan 3 tipos de contraste: si el signo es \neq (diferente de); el planteamiento es Bilateral o a dos colas, sería unilateral superior o de cola derecha si el signo es $>$ (mayor que) y unilateral inferior o de cola izquierda, si el signo de desigualdad es $<$ (menor que).

3. **Elección del Nivel de Significación apropiado para hacer la prueba.**- Por lo general, la mayoría de las pruebas se trabajan bajo la suposición de riesgo medio.

- **Riesgo medio ($\alpha = 0.05$)**

Este nivel se utiliza cuando el riesgo, en la toma de decisiones para la cual se realiza el proceso de contraste, no es excesivo y podemos permitirnos una probabilidad de equivocarnos, del 5%, manteniendo una confiabilidad del 95% en la decisión.

- **Riesgo bajo ($\alpha = 0.01$)**

Este nivel se utiliza cuando la decisión que se va a tomar, afecta de manera drástica a la población afectada, por ejemplo, estudios de nuevos fármacos, métodos quirúrgicos, etc., donde se pone en riesgo la salud de las personas involucradas; por lo tanto, la probabilidad de equivocarnos debe ser pequeña, y se trabajaría con una confiabilidad del 99%.

- **Riesgo alto ($\alpha = 0.10$)**

Este nivel se utiliza cuando sólo se intenta obtener una ligera noción de cómo andan las cosas y aún no se va a tomar una decisión al respecto, por lo que podemos correr un poco de mayor riesgo de equivocarnos. Se trabajaría con una confianza del 90%.

4. **Elección del estadístico de contraste correspondiente al parámetro y a los supuestos de la prueba** Cuando se realiza un contraste de hipótesis se debe recordarse que se está probando el valor de un parámetro y que cada estimador se distribuye de una manera específica y esto influye en los cálculos del proceso y en el modelo elegido para la comparación.
5. **Establecimiento de la región crítica o de rechazo de la hipótesis nula. Regla de Decisión.**- La regla de decisión es una gráfica, que representa a la distribución del estimador elegido, donde se habrán marcado las regiones de rechazo de la hipótesis nula, con base en las tablas probabilísticas de la misma distribución. Esta gráfica se utiliza como patrón de referencia para tomar una decisión acerca de la hipótesis planteada.

Un ejemplo de Regla de decisión bilateral se vería así:

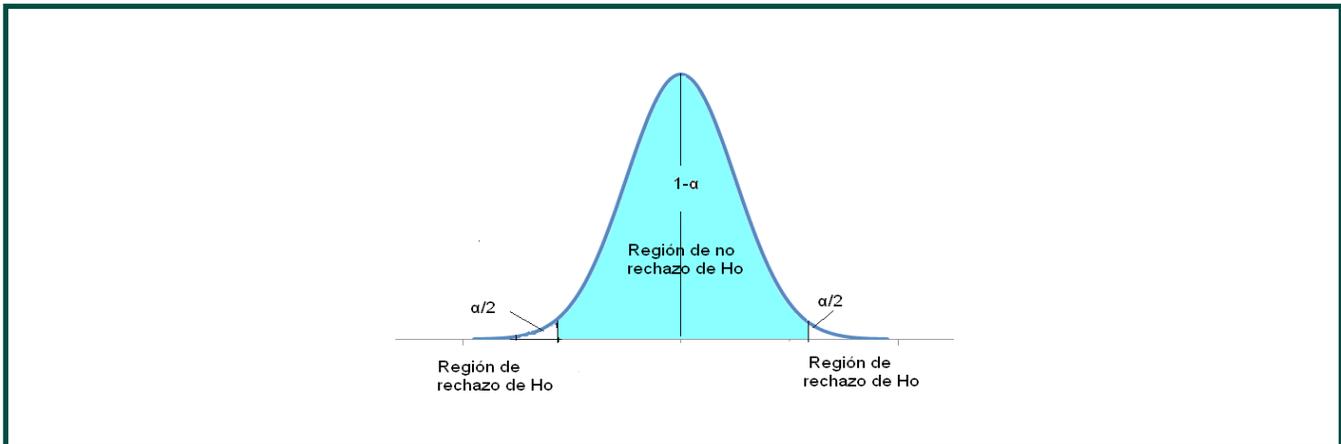


Figura 4.4.-Regla de decisión para un análisis bilateral.

Un ejemplo de Regla de decisión unilateral superior se vería así:

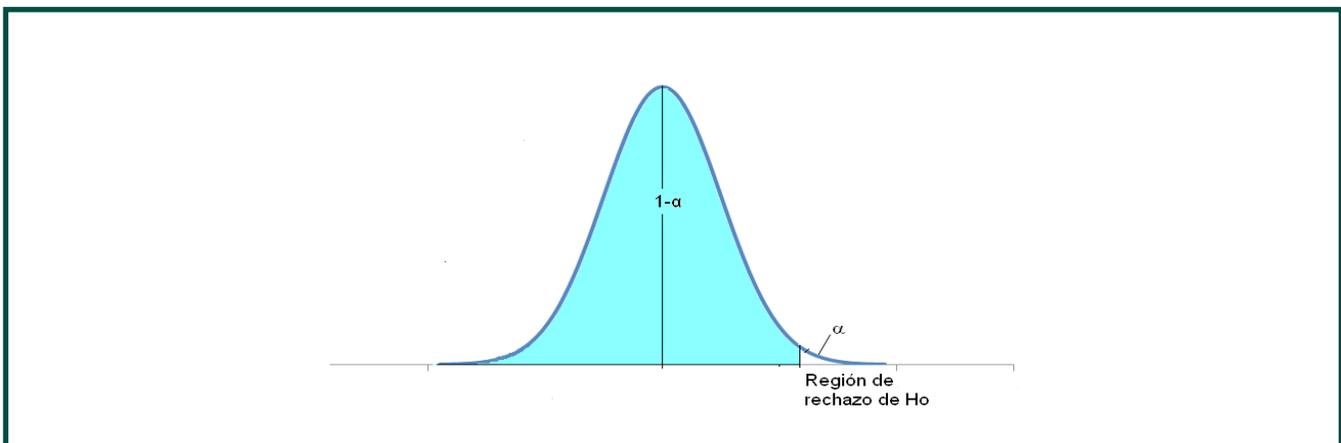


Figura 4.5.-Regla de decisión para un análisis Unilateral superior.

La región de rechazo de la hipótesis nula es el espacio definido como α cuando el planteamiento de las hipótesis es unilateral o de una cola, o como $\alpha/2$ cuando el planteamiento es bilateral.

6. **Cálculo del Estadístico de Contraste.-** Es una fórmula que define un valor calculado para el parámetro de la distribución de probabilidad elegida en el paso 4 y que se contrastará con las áreas marcadas en la regla de decisión.

7. **Toma de Decisión.**- Dependiendo del contraste entre el estadístico y la regla de decisión, se decide **rechazar la hipótesis nula** si el estadístico cae en región de rechazo (colas de la distribución) y **no rechazar H_0** si el estadístico cae fuera de las zonas de rechazo.
8. **Conclusión y comentario.**-En este rubro se comenta acerca de la decisión tomada y su significado, con respecto al problema que originó el proceso de prueba.

4.3.2 Aplicación del Proceso de Contraste

EJEMPLO 4.5. Las estimaciones de biomasa de la Tierra (cantidad de vegetación que hay en los bosques terrestres), son importantes para determinar cuánto bióxido de carbono es posible que no absorba la atmósfera de la Tierra. Suponga que una muestra de 75 parcelas, de un metro cuadrado, seleccionadas aleatoriamente en los bosques del norte de Estados Unidos, dan una biomasa promedio de 4.2 Kg/m² y una desviación estándar de 1.5 Kg/m².

- a) ¿Se podría asegurar, con una significación del 5% que la biomasa promedio real es como máximo 5 Kg/m²?
- b) ¿La varianza poblacional es menor que 2 (Kg/m²)²? Use $\alpha = 0.05$.

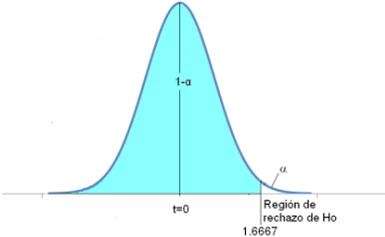
Solución:

- a) Los datos con que contamos, no incluyen la desviación estándar poblacional, por lo tanto tendremos que utilizar la distribución t de student para realizar el contraste de hipótesis.

Nos piden demostrar que la media poblacional, μ , es como máximo 5, esto es, menor o igual a 5, por lo que la hipótesis alterna (H_a) defenderá valores arriba de 5 y entonces, estará colocada en la cola derecha de la distribución “ t ”, esto significa que la región de rechazo de la hipótesis nula estará a la derecha, con valor positivo de “ t ”.

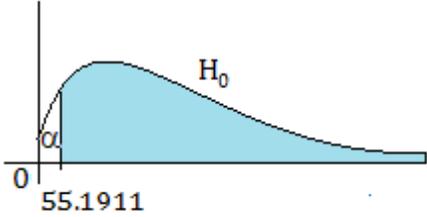
Como sólo hay una región de rechazo, el nivel de significación α se encuentra en la cola derecha de la distribución, leyéndose el valor de la t teórica o de tablas (valor crítico) y asignándole signo positivo.

1) Datos	2) Planteamiento de las hipótesis.	3) Nivel de Significación
$\bar{x} = 4.2 \text{ Kg/m}_2$ $n = 75 \text{ } s = 1.5 \text{ Kg/m}_2$ $\alpha = 0.05$ ¿Media de biomasa es como máximo 5? ¿ $\mu \leq 5$?	Unilateral Superior $H_0: \mu \leq 5$ $H_a: \mu > 5$	$\alpha = 0.05$

<p>4) Distribución utilizada y elección del estadístico de contraste.</p> $\bar{X} \sim t_{(1-\alpha, n-1)}$ $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	<p>5) Regla de Decisión o Región crítica.</p> <p>Calculando el percentil $1 - \alpha$ y leyendo en las tablas probabilísticas tenemos: $t_{(1-\alpha, n-1)} = t_{(0.95, 74)} = 1.6667$ Este valor se obtiene por interpolación entre 70 y 75, para 74 grados de libertad.</p> 	<p>6) Cálculo del Estadístico de Contraste</p> $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{4.2 - 5}{\frac{1.5}{\sqrt{75}}} = -4.62$
<p>7) Decisión</p> <p>No se rechaza H_0, porque el estadístico no está en región de rechazo.</p>	<p>8) Conclusión:</p> <p>No hay suficiente evidencia para concluir que la biomasa promedio real sea como máximo de 5 kg/m^2</p>	

b) Nos piden probar que la varianza verdadera, σ^2 , es menor que $2 \text{ (Kg/m}^2\text{)}^2$, por lo que el análisis será unilateral inferior, esto es, H_a se encontrará en la cola izquierda de la distribución de probabilidad Ji-cuadrada. Debe recordarse que esta distribución no es simétrica por lo que se buscará el valor teórico de la misma, dependiendo del lugar donde se ubique la región de rechazo, a la izquierda o a la derecha. En este caso específico, utilizaremos el lado izquierdo de la distribución.

<p>1) Datos</p> $\bar{x} = 4.2 \text{ Kg/m}^2$ $n = 75$ $s = 1.5 \text{ Kg/m}^2$ $\alpha = 0.05$ <p>¿Varianza de biomasa menor que 2? $\sigma^2 < 2?$</p>	<p>2) Planteamiento de las hipótesis</p> <p>Unilateral Inferior</p> $H_0: \sigma^2 \geq 2$ $H_a: \sigma^2 < 2$	<p>3) Nivel de Significación</p> $\alpha = 0.05$
--	--	---

<p>4) Distribución utilizada y elección del estadístico de contraste.</p> $s^2 \sim \chi^2_{(\alpha, n-1)}$ $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	<p>5) Regla de Decisión o Región crítica:</p>  <p>leyendo en las tablas probabilísticas con $\alpha = 0.05$ tenemos: $\chi^2_{(\alpha, n-1)} = \chi^2_{(0.05, 74)} = 55.19114$ Este valor sale por interpolación entre 70 y 75, para 74 g.l.</p>	<p>6) Cálculo del Estadístico de Contraste</p> $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{74(1.5^2)}{2}$ $= 83.25$
<p>7) Decisión: No se rechaza H_0, porque el estadístico no está en región de rechazo.</p>	<p>8) Conclusión: No hay suficiente evidencia para afirmar que la varianza real sea menor que 2 (Kg/m²)²</p>	

EJEMPLO 4.6. Un ingeniero químico realizó 36 mediciones de la profundidad, a la que un roto martillo puede introducir clavos para concreto en un muro, en centímetros, obteniendo los datos siguientes:

2.77	2.28	2.40	2.46	2.76	2.73	2.53	2.65	2.47
2.68	2.71	2.34	2.50	2.32	2.50	2.51	2.55	2.67
2.43	2.91	2.63	2.40	2.65	2.60	2.33	2.62	2.35
2.25	2.52	2.41	2.74	2.47	2.27	2.64	2.54	2.50

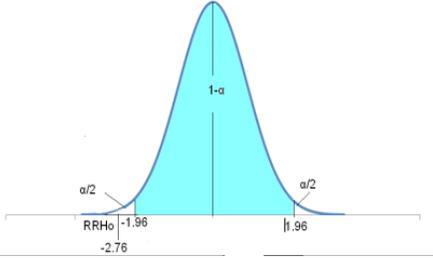
Si estos datos se consideran una muestra aleatoria, representativa de la capacidad del instrumento, y la varianza verdadera en la profundidad, es de 0.23 cm².

- a) Podríamos asegurar que la media verdadera es de 2.6 centímetros?
- b) Podríamos decir, con una significación del 5%, que la proporción de profundidades de al menos 2.6 centímetros es como mínimo 0.4?

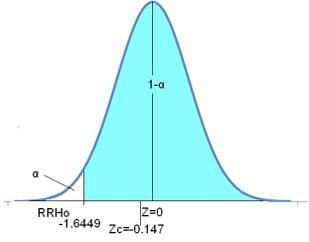
Solución:

- a) Nos preguntan si la media poblacional es 2.6, esto implica que el contraste será bilateral o de dos colas, pero además nos dan el dato de la varianza poblacional, lo que nos permite usar una distribución Normal, con parámetro Z.

Como nos están dando los datos crudos, hay que obtener la media muestral.

<p>1) Datos $\bar{x} = 2.53 \text{ cm}^2$ $n = 36$ $s = 0.23 \text{ cm}^2$ $\alpha = 0.05$</p> <p>¿Mediade profundidad es 2.6? ¿$\mu = 2.6$?</p>	<p>2) Planteamiento de las hipótesis</p> <p>Bilateral</p> <p>$H_0: \mu = 2.6$ $H_a: \mu \neq 2.6$</p>	<p>3) Nivel de Significación</p> <p>$\alpha = 0.05$</p> <p>Se divide α entre 2 para crear 2 regiones de rechazo.</p>
<p>4) Distribución utilizada y elección del estadístico de contraste.</p> <p>$\bar{x} \sim Z_{(1-\alpha/2)}$ $\alpha = 0.05; \alpha/2 = 0.025$ $1 - \alpha/2 = 0.975$</p> <p>Leyendo en la columna E de la Normal, tenemos que: $Z_{(0.975)} = \pm 1.96$</p>	<p>5) Regla de Decisión o Región crítica:</p> 	<p>6) Cálculo del Estadístico de Contraste</p> $Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{2.53 - 2.6}{\frac{0.1516575}{\sqrt{36}}} = -2.7694$
<p>7) Decisión: Se rechaza H_0, porque el valor del estadístico calculado es -2.7694 cae en la región de rechazo del lado izquierdo.</p>	<p>8) Conclusión: Con una significación del 5% se establece, con una significación del 5% que la media es significativamente diferente de 2.6.</p>	

- c) Ahora, el parámetro a contrastar es la proporción poblacional π . Nos preguntan si la proporción de profundidades de al menos 2.6 es como mínimo 0.4, o lo que es lo mismo, mayor o igual a 0.4. Por lo que nuestro análisis será unilateral inferior. Para obtener la proporción muestral de las profundidades de al menos 2.6, contamos los casos favorables para este evento, en la muestra y vemos que son 14

<p>1) Datos $n = 36$ $p = \frac{14}{36} = 0.38\bar{8}$ $q = 1 - p = 1 - 0.38\bar{8} = 0.61\bar{1}$ $\alpha = 0.05$ ¿Proporción es como mínimo 0.4? ¿ $\pi \geq 0.4$?</p>	<p>2) Planteamiento de las hipótesis Unilateral inferior $H_0: \pi \geq 0.4$ $H_a: \mu < 0.4$</p>	<p>3) Nivel de Significación $\alpha = 0.05$ Por ser una hipótesis unilateral inferior no se divide α entre 2 y sólo habrá una región de rechazo del lado izquierdo, por lo que el valor crítico será negativo.</p>
<p>4) Distribución utilizada y elección del estadístico de contraste. $p \sim Z_{(1-\alpha)}$ $\alpha = 0.05 \Leftrightarrow 1 - \alpha = 0.95$ Leyendo en la columna E de la Normal, tenemos que: $-Z_{(0.95)} = -1.6449$</p>	<p>5) Regla de Decisión o Región crítica:</p> 	<p>6) Cálculo del Estadístico de Contraste</p> $Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.388 - 0.4}{\sqrt{\frac{0.4(0.6)}{36}}}$ <p>= -0.147</p>
<p>7) Decisión: No se rechaza H_0, porque el valor del estadístico calculado es -0.147 no cae en la región de rechazo.</p>	<p>8) Conclusión: Con una significación del 5% se establece que la proporción de las profundidades de al menos 2.6 cm no es significativamente menor que 0.4.</p>	

4.4 Inferencia en Comparación de 2 Grupos de datos

4.4.1 Estimación por Intervalo

Cuando se estima un parámetro surgido de la comparación de dos grupos, se crean parámetros como:

- **Diferencia de Medias poblacionales Independientes** $\mu_1 - \mu_2$
- **Diferencia de Proporciones poblacionales** $\pi_1 - \pi_2$
- **Razón o relación de Varianzas poblacionales** $\frac{\sigma_1^2}{\sigma_2^2}$
- **Análisis de Datos Pareados o Diferencia de Medias Dependientes:** $\mu_{1-2} - \mu_D$

Esta forma de establecer los parámetros obedece a la situación de que la comparación se basa en operaciones básicas de la aritmética simple y llana. De tal manera que el cálculo de los intervalos respectivos se verá afectado por las reglas de la aritmética.

- Si se comparan 2 grupos por sus medias o por sus proporciones, los intervalos resultantes pueden ser positivos por ambos extremos, negativos por ambos extremos o negativos por un extremo y positivos por el otro, dependiendo del sentido en que se realiza la resta, esto es, cuál de los grupos ocupa el lugar del minuendo y cual ocupa el del sustraendo.

Por ejemplo: en la resta $17 - 5 = 12 \rightarrow$ minuendo - sustraendo = diferencia, el resultado es positivo, mientras que si la resta se efectúa de esta forma: $5 - 17 = -12$, el resultado es negativo. De la misma forma, si no hay una diferencia significativa en las medidas comparadas, el intervalo incluirá el cero.

- Las variaciones se comparan como un cociente porque la variación puede ser de igual valor pero de sentido contrario y la resta no reflejaría esta situación ya que +3 menos -3 daría cero, dando la idea de que no existe variación.

El resultado del cociente puede ser mayor que 1 o menor que 1 dependiendo que grupo ocupe el lugar del numerador y cual ocupe el denominador.

Por ejemplo: $\frac{8}{4} = 2 \rightarrow \frac{\text{numerador}}{\text{denominador}} = \text{cociente}$, en este caso el cociente es mayor que 1, mientras que si

el 8 ocupa el lugar del denominador, el cociente sería 0.5 que es menor que la unidad.

Sin importar la forma como se tomen las restas y los cocientes, la conclusión del análisis deberá ser la misma.

Cuando se trabaja un análisis para la relación de las varianzas poblacionales, se debe tomar en cuenta que cada varianza se distribuye de acuerdo con una Ji-cuadrada, entonces tendremos la relación de 2 distribuciones χ^2 , una en el numerador, cuyos grados de libertad corresponden al tamaño de la muestra menos 1, y otra en el denominador, cuyos grados libres también serán n-1. ***No es obligatorio que la comparaciones realicen usando muestras del mismo tamaño por lo que los grados libres de ambas distribuciones pueden diferir.***

La relación o cociente entre dos distribuciones χ^2 , genera una nueva distribución llamada Distribución F de Fisher.

4.4.2 Distribución F de Fisher

Esta distribución mide la relación entre 2 varianzas, es asimétrica positiva, semejante a la χ^2 , pero más esbelta. Tiene área unitaria y toma su forma dependiendo de 2 tipos de grados libres, los de la muestra en el numerador y los de la muestra en el denominador. Por esta razón, se generarán un número muy grande de curvas F, tantas como parejas de grados libres diferentes se puedan generar. Cuando se estima

por intervalo a la razón de varianzas poblacionales o cuando se prueba una hipótesis bilateral acerca de la razón de 2 varianzas la distribución presenta 2 regiones, $\alpha/2$, pero si se trabaja una hipótesis unilateral se tendrá una región derecha que vale $1 - \alpha$ o izquierda, que vale α , según sea el caso. Gráficamente, la distribución **F** bilateral, se ve así:

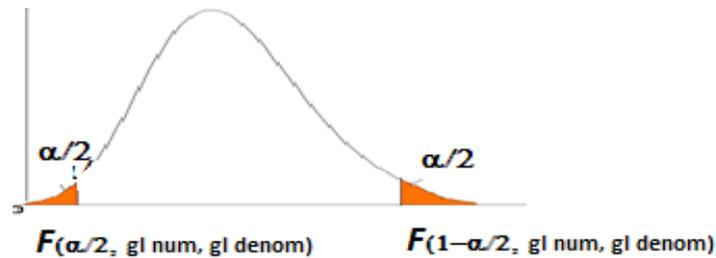


Figura 4.8.- Distribución **F** de Fisher, bilateral.

La distancia al primer valor crítico sería $\alpha/2$ (pegado al eje de las ordenadas) y la distancia al valor crítico de cola derecha sería $1-\alpha/2$, cuando se dibuja un intervalo para razón de varianzas o cuando se hace una prueba bilateral para el mismo parámetro.

Hay una tabla F diferente para cada percentil, los grados libres del numerador se localizan horizontalmente, mientras que los grados libres del denominador se localizan verticalmente en la misma tabla, de tal manera que el valor **F** adecuado para los cálculos será aquel donde se cruzan la línea horizontal con la vertical respectivas. Es importante señalar que las tablas de esta distribución permiten ubicar directamente los valores de la cola derecha de la distribución. Sin embargo, para localizar el valor de la cola izquierda, es necesario tomar el inverso del valor de cola derecha pero con los grados libres cambiados.

Por ejemplo, si la muestra en el numerador es $n_1 = 10$ y la muestra en el denominador es $n_2 = 13$ y deseáramos calcular el intervalo de 99% para la razón de 2 varianzas poblacionales, tendríamos:

$$1 - \alpha = 0.99 \Leftrightarrow \alpha = 1 - 0.99 = 0.01, \text{ entonces } \alpha/2 = 0.01/2 = 0.005 \text{ y } 1 - \alpha/2 = 0.995$$

Con base en lo anterior, tendremos que buscar el valor crítico, en la tabla T-7, percentiles de la distribución **F** con $1-\alpha/2$ igual a 0.995 y 9 grados libres en el numerador y 12 grados libres en el denominador, del Cuaderno de Problemas de Probabilidad y Estadística de Guerra, T., Marques, M.J. y López, J.M.,

Para $F_{(1-\alpha/2, n, d)} = F_{(0.995, 9, 12)} = 5.202$ que corresponde al valor crítico del lado derecho de la distribución.

Para buscar el valor crítico del lado izquierdo tenemos:

$$F_{(\alpha/2, 9, 12)} = F_{(0.005, 9, 12)} = \frac{1}{F_{(0.995, 12, 9)}} = \frac{1}{6.227} = 0.16059$$

Si dibujamos la distribución **F**, con estos límites, se vería así:

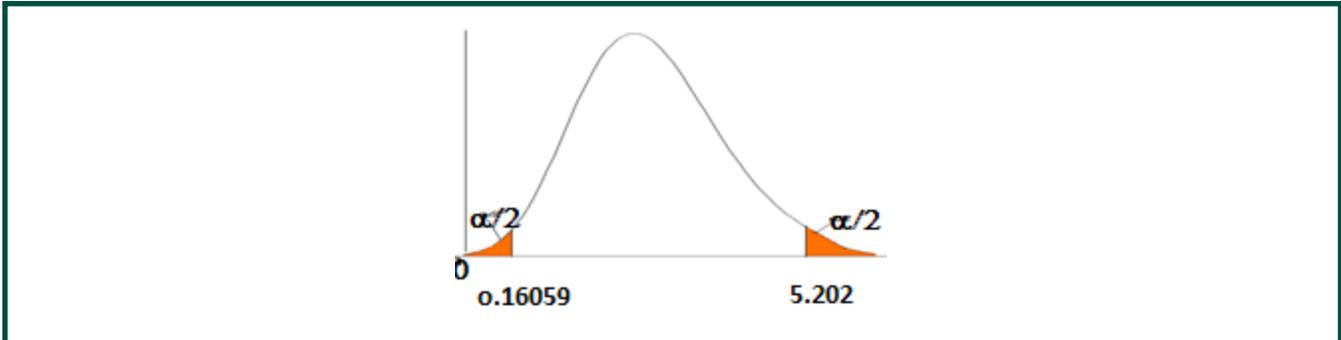


Figura 4.9.- Valores críticos de la distribución F.

4.4.3 Aplicación del Proceso de Estimación en Comparación de Grupos

EJEMPLO 4.7. Durante más de 15 años se realizaron encuestas entre adultos estadounidenses, respecto a sus hábitos relacionados con la salud. En 1991 se encuestó a 1251 adultos y en 2006 se entrevistó a 1200 adultos. La tabla siguiente muestra algunos de los resultados, que representan las proporciones respectivas para cada característica:

Hábitos	1991	2006
Consume la cantidad prescrita de fibra	$p_1 = 0.60$	$p_2 = 0.51$
Evitan las grasas	$p_1 = 0.56$	$p_2 = 0.49$
Evitan el exceso de sal	$p_1 = 0.54$	$p_2 = 0.44$

De acuerdo con estos resultados, ¿podríamos decir que los adultos actuales tienen menos cuidado con su salud, si tomamos como referencia el consumo de fibra? Use $1 - \alpha = 0.95$.

Solución:

De acuerdo con la pregunta, tendremos que comparar la diferencia entre las proporciones, de ambas poblaciones, sobre el consumo de fibra. Como no tenemos un valor supuesto a probar para la diferencia entre los grupos, realizaremos un intervalo de confianza para el parámetro diferencia de proporciones poblacionales $(\pi_{1991} - \pi_{2006})$.

$$(p_1 - p_2) - E < \pi_1 - \pi_2 < (p_1 - p_2) + E \quad \text{con nivel de confianza } 1 - \alpha = 0.95$$

Cálculo del error máximo de estimación:

$$E = Z_{1-\alpha/2} \sqrt{\frac{p_1 q_1 + p_2 q_2}{n_1 + n_2}}$$

Tomaremos los datos de 1991 como muestra 1 y los datos de 2006 como muestra 2.

$$E = Z_{1-\alpha/2} \sqrt{\frac{p_1 q_1 + p_2 q_2}{n_1 + n_2}} = 1.96 \sqrt{\frac{(0.60)(0.40)}{1251} + \frac{(0.51)(0.49)}{1200}} = 0.0392$$

Sustituyendo en la ecuación que define el intervalo, tenemos:

$$(0.6 - 0.51) - 0.0392 < \pi_{1991} - \pi_{2006} < (0.6 - 0.51) + 0.0392 \quad \text{con nivel de confianza de } 0.95\%$$

$$0.09 - 0.0392 < \pi_{1991} - \pi_{2006} < 0.09 + 0.0392$$

$$(0.0508, 0.1292)$$

Interpretación.- Como el intervalo para la diferencia de proporciones es positivo por ambos lados, concluimos que la tendencia favorece a las personas encuestadas en 1991, lo que sería indicativo de que, con un 95% de confianza, los adultos, en 1991 cuidaban más sus hábitos para la salud que los adultos actuales.

EJEMPLO 4.8.- En un estudio, cuyo objetivo es evaluar si las calificaciones que se obtienen en un examen general de conocimientos, difieren de acuerdo con el área de especialización de los estudiantes, se registraron las calificaciones obtenidas por 15 estudiantes de ingeniería y 18 estudiantes de filosofía, como sigue:

Área	Habilidad verbal		Matemáticas	
	\bar{x}	s	\bar{x}	s
Ingeniería	446	42	548	57
Filosofía	534	40	517	52

- Con una confianza de 95%, estime por intervalo, la razón de varianzas para las calificaciones en habilidad verbal, entre estudiantes de ambas áreas.
- Estime con 95% de confianza a la diferencia de medias en habilidad verbal para estudiantes de ambas áreas.
- Estime la razón de varianzas para las calificaciones en matemáticas, de los estudiantes de ambas áreas, con una confianza de 90%.

- d) Estime la diferencia de medias en las calificaciones de matemáticas para los estudiantes de ambas áreas, con una confianza de 90%.

Solución:

Al revisar los datos, podemos darnos cuenta que son datos muestrales y por lo tanto, las varianzas poblacionales son desconocidas y esto nos obliga a utilizar la distribución t de student. Aun sabiendo esto, para elegir la fórmula apropiada para estimar la diferencia de medias, cuando las varianzas poblacionales no son dato, es necesario analizar si las varianzas poblacionales se podrían considerar semejantes o diferentes.

Por esta razón, se pide primero la estimación para la razón de varianzas poblacionales de ambos grupos.

a) Puesto que el intervalo pedido es para la razón de varianzas poblacionales, primero obtenemos las varianzas muestrales, elevando al cuadrado las desviaciones estándar respectivas.

La fórmula del intervalo para la relación de varianzas es:

$$\frac{s_{Ing}^2}{(s_{Fil}^2)(Z_{(1-\alpha/2, n, d)})} < \frac{\sigma_{Ing}^2}{\sigma_{Fil}^2} < \frac{s_{Ing}^2}{(s_{Fil}^2)(Z_{(\alpha/2, n, d)})} \quad \text{con nivel de confianza } 1 - \alpha$$

Datos con que se cuenta:

$$\begin{array}{lll} n_{Ing} = 15 & s_{Ing}^2 = 42^2 = 1764 & 1 - \alpha = 0.95 \Leftrightarrow \alpha = 0.05 \\ & & \alpha/2 = 0.05/2 = 0.025 \\ n_{Fil} = 18 & s_{Fil}^2 = 40^2 = 1600 & 1 - \alpha/2 = 1 - 0.025 = 0.975 \end{array}$$

El valor crítico para 14 grados libres en el numerador, se obtiene por interpolación lineal entre 12 y 15 grados, como sigue:

$$F_{(0.975, 12, 17)} = 2.825 \quad \text{y} \quad F_{(0.95, 15, 17)} = 2.723$$

$$\text{Entonces: } \begin{cases} 12 \rightarrow 2.825 \\ 14 \rightarrow X = 2.757, \text{ por lo que: } F_{(0.975, 14, 17)} = 2.757 \\ 15 \rightarrow 2.723 \end{cases}$$

Para el valor crítico de la izquierda se hace lo mismo, interpolando ahora entre 15 y 20 para obtener el valor F con 17 grados en el numerador.

$$F_{(0.975, 15, 14)} = 2.949 \quad \text{y} \quad F_{(0.975, 20, 14)} = 2.844$$

$$\text{Entonces: } \begin{cases} 15 \rightarrow 2.949 \\ 17 \rightarrow X = 2.907 \rightarrow F_{(0.975, 17, 14)} = 2.907 \\ 20 \rightarrow 2.844 \end{cases}$$

$$F_{(0.025, 14, 17)} = \frac{1}{F_{(1-\alpha/2, 17, 14)}} = \frac{1}{F_{(0.975, 17, 14)}} = \frac{1}{2.907} = 0.3440$$

Teniendo los valores críticos F, sustituimos el intervalo.

$$\frac{1764}{(1600)(2.757)} < \frac{\sigma_{Ing}^2}{\sigma_{Fil}^2} < \frac{1764}{(1600)(0.2440)} \Rightarrow (0.3999, 3.2049)$$

Interpretación.- Como el intervalo va de 0.3999 a 3.2049, podemos afirmar que este intervalo contiene a la unidad, por lo que las varianzas pueden considerarse semejantes al 95% de confianza. Esto significa que el comportamiento de los alumnos de ingeniería muestra una **variación semejante** a la de los alumnos de filosofía en lo referente a las calificaciones de habilidad verbal.

- b) Se pide comparar el comportamiento promedio en habilidad verbal, para los grupos analizados. Usaremos una distribución t de student, con grados libres $n_1 + n_2 - 2 = 15 + 18 - 2 = 31$ y una confianza de 95%.

$$1 - \alpha = 0.95 \Leftrightarrow \alpha = 1 - 0.95 = 0.05 \Leftrightarrow \alpha/2 = 0.025 \quad y \quad 1 - \alpha/2 = 1 - 0.025 = 0.975$$

Así, el valor crítico para la distribución t se busca en tablas con 0.975 y 31 grados libres:

$$t_{(0.975, 31)} = 2.0395$$

La fórmula para el cálculo del intervalo para la diferencia de medias poblacionales independientes, cuando no se conocen las varianzas poblacionales pero se consideran semejantes (resultado del inciso anterior) es:

$$(\bar{x}_{Ing} - \bar{x}_{Fil}) - E < \mu_1 - \mu_2 < (\bar{x}_{Ing} - \bar{x}_{Fil}) + E \quad \text{con nivel de confianza de } 1 - \alpha$$

$$E = t_{(1-\alpha/2, (n_{Ing}+n_{Fil}-2))} S_p \sqrt{\frac{1}{n_{Ing}} + \frac{1}{n_{Fil}}}$$

Donde S_p es la desviación estándar, ponderada, para las 2 muestras, y se calcula así:

$$S_p = \sqrt{\frac{(n_{Ing} - 1)s_{Ing}^2 + (n_{Fil} - 1)s_{Fil}^2}{n_{Ing} + n_{Fil} - 2}}$$

Calculando la desviación ponderada:

$$S_p = \sqrt{\frac{(15 - 1)(1764) + (18 - 1)(1600)}{15 + 18 - 2}} = 40.915$$

Calculando el estimador:

$$\bar{x}_{Ing} - \bar{x}_{Fil} = 446 - 534 = -88$$

Calculando el Error máximo de estimación:

$$E = (2.0395)(40.915) \sqrt{\frac{1}{15} + \frac{1}{18}} = 29.172$$

Sustituyendo en el intervalo tenemos:

$$-88 - 29.173 < \mu_1 - \mu_3 < -88 + 29.173$$

$$(-117.173, -58.827)$$

Interpretación: Como ambos valores son negativos y la diferencia fue media de ingenieros menos media de filósofos, concluimos que de cada 100 intervalos que se calculen, comparando estudiantes de estas áreas, en 95 de ellos encontraremos que los estudiantes de filosofía parecen tener mayor habilidad verbal promedio que los estudiantes de ingeniería. Sin embargo, para confirmar lo anterior, sería necesario realizar un contraste unilateral.

- c) Nos piden estimar la razón de varianzas de las calificaciones de los estudiantes, en matemáticas, por lo que tendremos que construir el intervalo con una confianza de 90%.

Datos con que se cuenta:

$$\begin{array}{ll} n_{Ing} = 15 & s_{Ing}^2 = (57)^2 = 3249 \\ n_{Fil} = 18 & s_{Fil}^2 = (52)^2 = 2704 \end{array} \quad \left\{ \begin{array}{l} 1 - \alpha = 0.90 \therefore \alpha = 0.10 \\ \alpha/2 = 0.10/2 = 0.05 \\ 1 - \alpha/2 = 1 - 0.05 = 0.95 \end{array} \right.$$

El valor crítico para 14 grados libres en el numerador, se obtiene por interpolación lineal entre 12 y 15 grados, como sigue:

$$F_{(0.95, 12, 17)} = 2.381 \text{ y } F_{(0.95, 15, 17)} = 2.308$$

$$\text{Entonces: } \begin{cases} 12 \rightarrow 2.381 \\ 14 \rightarrow X = 2.3323 \\ 15 \rightarrow 2.308 \end{cases}$$

$$\text{por lo que: } F_{(0.95, 14, 17)} = 2.3323$$

Para el valor crítico de la izquierda se hace lo mismo, interpolando ahora entre 15 y 20 para obtener el valor F con 17 grados en el numerador.

$$F_{(0.95, 15, 14)} = 2.463 \text{ y } F_{(0.95, 20, 14)} = 2.388$$

$$\begin{cases} 15 \rightarrow 2.463 \\ 17 \rightarrow X = 2.433 \rightarrow F_{(0.95, 17, 14)} = 2.433 \\ 20 \rightarrow 2.388 \end{cases}$$

$$F_{(0.05, 14, 17)} = \frac{1}{F_{(1-\alpha/2, 17, 14)}} = \frac{1}{2.433} = 0.411$$

Sustituyendo en la fórmula del intervalo para razón de varianzas tenemos:

$$\frac{3249}{(2704)(2.5878)} < \frac{\sigma_{Ing}^2}{\sigma_{Fil}^2} < \frac{3249}{(2704)(0.411)} \Rightarrow (0.4643, 2.9234)$$

Interpretación.- La razón de varianzas en las calificaciones de matemáticas se encuentra entre 0.4643 y 2.9234, por lo que podemos concluir que de cada cien veces que se realice el muestreo entre estos dos grupos poblacionales, en 90 de ellos la variación en las calificaciones de matemáticas será semejante, ya que la unidad está incluida en el intervalo.

- d) Se pide estimar la diferencia en las calificaciones medias de los estudiantes de ingeniería y filosofía, en matemáticas, por lo que utilizaremos la fórmula del intervalo para el parámetro $\mu_{Ing} - \mu_{Fil}$ con una confianza de 90%, cuando las varianzas poblacionales son desconocidas pero se consideran no diferentes, como se demuestra en el inciso c. Usaremos la distribución t de student con $n_1 + n_2 - 2$.

$$1 - \alpha = 0.90 \Rightarrow \alpha = 1 - 0.90 = 0.10 \Rightarrow \alpha/2 = 0.05 \text{ y } 1 - \alpha/2 = 1 - 0.05 = 0.95$$

Así, el valor crítico para la distribución t se busca en tablas con percentil de 0.95 y 31 grados libres:

$$t_{(0.95, 31)} = 1.6955$$

Con base en las fórmulas para el cálculo del intervalo para la diferencia de medias poblacionales independientes, cuando no se conocen las varianzas poblacionales pero se consideran semejantes (resultado del inciso anterior), se calcula la desviación estándar ponderada para las 2 muestras así:

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Calculando la desviación ponderada:

$$S_p = \sqrt{\frac{(15-1)(3249) + (18-1)(2704)}{15+18-2}} = 54.315$$

Calculando el estimador:

$$\bar{x}_{Ing} - \bar{x}_{Fil} = 548 - 517 = 31$$

Calculando el Error máximo de estimación:

$$E = (1.6955)(54.315) \sqrt{\frac{1}{15} + \frac{1}{18}} = 32.1953$$

Sustituyendo en el intervalo tenemos:

$$31 - 32.1953 < \mu_{Ing} - \mu_{Fil} < 31 + 32.1953$$

$$(-1.1953, 63.1953)$$

Interpretación.- El intervalo para la diferencia entre las medias de calificación en matemáticas, va de un valor negativo a uno positivo incluyendo al cero, por lo que podríamos decir que, con 90% de confianza, no hay diferencia en el comportamiento promedio de ambos grupos de estudiantes, en esta materia. Sin embargo, es notorio el corrimiento del intervalo hacia la derecha, lo que indicaría que hay tendencia de los estudiantes de ingeniería a tener puntajes más altos en matemáticas. Idea que tendría que verificarse con un contraste de hipótesis unilateral.

EJEMPLO 4.9. En un estudio comparativo de tubos capilares para uso experimental, se midió el diámetro de los tubos capilares. Las muestras se obtuvieron al azar de dos diferentes líneas de proceso. Se sabe, por estudios anteriores, que la variación verdadera, en el diámetro de estos capilares, es de 9.77 micras para la línea 1 y de 13.26 micras para la línea 2. Los datos obtenidos, en micras, son:

Línea 1					Línea 2				
309	311	327	317	326	349	344	326	336	347
332	321	317	316	334	311	335	337	336	307
323	309	324	325	315	329	320	325	346	325
315	311	326	320	312	328	325	335	331	330
315	306	334	325	316	342	325	334	339	339

Estime, con una confianza de 99% a la diferencia media entre los diámetros de los capilares producidos en cada línea.

Solución:

Leyendo cuidadosamente el texto del problema, podremos darnos cuenta que nos están dando las desviaciones estándar poblacionales para el diámetro de los capilares, en ambas líneas, por lo que para estimar la diferencia de las medias nos basaremos en la distribución normal, con variable Z .

Calculamos el percentil para el valor Z

$$1 - \alpha = 0.99 \Rightarrow \alpha = 1 - 0.99 = 0.01 \Rightarrow \alpha/2 = 0.005 \quad y \quad 1 - \alpha/2 = 0.995$$

Leyendo en la columna E de las tablas de la normal, tenemos:

$$Z_{(1-\alpha/2)} = Z_{(0.995)} = 2.5758$$

De la ecuación apropiada para calcular el intervalo solicitado, E se calcula como:

$$E = Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_{L1}^2 + \sigma_{L2}^2}{n_{L1} + n_{L2}}} = 2.5758 \sqrt{\frac{(9.77)^2}{25} + \frac{(13.26)^2}{25}} = 8.485$$

En seguida se usa la calculadora en formato estadístico y se introducen los datos para obtener las medias, de ambas regiones.

$$\bar{x}_{L1} = 319.44 \quad y \quad \bar{x}_{L2} = 322.04$$

Calculando el estimador:

$$\bar{x}_{L1} - \bar{x}_{L2} = 319.44 - 322.04 = -12.6$$

Sustituyendo en la fórmula del intervalo.

$$-12.6 - 8.485 < \mu_{L1} - \mu_{L2} < -12.6 + 8.485$$

$$(-21.085, -4.115) \quad \text{con una confianza de 99\%}$$

Interpretación.- El resultado nos dice que, de cada 100 veces que se realice el proceso, en 99 de ellos, la diferencia entre las medias estará entre -21.085 y -4.115.

Como el intervalo es negativo por ambos lados, se podría deducir que la media de la Línea 2 es más grande, lo que significaría que el diámetro de los capilares en la Línea 2 pudiera ser mayor. Sin embargo, no se tiene evidencia suficiente para afirmarlo. En todo caso, si se quiere demostrar que existe mayor

diámetro para los capilares en la Línea 2, será necesario hacer un contraste unilateral para el parámetro diferencia de medias.

EJEMPLO 4.10. Un investigador desea saber el contenido medio de sacarosa en una concentración dada de jugo de remolacha, obtenido del fruto de diferentes cosechas. Para hacer la medición, él cuenta con 2 métodos y desea saber si ambos, dan la misma concentración media. Por esta razón, analiza la mitad del jugo con el método **A** y la otra mitad con el método **B**. Los datos obtenidos se registran en la tabla siguiente:

Cosecha	1	2	3	4	5	6	7	8	9	10
Método A	11	5.0	9.8	5.7	6.5	8.2	5.9	6.0	7.5	5.4
Método B	11.2	5.0	9.7	5.3	6.7	8.5	5.6	5.8	7.1	5.5
$d_i = A - B$	-0.2	0	0.1	0.4	-0.2	-0.3	0.3	0.2	0.4	-0.1

Estime, con una confianza de 99%, la diferencia media entre las mediciones de ambos métodos.

Solución:

El proceso consiste en comparar una muestra de 10 cosechas, en cuanto a su contenido de sacarosa, utilizando dos métodos, se puede decir que se está trabajando una muestra medida dos veces y lo que se quiere ver es la diferencia debida al método utilizado. Entonces se tiene un apareamiento de datos, por lo que se deben calcular las diferencias por pareja (método A menos método B) y obtener la media muestral de esas diferencias, por lo que este valor \bar{x}_D servirá de estimador para el parámetro media de las diferencias μ_D .

En formato estadístico, utilizamos la calculadora para obtener la media muestral de las diferencias observadas y la desviación estándar.

$$\bar{x}_D = 0.06 \quad \text{y} \quad s_d = 0.25906$$

Como todos los datos son muestrales y no se conoce la varianza poblacional, se trabajará con una distribución t de student, con $n-1$ grados de libertad.

$$1 - \alpha = 0.99 \Rightarrow \alpha = 1 - 0.99 = 0.01$$

$$\alpha/2 = 0.01 \Rightarrow 1 - \alpha/2 = 0.995$$

Buscando en las tablas estadísticas:

$$t_{(1-\alpha/2, n-1)} = t_{(0.995, 9)} = 3.2498$$

El error máximo de estimación para el intervalo de media de las diferencias es:

$$E = t_{(1-\alpha/2, n-1)} \frac{s_d}{\sqrt{n}} = 3.2498 \left(\frac{0.25906}{\sqrt{10}} \right) = 0.266228$$

Sustituyendo en la fórmula para el intervalo de estimación tenemos:

$$\bar{x}_D - E < \mu_D < \bar{x}_D + E$$

$$0.06 - 0.266228 < \mu_D < 0.06 + 0.266228 \quad \text{con una confianza de 99\%}$$

Entonces el intervalo queda:

$$(-0.206228, 0.326228)$$

Interpretación.- De cada 100 veces que se realice el estudio, en las mismas condiciones, en 99 de ellas los métodos darán resultados semejantes. Note que el intervalo va de negativo a positivo, entonces el cero está incluido, lo que significa que no hay diferencia estadística real entre los métodos.

4.5 Contraste de hipótesis para la comparación de dos Grupos de datos

Al realizar un contraste de hipótesis sobre comparación de grupos, se procede de la misma manera que al trabajar con una sola población. Esto es, deberán seguirse los 8 pasos establecidos anteriormente, la única diferencia radica en que el parámetro será una diferencia o una razón o cociente y el valor supuesto debe indicar la diferencia supuesta entre los grupos poblacionales comparados o la razón entre los mismos.

Nota.- Cuando no se da la diferencia supuesta, se establece como cero, porque para 2 cantidades semejantes la diferencia es cero y cuando no se da la razón supuesta se establece como 1 porque para dos cantidades iguales la razón o cociente debe ser 1.

4.5.1. Aplicación del Proceso de Contraste de Hipótesis en Comparación de grupos.

EJEMPLO 4.11. La industria China fabrica telas sintéticas a partir de la resina de politereftalato de etileno, obtenida del reciclado de envases plásticos de refrescos, alimentos etc. Para comparar la cantidad de fibra sintética obtenida en dos procesos de fabricación se tomaron muestras en peso (kg), de fibra obtenida al utilizar misma cantidad de envases reciclados y los registros para cada proceso son:

PROCESO A							PROCESO B				
26.5	25.2	25.3	26.5	30.1	27.9	26.9	35.6	28.3	25.0	27.1	26.9
25.7	33.0	25.3	29.3	25.7	30.7	39.7	25.0	25.1	26.5	25.7	25.6
27.3	25.5	25.4	29.0	36.3	25.3	25.0	26.3	25.1	27.6	26.0	25.0
25.2	26.6	25.9	31.7	31.9	32.4	25.2	25.9	27.3	30.6	25.1	25.1
28.5	27.0	25.0	26.3	25.0	33.0	27.2	25.1	28.3	30.2	28.5	25.2
25.1	25.8	39.0	34.1	25.1	25.2	25.0	25.6	25.2	29.0	29.8	28.2
25.2	26.5	25.4	25.7	25.1	25.3	25.4	25.8	26.7	26.4	28.1	25.2
27.6	25.1	28.2	29.5	26.8	26.7	25.3	26.6	25.0	27.3	25.4	25.3
27.6	27.7	25.1	31.4	33.7	25.1	25.5	27.5	26.3	28.5	27.0	25.0
28.1	25.1	26.6	26.4	25.0	25.6	36.6	32.3	27.0	28.7	31.5	25.1

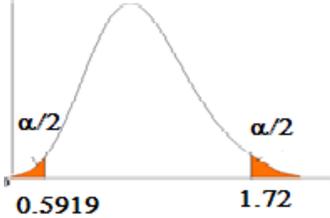
- a) Pruebe, con una significación del 5% si ambos procesos presentan la misma variación en cuanto a la cantidad de kg de fibra sintética obtenida.
- b) ¿Podría considerarse, al 5% de significación, que el proceso B produce una cantidad promedio menor de fibra que el proceso A?

Solución:

Primero tendremos que calcular la media y la varianza maestras, para los kilogramos de fibra obtenidos en cada proceso, pues estos estimadores nos servirán de base para los contrastes pedidos. Por lo tanto, introducimos los datos del proceso A, en formato estadístico para una sola variable y obtenemos las medidas descriptivas, después hacemos lo mismo con los datos del proceso B.

- a) Nos piden probar si ambos procesos son igualmente variables en cuanto a la cantidad de fibra sintética producida por lo que utilizaremos la distribución **F** de Fisher para contrastar usando la razón de varianzas.

1) Datos		2) Planteamiento de las hipótesis	3) Nivel de Significación
$\bar{x}_A = 27.7$ $s_A = 3.5256$ $s_A^2 = 12.43$ $n_A = 70$	$\bar{x}_B = 27.012$ $s_B = 2.21$ $s_B^2 = 4.888$ $n_B = 50$	Bilateral $H_0: \sigma_A^2 / \sigma_B^2 = 1$ $H_a: \sigma_A^2 / \sigma_B^2 \neq 1$	$\alpha = 0.05$ α se divide entre 2 y se generan 2 regiones de rechazo para la distribución. $\alpha/2 = 0.025 \rightarrow 1 - \alpha/2 = 0.975$

<p>4) Distribución utilizada y elección del estadístico de contraste.</p> $s_A^2 / s_B^2 \sim F$ $F_{(1-\alpha/2, 69, 49)} \text{ y } F_{(\alpha/2, 69, 49)}$ $F_{(\alpha/2, 69, 49)} = \frac{1}{F_{(1-\alpha/2, 49, 69)}}$ <p>Para obtener el valor de tablas hay que interpolar 3 veces dado que, los grados libres del numerador, 69 y los del denominador 49 no están registrados directamente en las tablas. Ver interpolaciones después de este cuadro</p>	<p>5) Regla de Decisión o Región crítica:</p> $F_{(0.025, 69, 49)} = \frac{1}{1.6894} = 0.5919$ $F_{(0.975, 69, 49)} = 1.72$ 	<p>6) Cálculo del Estadístico de Contraste</p> $F_{Calc} = \frac{12.42}{4.888} = 2.5409$
<p>7) Decisión:</p> <p>Se rechaza H_0, porque el estadístico 2.5409 está en la región derecha de rechazo.</p> <p>(2.5409 > 1.72)</p>	<p>8) Conclusión:</p> <p>Se establece, con una significación del 5% que las varianzas poblacionales de ambos procesos son significativamente diferentes.</p>	

Nota: Las interpolaciones para los límites de las regiones críticas de la distribución F , se encuentran abajo.

Para el lado derecho, primera interpolación.

Tomamos como referencia 45 grados libres en el denominador y 60 gl en el numerador, cuyo valor es 1.757. Después, con 45 en el denominador y 120 en el numerador tenemos el valor 1.677 e interpolamos para 69:

$$\left\{ \begin{array}{l} 60 \rightarrow 1.757 \\ 60 \rightarrow X = 1.745 \Rightarrow F_{(0.975, 69, 45)} = 1.745 \\ 60 \rightarrow 1.757 \end{array} \right.$$

Para el lado derecho, segunda interpolación.

Tomamos como referencia 60 grados libres en el denominador y 60 gl en el numerador, cuyo valor es 1.667. Después, con 60 en el denominador y 120 en el numerador tenemos 1.581 e interpolamos para 69:

$$\left\{ \begin{array}{l} 60 \rightarrow 1.667 \\ 69 \rightarrow X = 1.6541 \Rightarrow F_{(0.9785, 69, 60)} = 1.6541 \\ 120 \rightarrow 1.581 \end{array} \right.$$

Para el lado derecho, tercera interpolación.

Como ya tenemos los valores de 69 numerador con 45 denominador, 1.745 y 69 numerador 60 denominador 1.6541, interpolamos entre ellos para obtener 69 numerador con 49 denominador:

$$\begin{cases} 45 \rightarrow 1.745 \\ 49 \rightarrow X = 1.7208 \Rightarrow F_{(0.975, 69, 49)} = 1.7208 \\ 60 \rightarrow 1.6541 \end{cases}$$

Para el lado izquierdo, primera interpolación.

Tomamos como referencia 60 gl. en el denominador y 40 en el numerador, el valor es 1.744, Después, con 60 en el denominador y 60 en el numerador, tenemos 1.667 e interpolamos para 49:

$$\begin{cases} 40 \rightarrow 1.744 \\ 49 \rightarrow X = 1.70935 \Rightarrow F_{(0.975, 49, 60)} = 1.70935 \\ 60 \rightarrow 1.667 \end{cases}$$

Para lado izquierdo, segunda interpolación.

Tomando como referencia 120 gl. en el denominador y 40 en el numerador, el valor es 1.614. Después con 120 en el denominador y 60 en el numerador, tenemos 1.53 e interpolamos para 49:

$$\begin{cases} 40 \rightarrow 1.614 \\ 49 \rightarrow X = 1.5762 \Rightarrow F_{(0.975, 49, 120)} = 1.5762 \\ 60 \rightarrow 1.53 \end{cases}$$

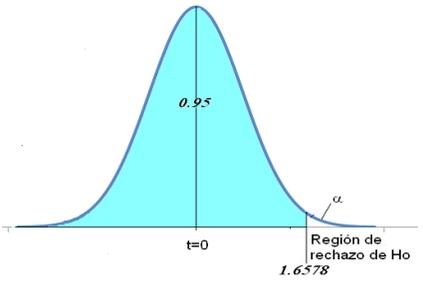
Para el lado izquierdo, tercera interpolación.

Como ya tenemos los valores de 49 numerador con 60 denominador, 1.70935 y 49 numerador 120 denominador, 1.5762, interpolamos entre ellos para obtener 49 numerador con 69 denominador.

$$\begin{cases} 40 \rightarrow 1.70935 \\ 49 \rightarrow X = 1.6894 \Rightarrow F_{(0.975, 49, 69)} = 1.6894 \\ 120 \rightarrow 1.5762 \end{cases}$$

$$\therefore F_{(0.025, 69, 49)} = \frac{1}{1.6894} = 0.5919$$

- b) Nos piden probar si la media poblacional del proceso B es inferior a la media poblacional del proceso A. Como las varianzas poblacionales son desconocidas pero diferentes, tendremos que usar una distribución t con grados libres calculados.

<p>1) Datos</p> $\bar{x}_A = 27.7$ $s_A = 3.5246$ $s_A^2 = 12.42$ $n_A = 70$	$\bar{x}_B = 27.012$ $s_B = 2.21$ $s_B^2 = 4.888$ $n_B = 50$	<p>2) Planteamiento de las hipótesis</p> <p>Unilateral superior, porque si B es menor que A la diferencia será positiva.</p> $H_0: \mu_A^2 - \mu_B^2 \leq 0$ $H_a: \mu_A^2 - \mu_B^2 > 0$	<p>3) Nivel de Significación</p> $\alpha = 0.05$ <p>α no se divide entre 2 pues sólo hay una región de rechazo, colocada a la derecha de la distribución.</p> $1 - \alpha = 0.95$
<p>4) Distribución utilizada y elección del estadístico de contraste.</p> $\bar{x}_A - \bar{x}_B \sim t_{(0.95, gl_{calc})} = t_{(0.95, 118)}$ $= 1.65784$ <p>* sale por interpolación entre 115 y 120</p> $\bar{x}_A - \bar{x}_B = 27.7 - 27.012$ $= 0.688$ $gl = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A + 1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B + 1}} - 2$ $gl = \frac{\left(\frac{12.43}{70} + \frac{4.888}{50}\right)^2}{\frac{\left(\frac{12.43}{70}\right)^2}{71} + \frac{\left(\frac{4.888}{50}\right)^2}{51}} - 2 = 118$	<p>5) Regla de Decisión o Región crítica:</p> 	<p>6) Cálculo del Estadístico de Contraste</p> $t_{calc} = \frac{(\bar{x}_A - \bar{x}_B) - \Delta_0}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$ $t_{calc} = \frac{0.688 - 0}{\sqrt{\frac{12.43}{70} + \frac{4.888}{50}}} = 1.311$ <p>Δ_0 es la diferencia supuesta entre las medias poblacionales. Como no nos dan un dato específico, comparamos contra cero.</p>	
<p>7) Decisión:</p> <p>No se rechaza H_0 pues $1.311 < 1.65784$ por lo que la t calculada no está en la región de rechazo.</p>	<p>8) Conclusión:</p> <p>Se concluye que, con una significación del 5%, la media poblacional del proceso B no es significativamente inferior a la media poblacional de A.</p>		

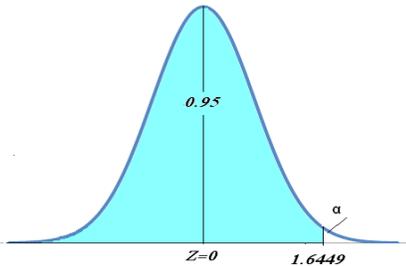
EJEMPLO 4.12. Se realizó un experimento para determinar si existe diferencia en la proporción de defectuosos encontrados en placas de acrílico manufacturadas en dos procesos en donde la temperatura de reacción cambia. De 50 placas obtenidas del proceso con temperatura 1, treinta y dos no presentan

defectos, mientras que de 50 placas fabricadas con el proceso a temperatura 2, veinte no presentan defectos. Con base en estos datos, ¿podríamos asegurar, con una significación del 5%, que la proporción de placas defectuosas del proceso con la temperatura 2 es 20% mayor que la proporción de placas defectuosas del proceso con la temperatura 1?

Solución:

Nos están pidiendo un contraste para la diferencia de proporciones poblacionales, $\pi_2 - \pi_1$, donde la diferencia supuesta es mayor que 20%. Es un análisis unilateral y será superior o inferior dependiendo del sentido en que tomemos la diferencia.

Se utilizará una distribución normal.

<p>1) Datos</p> $n_1 = 50$ $X_1 = 18$ $p_1 = \frac{18}{50} = 0.36$ $n_2 = 50$ $X_2 = 30$ $p_2 = \frac{30}{50} = 0.6$	<p>2) Planteamiento de las hipótesis</p> <p>Unilateral superior</p> $H_0: \pi_2 - \pi_1 \leq 0.20$ $H_a: \pi_2 - \pi_1 > 0.20$	<p>3) Nivel de Significación</p> $\alpha = 0.05$ <p>Como el proceso es unilateral superior, se tendrá una región de rechazo del lado derecho</p> $1 - \alpha = 1 - 0.05 = 0.95$
<p>4) Distribución utilizada y elección del estadístico de contraste.</p> $p_2 - p_1 \sim Z_{(1-\alpha)}$ $Z_{(0.955)} = 1.6449$ <p>Este valor se obtiene leyendo en tablas de la distribución Normal con percentil 0.95 en la columna E.</p>	<p>5) Regla de Decisión o Región crítica:</p> 	<p>6) Cálculo del Estadístico de Contraste</p> $(p_2 - p_1) = 0.6 - 0.36 = 0.24$ $Z_{calc} = \frac{(p_2 - p_1) - \Delta_0}{\sqrt{\bar{p} * \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ $Z_{calc} = \frac{0.24 - 0.20}{\sqrt{(0.48)0.52 \left(\frac{1}{50} + \frac{1}{50} \right)}} = 0.4$ <p>Δ_0 es la diferencia supuesta entre las proporciones poblacionales. Como nos dan el dato específico, comparamos con 0.20.</p>

<p>7) Decisión:</p> <p>No se rechaza H_0, pues $0.4 < 1.6449$ entonces, la Z calculada se encuentra en la región de no rechazo de H_0.</p>	<p>8) Conclusión:</p> <p>Con una significación del 5%, la diferencia de proporciones poblacionales de placas defectuosas del proceso 2 no supera en más del 20% a la proporción de defectuosos del proceso 1</p>	
--	---	--

EJEMPLO 4.13. En un estudio realizado para comparar las mediciones realizadas por dos máquinas, se utilizaron 12 tipos de alambre de acero. Cada tipo de alambre se dividió a la mitad y se midió la torsión de una mitad en la máquina A y la otra mitad en la máquina B. Los resultados registrados, como ángulo de ruptura, son los que aparecen en la tabla siguiente:

Tipo de Alambre	1	2	3	4	5	6	7	8	9	10	11	12
Máq. A	32	35	38	28	40	42	36	29	33	37	22	42
Máq. B	30	34	39	26	37	42	35	30	30	32	20	41

¿Con base en esta información, se podría considerar, al 1% de significación, que máquina A tiende a reportar mediciones de torsión más altas que la máquina B?

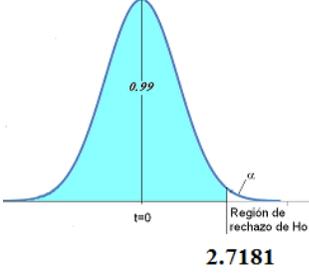
Solución:

El proceso consiste en dividir cada alambre en dos partes y medir la torsión soportada midiendo una mitad con A y la otra mitad con B. Tenemos entonces un caso de muestras pareadas esto es, es una muestra de elementos medida dos veces, por lo que debemos calcular las diferencias por pareja. Si $d_i = Máq_A - Máq_B$ la diferencia será positiva y el planteamiento será unilateral superior.

Al estar trabajando directamente con datos experimentales, no conocemos la varianza poblacional de las diferencias, por lo que usaremos la distribución t de student.

Tipo de Alambre	1	2	3	4	5	6	7	8	9	10	11	12
Máq. A	32	35	38	28	40	42	36	29	33	37	22	42
Máq. B	30	34	39	26	37	42	35	30	30	32	20	41
$d_i = Máq_A - Máq_B$	2	1	-1	2	3	0	1	-1	3	5	2	1

Metemos todas estas diferencias a la calculadora, en formato estadístico para una variable, respetando el signo y se calculan, la media de las diferencias, \bar{x}_d y la desviación estándar de estas diferencias, S_d .

1) Datos $n = 12$ $\bar{x}_d = 1.5$ $s_d = 1.732051$	2) Planteamiento de las hipótesis Unilateral superior $H_0: \mu_D \leq 0.20$ $H_a: \mu_D > 0.20$	3) Nivel de Significación $\alpha = 0.01$ $1 - \alpha = 1 - 0.01 = 0.99$
4) Distribución utilizada y elección del estadístico de contraste. $\bar{x} \sim t_{(1-\alpha, n-1)} = t_{(0.99, 11)} = 2.6810$ Este valor se obtiene leyendo en tablas de la distribución “t” con percentil 0.99 y 11 grados libres.	5) Regla de Decisión o Región crítica: 	6) Cálculo del Estadístico de Contraste $t_{calc} = \frac{\bar{x}_d - \Delta_0}{\frac{S_d}{\sqrt{n}}} = \frac{1.5 - 0}{\frac{1.732051}{\sqrt{12}}} = 2.9999$ <p>Δ_0, es la diferencia supuesta para la media de las diferencias poblacionales. Como no nos dan un dato específico, comparamos contra cero.</p>
7) Decisión: Se rechaza H_0 , $2.9999 > 2.7181$ por lo que el valor de la t calculada está en la región de rechazo.	8) Conclusión: Con una significación del 1%, no podemos rechazar que la máquina A registra mediciones más altas que la máquina B.	

4.6 Pruebas con datos categóricos: Pruebas de Independencia y Pruebas de Bondad de Ajuste

Tanto las pruebas de independencia como las de bondad de ajuste se basan en la distribución Ji cuadrada, que mide la variabilidad que presentan las frecuencias observadas dentro de un experimento aleatorio con respecto a las frecuencias esperadas dada una suposición estadística.

4.6.1 Pruebas de Independencia

En las pruebas de independencia, las frecuencias observadas (O_i) en un experimento, se encuentran clasificadas en una tabla de doble entrada con dos factores de clasificación, con 2 o más niveles de cada factor, uno en las filas o renglones y otro en las columnas. Se trata de demostrar que los factores de clasificación son estadísticamente independientes. Esto significa que la probabilidad de ocurrencia de un resultado o frecuencia esperada (E_i) se establece bajo el supuesto de que los eventos son independientes:

Factor A	Factor B			Totales de fila
	a	b	c	
1	O_{1a}	O_{1b}	O_{1c}	r_1
2	O_{2a}	O_{2b}	O_{2c}	r_2
3	O_{3a}	O_{3b}	O_{3c}	r_3
Totales de columna	c_1	c_2	c_3	n

$$E_{ij} = n \times p_{ij} = n \left(\frac{r_i}{n} \right) \left(\frac{c_j}{n} \right) = \frac{r_i c_j}{n}$$

En este tipo de pruebas, el planteamiento de hipótesis se establece de la siguiente manera:

H_0 establece la independencia, puesto que el cálculo de las frecuencias esperadas se basa en la probabilidad para eventos independientes, mientras que H_a niega la independencia, esto es, establece que hay una dependencia o asociación entre los factores de clasificación y ésta incide sobre las frecuencias observadas.

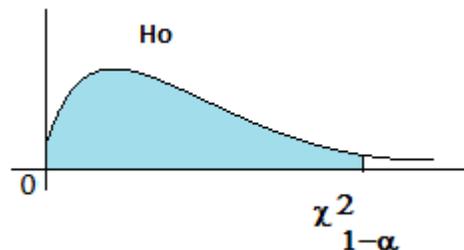
H_0 : Los factores de clasificación son independientes.

H_a : Los factores de clasificación son dependientes (o están asociados).

El contraste de hipótesis se basa en la distribución Ji cuadrada y se trabaja en forma unilateral superior, dado que sólo el exceso de variación implicará dependencia entre factores. Los grados de libertad para la distribución se calculan como el número de renglones menos 1 por número de columnas menos 1, Una de las restricciones de esta prueba es que las frecuencias esperadas deben ser todas mayores o iguales a 5 o si no es posible lograrlo, no más de 20% de las celdas deben tener frecuencias esperadas < 5 , entonces:

$$\chi_{[(1-\alpha/2), (r-1)(c-1)]}$$

Con el valor crítico de la distribución, obtenido de las tablas probabilísticas se construye la regla de decisión teórica o patrón de referencia para contrastar las hipótesis planteadas, como sigue:



Una vez establecida la gráfica de la distribución se realiza el cálculo del estadístico de contraste como sigue:

$$\mu^2_{calc} = \sum_{i=1}^n \frac{O_{ij}^2}{E_{ij}} - n$$

Para tomar una decisión, se compara el valor del estadístico de contraste con el valor crítico en la gráfica. Si el estadístico de contraste es mayor que el valor crítico, se rechaza H_0 . Esto es:

$$\chi^2_{calc} > \chi^2_{1-\alpha} \Rightarrow \text{Se rechaza } H_0$$

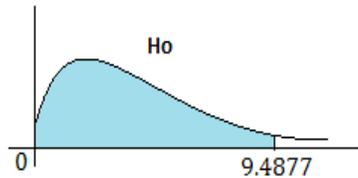
EJEMPLO 4.6.1.1. En una revista de mercadotecnia se publicó el siguiente estudio referente a la posible relación entre las condiciones de las instalaciones en tiendas de autoservicio y la política de precios aplicada a los productos. Se tomó una muestra de 484 tiendas y los resultados se clasificaron como se muestra en la tabla siguiente:

Condición	Política de precios			Total de Renglón
	Agresiva	Normal	No Agresiva	
Anticuada	29	32	22	83
Estándar	57	73	80	210
Moderna	63	84	44	191
Total de columna	149	189	146	484

De acuerdo con estos datos ¿se podría suponer que existe una relación entre la política de precios y las condiciones del establecimiento? Use $\alpha=0.05$.

Solución:

En el cuadro a continuación se realizarán los pasos que conduzcan a la solución del problema:

1) Planteamiento de Hipótesis:	2) Nivel de significación $\alpha = 0.05$	4) Regla de Decisión
H_0: No existe asociación entre la política de precios y la condición del establecimiento. H_a: Existe asociación entre la Política de precios y la condición del establecimiento.	3) Distribución utilizada Ji-cuadrada unilateral superior. Leyendo en tablas se tiene: $\chi^2_{[(1-\alpha),(r-1)(c-1)]} = \chi^2_{[0.95,(3-1)(3-1)]} =$ $\chi^2_{(0.95,(2 \times 2))} = \chi^2_{(0.95,4)} = 9.4877$	

Ahora se calculan las frecuencias esperadas correspondientes a cada celda de frecuencias observadas

$$E_{11} = \frac{83 \times 149}{484} = 25.55 \quad E_{12} = \frac{83 \times 189}{484} = 32.41 \quad E_{13} = \frac{83 \times 146}{484} = 25.04$$

$$E_{21} = \frac{210 \times 149}{484} = 64.65 \quad E_{22} = \frac{210 \times 189}{484} = 82 \quad E_{23} = \frac{210 \times 146}{484} = 63.35$$

Con las frecuencias observadas y las esperadas se calcula el estadístico de contraste como sigue:

Condición	Política de precios						Total de Renglón
	Agresiva		Normal		No Agresiva		
Anticuada	29	25.55	32	32.41	22	25.04	83
Estándar	57	64.65	73	82	80	63.35	210
Moderna	63	58.8	84	74.59	44	57.61	191
Total de columna	149		189		146		484

Otra forma es la siguiente:

O_{ij}	29	57	63	32	73	84	22	80	44
E_{ij}	25.25	64.65	58.8	32.41	82	74.59	25.04	63.35	57.61
O_{ij}^2 / E_{ij}	33.31	50.25	67.5	31.6	64.99	94.6	19.33	101.02	33.60

La suma de los cocientes de la última fila o renglón en el cuadro anterior es:

$$\sum_{i=1}^3 \sum_{j=1}^3 \frac{O_{ij}^2}{E_{ij}} = 495.81$$

Entonces, el estadístico de contraste queda:

$$\chi_{calc}^2 = \sum_{i=1}^3 \sum_{j=1}^3 \frac{O_{ij}^2}{e_{ij}} - n = 495.81 - 484 = 11.81$$

Que comparado con el valor crítico igual a 9.4877 es mayor, por lo que se rechaza H_0 y se concluye que, al 5% de significación, existe una asociación entre la política de precios y la condición de las instalaciones de las tiendas.

4.6.1.2 Pruebas de Independencia con tablas 2 por 2

En el caso específico de que los datos para una prueba de independencia estén clasificados en una tabla con dos filas y 2 columnas no es conveniente utilizar el método de cálculo de frecuencias esperadas para el estadístico de contraste porque la cantidad de información con que se cuenta no es suficiente. Las celdas de datos se marcan como se muestra en la tabla para identificar los elementos correspondientes a la sustitución del estadístico.

Factor fila	Factor columna		Total de filas
	A	B	
1	a	b	r_1
2	c	d	r_2
Total de columnas	c_1	c_2	n

En este caso es recomendable usar el cálculo directo con el estadístico

$$\chi_{Yates}^2$$

Este estadístico permite hacer la corrección del cálculo como sigue:

$$\chi_{Yates}^2 = \frac{n(|ad - bc| - \frac{n}{2})^2}{(r_1 \times r_2 \times c_1 \times c_2)}, \quad gl = 1$$

Como puede verse en la fórmula del estadístico de contraste, la diferencia de los productos entre las celdas se toma en valor absoluto

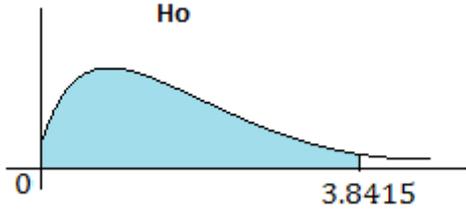
EJEMPLO 4.6.2.1. Se cree que las personas con carácter fuerte son más propensas a presentar problemas de hipertensión que las personas con carácter apacible. Con el fin de probar si esto es real se clasificó la información de 318 personas respecto al padecimiento en relación con el tipo de carácter presentado. Los resultados se registran en la tabla:

Carácter	Hipertensión	No Hipertensión	Total de filas
Fuerte	$60 = a$	$72 = b$	$132 = r_1$
Apacible	$46 = c$	$140 = d$	$186 = r_2$
Total de columnas	$106 = c_1$	$212 = c_2$	$318 = n$

Con base en la información anterior se podría asegurar, al 5% de significación que esta creencia está justificada?

Solución:

La clasificación de los datos se encuentra concentrada en una tabla con 2 renglones o filas y 2 columnas por lo que es conveniente realizar el contraste utilizando la corrección de Yates.

<p>Planteamiento de las hipótesis Ho: Los factores son independientes Ha: Los factores están asociados</p> <p>Nivel de significación $\alpha = 0.05$ El valor crítico se busca con $gl = (2-1)(2-1)=1$</p> <p>$\chi^2_{(1-\alpha, 1)}$ $\chi^2_{(0.95, 1)} = 3.8415$</p>	<p>Regla de decisión</p>  <p>Cálculo del estadístico de contraste</p> $\chi^2_{Yates} = \frac{318(60 \times 140 - 72 \times 46 - \frac{318}{2})^2}{132 \times 186 \times 106 \times 212}$ $\chi^2_{Yates} = 14.0028$	<p>Decisión: como χ^2_{Yates} es mayor que el valor crítico 3.8415, se rechaza Ho</p> <p>Conclusión: Al 5% de significación se infiere que los factores carácter e hipertensión están relacionados.</p>
--	--	--

4.6.2 Pruebas de Bondad de Ajuste

Estas pruebas se aplican para demostrar estadísticamente que un experimento aleatorio se comporta o no de acuerdo con un modelo de distribución de probabilidad conocido, sea de variable discreta, como por ejemplo, Binomial, Poisson, Multinomial y uniforme, o de variable continua como la distribución normal. Se mide el posible exceso de variación que indicaría falta de ajuste al modelo propuesto usando la distribución χ^2 .

El proceso consiste en establecer y probar una hipótesis respecto al modelo probabilístico que se cree explica el comportamiento del fenómeno aleatorio observado. Una vez planteadas las hipótesis estadísticas se calculan las frecuencias esperadas que se comparan con las frecuencias observadas mediante el estadístico de prueba, que se contrasta con la regla de decisión teórica –gráfica de la distribución χ^2 cuyo valor de tablas define el límite para la región de rechazo de la hipótesis nula- con objeto de tomar una decisión y concluir si los datos del experimento se ajustan o provienen de la distribución supuesta.

- **Planteamiento de hipótesis:** La hipótesis nula Ho, defiende el ajuste al modelo propuesto ya que el cálculo de las frecuencias esperadas se basa en dicho modelo, mientras que la hipótesis alterna Ha, niega el ajuste a dicho modelo.
- **Elección del nivel de significación α ,** apropiado para realizar el contraste.
- **Establecimiento de la regla de decisión teórica** para hacer la prueba, en donde el límite para la región de rechazo se busca en las tablas de la distribución en forma unilateral superior con un nivel de $1-\alpha$ y grados libres **k-m-1**, donde **k** es el número de categorías con frecuencia esperada

mayor o igual a 5; m es el número de parámetros desconocidos que se sustituyen por estimadores para calcular las probabilidades de ocurrencia de los resultados del experimento y las frecuencias esperadas. Entonces en tablas se busca el valor:

$$\chi^2_{(1-\alpha, k-m-1)}$$

- **Cálculo del estadístico de contraste:**

$$\chi^2_{calc} = \sum_{i=1}^r \sum_{j=1}^c \frac{o_{ij}^2}{e_{ij}} - n$$

- **Toma de Decisión**
- **Conclusión**

EJEMPLO 4.6.3.1. El ingeniero de producción de una empresa fabricante de botones cree que el número de botones de plástico, defectuosos, se comporta como una distribución binomial. Con objeto de probarlo obtiene una muestra al azar de 8 botones durante 68 días consecutivos y registra el número de defectuosos encontrados como sigue:

Defectos (X_i)	0	1	2	3	4	5	6	7	8
Frecuencia (f_i)	20	16	12	8	5	4	2	1	0

Pruebe si la suposición del ingeniero es correcta usando $\alpha=0.05$.

Solución:

Partiendo de la suposición de que el comportamiento es binomial, se hace el planteamiento de H_0 en este sentido. Después se calculan las frecuencias esperadas con base en el cálculo de probabilidades binomiales. La probabilidad de éxito es desconocida, por lo que se despejará de la definición de la media de un proceso binomial. Como no tenemos el valor del parámetro media, usaremos la media aritmética de botones defectuosos por lo que perderemos un grado de libertad, esto es, $m=1$. Entonces, para calcular la proporción de defectuosos despejamos p de la definición de media de una distribución binomial.

$$N^{\circ} \text{ de días} = \sum f_i = 68$$

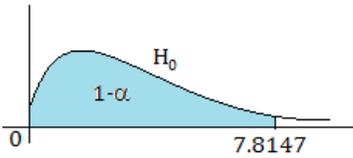
$$n = \text{número de botones} = \text{número de ensayos} = 8$$

$$\sum_{x=0}^8 f_i x_i = 143 \Rightarrow \bar{X} = \frac{143}{68} = 1.8088 \approx np$$

$$\Rightarrow p = \frac{\bar{X}}{n} = \frac{1.8088}{8} = 0.226$$

Probabilidad Binomial	$E_i = n \times p(x_i)$	O_i	O_i^2	$\frac{O_i^2}{E_i}$
$C_0^8(0.226)^0(0.774)^8 = 0.1288$	8.76	20	400	45.6621
$C_1^8(0.226)^1(0.774)^7 = 0.30087$	20.46	16	256	12.5122
$C_2^8(0.226)^2(0.774)^6 = 0.3075$	20.91	12	144	6.887
$C_3^8(0.226)^3(0.774)^5 = 0.1796$	12.21	8	64	5.2416
$C_4^8(0.226)^4(0.774)^4 = 0.0655$	**[4.45]5	144	25.46
$C_5^8(0.226)^5(0.774)^3 = 0.015309$	1.041	4	—	—
$C_6^8(0.226)^6(0.774)^2 = 2.235 \times 10^{-3}$	0.152	2	—	—
$C_7^8(0.226)^7(0.774)^1 = 1.864 \times 10^{-4}$	0.0127	1	—	—
$C_8^8(0.226)^8(0.774)^0 = 6.8056 \times 10^{-6}$	4.3x10⁻⁴] =5.65613	0] =12	—	—
	$\sum E_i = 68$	$\sum O_i = 68$	—	$\sum \frac{O_i^2}{E_i} = 95.76$

** Como las 5 últimas categorías tienen frecuencia esperada menor que 5, deben acumularse para formar una categoría con frecuencia esperada mayor o igual a 5, también se acumulan las 5 últimas categorías de las frecuencias observadas y se realiza el cociente entre ambos acumulados, entonces $k = 5$ y $m = 1$.

<u>Planteamiento de hipótesis</u>	<u>Nivel de significación</u>	<u>Estadístico de contraste</u>
<p>Ho: El número de botones defectuosos se comporta binomialmente</p> <p>Ha: El número de botones defectuosos no se comporta binomialmente</p>	<p>$\alpha = 0.05$</p> <p><u>Valor crítico</u> (unilateral superior)</p> <p>$\chi^2_{(1-\alpha, k-m-1)}$</p> <p>$\chi^2_{(0.95, 5-1-1)} = 7.8147$</p> <p><u>Regla de decisión</u></p> 	<p>$\chi^2_{calc} = \sum_{i=1}^5 \frac{O_i^2}{E_i} - n$</p> <p>De la tabla de cálculos:</p> <p>$\chi^2_{calc} = 95.76 - 68 = 27.76$</p> <p><u>Decisión</u></p> <p>Como $\chi^2_{calc} = \chi^2_{teo}$</p> <p>$27.76 > 7.8147$</p> <p>Se rechaza Ho</p> <p><u>Conclusión</u></p> <p>La distribución de defectos en los botones no se comporta binomialmente.</p>

EJEMPLO 4.6.3.2. Un ingeniero químico, jefe del laboratorio de control de calidad de una empresa fabricante de láminas de acrílico utilizadas en la construcción, supone que la cantidad de defectos por

metro presentados por las láminas se distribuye como una función de probabilidad de Poisson. Con objeto de probarlo toma una muestra aleatoria de 150 láminas de un metro de la producción del mes anterior y registra el número de defectos (opacidad, burbujas, rallado, etc.).

Defectos	0	1	2	3	4	5 o más	—
Frecuencia	72	50	20	6	1	1	n = 150

Con base en los datos registrados, ¿se puede considerar que la suposición es cierta? Use $\alpha=0.05$.

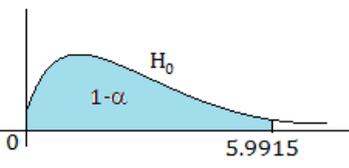
Solución:

Partiendo de la suposición de que el comportamiento es Poisson, se hace el planteamiento de H_0 en este sentido, después se calculan las frecuencias esperadas con base en el cálculo de probabilidades del modelo Poisson. La media λ de defectos por metro es desconocida por lo que se sustituirá por la media aritmética de los defectos registrados y por esta razón se pierde un grado de libertad, esto es, $m=1$.

$$\sum_{x=0}^5 f_i x_i = 117 \Rightarrow \bar{X} = \frac{117}{150} = 0.78 \Rightarrow \lambda \approx 0.78$$

Cálculos para el estadístico de contraste				
Probabilidad de Poisson	$E_i = n \times p(x_i)$	O_i	O_i^2	$\frac{O_i^2}{E_i}$
$P(x=0) = \frac{e^{-0.78} 0.78^0}{0!} = 0.4584$	68.76	72	5184	75.3927
$P(x=1) = \frac{e^{-0.78} 0.78^1}{1!} = 0.3576$	53.64	50	2500	46.6070
$P(x=2) = \frac{e^{-0.78} 0.78^2}{2!} = 0.1394$	20.91	20	400	19.1296
$P(x=3) = \frac{e^{-0.78} 0.78^3}{3!} = 0.0363$	[5.445	[6	64	9.5665
$P(x=4) = \frac{e^{-0.78} 0.78^4}{4!} = 7.1 \times 10^{-3}$	1.065	1	—	—
$P(x \geq 5) = 1 - 0.9988 = 1.2 \times 10^{-3}$	0.18]= 6.69	1]= 8	—	—
	$\sum E_i = 150$	$\sum O_i = 150$	—	$\sum \frac{O_i^2}{E_i} = 150.6958$

NOTA: Se acumularon las frecuencias esperadas entre corchetes para ajustar las frecuencias esperadas a 5 o más y al mismo tiempo deben acumularse las frecuencias observadas pertenecientes a las mismas categorías.

<p><u>Planteamiento de hipótesis</u> H_0: El número defectos se comporta como una Poisson H_a: El número de defectos no se comporta como una Poisson</p> <p><u>Nivel de significación</u> $\alpha=0.05$</p> <p><u>Valor crítico</u> (unilateral superior)</p> <p>$\chi^2_{(1-\alpha, k-m-1)}$</p> <p>$\chi^2_{(0.95, 4-1-1)} = 5.9915$</p>	<p><u>Regla de decisión</u></p>  <p><u>Estadístico de contraste</u></p> $\chi^2_{calc} = \sum_{i=1}^4 \frac{O_i^2}{E_i} - n$ <p>De la tabla de cálculos: $\chi^2_{calc} = 150.6958 - 150 = 0.6958$</p>	<p><u>Decisión</u></p> <p>Como $\chi^2_{calc} < \chi^2_{teo}$</p> <p>$0.6958 < 5.9915$</p> <p>No se rechaza H_0</p> <p><u>Conclusión</u></p> <p>Por lo tanto, la distribución de defectos en los láminas de acrílico se comporta como Poisson</p>
--	---	--

EJEMPLO 4.6.3.3. En un estudio llamado Contribución al estudio de la radiación solar y el clima en la ciudad de Bagdad, realizado en 1990, se determinó el índice de claridad del cielo de Bagdad durante 365 días, los resultados se agruparon como sigue

Intervalo de Clase	Frecuencia ($f_i = O_i$)
(0.15, 0.25]	8
(0.25, 0.35]	14
(0.35, 0.45]	28
(0.45, 0.55]	24
(0.55, 0.65]	39
(0.65, 0.75]	51
(0.75, 0.85]	106
(0.85, 0.95]	84
(0.95, 1.05]	11
	$\sum f_i = \sum O_i = 365$

Con base en esta información y al 5% de significación, ¿se podría afirmar que los índices de claridad se comportan de acuerdo con la distribución Normal?

Solución:

Puesto que se desea probar el ajuste a un modelo normal, es conveniente calcular las frecuencias esperadas sobre la suposición de normalidad, por lo que primero calcularemos las probabilidades o áreas bajo la curva normal, estandarizando sobre los límites reales superiores. Entonces usamos como sustitutos de los parámetros de la distribución Normal, la media aritmética y la desviación estándar, completando la tabla de datos con las marcas de clase como sigue:

Intervalo de clase	Frecuencia ($f_i = O_i$)	Marcas (m_i)	$f_i(m_i)$	$f_i(m_i^2)$
(0.15, 0.25]	8	0.2	1.6	0.32
(0.25, 0.35]	14	0.3	4.2	1.26
(0.35, 0.45]	28	0.4	11.2	4.48
(0.45, 0.55]	24	0.5	12	6.0
(0.55, 0.65]	39	0.6	23.4	14.04
(0.65, 0.75]	51	0.7	35.7	24.99
(0.75, 0.85]	106	0.8	84.8	67.84
(0.85, 0.95]	84	0.9	75.6	68.04
(0.95, 1.05]	11	1.0	11	11
	$n = \sum O_i = 365$		$\sum f_i m_i = 259.5$	$\sum f_i m_i^2 = 197.97$

$$\bar{X} = \frac{\sum f_i m_i}{n} = \frac{259.5}{365} = 0.711$$

$$s = \frac{\sum f_i(m_i^2) - n(\bar{X})^2}{n - 1} = \frac{197.97 - 365(0.711)^2}{364} = 0.1924$$

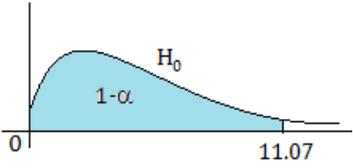
Con los valores críticos de la normal (z_i), se calculan las áreas acumuladas y después se restan el área inmediata anterior acumulada para obtener las probabilidades correspondientes a cada intervalo. Para las zetas negativas, las áreas se buscan en la columna B (porque son áreas acumuladas menores que 50%) y para las zetas positivas, las áreas se buscan en la columna E (porque son áreas acumuladas mayores que 50%).

Intervalo de clase	Frecuencia ($f_i = O_i$)	Z_i	Área acum	$P(X)$
(0.15, 0.25]	8	-2.39	0.0084	0.0084
(0.25, 0.35]	14	-1.88	0.0301	0.0217
(0.35, 0.45]	28	-1.36	0.0869	0.0568
(0.45, 0.55]	24	-0.84	0.2005	0.1136
(0.55, 0.65]	39	-0.32	0.3745	0.174
(0.65, 0.75]	51	0.20	0.5793	0.2048
(0.75, 0.85]	106	0.72	0.7642	0.1849
(0.85, 0.95]	84	1.24	0.8925	0.1283
(0.95, 1.05]	11		1.00	0.1075

$z = \frac{LRSup - \bar{X}}{s}$	$z = \frac{0.45 - 0.711}{0.1924} = -1.36$	$z = \frac{0.75 - 0.711}{0.1924} = 0.20$
$z = \frac{0.25 - 0.711}{0.1924} = -2.39$	$z = \frac{0.55 - 0.711}{0.1924} = -0.84$	$z = \frac{0.85 - 0.711}{0.1924} = 0.72$
$z = \frac{0.35 - 0.711}{0.1924} = -1.88$	$z = \frac{0.65 - 0.711}{0.1924} = -0.32$	$z = \frac{0.95 - 0.711}{0.1924} = 1.24$

Cálculo de frecuencias esperadas y cálculos adicionales		
O_i	$E_i = n \times p(x_i)$	$\frac{O_i^2}{E_i}$
[8	$365(0.0084) = 3.066$	
14] = 22	$365(0.0217) = 7.92 = 10.986$	44.06
28	$365(0.0568) = 20.73$	37.82
24	$365(0.1136) = 41.46$	13.89
39	$365(0.174) = 63.51$	23.95
51	$365(0.2048) = 74.75$	37.8
106	$365(0.1849) = 67.49$	166.48
84	$365(0.1283) = 46.83$	150.67
11	$365(0.1075) = 39.24$	3.083
$n = \sum O_i = 365$	$n = \sum E_i = 365$	$\sum \frac{O_i^2}{E_i} = 477.753$

NOTA: Las 2 primeras categorías de frecuencias esperadas se acumulan para lograr valores esperados mayores o iguales a 5 y lo mismo se hace con las observadas.

<p><u>Planteamiento de las hipótesis</u> Ho: El índice de claridad de cielo se distribuye normalmente. Ha: El índice de claridad de cielo no se distribuye normalmente</p> <p><u>Nivel de significación</u> $\alpha=0.05$</p> <p><u>Estadístico de contraste</u></p> $\chi^2_{calc} = \sum_{i=1}^8 \frac{O_i^2}{E_i} - n$	<p><u>Regla de decisión</u> (unilateral superior)</p> $\chi^2_{(1-\alpha, k-m-1)}$ $\chi^2_{(0.95, 8-2-1)} = 11.07$  <p><u>Cálculos</u></p> <p>De la tabla de cálculos: $\chi^2_{calc} = 477.753 - 365 = 112.753$</p> <p>Como $\chi^2_{calc} > \chi^2_{teo}$ $112.753 > 11.07$</p>	<p><u>Decisión:</u> Se rechaza Ho</p> <p><u>Conclusión:</u> Por lo tanto la distribución de índices de claridad no se comporta normalmente.</p>
--	---	--

Nota: El número de categorías con frecuencia esperada mayor o igual 5 es $k = 8$ y se pierden 2 grados libres al substituir la media y la desviación muestrales en lugar de los parámetros respectivos (**m**).

EJEMPLO 4.6.3.4. Una empresa fabricante de aparatos electrónicos distribuye I-pods en colores rojo, negro, blanco, azul y violeta, de acuerdo con las preferencias del consumidor, las políticas de venta la producción de estos aparatos está de acuerdo con la proporción: 8:6:4:3:1 respectivamente. Al revisar un lote de 200 aparatos que se enviarán a una tienda se encuentran 78 rojos, 50 negros, 40 blancos, 22 azules y 10 violeta. De acuerdo con estos resultados ¿se podría afirmar, al 5% de significación que en este lote se cumple la proporcionalidad por color?

Solución:

En este problema se tiene que probar si la proporción de colores se ajusta al modelo 8:6:4:3:1.

Esto es, se trata de probar un ajuste multinomial de los colores. En este caso, las frecuencias esperadas se calculan obteniendo la proporcionalidad por color con respecto al total por lo que no es necesario substituir ningún parámetro para fundamentar el cálculo, entonces **m** vale cero y los grados libres para la distribución Ji cuadrada teórica serán **k-1**.

Cálculo de frecuencias esperadas y cálculos adicionales			
O_i	$p(x_i)$	$E_i = n \times p(x_i)$	$\frac{O_i^2}{E_i}$
78	$\frac{8}{22}$	$E_1 = 200 \times \frac{8}{22} = 72.73$	$\frac{78^2}{72.73} = 83.655$
50	$\frac{6}{22}$	$E_2 = 200 \times \frac{6}{22} = 54.54$	$\frac{50^2}{54.54} = 45.833$
40	$\frac{4}{22}$	$E_3 = 200 \times \frac{4}{22} = 36.36$	$\frac{40^2}{36.36} = 44.00$
22	$\frac{3}{22}$	$E_4 = 200 \times \frac{3}{22} = 27.27$	$\frac{22^2}{27.27} = 17.75$
10	$\frac{1}{22}$	$E_5 = 200 \times \frac{1}{22} = 10.99$	$\frac{10^2}{10.99} = 9.099$
$\sum O_i = 200$	$\sum p(x_i) = 1$	$\sum E_i = 200$	$\sum \frac{O_i^2}{E_i} = 202.228$

Planteamiento de hipótesis

H₀: La distribución por color se ajusta al modelo 8:6:4:3:1

H_a: La distribución por color **no** se ajusta al modelo 8:6:4:3:1

Nivel de significación

$\alpha = 0.05$

Estadístico de contraste

$$\chi^2_{calc} = \sum_{i=1}^5 \frac{O_i^2}{E_i} - n$$

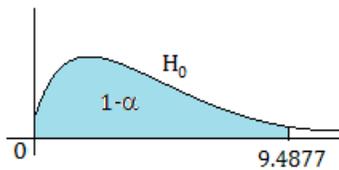
g.l. = 5 - 1 = 4

Regla de decisión

Valor crítico (unilateral superior)

$$\chi^2_{(1-\alpha, k-1)}$$

$$\chi^2_{(0.95, 5-1)} = 9.4877$$



Cálculos

$$\chi^2_{calc} = 202.228 - 200 = 2.228$$

Decisión:

Como $\chi^2_{calc} < \chi^2_{teo}$

$$2.228 < 9.4877$$

No se rechaza H₀

Conclusión:

La distribución por color se ajusta al modelo 8:6:4:3:1

EJEMPLO 4.6.3.5. En una fábrica de dulces se venden caramelos de 5 sabores: Limón, Piña, Uva, Fresa y Naranja. Los dulces son empacados en sobres de 100 gramos, de un solo sabor y después una máquina es programada para llenar cajas con 500 sobres, en donde debe haber el mismo número de sobres de cada sabor.

Los clientes se han quejado de que las cajas no vienen surtidas como debería. Por esta razón, el supervisor del área de empaqueo revisa 10 cajas, elegidas aleatoriamente del almacén y encuentra lo siguiente:

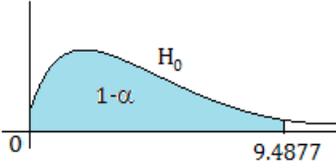
Sabor	Limón	Piña	Uva	Fresa	Naranja
Número de sobres en las 10 cajas	872	1002	978	1010	1138

Con base en estos resultados ¿se podría asegurar que la máquina empaquera está mal programada? Haga la prueba a una significación del 5%

Solución:

Este problema se refiere a comprobar si la máquina está empacando de manera uniforme, esto es, el mismo número de sobres de cada sabor. Entonces las frecuencias esperadas se calculan sumando el total de sobres y dividiendo entre los 5 sabores, por lo que todas las categorías de sabor deberían presentar 1000 sobres.

Sabor	O_i	E_i	$\frac{O_i^2}{E_i}$
Limón	872	1000	760.384
Piña	1002	1000	1004.004
Uva	978	1000	956.484
Fresa	1010	1000	1020.1
Naranja	1138	1000	1295.044
	$\sum O_i = 5000$	$\sum E_i = 5000$	$\sum \frac{O_i^2}{E_i} = 5036.016$

<p><u>Planteamiento de hipótesis</u> H₀: La distribución por sabor es semejante (uniforme) H_a: La distribución por sabor no es uniforme <u>Nivel de significación</u> $\alpha=0.05$ <u>Estadístico de contraste</u></p> $\chi^2_{calc} = \sum_{i=1}^5 \frac{O_i^2}{E_i} - n$ <p>g.l. = 5 - 1 = 4</p>	<p><u>Regla de decisión</u> Valor crítico (unilateral superior) $\chi^2_{(1-\alpha, k-1)}$ $\chi^2_{(0.95, 5-1)} = 9.4877$</p>  <p>El diagrama muestra una curva de la distribución chi-cuadrada. El eje horizontal está etiquetado con 0 y 9.4877. El área bajo la curva a la izquierda de 9.4877 está sombreada en azul y etiquetada como 1-α. La curva misma está etiquetada como H₀.</p>	<p><u>Cálculos</u> De la tabla de cálculos: $\chi^2_{calc} = 5036.016 - 5000 = 36.016$ Como $\chi^2_{calc} < \chi^2_{teo}$ $36.016 > 9.4877$</p> <p><u>Decisión:</u> Se rechaza H₀</p> <p><u>Conclusión:</u> La distribución por sabor no es uniforme.</p>
---	--	---

Diseño Experimental y Regresión

5.1. Relación entre Diseño de Experimentos y Análisis de Varianza

Cuando un investigador desea conocer el comportamiento de ciertas variables o factores involucrados en un proceso, es necesario que diseñe el experimento a realizar eligiendo, dentro de todas las variables del mismo, aquellas que ejercen la mayor influencia sobre los resultados esperados, la forma de seleccionar la(s) muestra(s), la forma de manejar y medir los atos y el método estadístico adecuado para hacer el análisis, de tal manera que pueda obtener resultados pertinentes, con el menor porcentaje de error y con una alta confiabilidad.

Lo anterior, implica que para él sea muy importante definir las fuentes de variación del proceso y seleccionar el modelo de análisis de varianza adecuado, para medir cuanto contribuyen esas fuentes a la variación total, de tal manera que se le facilite eliminar del estudio aquellas variables que no influyen de manera directa en los resultados esperados y realizar un experimento más sencillo, sin mermar la calidad de los resultados.

El análisis de varianza se utiliza para comparar más de 2 poblaciones o tratamientos dentro de un experimento. Este análisis nos permite dividir la variación total presente en una muestra, en sus diferentes componentes y medir la magnitud de las contribuciones. Al investigador le interesará conocer si los sujetos experimentales manifiestan una reacción diferente dependiendo del tipo de tratamiento que se aplique y la existencia de posibles interacciones entre los factores que posiblemente influyan sobre la variable respuesta.

Para aplicar un análisis de varianza, es obligatorio que los datos de la variable respuesta sean cuantitativos, aleatorios, que la distribución de las poblaciones comparadas sea normal y que exista homogeneidad en el valor de varianzas. En caso de no cumplirse estas condiciones habría que recurrir a la transformación de datos y en caso de no hacerse esta transformación no se podrá aplicar el análisis de varianza.

El objeto de este análisis es determinar si existen diferencias entre los resultados de los distintos tratamientos y en consecuencia, definir cuál tratamiento es el óptimo para mejorar u obtener un resultado deseado, por ejemplo: ¿Con qué tratamiento se evita la aparición de caries dentales?

- 1) Cepillado de dientes 3 veces al día.
- 2) Cepillado y uso de enjuague bucal 3 veces al día.
- 3) Cepillado, uso de enjuague bucal e hilo dental 3 veces al día.

Existen varias alternativas para diseñar un experimento, pero en este caso, el experimento se organizaría eligiendo al azar a las personas para formar parte de cada grupo experimental, que recibirá un tipo de tratamiento, elegido también al azar. Después de terminado el período de aplicación del tratamiento respectivo se revisaría a las personas para definir el estado de salud de su boca y se cuantificaría y registraría el número de caries encontradas. Con los datos obtenidos se aplicaría un modelo de análisis de varianza adecuado para establecer el mejor tratamiento para el problema planteado.

Al plantear un diseño de este tipo para el análisis estaremos hablando de un **análisis de varianza de un factor completamente al azar**. El factor de estudio sería la salud dental, los diferentes niveles del factor serían los diferentes tratamientos para la higiene bucal enumerados arriba y la variable de respuesta sería el número de caries encontradas.

5.1.1 Análisis de Varianza (ANDEVA) de un Factor Completamente al azar

Los datos deberán clasificarse ya sea en columnas o en renglones, en donde cada columna o renglón identificará un grupo o tratamiento a comparar. El número de observaciones por grupo puede ser el mismo o no. Esto es, no es obligatorio que las muestras sean del mismo tamaño para ser comparadas, aunque si es deseable porque facilita los cálculos.

Tabla 5.1 Clasificación por columnas para un análisis de varianza de un factor completamente al azar

Factor de estudio			
Tratamiento 1	Tratamiento 2	Tratamiento 3	Tratamiento 4
Y_{11}	Y_{12}	Y_{13}	Y_{14}
Y_{21}	Y_{22}	Y_{23}	Y_{24}
Y_{31}	Y_{32}	Y_{33}	Y_{34}
Y_{41}	Y_{42}	Y_{43}	Y_{44}

Y_{ij} representa cada valor de la variable de respuesta, ubicado en un renglón i , determinado y una columna j , específica. (v.g. Y_{32} representa el valor cuantitativo de la variable colocada en el tercer renglón y segunda columna).

NOTA: La variable respuesta la identificamos con la letra Y porque la X identifica a los tratamientos o variable independiente.

5.1.1.1 Modelo de un Factor completamente al azar

$$Y_{ij} = \mu + \tau_{\cdot j} + \varepsilon_{ij}$$

Desglosando cada uno de estos efectos se tiene:

$$Y_{ij} - \mu(\mu_{\cdot j} - \mu) + (Y_{ij} - \mu_{\cdot j})$$

Esta fórmula nos está indicando que cada valor de la variable respuesta dentro del experimento, se ve afectada por la media general del proceso, sufre los efectos del tratamiento aplicado y las fluctuaciones usuales debidas a la aleatoriedad del muestreo o efecto dentro de su propio grupo o tratamiento, conocidas como error residual.

Analizando el modelo anterior vemos que sólo hay 2 fuentes de variación en el diseño de un factor completamente al azar, la variación debida al tratamiento aplicado y la variación debida al error aleatorio.

5.1.1.2 Proceso de contraste de Hipótesis en el Análisis de Varianza de un Factor

1) Planteamiento de Hipótesis

Para plantear el par de hipótesis, Nula y Alternativa, se parte de la suposición de que los diversos tratamientos no conducen a resultados diferentes y entonces la hipótesis nula establecería que todos los tratamientos aplicados funcionan igual, en promedio. Mientras que la hipótesis alternativa establecería posibles diferencias, en promedio, parciales o totales.

$$H_0: \mu_{T1} = \mu_{T2} = \mu_{T3} = \mu_{T4}$$

H_A : *Al menos un par de medias es diferente*

2) Selección del Nivel de Significación

La selección del nivel de significación depende del riesgo que el investigador esté dispuesto a aceptar en sus conclusiones.

$$\alpha=0.01; \alpha=0.05 \text{ o } \alpha=0.1$$

Por lo general, la mayoría de los procesos se prueban al 5% de significación porque los paquetes de cómputo estadístico lo dan por default.

Aunque en el proceso de prueba se contrastan las medias poblacionales, el método de contraste consiste en desglosar la variación total en sus diferentes componentes, con objeto de demostrar que la variación entre tratamientos no es alta y por lo tanto éstos, no se consideran significativamente diferentes, lo que implicaría

cumplir con la hipótesis nula, en caso contrario, al menos un par de los tratamientos se consideraría diferente.

3) Distribución de Probabilidad utilizada

Se utiliza una distribución F de Fisher, para medir la naturaleza de las variaciones del experimento, mediante un contraste para la relación de varianzas. Se compara la variación entre grupos con la variación dentro de grupos en donde, los grados libres del numerador corresponderán a los de la varianza entre tratamientos y los del denominador, a los de la varianza dentro de tratamientos. El contraste se maneja en forma unilateral superior porque lo que nos interesa medir es el exceso de variación, esto es, la región de rechazo de la hipótesis nula se localiza en el lado derecho de la distribución. En caso de rechazar la hipótesis nula concluiríamos que las medias y por lo tanto los resultados de los distintos tratamientos no son semejantes o que hay efecto de tratamiento.

4) Estadístico de Contraste

El estadístico es una distribución F calculada mediante la relación de la varianza o cuadrado medio entre tratamientos y la varianza o cuadrado medio dentro de tratamientos.

$$F = \frac{CM_{Trat}}{CM_{Error}}$$

Para obtener los cuadrados medios o varianzas que se relacionan en el estadístico es necesario construir una tabla de Análisis de Varianza, (ANDEVA) que nos ayude a obtener, paso a paso, los elementos necesarios para nuestro análisis, como sigue:

TABLA DE ANDEVA DE UN FACTOR AL AZAR					
Fuente de Variación	Grados Libre (gl)	Suma de Cuadrados (SC)	Cuadrado Medio (CM)	Estadístico de Contraste	F de tablas
Entre Tratamientos	*K-1	SC _{Trat}	CM _{Trat}	$F = \frac{CM_{Trat}}{CM_{Error}}$	F(1- α , K-1, N-K)
Dentro de Tratamientos Error	**N-K	SC _{Error}	CM _{Error}		
Total	N-1	SC _{Total}	—		

*K es el número de tratamientos o grupos comparados.

** N es el total de observaciones en el experimento.

La nomenclatura utilizada en la tabla de ANDEVA de un factor completamente al azar, corresponde a una clasificación por columnas y entonces, el subíndice principal es la letra j . En el caso de que los tratamientos estén clasificados por fila o renglón, el subíndice principal será la letra i .

5.1.1.3 Definición matemática de las Sumas de Cuadrados (clasificación por columna)

- **Suma de Cuadrados Entre Tratamientos SC_{Trat}**

$$SC_{Trat} = \sum_{j=1}^n \frac{Y_{.j}^2}{n_j} - \frac{Y_{..}^2}{N}$$

Donde $Y_{.j}^2$ es el cuadrado de la suma de cada columna o tratamiento j .

n_j es el número de observaciones por columna o tratamiento.

$Y_{..}^2$ es el cuadrado de la suma del total de observaciones (Los puntos en el subíndice indican que se suman todas las observaciones tomando en cuenta su ubicación por renglón y por columna).

- **Suma de Cuadrados Total SC_{Total}**

$$SC_{Total} = \sum_{i=1}^r \sum_{j=1}^c Y_{ij}^2 - \frac{Y_{..}^2}{N}$$

Donde Y_{ij}^2 , es el cuadrado de cada observación en el experimento.

- **Suma de Cuadrados Dentro de Tratamientos o del Error SC_{Error}**

$$SC_{Error} = SC_{Total} - SC_{Trat}$$

EJEMPLO 5.1.1. Se analizaron 4 tipos de cereal, producidos en cierta región, para determinar el contenido de Tiamina y verificar si estos cereales presentan un contenido diferente de esta vitamina. El experimento consistió en tomar muestras al azar, de tamaño 6, de cada variedad de cereal y medir la cantidad de Tiamina, en miligramos por gramo de cereal, los resultados obtenidos aparecen registrados en la siguiente tabla:

Trigo	5.2	4.5	6.0	6.1	6.7	5.8	$Y_{1.} = 34.3$
Cebada	6.5	8.0	6.1	7.5	5.9	5.6	$Y_{2.} = 39.6$
Maíz	5.8	4.7	6.4	4.9	6.0	5.2	$Y_{3.} = 33.0$
Avena	8.3	6.1	7.8	7.0	5.5	7.2	$Y_{4.} = 41.9$
							$Y_{..} = 148.8$

- a) ¿Esta información sugiere que el contenido de Tiamina es diferente entre los cereales comparados? Use $\alpha = 0.05$.
- b) En caso de que el resultado del inciso anterior sea afirmativo, defina los pares de medias que son diferentes.

Solución:

a) Antes de iniciar el proceso de cálculo, debemos analizar cómo están clasificados los tratamientos, que en este caso, son los diferentes tipos de cereales. Vemos que los datos se acomodaron por renglón, esto es, todos los datos correspondientes al trigo se encuentran en el primer renglón, los de cebada en el segundo y así sucesivamente, por lo tanto, la nomenclatura en las fórmulas, deberá corresponder a la clasificación por renglón o fila, utilizando el subíndice i . Hay 4 tratamientos, $K=4$ y un total de 24 observaciones, $N=24$.

Planteamiento de Hipótesis

$$H_0: \mu_T = \mu_C = \mu_M = \mu_A.$$

H_a : Al menos un par de medias es diferente.

Nivel de Significación: $\alpha = 0.05$.

Distribución utilizada y Estadístico de contraste.- Se usa la distribución F con percentil de 95%, los grados de libertad correspondientes se obtienen de la tabla de ANDEVA. Para este ejemplo, se tiene:

TABLA DE ANDEVA DE UN FACTOR AL AZAR					
Fuente de Variación	Grados Libres (gl)	Suma de Cuadrados (SC)	Cuadrados Medios (CM)	Estadístico de Contraste	F de tablas
Entre Tratamientos (Cereales)	$K-1=4-1=3$	8.9833	$\frac{8.9833}{3} = 2.9944$	$F_{calc} = \frac{2.9444}{0.7568}$	$F_{(1-\alpha, K-1, N-K)} = F_{(0.95, 3, 20)} = 2.86$
*Dentro de Cereales o Error	$N-K=24-4=20$	15.1367	$\frac{15.1367}{20} = 0.7568$		
Total	$N-1=24-1=23$	24.12	—	$F_{calc} = 3.9565$	

*La fuente de variación dentro, también se conoce como Error.

Cálculo de las Sumas de Cuadrados

Primero obtenemos las sumas parciales por renglón para obtener las Y_1 . Después obtenemos la suma total de los renglones para obtener $Y_{..}$.

$$SC_{Trat} = \sum_{j=1}^n \frac{Y_{\cdot j}^2}{n_j} - \frac{Y_{\cdot\cdot}^2}{N}$$

Sustituyendo la ecuación tenemos:

$$SC_{Trat} = \frac{(34.3)^2 + (39.6)^2 + (33)^2 + (41.9)^2}{6} - \frac{(148.8)^2}{24} = 931.5433 - 922.56 = 8.9833$$

- ❖ Se utilizó como denominador común el 6 porque todos los tratamientos tienen seis observaciones. Si cada tratamiento tuviera diferente número de observaciones se tendrían que obtener los cocientes de cada tratamiento y después sumarlos.

Para la suma de cuadrados total, se suman los cuadrados de cada observación en el experimento y se resta el total de totales al cuadrado dividido por el tamaño de la muestra.

$$SC_{Total} = \sum_{i=1}^r \sum_{j=1}^c Y_{ij}^2 - \frac{Y_{\cdot\cdot}^2}{N} = (5.2)^2 + (4.5)^2 + (6.0)^2 + \dots + (5.5)^2 + (7.2)^2 - \frac{(148.8)^2}{24}$$

$$= 946.68 - 922.56 = 24.12$$

$$SC_{Error} = SC_{Total} - SC_{Trat} = 24.12 - 8.9833 = 15.1367$$

Decisión: Tomando como base los resultados mostrados en la tabla de ANDEVA, vemos que $F_c > F_{(0.95, 3, 20)}$ por lo que el estadístico de contraste se ubica en la región de rechazo de la hipótesis nula, entonces, rechazamos la suposición de que las medias poblacionales son semejantes.

Conclusión: Con una significación del 5% podemos afirmar que en al menos un par de cereales, el contenido medio de Tiamina es diferente.

5.1.1.4 Prueba de la Diferencia Significativa Honesta de Tukey

Cuando se realiza el contraste para la diferencia de medias, el análisis de varianza nos puede indicar que existen diferencias pero no nos dice cuales tratamientos o grupos poblacionales son diferentes. Por esta razón, es necesario aplicar la prueba de Tukey para identificar pares diferentes.

La prueba de Tukey, de la Diferencia Significativa Honesta, consiste en definir una diferencia probabilística, límite, que se compara con todas y cada una de las diferencias de medias, por pareja, de tal manera que todas aquellas diferencias entre medias muestrales que sean mayores que la DSH identificarán parejas de medias poblacionales diferentes.

$$DSH = q_{(\alpha, K, gl_{Error})} \sqrt{\frac{CM_{Error}}{n_j}}$$

Donde:

$q_{(\alpha, K, gl_{Error})}$, es el valor del porcentaje de rango estudentizado leído en la tabla T-9 del Cuaderno de Problemas de Probabilidad y Estadística.

CM_{Error} , es el cuadrado medio del error o varianza del error.

n_j , número de observaciones por tratamiento, ordenadas por columna. Si los tratamientos estuvieran ordenados por renglón o fila sería n_i

Cuando el modelo está desbalanceado, diferente número de observaciones por tratamiento, n_j se calcula como la media armónica de los diferentes tamaños de muestra.

b) Aplicaremos la prueba de Tukey para definir las parejas de medias poblacionales diferentes.

Cálculo de la DSH

$$q_{(\alpha, K, gl_{Error})} = q_{(0.05, 4, 20)} = 3.96 \quad CM_{Error} = 0.7568 \quad n_i = 6$$

Sustituyendo

$$DSH = q_{(\alpha, K, gl_{Error})} \sqrt{\frac{CM_{Error}}{n_i}} = 3.96 \sqrt{\frac{0.7568}{6}} = 1.4061$$

Así toda diferencia entre parejas de medias muestrales que sean mayores a este valor identificarán a parejas de medias poblacionales diferentes.

Ahora calculamos las medias de los tratamientos y obtenemos los valores absolutos de las diferencias por pareja.

	$\bar{Y}_T = 5.7166$	$\bar{Y}_C = 6.6$	$\bar{Y}_M = 5.5$	$\bar{Y}_A = 6.983$
$\bar{Y}_T = 5.7166$	—	$ \bar{Y}_T - \bar{Y}_C = 0.8834$	$ \bar{Y}_T - \bar{Y}_M = 0.2166$	$ \bar{Y}_T - \bar{Y}_A = 1.2664$
$\bar{Y}_C = 6.6$	—	—	$ \bar{Y}_C - \bar{Y}_M = 1.1$	$ \bar{Y}_C - \bar{Y}_A = 0.383$
$\bar{Y}_M = 5.5$	—	—	—	$ \bar{Y}_M - \bar{Y}_A = 1.483$

De estas 6 diferencias, sólo la $\bar{X}_M - \bar{X}_A$ es mayor que DSH, por lo tanto el contenido medio de tiamina del maíz es diferente del contenido medio de tiamina de la avena en la población y todas las demás parejas poblacionales no son significativamente diferentes.

$$(\mu_M \neq \mu_A)$$

5.1.2 Análisis de varianza de un factor con bloques al azar

En este modelo, los datos se encuentran clasificados en un cuadro de doble entrada porque el diseño incluye dos criterios de clasificación Sin embargo, al investigador sólo le interesa analizar efecto de uno de ellos y el otro criterio se maneja como variable de ruido cuyos efectos se miden para eliminar del error, esa fuente de variación, ya que todos los bloques se consideran de antemano diferentes.

Tabla 5.2 Clasificación de 1 factor con bloques al azar donde uno de ellos se bloquea para no interferir en el análisis del factor de interés.

	Tratamientos			
Bloques	1	2	3	4
A	Y_{A1}	Y_{A2}	Y_{A3}	Y_{A4}
B	Y_{B1}	Y_{B2}	Y_{B3}	Y_{B4}
C	Y_{C1}	Y_{C2}	Y_{C3}	Y_{C4}
D	Y_{D1}	Y_{D2}	Y_{D3}	Y_{D4}

Y_{ij} representa cada valor de la variable de respuesta, ubicado en un renglón **i**, determinado y una columna **j**, específica.

5.1.2.1 Modelo de un factor con bloques al azar

$$Y_{ij} = \mu + \tau_{\cdot j} + \beta_{i\cdot} + \varepsilon_{ij}$$

En este modelo existen 3 fuentes de variación, una debida al tratamiento aplicado o factor de interés, otra debida a los bloques, que no nos interesa analizar pero, como no se puede desaparecer es necesario medir sus efectos y la última debida al error residual del diseño.

Bloquear los efectos de un factor, significa contabilizar los efectos y separarlos para que no formen parte del error residual.

Al igual que en el diseño de un factor, deben plantearse la hipótesis nula y la alternativa, definir el nivel de significación, establecer la regla de decisión y presentar los cálculos en una tabla de ANDEVA.

Es muy importante definir cuál es el factor de interés antes de iniciar el proceso de prueba para evitar errores en la toma de decisiones.

Tabla 5.2 Clasificación de 1 factor con bloques al azar donde uno de ellos se bloquea para no interferir en el análisis del factor de interés.

TABLA DE ANDEVA DE UN FACTOR CON BLOQUES AL AZAR					
Fuente de variación (fv)	Grados libres (gl)	Suma de Cuadrados (SC)	Cuadrados Medios (CM)	Estadístico de Contraste	F de tablas
Tratamiento	*T-1	SC_{Trat}	CM_{Trat}	$F_{Trat} = \frac{CM_{Trat}}{CM_{Error}}$	$F_{[1-\alpha, T-1, (T-1) \times (B-1)]}$
Bloques	**B-1	SC_{Bloq}	---		
Error	(T-1)(B-1)	SC_{Error}	CM_{Error}		
Total	N-1	SC_{Total}			

*T = Número de Tratamientos **B = Número de Bloques ***N= Total de observaciones

5.1.2.2 Definición matemática de las Sumas de Cuadrados (Tratamientos en columna)

▪ Suma de Cuadrados de Tratamientos

$$SC_{Trat} = \sum_{j=1}^c \frac{Y_{.j}^2}{n_j} - \frac{Y_{..}^2}{N}$$

Donde $Y_{.j}^2$ es el cuadrado de la suma de cada columna o tratamiento.

n_j es el número de observaciones por columna o tratamiento.

$Y_{..}^2$, es el cuadrado de la suma del total de observaciones (Los puntos en el subíndice indican que se suman todas las observaciones tomando en cuenta su ubicación por renglón y por columna.

▪ **Suma de Cuadrados de Bloques**, cuando el criterio de clasificación bloqueado es el de los renglones o filas.

$$SC_{Bloq} = \sum_{i=1}^r \frac{Y_{i\cdot}^2}{n_i} - \frac{Y_{\cdot\cdot}^2}{N}$$

Donde $Y_{i\cdot}^2$ es el cuadrado de la suma de cada bloque, en este caso, filas.

▪ **Suma de Cuadrados Total (S.C. Total)**

$$SC_{Total} = \sum_{i=1}^r \sum_{j=1}^c Y_{ij}^2 - \frac{Y_{\cdot\cdot}^2}{N}$$

Donde Y_{ij}^2 , es el cuadrado de cada observación en el experimento.

Suma de Cuadrados del Error

$$SC_{Error} = SC_{Total} - SC_{Trat} - SC_{Bloq}$$

EJEMPLO 5.1.2. Se desea probar la resistencia de las telas a diferentes sustancias químicas que se utilizan para lograr el planchado permanente, por esta razón, se eligen 5 diferentes tipos de tela para probar 4 sustancias químicas y se mide la resistencia resultante como sigue:

Sustancia Química	Tipo de Tela					Total de Fila (Trat) $Y_{i\cdot}$
	1	2	3	4	5	
A	1.3	1.6	0.5	1.2	1.1	5.7
B	2.2	2.4	0.4	2.0	1.8	8.8
C	1.8	1.7	0.6	1.5	1.3	6.9
D	3.9	4.4	2.0	4.1	3.4	17.8
Total de Columna (Bloq) $Y_{\cdot j}$	9.2	10.1	3.5	8.8	7.6	39.2
						$Y_{\cdot\cdot}$

¿Se puede considerar, al 5% de significación, que las sustancias químicas afectan de igual manera la resistencia de las telas?

Solución:

Los resultados se encuentran clasificados en un cuadro de doble entrada: Sustancia Química en las filas y Tipo de Tela en las columnas. De acuerdo con el texto del problema se requiere probar el efecto de las sustancias químicas sobre la resistencia de las telas por lo que el tratamiento de interés son las sustancias

químicas, entonces se trata de un modelo de análisis de un factor con bloques al azar, en donde los bloques son los tipos de tela. Para comenzar el proceso de contraste es importante plantear la hipótesis nula y la alternativa para este ejemplo, como sigue:

Planteamiento de Hipótesis:

$$H_0: \mu_A = \mu_B = \mu_C = \mu_D.$$

H_a : Al menos un par de medias es diferente.

Nivel de significación: $\alpha=0.05$

Los resultados de los cálculos pertinentes se establecerán dentro de la tabla de ANDEVA siguiente de acuerdo con el modelo:

$$Y_{ij} = \mu + \tau_{i\cdot} + \beta_{\cdot j} + \varepsilon_{ij}$$

TABLA DE ANDEVA DE UN FACTOR CON BLOQUES AL AZAR					
Fuente de Variación	Grados Libres	Suma de Cuadrados	Cuadrado Medio	Estadístico de Contraste	F de tablas
Tratamiento Sustancias $\tau_{i\cdot}$	*T-1 4-1=3	18.044	$\frac{18.044}{3} = 6.01467$	$F_{Trat} = \frac{6.01467}{0.07925} = 75.895$	$F_{(0.95, 3, 12)} = 3.49$
Bloques Telas $\beta_{\cdot j}$	**B-1 5-1=4	6.693	---		
Error ε_{ij}	(T-1)(B-1) 3×4=12	0.951	$\frac{0.951}{12} = 0.07925$		
Total	N-1 20-1=19	25.688			

*T = Número de Tratamientos **B = Número de Bloques ***N= Total de observaciones

Cálculo de las Sumas de Cuadrados (los tratamientos están en fila y los bloques en columna, por lo que:

Suma de cuadrados de tratamientos (sustancias)

$$SC_{Trat} = \sum_{i=1}^r \frac{Y_i^2}{n_i} - \frac{Y_{..}^2}{N} = \frac{(5.7)^2 + (8.8)^2 + (6.9)^2 + (17.8)^2}{5} - \frac{(39.2)^2}{20} = 94.876 - 76.832 = 18.044$$

Suma de cuadrados de Bloques (telas)

$$SC_{Bloq} = \sum_{j=1}^c \frac{Y_{.j}^2}{n_j} - \frac{Y_{..}^2}{N} = \frac{(9.2)^2 + (10.1)^2 + (3.5)^2 + (8.8)^2 + (7.6)^2}{4} - \frac{(39.2)^2}{20} =$$

$$= 83.525 - 76.832 = 6.693$$

Suma de cuadrados Total

$$SC_{Total} = \sum_{i=1}^r \sum_{j=1}^c Y_{ij}^2 - \frac{Y_{..}^2}{N} = (1.3)^2 + (1.6)^2 + (0.5)^2 + (1.2)^2 + \dots + (4.1)^2 + (3.4)^2 - \frac{(39.2)^2}{20} =$$

$$= 102.52 - 76.832 = 25.688$$

Suma de cuadrados del Error Residual

$$SC_{Error} = 25.688 - 18.044 - 6.693 = 0.951$$

Decisión: Al comparar el estadístico de contraste con la regla de decisión se ve que:

$$f_{calc} > f_{(0.95, 3, 12)}$$

$$75.895 > 3.49$$

Por lo tanto, se rechaza H_0 y se concluye que hay efecto de las sustancias químicas en el comportamiento medio de la resistencia de las telas.

Para saber que parejas de sustancias hacen la diferencia es necesario hacer la prueba de Tukey, de la diferencia significativa honesta.

Cálculo de la DSH

$$q_{(\alpha, T, gl_{Error})} = q_{(0.05, 4, 12)} = 4.20 \quad CM_{Error} = 0.07925$$

Sustituyendo

$$DSH = q_{(\alpha, T, gl_{Error})} \sqrt{\frac{CM_{Error}}{n_i}} = 4.20 \sqrt{\frac{0.07925}{5}} = 0.52877$$

Realizando las diferencias por pareja, se tiene:

	$\bar{Y}_A = 1.14$	$\bar{Y}_B = 1.76$	$\bar{Y}_C = 1.38$	$\bar{Y}_D = 2.42$
$\bar{Y}_A = 1.14$	—	$ \bar{Y}_A - \bar{Y}_B = 0.62$	$ \bar{Y}_A - \bar{Y}_C = 0.24$	$ \bar{Y}_A - \bar{Y}_D = 2.42$
$\bar{Y}_B = 1.76$	—	—	$ \bar{Y}_B - \bar{Y}_C = 0.38$	$ \bar{Y}_B - \bar{Y}_D = 1.8$
$\bar{Y}_C = 1.38$	—	—	—	$ \bar{Y}_C - \bar{Y}_D = 2.18$

Comparando estas diferencias con la DSH, se concluye que las parejas de sustancias diferentes son:

$$\mu_A \neq \mu_B, \quad \mu_A \neq \mu_D, \quad \mu_B \neq \mu_D, \quad \mu_C \neq \mu_D$$

5.1.3 Análisis de varianza factorial de dos factores, completamente al azar, con repetición

En este diseño, los datos se encuentran clasificados en un cuadro de doble entrada con un factor en las filas o renglones y otro en las columnas. En cada celda formada por la intersección de los dos factores hay más de una observación. Todas las celdas o interacciones deben tener el mismo número de observaciones, es decir debe ser un diseño balanceado. En este diseño, el interés principal consiste en probar si existe interacción entre los factores de clasificación que cause efecto sobre la variable respuesta. Aunque también se puede analizar el comportamiento debido a cada factor, por lo que se plantean 3 hipótesis: de filas, de columnas y de interacción.

Tabla 5.3 Clasificación de 2 factores con repetición, el tercer subíndice en cada observación representa el número de repetición.

Factor 1	Factor 2			
	1	2	3	4
A	Y_{A11} Y_{A12}	Y_{A21} Y_{A22}	Y_{A31} Y_{A32}	Y_{A41} Y_{A42}
B	Y_{B11} Y_{B12}	Y_{B21} Y_{B22}	Y_{B31} Y_{B32}	Y_{B41} Y_{B42}
C	Y_{C11} Y_{C12}	Y_{C21} Y_{C22}	Y_{C31} Y_{C32}	Y_{C41} Y_{C42}
D	Y_{D11} Y_{D12}	Y_{D21} Y_{D22}	Y_{D31} Y_{D32}	Y_{D41} Y_{D42}

5.1.3.1 Modelo de 2 factores con repetición

$$Y_{ijk} = \mu + \alpha_{i..} + \beta_{.j.} + (\alpha\beta_{ij.}) + \varepsilon_{ijk}$$

En este modelo hay 4 fuentes de variación: una debida al factor de filas, otra al factor de columnas, otra debida a la posible interacción y otra a la del error residual.

Los pasos para resolver problemas de este tipo serán semejantes a los modelos anteriores pero la tabla de ANDEVA se modifica de acuerdo con las fuentes de variación del modelo.

5.1.3.2 Cálculo de las sumas de cuadrados

Suma de cuadrados de fila

$$SC_F = \sum_{i=1}^r \frac{Y_{i..}^2}{n_i} - \frac{Y_{...}^2}{N}$$

Suma de cuadrados de columna

$$SC_C = \sum_{j=1}^c \frac{Y_{.j.}^2}{n_j} - \frac{Y_{...}^2}{N}$$

Suma de cuadrados subtotal

$$SC_{Subt} = \sum_{k=1}^K \frac{Y_{...k}^2}{n_k} - \frac{Y_{...}^2}{N}$$

Suma de cuadrados de interacción

$$SC_{Inter} = SC_{Subt} - SC_F - SC_C$$

Suma de cuadrados total

$$SC_{Tot} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^K Y_{ijk}^2 - \frac{Y_{...}^2}{N}$$

Suma de cuadrados del error residual

$$SC_{Error} = SC_{Tot} - SC_{Subt}$$

TABLA DE ANDEVA DE 2 FACTORES CON REPETICIÓN					
Fuente de Variación	Grados Libres	Suma de Cuadrados	Cuadrado Medio	Estadístico de Contraste	F de tablas
Filas	$F-1$	SC_{Fila}	CM_{Fila}	$*F_{Fila} = \frac{CM_{Fila}}{CM_{Error}}$	$F_{(1-\alpha, F-1, gI_{Error})}$
Columnas	$C-1$	SC_{Col}	CM_{Col}	$*F_{Col} = \frac{CM_{Col}}{CM_{Error}}$	$F_{(1-\alpha, C-1, gI_{Error})}$
Interacción	$(F-1)(C-1)$	SC_{Inter}	CM_{Inter}	$*F_{Inter} = \frac{CM_{Inter}}{CM_{Error}}$	$F_{[1-\alpha, (F-1)(C-1), gI_{Error}]}$
Subtotal	$(F \times C)-1$	SC_{Subt}	—	—	—
Error	$N-(F \times C)$	SC_{Error}	CM_{Error}	—	—
Total	$N-1$	SC_{Total}	—	—	—

*La definición de los estadísticos de contraste y de la regla de decisión de cada prueba anotados aquí son los adecuados para un modelo con factores fijos.

Nota Importante.- Un diseño de 2 factores con repetición se considera **a)** de factores fijos cuando desde antes de realizar el experimento se ha decidido probar determinados niveles o tratamientos para cada factor, ya sea porque son de interés particular del investigador o son los únicos disponibles; **b)** de factores aleatorios cuando de todos los niveles con que se cuenta en cada factor, se decide aleatoriamente cuales niveles probar **c)** será de factores mixtos cuando uno de los factores sea fijo y el otro sea aleatorio.

En los 2 últimos casos anteriores los estadísticos de contraste serán las siguientes:

Factores Aleatorios	Factores Mixtos Filas fijo, columnas aleatorio	Factores Mixtos Filas aleatorio, columnas fijo
$F_{Fila} = \frac{CM_{Fila}}{CM_{Inter}}$	$F_{Fila} = \frac{CM_{Fila}}{CM_{Inter}}$	$F_{Fila} = \frac{CM_{Fila}}{CM_{Inter}}$
$F_{Col} = \frac{CM_{Col}}{CM_{Inter}}$	$F_{Col} = \frac{CM_{Col}}{CM_{Error}}$	$F_{Col} = \frac{CM_{Col}}{CM_{Inter}}$
$F_{Inter} = \frac{CM_{Inter}}{CM_{Error}}$	$F_{Inter} = \frac{CM_{Inter}}{CM_{Error}}$	$F_{Inter} = \frac{CM_{Inter}}{CM_{Error}}$

EJEMPLO 5.1.3. Un ingeniero desea probar si hay cambios en el rendimiento de cierto tipo de motor por usar gasolina comprada en el D.F., en el Estado de México o en el Estado de Morelos, mezclada con aditivos de tres diferentes fabricantes. Para ello diseñó un experimento donde se probaron 36 motores idénticos, 4 en cada combinación de gasolina-aditivo. Se midió el rendimiento, en unidades estándar, y los registros aparecen en la siguiente tabla:

	Aditivo A		Aditivo B		Aditivo C		Sumas de filas
D.F.	126.2	124.8	130.4	131.6	127.0	126.6	1539.5
	125.3	127.0	132.5	128.6	129.4	130.1	
Edo. Méx	127.2	126.6	142.1	132.6	129.5	142.6	1593.7
	125.8	128.4	128.5	131.2	140.5	138.7	
Edo. Mor.	127.1	128.3	132.3	134.1	125.2	123.3	1527.4
	125.1	124.9	130.6	133.0	122.6	120.9	
Sumas de columnas	1516.7		1587.5		1556.4		4660.6

- Especifique el modelo de análisis de varianza de este diseño.
- De acuerdo con el modelo elegido, reporte sus cálculos en una tabla de ANDEVA pertinente.
- ¿Dan todas las gasolinas el mismo rendimiento medio? Use $\alpha = 0.05$
- ¿Los diferentes aditivos funcionan semejante, en promedio? Use $\alpha = 0.05$
- ¿Indican los datos algún efecto de interacción? Use $\alpha = 0.05$

Solución:

a) Este es un diseño factorial de dos factores con repetición y está balanceado porque hay el mismo número de observaciones, 4, por celda. Dado que en el texto del problema no se especifica nada respecto a la forma en que se eligieron los factores, se considera un modelo de factores fijos.

b) La tabla de ANDEVA queda de la siguiente manera, de acuerdo con los siguientes cálculos

Cálculo de las sumas de cuadrados:

Suma de cuadrados de fila:

$$\begin{aligned}
 SC_{Fila} &= \frac{(1539.5)^2 + (1593.7)^2 + (1527.4)^2}{12} - \frac{(4660.6)^2}{36} = \\
 &= 603574.225 - 603366.4544 = 207.77
 \end{aligned}$$

Suma de cuadrados de columna:

$$SC_{Col} = \frac{(1516.7)^2 + (1587.5)^2 + (1556.4)^2}{12} - \frac{(4660.6)^2}{36} =$$

$$= 603576.3417 - 603366.4544 = 209.887$$

Suma de cuadrados subtotal:

$$SC_{Subt} = \frac{(503.3)^2 + (523.1)^2 + (513.1)^2 + (508)^2 + (534.4)^2 + (551.3)^2 + (505.4)^2 + (530)^2 + (492)^2}{4} - \frac{(4660.6)^2}{36} =$$

$$= 604047.08 - 603366.4544 = 680.6256$$

Suma de cuadrados de interacción:

$$SC_{Inter} = SC_{Subt} - SC_{Fila} - SC_{Col} = 608.6256 - 207.77 - 209.87 = 262.9686$$

Suma de cuadrados total:

$$SC_{Total} = (126.2)^2 + (124.8)^2 + (130.4)^2 + (131.6)^2 + \dots + (133)^2 + (122.6)^2 + (120.9)^2 - \frac{(4660.6)^2}{36} =$$

$$= 604300.12 - 603366.4544 = 933.666$$

Suma de cuadrados del error

$$SC_{Error} = SC_{Total} - SC_{Subt} = 933.666 - 680.6256 = 2536.04$$

TABLA DE ANDEVA DE 2 FACTORES CON REPETICIÓN					
Fuente de Variación	Grados libres	Suma de Cuadrados	Cuadrado Medio	Estadístico de Contraste	F de tablas
Filas (gasolinas)	3-1=2	207.77	103.885	$*F_{Fila} = \frac{103.885}{9.3719} = 11.0847$	$F_{(0.95, 2, 27)}$ 3.354
Columnas (aditivos)	3-1=2	209.887	104.9435	$F_{Col} = \frac{104.9435}{9.3719} = 11.19767$	$F_{(0.95, 2, 27)}$ 3.327
Interacción (gasol-adit)	(2)(2)=4	262.9686.	65.74215	$F_{Inter} = \frac{65.74215}{9.3719} = 7.0148$	$F_{(0.95, 4, 27)}$ 2.728
Subtotal	(3)(3)-1=8	680.6256	—	—	
Error Residual	36-(3)(3)=27	253.0404	9.3719		
Total	36-1=35	933.666	—		

* La definición de los estadísticos de contraste y de la regla de decisión de cada prueba anotados aquí son los adecuados para un modelo con factores fijos.

Para cada una de las preguntas se plantean las hipótesis respectivas, como sigue:

c) ¿Dan todas las gasolinas el mismo rendimiento medio? Use $\alpha = 0.05$

$$H_o: \mu_{DF} = \mu_{E.Mex} = \mu_{Mor}$$

$$H_a: \text{Al menos un par de } \mu \text{ es diferente}$$

$$\alpha = 0.05$$

Al comparar el estadístico de prueba $F_{Fila} = 11.0847$ con la F teórica $F_{(0.95, 2, 27)} = 3.354$; se rechaza H_o y se concluye que al menos un par de medias de los rendimientos para las distintas gasolinas es diferente.

Haciendo la prueba de Tukey:

$$DSH = q_{(0.05, 3, 27)} = \sqrt{\frac{CM_{Error}}{n_i}} = 3.51 \sqrt{\frac{9.3719}{12}} = 3.1019$$

Obteniendo las diferencias de medias muestrales, de gasolinas, con valor absoluto:

$$|\bar{Y}_{DF} - \bar{Y}_{EMex}| = 4.52, \quad |\bar{Y}_{DF} - \bar{Y}_{Mor}| = 1.01, \quad |\bar{Y}_{EMex} - \bar{Y}_{Mor}| = 5.53$$

Se puede ver que las medias poblacionales, de gasolinas, que son diferentes son:

$$\mu_{DF} \neq \mu_{EMex} \quad y \quad \mu_{EMex} \neq \mu_{Mor}$$

d) ¿Los diferentes aditivos funcionan semejante, en promedio? Use $\alpha = 0.05$

$$H_0: \mu_A = \mu_B = \mu_C$$

$$H_a: \text{Al menos un par de } \mu \text{ es diferente}$$

$$\alpha = 0.05$$

Al comparar el estadístico de prueba $F_{Col} = 22.395$ con la F teórica $F_{(0.95, 2, 27)} = 3.354$; se rechaza H_0 y se concluye que al menos un par de medias de aditivos es diferente.

Haciendo la prueba de Tukey:

$$DSH = q_{(0.05, 3, 27)} = \sqrt{\frac{CM_{Error}}{n_j}} = 3.51 \sqrt{\frac{9.3719}{12}} = 3.1019$$

$$\bar{Y}_A = 126.39, \quad \bar{Y}_B = 132.291, \quad \bar{Y}_C = 129.7$$

Obteniendo las diferencias de medias muestrales, de aditivos, con valor absoluto:

$$|\bar{Y}_A - \bar{Y}_B| = 5.9, \quad |\bar{Y}_A - \bar{Y}_C| = 3.31, \quad |\bar{Y}_B - \bar{Y}_C| = 2.59$$

Se puede ver que las medias poblacionales, de aditivos, que son significativamente diferentes son:

$$\mu_A \neq \mu_B, \quad \mu_A \neq \mu_C$$

e) ¿Existe efecto de interacción entre gasolinas y aditivos?

$$H_0: \alpha_{i..} \times \beta_{.j.} = 0 \text{ (no hay interacción)}$$

$$H_a: \alpha_{i..} \times \beta_{.j.} \neq 0 \text{ (hay interacción)}$$

Al comparar el estadístico de interacción $F_{Inter} = 7.0158$ con la F teórica, $F_{(0.95, 4, 27)} = 2.728$ se rechaza H_0 , por lo que se concluye que si existe efecto de interacción entre las gasolinas y los aditivos.

5.2 Análisis de Regresión

Por regresión se entiende, en estadística, una relación causa-efecto entre 2 o más variables cuantitativas independientes y una variable dependiente.

El análisis de regresión se utiliza cuando un investigador desea:

- a) conocer la posible relación entre 2 o más variables cuantitativas de un proceso aleatorio.
- b) definir el tipo de relación expresándolo como un modelo matemático que explique la naturaleza de dicha relación.

Por ejemplo, la cantidad de lluvia, la naturaleza del suelo y la cantidad cosechada en parcelas de temporal; la cantidad de fertilizante y el crecimiento de las plantas; la cantidad de oxígeno disuelto y el tipo de organismos en un cuerpo de agua, etc.

5.2.1 Análisis de Regresión Lineal Simple

Un análisis de regresión lineal simple, se establece cuando se estudia la relación entre dos variables, en donde una de ellas depende en forma lineal, de la otra. Esto es, el tipo de relación que guardan, entre sí estas variables se explica mediante un modelo lineal, aquel que se representa mediante una ecuación de primer orden.

Cuando se hace un análisis de regresión simple, buscamos aquel modelo lineal que mejor explique la relación entre las variables, aquel que haga mínima la diferencia entre los valores Y observados en el experimento y los valores \hat{Y} esperados, calculados con el modelo ajustado. Para considerar adecuado el modelo matemático es necesario cumplir con algunos supuestos importantes.

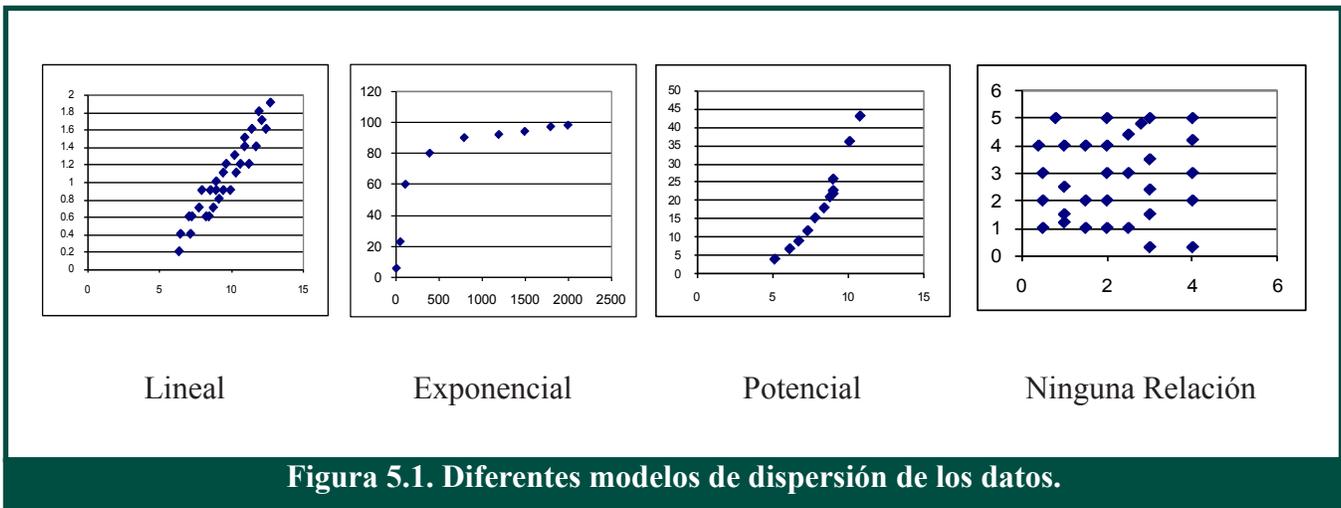
5.2.1.1 Supuestos en el Análisis de Regresión

- Los valores de la variable independiente, X , son fijos, y manejados por el investigador.
- Para cada valor fijo X , habrá una subpoblación de valores Y , cuya distribución es normal.
- Las subpoblaciones de valores Y , presentan variabilidad semejante.
- Las medias de las subpoblaciones Y , se ubican sobre la misma línea.
- Los valores de Y son estadísticamente independientes.

Antes de poder proponer un modelo de ajuste para la población estudiada, es necesario graficar los valores (X, Y) de la muestra y observar la tendencia que presentan, con el fin de definir si es o no pertinente asociar un modelo lineal a los valores experimentales.

5.2.1.2 Diagrama de Dispersión

Es un plano, donde se dispersan los pares ordenados (x, y) , obtenidos de un experimento. Al graficar los pares, se va formando una nube de puntos que nos permite observar la tendencia en la dispersión de los datos. Este gráfico nos permite ver el tipo de modelo matemático al que se ajustan más los datos, por ejemplo:



El modelo lineal simple es el modelo más sencillo, la relación puede ser directa, esto es al crecer x también crece y , o puede ser inversa, cuando al crecer x disminuye y .

Al hacer un análisis de regresión, lo que nos interesa conocer es la media de la población de valores de y , para un valor particular x , cuya definición lineal es:

$$\mu_{y/x} = \alpha + \beta x$$

Donde:

α es la ordenada al origen poblacional, esto es, el punto donde la recta ajustada corta al eje de las Y .

β , es la pendiente poblacional, representa la inclinación de la recta ajustada con respecto al eje X (variable independiente)

x , son valores fijos, elegidos por el investigador, para la variable independiente.

Para definir probabilísticamente a la media poblacional $\mu_{y/x}$, debemos partir de su estimador, que es la media muestral de valores de y :

$$\hat{y}_i = a + bx_i + e_i$$

Esta media muestral se obtiene de una muestra aleatoria de tamaño n . Puesto que la recta obtenida de la muestra es aleatoria, se genera un error probabilístico identificado con la letra griega e_i .

A partir de la media muestral podemos inferir el valor de los parámetros para la población, $\mu_{y/x}$, α y β , para lo cual, es necesario obtener primero el valor de las constantes de regresión a y b .

5.2.1.3 Método de mínimos cuadrados para calcular las constantes de la regresión

Este método se basa en el hecho de que el mejor modelo lineal, es aquel que hace mínima la suma de los cuadrados de los errores, esto es, $\sum_{i=1}^n e_i^2 \rightarrow 0$, para lograrlo es necesario obtener el mínimo, derivando parcialmente la suma con respecto a las constantes de regresión a y b , e igualando a cero.

Por definición, $e_i = y_i - \hat{y}_i$, entonces, $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, sustituyendo la definición de \hat{Y} estimada, tenemos:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

$$\frac{\partial \sum_{i=1}^n [y_i - (a + bx_i)]^2}{\partial a} = -2 \sum_{i=1}^n (y_i - a + bx_i) = 0$$

$$\frac{\partial \sum_{i=1}^n [y_i - (a + bx_i)]^2}{\partial b} = -2 \sum_{i=1}^n (y_i - a - bx_i)x_i = 0$$

originándose así, las ecuaciones normales:

$$\sum_{i=1}^n y_i = na - b \sum_{i=1}^n x_i$$

$$\sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2$$

De la primera ecuación normal, despejamos a la constante a , multiplicando toda esta ecuación por n ; y en la segunda sustituimos la definición algebraica de a y despejamos la constante b quedando como sigue:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x^2 (n-1)}$$

Una vez que se han obtenido los valores de las constantes de regresión, se sustituyen en el modelo de la ecuación lineal que representa a la media muestral de la regresión.

5.2.1.4 Evaluación del Modelo Ajustado

Con objeto de evaluar la bondad de nuestro ajuste, esto es, que tanto acercan los valores de la variable *y estimada* con el modelo propuesto (\hat{y}), a los valores *y observados*, se utiliza el **Coefficiente de Determinación r^2** .

Este coeficiente toma valores entre cero y uno, de tal manera que entre más tienda a cero la diferencia entre las *y observadas* y las *y ajustadas*, más cercano a 1 es el valor de r^2 .

El coeficiente de determinación r^2 , se define como la relación entre la suma de cuadrados explicada por la regresión y la suma de cuadrados total:

$$r^2 = \frac{SC_{Explicada}}{SC_{Total}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Usar esta definición implica sustituir todos los valores x en la ecuación ajustada para obtener las \hat{y} ajustadas y realizar las diferencias cuadráticas con respecto a la media de valores \bar{y} observados. De igual manera, realizar la suma de los cuadrados de todas las diferencias $(y_i - \bar{y})^2$. Esta forma de cálculo es muy tediosa y tardada, sin embargo, si ya tenemos calculadas las constantes de regresión, podemos definir algebraicamente al coeficiente de determinación como sigue:

$$r^2 = b^2 \left[\frac{s_x^2}{s_y^2} \right]$$

5.2.2 Inferencia en el análisis de regresión

Hasta ahora, hemos estado trabajando con los datos obtenidos del muestreo en una población, sin embargo, la finalidad del análisis es conocer, probabilísticamente hablando, el comportamiento de la población de origen. Por esta razón, es importante usar métodos inferenciales para conocer los parámetros de la población objeto de estudio.

En regresión, es válido inferir sobre las medidas poblacionales siempre y cuando, la regresión sea lineal o, cuando no siendo lineal, se convierta a un modelo linealizado, usando transformaciones en los datos originales, como por ejemplo usar los logaritmos de estos datos.

Recalcando, como el motivo real del análisis de regresión es predecir el comportamiento de la población de donde se obtuvo la muestra, es necesario usar métodos inferenciales que permitan definir probabilísticamente a la pendiente poblacional, la media o valor esperado de la población, la ordenada al origen poblacional, la varianza del error de la regresión poblacional, etc. Los cálculos realizados para la inferencia en regresión incluyen intervalos de confianza y contrastes de hipótesis para los parámetros de la regresión.

Es necesario hacer hincapié en que la inferencia en regresión sólo es válida si el modelo es lineal o es un modelo no lineal que ha sido linealizado matemáticamente.

5.2.2.1 Estimación por Intervalo para los parámetros de la regresión

5.2.2.1.1 Intervalo de Confianza para la Varianza del Error de la Regresión

Este intervalo nos permite evaluar probabilísticamente la variabilidad en el error del análisis de regresión, con una confiabilidad $1 - \alpha$. Se utiliza la distribución χ^2 con $n-2$ grados libres. Se pierden 2 grados libres porque para estimar a la varianza del error hay que estimar primero las constantes de regresión a y b .

$$P \left(\sqrt{\frac{(n-2) s_{y/x}^2}{\chi_{(1-\frac{\alpha}{2}, n-2)}^2}} < \sigma_{y/x} < \sqrt{\frac{(n-2) s_{y/x}^2}{\chi_{(1-\frac{\alpha}{2}, n-2)}^2}} \right) = 1 - \alpha$$

$s_{y/x}^2$ es la varianza del error de la regresión en la muestra obtenida aleatoriamente, que se define así:

$$s_{y/x}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \left(\frac{n-1}{n-2} \right) (s_y^2 - b^2 s_x^2)$$

5.2.2.1.2 Intervalo de Confianza para el Error Estándar de la Regresión

Si deseamos el intervalo del error estándar de la regresión, es necesario obtener la raíz cuadrada de ambos extremos del intervalo para la varianza y eliminar el cuadrado en el símbolo de la varianza:

$$P \left(\sqrt{\frac{(n-2) s_{y/x}^2}{\chi_{(1-\frac{\alpha}{2}, n-2)}^2}} < \sigma_{y/x} < \sqrt{\frac{(n-2) s_{y/x}^2}{\chi_{(1-\frac{\alpha}{2}, n-2)}^2}} \right) = 1 - \alpha$$

5.2.2.1.3 Estimación por Intervalo para la Pendiente Poblacional β

Este intervalo nos permite evaluar probabilísticamente a la pendiente poblacional o inclinación de la recta con respecto al eje X . Se utiliza el valor de la pendiente muestral b como estimador para β y el error estándar de la pendiente $\sqrt{s_b^2}$

$$P(b - t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{s_b^2} < b < b + t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{s_b^2} = 1 - \alpha$$

Donde el error estándar de la pendiente $\sqrt{s_b^2}$ se calcula así:

$$\sqrt{s_b^2} = \sqrt{\frac{s_{y/x}^2}{(n-1) s_x^2}}$$

5.2.2.1.4 Estimación por Intervalo para la Ordenada al Origen, Poblacional

Se utiliza la distribución t con $n-2$ grados de libertad, el estimador de la ordenada a , y el error estándar de la ordenada al origen $\sqrt{s_a^2}$:

$$P(a - t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{s_a^2} < A < a + t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{s_a^2} = 1 - \alpha$$

Donde:

$$\sqrt{s_a^2} = \sqrt{s_{y/x}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1) s_x^2} \right)}$$

****Note que usamos la letra A para identificar al parámetro, ordenada al origen, para evitar que se confunda con el símbolo de nivel de significación α .**

5.2.2.1.5 Estimación por Intervalo para el Valor Esperado o Media Poblacional en la Regresión $\mu_{y/x}$.

Cuando se estima por intervalo a la media de la población bivariada, el estimador para un punto dado en la recta media verdadera, será el valor obtenido al sustituir el valor x elegido, en la ecuación ajustada $\hat{y} = a + bx$. Se utiliza la distribución t con $n-2$ grados libres y el error estándar del valor medio $\sqrt{s_{\mu_{y/x}}^2}$

$$P(\hat{y}_0 - t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{s_{\mu_{y/x}}^2} < \mu_{y/x} < \hat{y}_0 + t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{s_{\mu_{y/x}}^2} = 1 - \alpha$$

Donde:

$$\sqrt{s_{\mu_{y/x}}^2} = \sqrt{s_{y/x}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1) s_x^2} \right)}$$

Es importante comentar que al sustituir en el intervalo para un punto x , sólo se estará obteniendo el valor esperado para la población de valores y , específicos para ese valor x . Pero como a medida que nos alejamos del valor medio \bar{x} del intervalo experimental, la incertidumbre crece, el intervalo de valor medio es curvo. Por lo que si deseamos construirlo, es necesario sustituir, al menos 3 veces en la fórmula, eligiendo los dos valores extremos de y y un valor central de la misma variable.

5.2.2.1.6 Estimación por Intervalo para predecir el valor de una subpoblación individual \hat{y} dado un valor x_0 .

Este intervalo se utiliza para predecir valores de las subpoblaciones y , cuando se fija un valor determinado de la variable x . Al igual que en el caso del intervalo para el valor medio, el intervalo de predicción es curvo, pero más amplio que el intervalo para la media, para los mismos valores x , dado que al predecir el comportamiento aleatorio de los sujetos experimentales, la incertidumbre es mayor.

Es importante enfatizar que la predicción será válida, siempre y cuando el investigador tenga la certeza de que el modelo de regresión se mantiene dentro del intervalo que incluye el valor x al que desea predecir.

Para construir un intervalo de predicción es necesario obtener el valor \hat{y} , dado un valor fijo x_0 y calcular el error estándar para la predicción de valores individuales $\sqrt{s_y^2}$.

$$P(\hat{y}_0 - t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{s_y^2} < \hat{y} < \hat{y}_0 + t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{s_y^2} = 1 - \alpha$$

Donde:

$$\sqrt{s_y^2} = \sqrt{s_{y/x}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1) s_x^2} \right]}$$

5.2.2.2 Contrastes de Hipótesis en Regresión Lineal

5.2.2.2.1 Contraste para la linealidad de la regresión

Uno de los contrastes más importantes en regresión es el de linealidad, porque nos sirve para demostrar, probabilísticamente, que la relación lineal entre las variables consideradas, se mantiene dentro de la población. Esta prueba, consiste en demostrar que la pendiente poblacional β es diferente de cero, ya que una pendiente cero, identifica a una recta paralela al eje de las X , esto es, tendrá una inclinación cero y al no haber inclinación tampoco existirá relación.

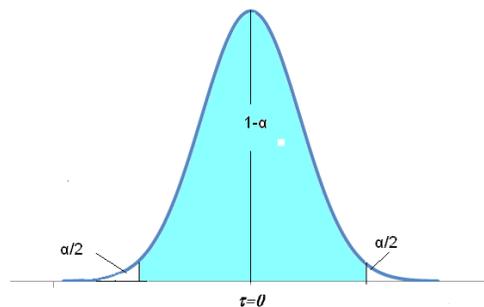
- **Planteamiento de Hipótesis**
 - $H_0: \beta = 0$ (no hay relación lineal entre las variables)
 - $H_a: \beta \neq 0$ (hay relación lineal entre las variables)
- **Elección del nivel de significación α**

- Elección del estimador apropiado y de su distribución de probabilidad

$$b = t_{(1-\frac{\alpha}{2}, n-2)}$$

- Establecimiento de la regla de decisión teórica

Esta prueba es bilateral, es decir, el error α se divide en 2 para generar 2 zonas de rechazo de la hipótesis nula, la de valores por abajo del cero y la de valores por arriba del cero. Entonces leyendo en las tablas probabilísticas de la distribución t se marcan los límites entre rechazo y no rechazo de la hipótesis nula **H₀**.



- Cálculo del estadístico de contraste

$$t = \frac{b - \beta_0}{\sqrt{s_b^2}}$$

Donde β_0 es el valor supuesto para la pendiente poblacional, que en este caso es cero según nuestro planteamiento de hipótesis.

- Toma de decisión

Se compara el valor del estadístico de contraste con el valor límite, en la gráfica de la regla de decisión teórica, obtenido de las tablas de la distribución *t de student* y se rechaza H_0 si el estadístico es mayor, en valor absoluto, que el valor límite, de tablas.

- Conclusión y comentario

5.2.2.2.2 Contraste de hipótesis para probar un valor específico

Aplicando el mismo estadístico de prueba pero estableciendo un valor supuesto para la pendiente poblacional, se puede demostrar, con un nivel de significación elegido previamente por el investigador, que la pendiente verdadera presenta un valor específico o no. Entonces se estaría realizando un contraste para la pendiente poblacional, con un valor que el investigador, por experiencia intuye que puede ser posible. Por

lo demás, el proceso es el mismo explicado en el punto 5.2.2.2.1, aunque en este caso, el contraste puede ser unilateral superior, unilateral inferior o bilateral según los requerimientos de la hipótesis planteada.

5.2.2.2.3 Contraste de Hipótesis para la Ordenada al Origen, α o A .

Se puede probar un valor específico para la ordenada el origen poblacional, usando un estadístico de prueba adecuado, aunque esta prueba no es tan importante como la de la pendiente. El proceso de contraste es semejante al de la prueba para la pendiente pero el planteamiento se refiere a la ordenada al origen y puede ser bilateral o unilateral, también se basa en la distribución *t de student*.

Planteamiento

Bilateral	Unilat. Inferior	Unilat. Superior
$H_0: A = A_0$	$H_0: A \geq A_0$	$H_0: A \leq A_0$
$H_a: A \neq A_0$	$H_a: A < A_0$	$H_a: A > A_0$

Note que la denominación inferior se refiere a que la hipótesis alterna H_a se encuentra en la cola izquierda de la distribución de probabilidad y la superior por el contrario, coloca a la hipótesis alterna del lado derecho en la distribución.

5.2.2.3 Aplicación de la Inferencia en el Análisis de Regresión

EJEMPLO 5.2.1.- Un agrónomo desea evaluar cuál es la producción de maíz, en diferentes condiciones de fertilización del terreno. Como sus fondos son limitados, usa un sólo fertilizante, que aplica a diferentes concentraciones. Los registros de su experimento son los siguientes:

Fertilizante (Lb/Ha)	100	200	300	400	500	600	700
Producción (Quintales/Ha)	40	45	50	65	70	70	80

- a) Dibuje un diagrama de dispersión de los datos.
- b) Establezca el modelo de mínimos cuadrados, que describa la relación entre las variables y trace la recta sobre su diagrama de dispersión.
- c) Evalúe el ajuste obtenido en el inciso **b** e interprete el resultado.
- d) Estime por intervalo la varianza del error de la regresión, con una confianza del 99%.
- e) Estime por intervalo la desviación estándar del error de la regresión, para el problema, con una confianza de 99%
- f) Estime por intervalo la pendiente poblacional β , con una confianza de 99%.
- g) Estime por intervalo la ordenada al origen poblacional **A**, con una confianza de 99%
- h) Estime por intervalo la media poblacional de producción de maíz, cuando la cantidad de fertilizante utilizado es 300 Lb/Ha. Usa un nivel de confianza de 99%.

- i) Haga una predicción para los valores de la subpoblación \hat{Y} , cuando la cantidad de fertilizante aplicada es de 630 Lb/Ha. Use $1 - \alpha = 0.99$
- j) Pruebe si la relación entre el fertilizante y la producción es lineal para la población estudiada. Use $\alpha = 0.05$
- k) ¿Podríamos considerar, al 5% de significación, que la pendiente poblacional es al menos de 0.04?

Solución:

- a) Nos piden trazar el diagrama de dispersión que nos permita visualizar la tendencia de los puntos sobre el plano, por lo que graficaremos los pares ordenados. De acuerdo con el texto del problema, la variable independiente es la cantidad de fertilizante utilizada y la producción lograda es la variable dependiente.

Nota. Es muy importante identificar cual es la variable independiente y cual la dependiente, con objeto de que el modelo que explica la relación sea el correcto.

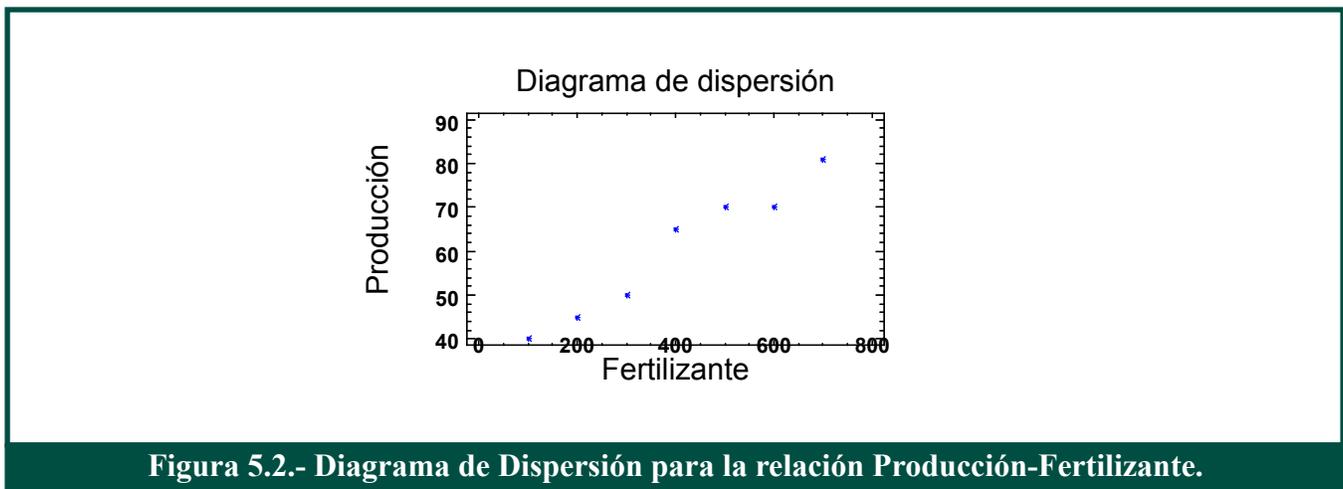


Figura 5.2.- Diagrama de Dispersión para la relación Producción-Fertilizante.

Podemos observar que la tendencia es aproximadamente lineal, por lo que podemos ajustarle un modelo lineal a los datos.

- b) Para establecer el modelo de ajuste, usando nuestra calculadora, en formato estadístico, obtenemos la media y la varianza de ambas variables, fertilizante y producción, además de calcular la suma de los productos X por Y .

Datos	Cálculo de la ordenada	Cálculo de la pendiente
$n = 7$ $\bar{x} = 400$ $s_x^2 = 46,666.66$ $\bar{y} = 60$ $s_y^2 = 225$ $\sum_{i=1}^7 x_i y_i = 187000$	$a = \bar{y} - b\bar{x}$ $a = 60 - 0.067857(400)$ $a = 32.8571$	$b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x^2 (n - 1)} =$ $b = \frac{187000 - 7(400)60}{(46,666.66)6} =$ $b = 0.067857$

Tomando en cuenta que la variable dependiente es la producción \hat{P} y que la variable independiente es la cantidad de fertilizante F , escribimos la ecuación ajustada, como sigue:

$$\hat{P} = 32.8571 + 0.067857F$$

Si sustituimos el valor **más pequeño** de la variable F , fertilizante, y el **más grande** en la ecuación ajustada, podremos trazar la recta que mejor se ajusta a los datos observados.

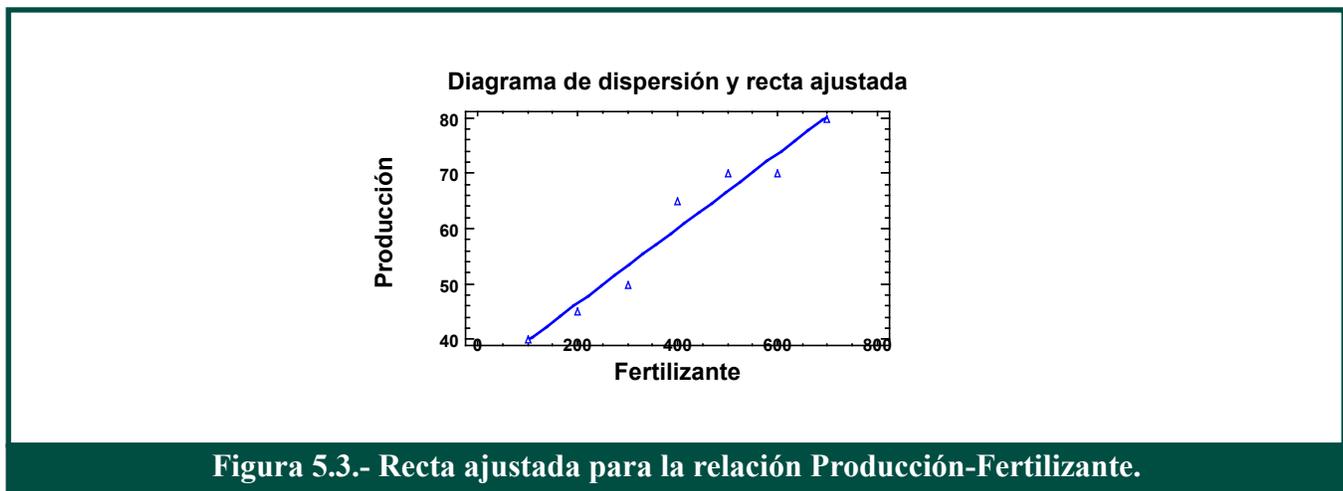


Figura 5.3.- Recta ajustada para la relación Producción-Fertilizante.

Nota. Si contamos con una calculadora que tenga programa de regresión lineal, es más sencillo obtener el modelo ajustado, basta introducir las parejas de datos y pedir los valores de las constantes de regresión.

- c) Para evaluar el ajuste del modelo lineal obtenido, calculamos el coeficiente de determinación muestral r^2 de acuerdo con su fórmula.

$$r^2 = b^2 \left[\frac{s_x^2}{s_y^2} \right] = 0.067857^2 \left[\frac{46,666.66}{225} \right] = 0.955$$

El valor obtenido nos indica que el ajuste es bueno pues 0.955 es cercano a 1. Este coeficiente se interpreta como sigue:

El 95.5% de la variación en la cantidad de maíz producida es debida o es explicada por la variación en la cantidad de fertilizante.

Nota: para realizar todas las estimaciones por intervalo y todos los contrastes se usarán las tablas probabilísticas contenidas en el *Cuaderno de Problemas de Probabilidad y Estadística, Guerra Dávila, T., Marques Dos Santos, M.J. y López Reynoso, J.M. (2009).*

d) Estime por intervalo a la varianza del error de la regresión, con una confianza del 99%.

Primero calculamos la varianza del error de la regresión muestral:

$$s_{y/x}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \left(\frac{n-1}{n-2} \right) (s_y^2 - b^2 s_x^2) = \left(\frac{7-1}{7-2} \right) [225 - (0.067857)^2 (46,666.66)] = 12.144$$

Obtenemos de tablas el valor de la χ^2 para ambos lados de la distribución:

$$1 - \alpha = 0.99 \rightarrow 1 - 0.99 = 0.01$$

$$\alpha = 0.01 \Rightarrow \frac{\alpha}{2} = 0.005 \Rightarrow 1 - \frac{\alpha}{2} = 1 - 0.005 = 0.995$$

$$\chi^2_{(1-\alpha/2, n-2)} = \chi^2_{(0.995, 5)} = 16.7496 \quad \chi^2_{(\alpha/2, n-2)} = \chi^2_{(0.005, 5)} = 0.4118$$

Sustituyendo en la ecuación que define al intervalo tenemos:

$$\frac{5(12.144)}{16.7496} < \sigma_{y/x}^2 < \frac{5(12.144)}{0.4118}$$

$$3.625 < \sigma_{y/x}^2 < 147.450, \text{ con } 99\% \text{ de confianza}$$

Interpretación: Este resultado nos indica que la varianza verdadera, del error de regresión se encontrará entre 3.625 y 147.45 (Quintales/Ha)², en noventa y nueve de cada 100 veces que hagamos la estimación, en las mismas condiciones.

- e) Estime por intervalo a la desviación estándar del error de la regresión, para el problema, con una confianza de 99%.

$$\sqrt{\frac{5(12.144)}{16.7496}} < \sigma_{y/x}^2 < \sqrt{\frac{5(12.144)}{0.4118}}$$

$$1.904 < \sigma_{y/x} < 12.143, \text{ con } 99\% \text{ de confianza}$$

Interpretación: En este caso, la interpretación del resultado nos dice que, de cada 100 intervalos calculados en estas condiciones, en 99 de ellos, el error estándar de la regresión caerá entre estos límites.

- f) Estime por intervalo a la pendiente poblacional β , con una confianza de 99%.

$$1 - \alpha = 0.99 \Rightarrow 1 - 0.99 = 0.01$$

$$\alpha = 0.01 \Rightarrow \alpha/2 = 0.005 \Rightarrow 1 - \alpha/2 = 1 - 0.005 = 0.995$$

$$t_{(1-\alpha/2, n-2)} = t_{(0.995, 5)} = 4.0321$$

Cálculo del error estándar de la pendiente:

$$\sqrt{s_b^2} = \sqrt{\frac{s_{y/x}^2}{(n-2)s_x^2}} = \sqrt{\frac{12.144}{6(46,666.66)}} = 6.5857 \times 10^{-3}$$

Sustituyendo en la definición del intervalo para la pendiente poblacional

$$0.067857 - 4.0321(6.5857 \times 10^{-3}) < \beta < 0.067857 + 4.0321(6.5857 \times 10^{-3})$$

$$0.067857 - 0.02655420 < \beta < 0.067857 + 0.0265542$$

$$0.0413 < \beta < 0.0944, \text{ con } 99\% \text{ de confianza}$$

Interpretación: Este resultado nos indica que, de cada 100 intervalos calculados en estas condiciones, la pendiente de la población muestreada caerá entre estos límites

- g) Estime por intervalo a la ordenada al origen poblacional \mathbf{A} , con una confianza de 99%.

Como vamos a estimar al mismo nivel de confianza que utilizamos para la pendiente poblacional, el valor de la distribución t es el mismo:

$$t_{(1-\alpha/2, n-2)} = t_{(0.995, 5)} = 4.0321$$

Calculando el error estándar de la ordenada al origen

$$\sqrt{s_a^2} = \sqrt{s_{y/x}^2 \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} = \sqrt{12.144 \frac{1}{7} + \frac{400^2}{6(46,666.66)}} = 2.9452$$

Sustituyendo en la ecuación que define al intervalo de la ordenada:

$$32.8571 - 4.0321(2.9452) < A < 32.8571 + 4.0321(2.9452)$$

$$32.8571 - 11.8753 < A < 32.8571 + 11.8753$$

$$20.9818 < A < 44.7324, \text{ con } 99\% \text{ de confianza}$$

Interpretación: De cada cien veces que se trabaje el muestreo en la población estudiada, en 99 de ellas, la ordenada al origen poblacional caerá entre estos límites.

- h) Estime por intervalo a la media poblacional de producción de maíz, cuando la cantidad de fertilizante utilizado es 300 Lb/Ha. Use un nivel de confianza de 99%.

Primero sustituimos el valor 400 en la ecuación ajustada para obtener la Y estimada en ese punto:

$$\hat{P} = 32.8571 + 0.067857(300) = 53.2142$$

Calculando el error estándar de valor medio:

$$\sqrt{s_{\mu y/x}^2} = \sqrt{s_{\mu y/x}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)} = \sqrt{12.144 \left(\frac{1}{7} + \frac{(300 - 400)^2}{6(46,666.66)} \right)} = \sqrt{2.16857} = 1.4726$$

Como el nivel de confianza es el mismo que el de los intervalos anteriores, usamos el mismo valor de la distribución t .

$$t_{(1-\alpha/2, n-2)} = t_{(0.995, 5)} = 4.0321$$

Sustituyendo en la ecuación del intervalo tenemos:

$$53.2142 - 4.0321(1.4726) < \mu_{y/x} < 53.2142 + 4.0321(1.4726)$$

$$53.2142 - 5.9377 < \mu_{y/x} < 53.2142 + 5.9377$$

$$47.2765 < \mu_{y/x} < 59.1519, \text{ con } 99\% \text{ de confianza}$$

Interpretación: Este resultado nos está indicando que de cada 100 intervalos calculados en estas condiciones en 99 de ellos el valor esperado para la producción de maíz, cuando X es 300, se encontrará entre 47.2765 y 59.1519 Quintales/Ha.

Los intervalos de confianza para valores esperados correspondientes a cada uno de los valores de x_i al ser dibujados sobre el diagrama de dispersión con la recta ajustada, determinarían la banda de confianza entre las líneas de color rojo que se muestra en el siguiente diagrama:

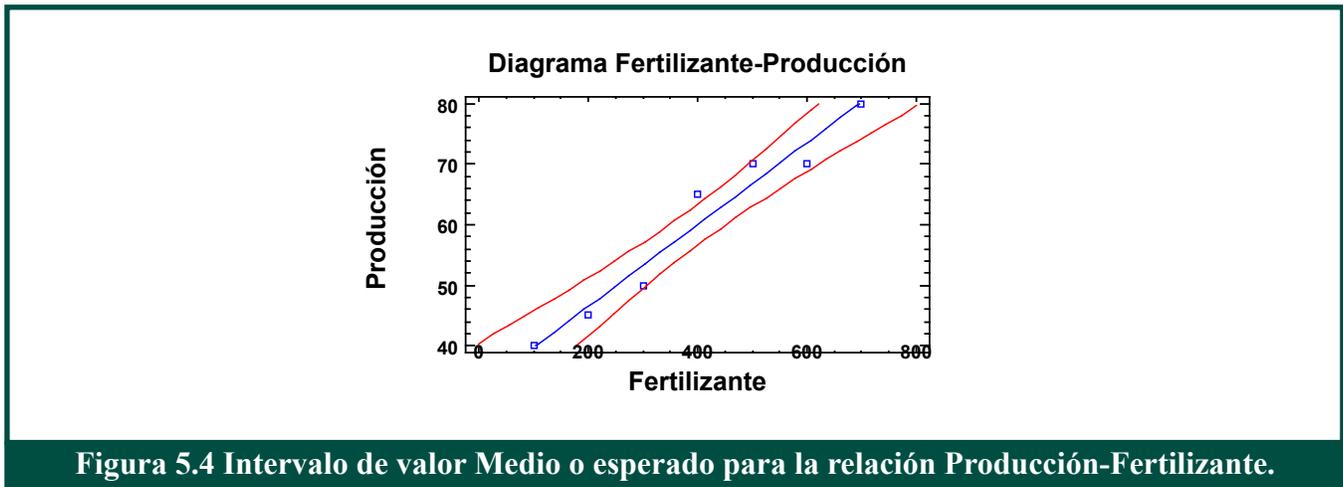


Figura 5.4 Intervalo de valor Medio o esperado para la relación Producción-Fertilizante.

Note que el intervalo es más estrecho en los valores centrales de la variable x y más amplio en los extremos, a medida que nos alejamos del centro del intervalo de observación.

- i) Haga una predicción para los valores de la subpoblación \hat{y} , cuando la cantidad de fertilizante aplicada es de 630 Lb/Ha. Use $1 - \alpha = 0.9$.

Primero calculamos el valor puntual de predicción \hat{y}_0 , sustituyendo la ecuación ajustada para el valor $x_0 = 630$.

$$\hat{P} = 32.8571 + 0.067857(630) = 75.607$$

Cálculo del error estándar para la predicción:

$$\sqrt{s_y^2} = \sqrt{s_{y/x}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)} = \sqrt{12.144 \left(1 + \frac{1}{7} + \frac{(630 - 400)^2}{6(46,666.66)} \right)} = \sqrt{16.1717} = 4.0214$$

El valor de la distribución t es el mismo utilizado en los intervalos anteriores, porque se trabaja con la misma confiabilidad: $t_{(1-\alpha/2, n-2)} = t_{(0.995, 5)} = 4.0321$

Sustituyendo en la definición por intervalo para predicción:

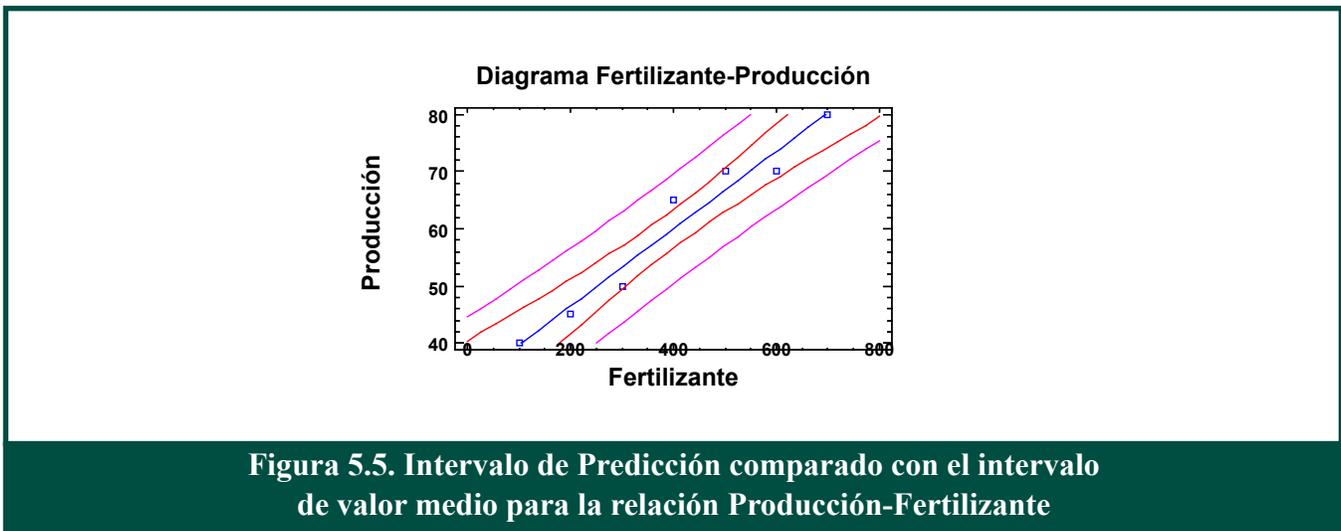
$$75.607 - 4.0321(4.0214) < \hat{y} < 75.607 + 4.0321(4.0214)$$

$$75.607 - 16.2147 < \hat{y} < 75.607 + 16.2147$$

$$59.3924 < \hat{y} < 91.8218 \text{ con una confianza de } 99\%$$

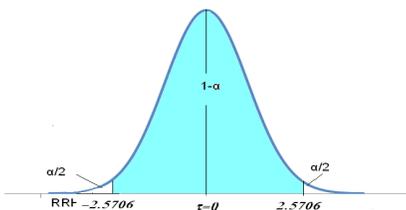
Interpretación: Este resultado nos indica que cuando se fija la cantidad de fertilizante utilizada en 630 Lb/Ha, la producción se encontrará entre 59.3924 y 91.8218 Lb/Ha, en 99 de cada 100 intervalos calculados en estas condiciones.

Cuando sobre el diagrama de dispersión se dibujan, tanto el intervalo de valor medio como el de valores de predicción, puede verse claramente que éste último es más amplio y que también es curvo, más estrecho en el centro y más amplio en los extremos.

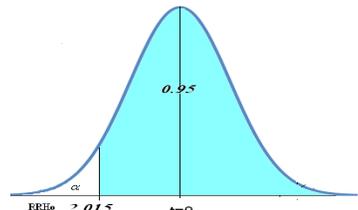


- j) Pruebe si la relación entre la producción y el fertilizante es lineal para la población estudiada. Use $\alpha = 0.05$

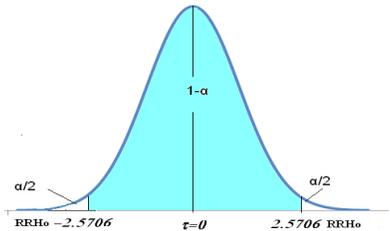
1) Planteamiento de la hipótesis	2) Nivel de significación
Bilateral $H_0: \beta = 0$ $H_a: \beta \neq 0$	$\alpha = 0.05$ $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$ $1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$

<p>3) Distribución utilizada y valor crítico</p> $b \sim t_{(1-\alpha/2, n-2)}$ $t_{(1-\alpha/2, n-2)} = t_{(0.975, 5)} = 2.5706$	<p>4) Regla de Decisión</p> 
<p>5) Estadístico de contraste</p> $t_{Calc} = \frac{b - \beta_0}{\sqrt{S_b^2}}$	<p>6) Cálculos</p> $t_{Calc} = \frac{0.067857 - 0}{6.5857 \times 10^{-3}} = 10.30$
<p>7) Decisión</p> <p>Se rechaza H_0 porque $10.3 > 2.5706$, el valor calculado está en la región de rechazo del lado derecho</p>	<p>8) Conclusión</p> <p>Con una significación del 5% podemos asegurar que la pendiente poblacional es diferente de cero y por lo tanto, se mantiene la relación entre las variables fertilizante y producción.</p>

k) ¿Podríamos considerar, al 5% de significación, que la pendiente poblacional es al menos de 0.04?

<p>1) Planteamiento de la hipótesis</p> <p>Unilateral Inferior</p> $H_0: \beta \geq 0.04$ $H_a: \beta < 0.04$	<p>2) Nivel de significación</p> <p>$\alpha = 0.05$</p> <p>Como el planteamiento es unilateral inferior, no se divide el α entre dos, por lo que la confianza será: $1 - \alpha = 0.95$</p> <p>La región de rechazo estará del lado izquierdo</p>
<p>3) Distribución utilizada y valor crítico</p> $b \sim -t_{(1-\alpha, n-2)}$ $-t_{(1-\alpha, n-2)} = -t_{(0.95, 5)} = 2.015$	<p>4) Regla de Decisión</p> 
<p>5) Estadístico de contraste</p> $t_{Calc} = \frac{b - \beta_0}{\sqrt{S_b^2}}$	<p>6) Cálculos</p> $t_{Calc} = \frac{0.067857 - 0.04}{6.5857 \times 10^{-3}} = 4.23$
<p>7) Decisión</p> <p>No se rechaza H_0 porque, el valor calculado 4.23 se encuentra en el lado derecho de la distribución t y por lo tanto no se encuentra en la región de rechazo.</p>	<p>8) Conclusión</p> <p>Con una significación del 5% podemos asegurar que la pendiente poblacional es al menos 0.04.</p>

1) ¿Es la ordenada al origen poblacional igual a 40? Use $\alpha = 0.05$.

1) Planteamiento de la hipótesis	2) Nivel de significación
Bilateral $H_0: A = 40$ $H_a: A \neq 40$	$\alpha = 0.05$ $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$ $1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$ Como el análisis es bilateral, habrá 2 regiones de rechazo
3) Distribución utilizada y valor crítico	4) Regla de Decisión
$\alpha \sim t_{(1-\alpha/2, n-2)}$ $t_{(1-\alpha/2, n-2)} = t_{(0.975, 5)} = 2.5706$	
5) Estadístico de prueba	6) Cálculos
$t_{Calc} = \frac{a - A_0}{\sqrt{S_a^2}}$	$t_{Calc} = \frac{32.8571 - 40}{2.9452} = 2.425$
7) Decisión	8) Conclusión
No se rechaza H_0 porque $2.425 < 2.5706$, en valor absoluto por lo tanto no toca la región de rechazo.	Con una significación del 5% no podemos rechazar que la ordenada al origen poblacional sea igual a 40.

5.2.3 Análisis de regresión no lineal

Cuando en un diagrama de dispersión la nube de puntos muestra una cierta curvatura, el modelo de regresión no es lineal y podría tratarse de un modelo de regresión exponencial o semi-logarítmico o de un modelo potencial o doble logarítmico.

5.2.3.1. Análisis comparativo de los modelos exponencial y potencial con el lineal

Con objeto de definir qué tipo de modelo se tiene, es conveniente analizar el comportamiento de las variables. Si se relaciona el logaritmo natural de la variable dependiente y con la variable independiente x , y el diagrama de dispersión muestra un comportamiento lineal, se tendrá una regresión exponencial. Si el diagrama, después de tomar logaritmos en y sigue mostrando tendencia curva, no se tratará de un modelo

exponencial. Entonces, será conveniente relacionar los logaritmos naturales de ambas variables y trazar el diagrama de dispersión, si éste muestra una tendencia lineal el modelo de ajuste será el potencial.

La ecuación real del modelo no lineal se establece obteniendo la exponencial de las constantes de regresión linealizadas. Con objeto de aclarar estas transformaciones revise el cuadro comparativo siguiente:

Regresión Lineal	Regresión Exponencial	Regresión Potencial
Pendiente muestral $b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{s_x^2 (n - 1)}$	Logaritmo de la pendiente muestral $\ln b = \frac{\sum_{i=1}^n x_i (\ln y_i) - n\bar{x}(\overline{\ln y})}{s_x^2 (n - 1)}$	Pendiente muestral $b = \frac{\sum_{i=1}^n x_i (\ln x)(\ln y) - n(\overline{\ln x})(\overline{\ln y})}{(n - 1) s_{\ln x}^2}$
Ordenada al origen muestral $a = \bar{y} - b\bar{x}$	Logaritmo de la ordenada muestral $\ln a = \overline{\ln y} - (\ln b)\bar{x}$	Logaritmo de la ordenada muestral $\ln a = \overline{\ln y} - b \overline{\ln x}$
Ecuación del modelo lineal $\hat{y} = a + bx$	Ecuación linealizada del modelo exponencial $\ln \hat{y} = \ln a + (\ln b)x$ Ecuación del modelo exponencial $\hat{y} = a \times b^x$	Ecuación linealizada del modelo potencial $\ln y = \ln a + b \ln x$ Ecuación del modelo potencial $\hat{y} = a x^b$

Como se puede observar en el cuadro, las ecuaciones para las constantes y para el modelo de regresión se ven afectadas por las modificaciones utilizadas para linealizar el modelo.

Para tener más claro cuál modelo se ajusta mejor a los datos observados, deben compararse los coeficientes de determinación r^2 y el mejor ajuste será el del modelo con el coeficiente de determinación más cercano a 1.

5.2.3.2 Inferencia en regresión no lineal

La inferencia sólo es válida para la regresión lineal, por lo que si el modelo de ajuste es no lineal deberá utilizarse la ecuación linealizada del modelo para poder realizar el análisis inferencial, desde luego, modificando las fórmulas de cálculo de intervalos y contrastes de hipótesis. Para que esto quede claro, se trabajará, paso a paso con un ejemplo de cada tipo.

EJEMPLO 5.2.3. Los siguientes datos corresponden al costo de producción de ciertos componentes electrónicos y el número de unidades que se producen:

Tamaño del lote	50	100	250	500	1000
Costo unitario (\$)	108	53	24	9	5

- Haga un diagrama de dispersión de los datos
- Determine el modelo que mejor se ajusta a los datos
- Utilice el modelo ajustado para pronosticar el costo unitario de un lote de 300 componentes, con una confianza de 95%.

Solución:

- En este caso, la variable dependiente y es el costo unitario.

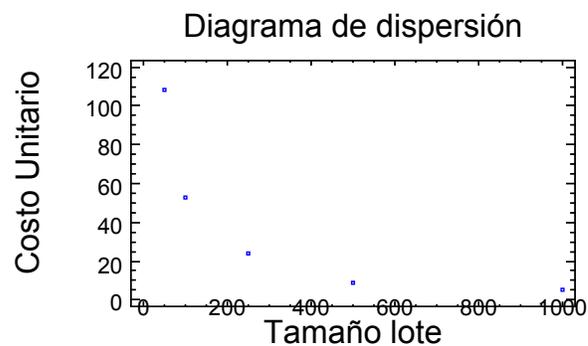


Figura 5.6 Diagrama de dispersión del ejemplo 5.2.3.

En este diagrama se observa que la tendencia de los puntos es curvilínea por lo que se modificaran los datos relacionando $\ln y$ contra x .

- Si se realiza el diagrama de dispersión modificado con el logaritmo natural de y , se tiene:

Tamaño del Lote	50	100	250	500	1000
Ln (costo unitario)	4.682131227	3.970291914	3.17805383	2.197224577	1.609437912

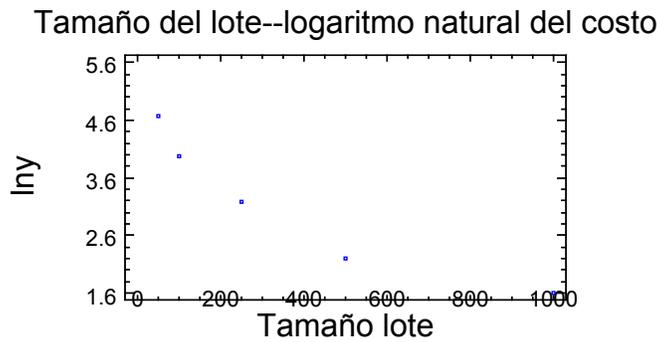


Figura 5.7 Diagrama de dispersión modificado del ejemplo 5.2.3.

Este diagrama nos muestra que al graficar el logaritmo natural del costo unitario contra el tamaño del lote, se sigue observando cierta curvatura, por lo que graficaremos tomando los logaritmos naturales de ambas variables.

Ln (Tamaño del lote)	3.912023	4.60517	5.52146	6.214608	6.907755
Ln (costo unitario)	4.682131227	3.970291914	3.17805383	2.197224577	1.609437912

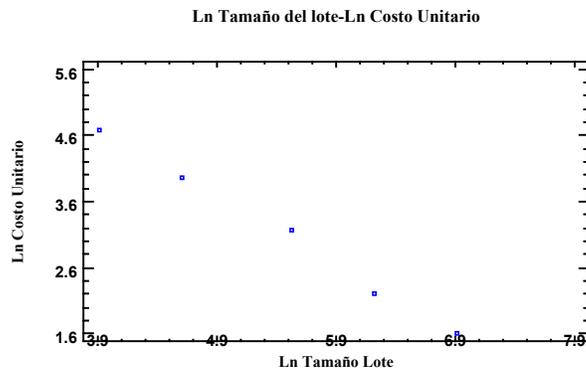


Figura 5.8 Diagrama del ejemplo 5.2.3 modificado tomando logaritmos en ambas variables.

En este diagrama puede observarse que la tendencia de los puntos es lineal por lo que podemos determinar que el modelo que mejor se ajusta a los datos es el modelo potencial.

Para establecer el modelo potencial que explique las variaciones del costo unitario respecto al tamaño del lote se harán los cálculos pertinentes de acuerdo con las fórmulas de la última columna del cuadro comparativo de la página 137.

$\overline{\ln x} = 5.432203497; \quad s_{\ln x} = 1.203360891$ $\overline{\ln y} = 3.127427892; \quad s_{\ln y} = 1.254955875$ $n = 5$	<p style="text-align: center;">Cálculo de la pendiente</p> $b = \frac{\sum_{i=1}^n x_i (\ln x) (\ln y) - n(\overline{\ln x})(\overline{\ln y})}{(n-1) s_{\ln x}^2} =$ $b = \frac{78.92046793 - 5(5.432203497)(3.127427892)}{1.448077433(5-1)} =$ $b = 1.039940198$
<p style="text-align: center;">Cálculo de la ordenada</p> $\ln a = \ln y - b \ln x =$ $3.127427892 - (-1.039940198)(5.432203497) =$ $\ln a = 8.776594675$	<p style="text-align: center;">Ecuación linealizada del modelo potencial</p> $\ln \hat{y} = \ln a + b \ln x$ <p style="text-align: center;">Valor puntual de predicción</p> $(\ln \hat{y})_0 = 8.776594675 + (-1.039940198)(\ln 300)$ $(\ln \hat{y})_0 = 8.776594675$

De acuerdo con el cuadro anterior, la ecuación del modelo potencial es:

$$\hat{y} = ax^b \Rightarrow \hat{y} = 6480.77 x^{-1.03994}$$

- c) Para hacer la predicción por intervalo del costo unitario cuando el lote es de 300, se estimará por intervalo al 95% de confianza.

<p>1) Estimación puntual del costo</p> $\ln \hat{y} = 8.776594675 + (-1.039940198) \ln x$ $\ln \hat{y} = 8.776594675 - 1.039940198(\ln 300) =$ $\ln \hat{y} = 2.845001999$	<p>2) Nivel de confianza para el cálculo</p> $1 - \alpha = 0.95$ <p>3) Valor de tablas para el nivel de confianza establecido</p> $t_{(1-\alpha/2, n-1)} = t_{(0.975, 3)} = 3.1825$
---	---

Cálculo del error estándar de regresión

$$s_{\ln y / \ln x} = \sqrt{\left(\frac{n-1}{n-2}\right)(s_{\ln y}^2 - b^2 s_{\ln x}^2)}$$

$$s_{\ln y/\ln x} = \sqrt{\frac{4}{3} (1.5749 - (-1.03994)^2) (1.4481)}$$

$$s_{\ln y/\ln x} = 0.1086512$$

Estimación por intervalo del costo unitario cuando el lote es de 300

$$\ln \hat{y}_0 - t_{(1-\alpha/2, n-1)} s_{\ln y/\ln x} \sqrt{1 + \frac{1}{n} + \frac{(\ln x_0 - \overline{\ln x})^2}{(n-1)s_{\ln x}^2}} < \ln \hat{y} < \ln \hat{y}_0 + t_{(1-\alpha/2, n-1)} s_{\ln y/\ln x} \sqrt{1 + \frac{1}{n} + \frac{(\ln x_0 - \overline{\ln x})^2}{(n-1)s_{\ln x}^2}}$$

$$2.845002 - (3.1825)(0.1086512) \sqrt{1 + \frac{1}{5} + \frac{(\ln 300 - 5.4322)^2}{(4)(1.4481)}} < \ln \hat{y} < 2.845002 + \dots$$

$$2.4642 < \ln \hat{y} < 3.2258$$

Para obtener el valor real del costo se debe obtener la exponencial a ambos lados del intervalo anterior

$$11.7542 < \hat{y} < 25.1735, \text{ con } 95\% \text{ de confianza}$$

Interpretación:

La relación Costo-tamaño del lote, se comporta de acuerdo con un modelo potencial y de cada 100 intervalos que se calculen con el mismo nivel de confianza, en 95 de ellos el costo unitario para un lote de 300 estará entre 11.7542 y 25.1735 pesos.

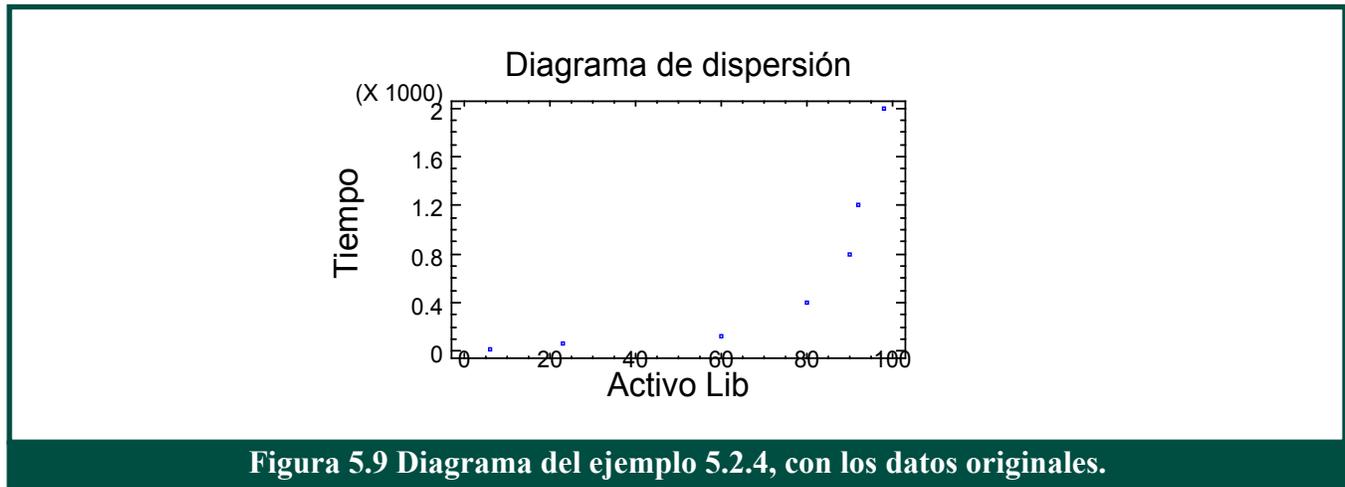
EJEMPLO 5.2.4. Hoech de México, encabezó un estudio para determinar el tiempo de disolución de cierta formulación de tabletas de liberación controlada. Cuando se usó el método de sales biliars de la USP XXI se obtuvieron los siguientes datos:

Tiempo (min)	15	60	120	400	800	1200	2000
Conc. de activo liberado (%)	6	23	60	80	90	92	98

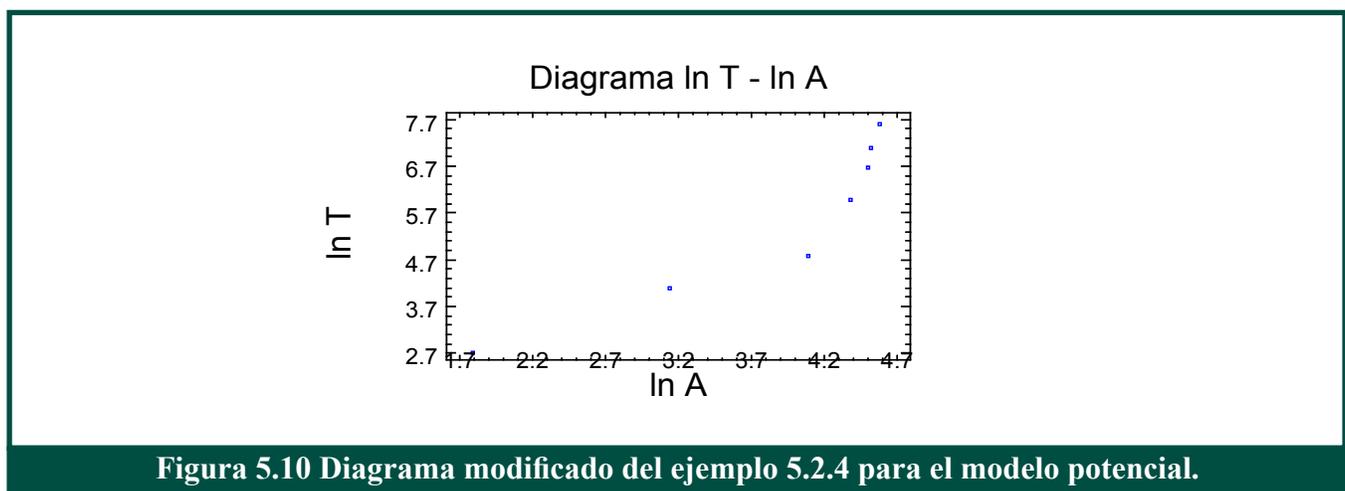
- a) Trace un diagrama de dispersión de los datos
- b) Establezca y escriba la ecuación del modelo de regresión adecuado para estos datos
- c) Estime, con una confianza de 99% la cantidad de minutos necesarios para que se libere el 95% del activo.

Solución:

a) Diagrama de Dispersión



Como puede observarse en este diagrama, la relación entre las variables no es lineal. Se probará la relación con los datos transformados, graficando los logaritmos naturales de ambas variables.



De acuerdo con este diagrama, puede concluirse que el modelo potencial tampoco es el adecuado para representar el comportamiento de las variables, tiempo- activo liberado porque los puntos no forman una línea recta.

Ahora se tratará de ajustar un modelo exponencial a los datos trazando el diagrama Activo liberado contra el logaritmo natural del tiempo.

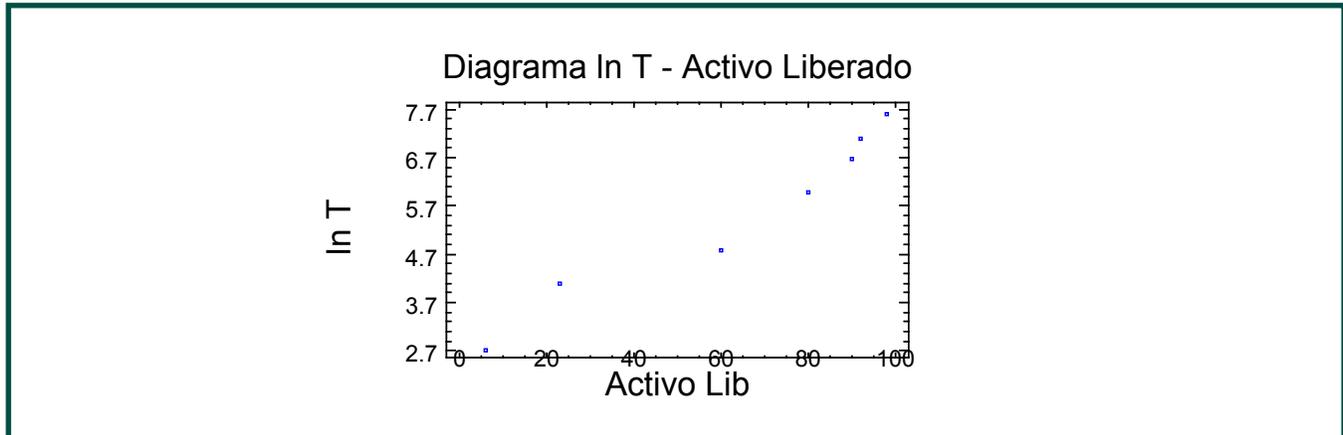


Figura 5.11 Diagrama modificado del ejemplo 5.2.4 para el modelo exponencial.

- b) En este último gráfico puede verse que los puntos se dispersan aproximadamente como una recta. Sin embargo, ante la duda será conveniente calcular los coeficientes de determinación r^2 para cada modelo. El modelo que más se ajusta es aquel que presenta el coeficiente más cercano a uno.

Lineal	Potencial	Exponencial
$n = 7$	$n = 7$	$n = 7$
$\bar{x} = 64.1429$	$\overline{\ln x} = 3.858599$	$\bar{x} = 64.1429$
$s_x^2 = 1322.1429$	$s_{\ln x}^2 = 1.08381$	$s_x^2 = 1322.1429$
$\bar{y} = 656.429$	$\overline{\ln y} = 5.5652775$	$\overline{\ln y} = 5.5652775$
$s_y^2 = 540322.619$	$s_{\ln y}^2 = 3.139278$	$s_{\ln y}^2 = 3.139278$
$a = 348.8978$	$\ln a = -0.04752$	$\ln a = 2.517237$
$b = 15.6732$	$b = 1.571061$	$\ln b = 0.04752$
$r^2 = 0.6011$	$r^2 = 0.8521$	$r^2 = 0.95103$
		$\ln \hat{y}_0 \big _{x_0=95} = 7.0315954$

Al revisar los valores del coeficiente de determinación muestral, r^2 , de cada modelo se puede observar que el modelo exponencial es el que presenta una r^2 más cercana a la unidad por lo que se concluye que el modelo más adecuado para representar la relación entre el tiempo y la cantidad de activo liberado es el exponencial.

La ecuación linealizada del modelo exponencial, para este problema es:

$$\ln T = \ln a + A \times \ln b$$

$$\ln T = 2.517237 + 0.04752 A$$

Donde T representa tiempo y A representa activo liberado.

Ahora bien, la ecuación del modelo exponencial se obtiene calculando y sustituyendo la exponencial de las constantes de regresión en la definición de este modelo como sigue:

$$\hat{Y} = a(b)^x \Rightarrow \hat{T} = (12.3943)(1.048667)^4$$

- c) Para estimar la cantidad de minutos necesarios para que se libere el 95% del activo, se realiza el intervalo de predicción para el tiempo requerido al 99% de confianza:

Debe recordarse que la inferencia sólo es válida para modelos lineales o modelos linealizados, por lo que se trabajará con la ecuación linealizada y al final se transformaran los valores al modelo real.

$$\ln \hat{y} \mp t_{(1-\alpha/2, n-2)} s_{\ln y/x} \sqrt{1 + \frac{1}{7} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}} < \ln \hat{Y} < \dots$$

$$\ln \hat{y} = \ln a + x \ln b$$

$$\ln \hat{y}_{x=95} = 2.517237 + 0.04752(95) = 7.0316$$

$$t_{(1-\alpha/2, n-2)} = t_{(0.995, 5)} = 4.0321$$

$$s_{\ln y/x} = \sqrt{\left(\frac{n-1}{n-2}\right)(s_{\ln y}^2 - (\ln b)^2 s_x^2)} = \sqrt{\frac{6}{5}(3.139278 - (0.04752)^2 1322.1429)}$$

$$s_{\ln y/x} = 0.429515$$

Sustituyendo los valores del logaritmo de la y estimada, el valor de la distribución t y todo lo que se pide en el intervalo de predicción se tiene:

$$7.0316 \mp 4.0321(0.429515) \sqrt{1 + \frac{1}{7} + \frac{(95 - 64.1429)^2}{(6)1322.1429}} < \ln \hat{y} < \dots$$

$$7.0316 \mp 1.946218 < \ln \hat{T} < \dots$$

$(5.085377 < \ln \hat{T} < 8.977813)$ con 99% de confianza

$(161.6409 < \hat{T} < 7925.2844)$ con 99% de confianza

Interpretación: De cada cien intervalos calculados en las mismas condiciones, 99 de ellos mostraran que el valor de predicción está entre 161.64 y 7925.28.

5.3 Análisis de Correlación Lineal

La asociación lineal entre 2 variables cuantitativas puede ocurrir cuando existe dependencia entre ellas o sin que las variables dependan una de la otra, esto significa que ambas variables están relacionadas sin que haya una relación de causa-efecto. Aunque estrictamente, en el análisis de correlación no hay una variable independiente x y una variable dependiente y , es conveniente utilizar la nomenclatura x_i, y_i para diferenciarlas, con objeto de facilitar los cálculos.

Por ejemplo, el llanto del bebé en una noche cualquiera y el insomnio de la madre están asociados, aunque puede que una señora tenga insomnio sin que el bebé llore. Es decir, no existe una relación causa-efecto. Sin embargo, la venta de pan y la criminalidad pueden presentar una asociación lineal sin que haya una dependencia entre estas 2 variables, ¿si se deja de vender pan se acaba con la criminalidad?, es claro que no, el efecto de asociación se debe a que ambas variables, independientemente, sufren el efecto del crecimiento poblacional. En el caso de la asociación lineal, ambas variables son medidas sin que una de ellas esté controlada, esto es, ambas son variables aleatorias.

La asociación lineal entre 2 variables cuantitativas y_1, y_2 se mide por medio de un coeficiente llamado Coeficiente de Correlación lineal y se representa con la letra r , cuando se habla de la correlación en una muestra. Cuando se trata de la población, utilizamos el símbolo griego ρ .

A un investigador puede interesarle saber si puede medir indirectamente el valor adquirido por una determinada variable utilizando como referencia otra variable, sabiendo que no son dependientes por lo que primero deberá trazar un diagrama de dispersión de los pares ordenados formados por los valores de las variables para observar la tendencia de la asociación, después, si es pertinente, deberá definir el grado de asociación lineal entre las mismas y después tal vez realizará contrastes de hipótesis para definir la validez de sus suposiciones estadísticas respecto a la asociación que ambas variables guardan en la población.

Tanto si se traza el diagrama como si se hace el cálculo del coeficiente de correlación, no importará cual variable se coloque en el eje de las abscisas y cual en el eje de las ordenadas, el resultado es equivalente.

5.3.1 Definición Matemática del Coeficiente de Correlación Muestral

Para calcular el coeficiente de correlación muestral se utiliza la fórmula siguiente:

$$r = \frac{\sum_{i=1}^n y_{1i}y_{2i} - n\bar{y}_1\bar{y}_2}{(n-1)s_{y_1}s_{y_2}}$$

5.3.2 Contrastes de Hipótesis Relacionados con la Correlación Lineal Poblacional

5.3.2.1 Contraste de Hipótesis para probar la existencia de asociación lineal verdadera entre las variables de un experimento aleatorio

Para un investigador es importante saber si la asociación entre las variables estudiadas se mantiene en la población de donde fue extraída la muestra, por esta razón recurre al análisis inferencial, aplicado a la correlación lineal entre variables.

La prueba más sencilla consiste en suponer que la correlación poblacional ρ vale cero, lo que indicaría una falta total de asociación entre las variables de la población, es una prueba bilateral. Para realizarla se elige un nivel de significación adecuado, se hace el planteamiento y utilizando la distribución *t de Student*, con $n-2$ grados libres como patrón de referencia teórica, se contrasta con el estadístico de prueba siguiente:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

5.3.2.2 Contraste de Hipótesis para probar si la población presenta un grado de asociación específico

ρ_0 .

Si el investigador deseara probar la factibilidad que el coeficiente de correlación poblacional ρ presente un valor determinado, deberá aplicar un contraste de hipótesis para este parámetro, bajo el supuesto de que la distribución del estadístico es normal, usando el parámetro Z de la normal.

Para que los resultados obtenidos sean confiables es necesario que la cantidad de datos en la muestra sea grande ($n \geq 50$).

El estadístico de contraste para un valor supuesto para el coeficiente de correlación poblacional es como sigue:

$$z^* = \frac{z_r - z_{\rho_0}}{\sigma_{z_r}}$$

Donde: z_r , es el valor estandarizado del coeficiente de correlación $z_r = \frac{1}{2} \ln \left| \frac{1+r}{1-r} \right|$

σ_{z_r} , es el error estándar del estimador $\sigma_{z_r} = \frac{1}{\sqrt{n-3}}$

z_{ρ_0} , es el valor estandarizado del valor supuesto para el coeficiente de correlación poblacional

$$z_{\rho_0} = \frac{1}{2} \ln \left| \frac{1 + \rho_0}{1 - \rho_0} \right|$$

5.3.3 Estimación por Intervalo para el Coeficiente de Correlación Poblacional

También puede estimarse, por intervalo, el valor probable del coeficiente de correlación poblacional, siempre bajo la suposición de normalidad y utilizando $n \geq 50$, como sigue:

$$\left(z_r - z_{(1-\alpha/2)} \frac{1}{\sqrt{n-3}} < z_\rho < z_r + z_{(1-\alpha/2)} \frac{1}{\sqrt{n-3}} \right)$$

$$z_{\rho_1} < z_\rho < z_{\rho_2}$$

Como el valor obtenido está estandarizado, hay que reconvertir los resultados de cada extremo del intervalo usando la fórmula:

$$\frac{e^{2z_{\rho_1}} - 1}{e^{2z_{\rho_1}} + 1} < \rho < \frac{e^{2z_{\rho_2}} - 1}{e^{2z_{\rho_2}} + 1}$$

5.3.4 Aplicación de la Inferencia en el Análisis de Correlación Lineal

EJEMPLO 5.3.1. Una muestra aleatoria de 15 niños aparentemente sanos con edades entre 6 meses y 15 años produjo los siguientes datos respecto a la edad y el volumen del hígado por unidad de peso corporal, en ml/kg

Edad	0.5	0.7	2.5	4.1	5.9	6.1	7.0	8.2
Volumen	41	55	41	39	50	32	41	42
Edad	10	10.1	10.9	11.5	12.1	14.1	15.0	
Volumen	26	35	25	31	31	29	23	

Con base en esta información:

- Elabore un diagrama de dispersión
- Calcule el coeficiente de correlación de la muestra

- c) Pruebe que existe una correlación lineal entre las variables de la población con una significación del 5%
- d) Determine el valor verdadero del coeficiente de correlación, con una confianza de 95%

Solución:

a) Diagrama de dispersión

Se realiza la gráfica para observar si los puntos se dispersan en forma lineal en el plano y definir la tendencia. En este caso, se puede ver que los puntos se dispersan de forma lineal e inversamente proporcional, por lo que el coeficiente de correlación o grado de asociación entre las variables es negativo.

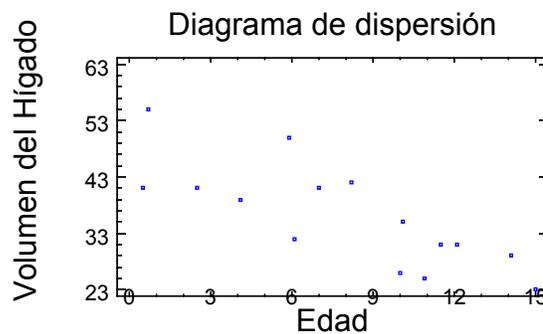
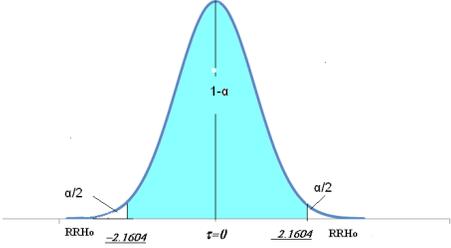


Figura 5.9 Diagrama de dispersión del ejemplo 5.3.1.

- b) Para el cálculo del coeficiente de correlación, con base en las medidas descriptivas de ambas variables, se sustituye la fórmula que lo define, como sigue:

Datos	Sustitución
$n = 15;$ $\bar{x} = 7.9133$ $s_x = 4.59843;$ $\bar{y} = 36.066$ $s_y = 9.1921;$ $xy = 3814.5$	$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$ $r = \frac{3814.5 - 15(7.9933)36.066}{14(4.5984)9.192} = -0.7885$

- c) Para probar la existencia de una asociación entre las variables de la población se siguen todos los pasos para un contraste de hipótesis como se muestra:

<p>1. Planteamiento de Hipótesis $H_0: \rho = 0$ $H_a: \rho \neq 0$</p>	<p>2. Nivel de significación $\alpha = 0.05$</p>
<p>3. Estimador y Distribución utilizada</p> $r \sim t_{(1-\alpha/2, n-2)} = t_{(0.975, 13)} = 2.1604$	<p>4. Regla de decisión</p> 
<p>5. Estadístico de contraste</p> $t = r \sqrt{\frac{n-2}{1-r^2}}$	<p>6. Cálculos</p> $t_{Calc} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.7885\sqrt{13}}{\sqrt{1-0.621738}} = -4.6225$
<p>7. Toma de decisión</p> <p>Como $-4.6225 < -2.1604$ se rechaza la hipótesis nula</p>	<p>8. Conclusión</p> <p>De acuerdo con la decisión, se puede afirmar, con una significación del 5% que existe un grado de asociación entre la edad y el volumen del hígado</p>

d) Intervalo de confianza para el coeficiente de correlación

Para obtener el valor verdadero del coeficiente de correlación poblacional es necesario calcularlo por intervalo, y el resultado se considerará válido siempre y cuando el tamaño de la muestra sea de 50 o más registros, sin embargo, con el objeto de ejemplificar los cálculos se realizaron las operaciones aun cuando el tamaño de la muestra solo es de 15 elementos, como sigue:

$$z_r - z_{(1-\alpha/2)} \frac{1}{\sqrt{n-3}} < z_\rho < z_r + z_{(1-\alpha/2)} \frac{1}{\sqrt{n-3}}$$

Donde: $z_r = \frac{1}{2} \ln \left| \frac{1+r}{1-r} \right|$

$$z_r = \frac{1}{2} \ln \left| \frac{1+(-0.7885)}{1-(-0.7885)} \right| = -1.067464$$

$$-1.067464 - 1.96 \frac{1}{\sqrt{12}} < z_\rho < -1.067464 + 1.96 \frac{1}{\sqrt{12}}$$

$$-1.633267 < z_\rho < -0.50166$$

Esta fórmula permite calcular por intervalo el coeficiente estandarizado z_ρ que deberá transformarse a ρ mediante la fórmula:

$$\frac{e^{2z_{\rho 1}} - 1}{e^{2z_{\rho 1}} + 1} < \rho < \frac{e^{2z_{\rho 2}} - 1}{e^{2z_{\rho 2}} + 1}$$

$$\frac{e^{2(-1.633267)} - 1}{e^{2(-1.633267)} + 1} < \rho < \frac{e^{2(-0.501660)} - 1}{e^{2(-1.633267)} + 1}$$

$$\frac{-0.9618616}{1.0381384} < \rho < \frac{-0.63334}{1.36666}$$

$$-0.926525 < \rho < -0.46342$$

Interpretación: De cada cien intervalos calculados en las mismas condiciones, en 95 de ellos el valor real de la correlación lineal se encontrará entre sus límites.

EJEMPLO 5.3.2. En un artículo especializado en agricultura se publicaron datos sobre el contenido de fibra de los espárragos y su resistencia, como determinantes de la calidad de este vegetal. Específicamente se midió la fuerza cortante y_1 en kilogramos y el porcentaje del peso seco de fibra, y_2 , los datos registrados aparecen en la tabla:

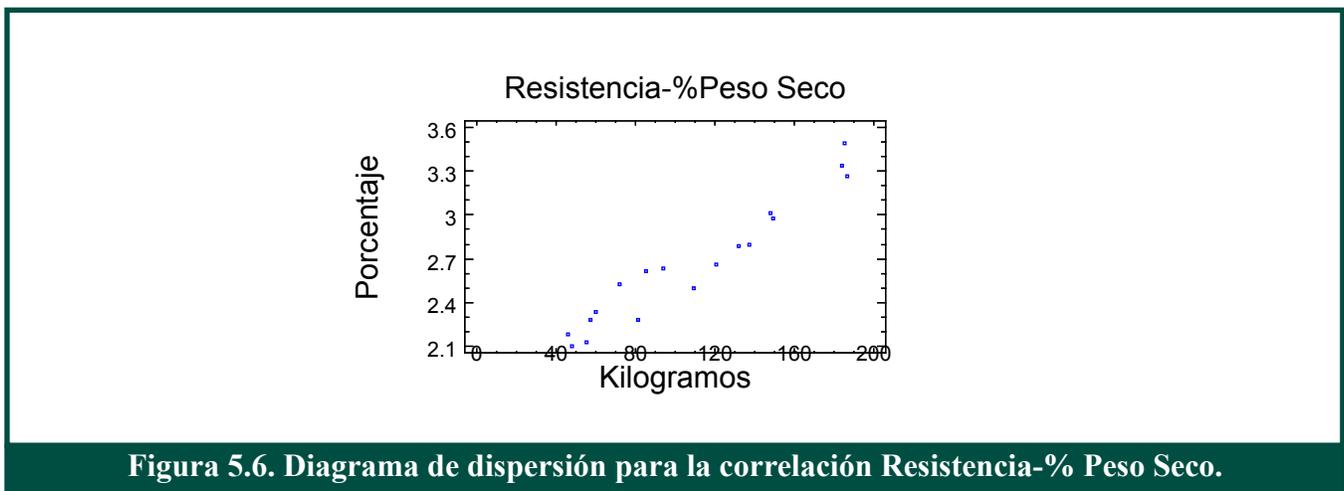
Var 1 y_1	46	48	55	57	60	72	81	85	94
Var 2 y_2	2.18	2.10	2.13	2.28	2.34	2.53	2.28	2.62	2.63
Var 1 y_1	109	121	132	137	148	149	184	185	187
Var 2 y_2	2.50	2.66	2.79	2.80	3.01	2.98	3.34	3.49	3.26

Con base en esta información, resuelva los incisos siguientes:

Solución:

- a) Trace un diagrama de dispersión de los datos

Para resolver este inciso, basta con marcar los pares ordenados (y_1, y_2) en un plano XY para obtener la tendencia con que se dispersan los puntos formados.



Al observar el diagrama nos podemos dar cuenta que es razonable suponer una asociación lineal entre la resistencia y el porcentaje en peso seco de fibra, en el vegetal.

- b) Calcule el coeficiente de correlación muestral r .

Para hacer el cálculo, es necesario sustituir los elementos establecidos en la definición de este coeficiente. Por lo mismo, necesitamos calcular los valores medios de ambas variables, las varianzas respectivas y la suma de los productos $y_1 y_2$

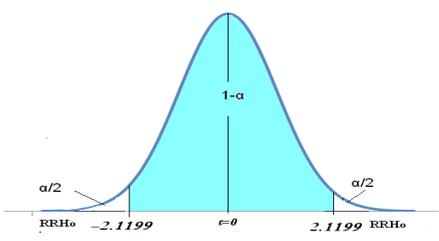
Si tenemos una calculadora con programa de regresión podemos introducir las parejas de datos y obtener las medidas necesarias.

Datos	Sustitución
$n = 18;$ $\bar{y}_1 = 108.333$ $\bar{y}_2 = 2.6622;$ $s_{y_1} = 48.9417$ $s_{y_2} = 0.422438;$ $\sum_{i=1}^n y_{1i}y_{2i} = 5530.92$	$r = \frac{\sum_{i=1}^n y_{1i}y_{2i} - n\bar{y}_1\bar{y}_2}{(n-1)s_{y_1}s_{y_2}}$ $r = \frac{5530.92 - 18(108.333)2.6622}{17(48.9417)0.422438} = 0.966$

Como podemos ver, el valor del coeficiente de correlación muestral es alto, muy cercano a 1, lo que significa que el grado de asociación entre las variables resistencia y porcentaje en peso seco de fibra es muy fuerte.

- c) ¿Podría asegurarse al 5% de significación que existe una verdadera asociación lineal entre la resistencia y el porcentaje en peso seco de la fibra, en los espárragos?

Aplicando la prueba con valor supuesto para ρ igual a cero, esto es, suponiendo que no existe grado de asociación entre las variables, tenemos:

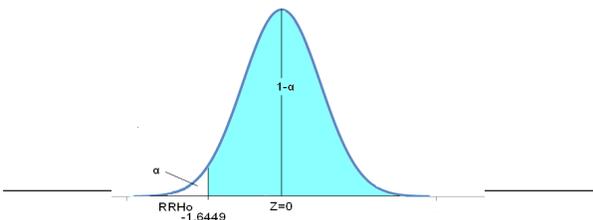
1) Planteamiento de las hipótesis	2) Nivel de Significación
Bilateral $H_0: \rho = 0$ $H_a: \rho \neq 0$	$\alpha = 0.05;$ $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$ $1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$
3) Distribución Utilizada y valor crítico	4) Regla de Decisión
$r \sim t_{(1-\alpha/2, n-2)}$ $t_{(1-\alpha/2, n-2)} = t_{(0.975, 16)} = 2.1199$	
5) Estadístico de contraste	6) Cálculo del Estadístico
$t = r \sqrt{\frac{n-2}{1-r^2}}, \quad gl = n-2$	$t = 0.966 \sqrt{\frac{18-2}{1-0.933}} = 14.945$

7) Decisión	8) Conclusión
Como el estadístico calculado, $t_c = 14.945$ es mayor que el valor crítico 2.1199 obtenido de tablas, se rechaza H_0 .	Con una confianza del 95% podemos afirmar que existe una asociación real entre la resistencia y el porcentaje de peso seco de fibra en los espárragos.

Nota: Aun cuando nuestro ejemplo no cuenta con una muestra, con un mínimo de 50 datos, aplicaremos el proceso para ejemplificar la secuencia de cálculo.

- d) Pruebe si es factible que el coeficiente de correlación poblacional tenga un valor de al menos 0.98, con una significación del 5%.

Como nos piden contrastar un valor para ρ de al menos 0.98, deberemos plantear un análisis unilateral inferior como se muestra en el cuadro siguiente:

1) Planteamiento de las hipótesis	2) Nivel de Significación
Unilateral Inferior $H_0: \rho \geq 0$ $H_a: \rho < 0$	$\alpha = 0.05; \quad 1 - \alpha = 0.95$
3) Distribución utilizada y valor crítico	4) Regla de Decisión
$z_r \sim z$ $z_{(1-\alpha)} = -z_{(0.95)} = -1.6449$	
5) Estadístico de contraste	6) Cálculo del Estadístico
$z^* = \frac{z_r - z_{\rho_0}}{\sigma_{z_r}}$	$z_r = \frac{1}{2} \ln \left \frac{1+r}{1-r} \right = \frac{1}{2} \ln \left \frac{1.966}{0.034} \right = 2.0287$ $z_{\rho_0} = \frac{1}{2} \ln \left \frac{1+\rho_0}{1-\rho_0} \right = \frac{1}{2} \ln \left \frac{1.98}{0.02} \right = 2.29756$ $\rightarrow \sigma_{z_r} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{15}} = 0.2582$ $z^* = \frac{z_r - z_{\rho_0}}{\sigma_{z_r}} = \frac{2.0287 - 2.29756}{0.2582} = -1.04129$

7) Decisión	8) Conclusión
Como el valor calculado de z^* es mayor que -1.6449 , no toca la región de rechazo que se encuentra en la cola inferior; por lo tanto, no se rechaza H_0 .	Por lo anterior concluimos, con una confianza del 95%, que el coeficiente de correlación poblacional no es significativamente menor que 0.98.

e) Estime, con una confianza de 99%, el valor verdadero para el coeficiente de correlación poblacional ρ .

Aplicando la fórmula que define al intervalo tenemos:

$$z_r - z_{(1-\alpha/2)} \frac{1}{\sqrt{n-3}} < z_\rho < z_r + z_{(1-\alpha/2)} \frac{1}{\sqrt{n-3}}$$

$$2.0287 - 2.5758 \frac{1}{\sqrt{15}} < z_\rho < 2.0287 + 2.5758 \frac{1}{\sqrt{15}}$$

$$2.0287 - 0.66507 < z_\rho < 2.0287 + 0.66507$$

$$1.36363 < z_\rho < 2.69397$$

$$\frac{e^{2z_{p1}} - 1}{e^{2z_{p1}} + 1} < \rho < \frac{e^{2z_{p2}} - 1}{e^{2z_{p2}} + 1}$$

$$\frac{e^{2z_{p1}} - 1}{e^{2z_{p1}} + 1} = \frac{14.290993241}{16.29093241} = 0.8772$$

$$\frac{e^{2z_{p2}} - 1}{e^{2z_{p2}} + 1} = \frac{217.7523}{219.7523} = 0.9909$$

$$0.8772 < \rho < 0.9909, \text{ con } 99\% \text{ de confianza}$$

Interpretación: El grado de asociación entre las variables resistencia y porcentaje de peso seco de fibra, se encuentra entre 98.5 y 99.9 %, con 99% de confianza.

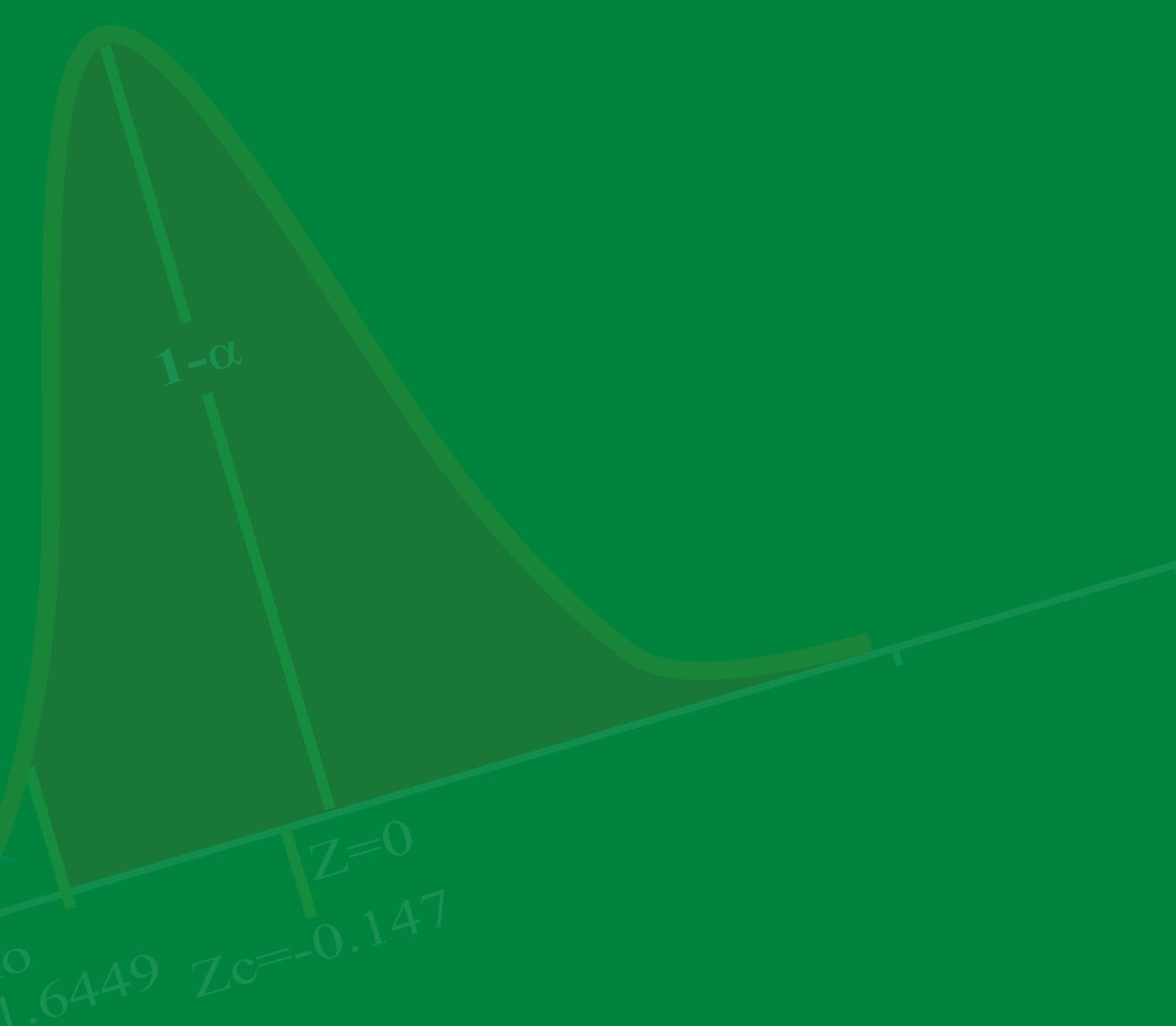
Debe recordarse que los resultados de los últimos 2 incisos se considerarían válidos si la muestra estuviera conformada de 50 o más elementos.

Referencias

- **Lind, D., Marchall, W., Wathen, S.**, 2012. Estadística aplicada a los negocios y economía, Mc Graw-Hill, 15ª ed, .México
- **Devore, J. L.**, 2011. Probabilidad y Estadística para Ingeniería y Ciencias, Thomson Learning, México.
- **Triola Mario**, 2013. Estadística, Ed. Pearson, 11ª ed. España
- **Walpole, R.E., Myers, R. H., Myers S. L. ,Ye Keying**, 2007. Probabilidad y estadística para ingeniería y ciencias, Prentice Hall, 8ª ed. México
- **Ross, S. M.**, 2007 Introducción a la Estadística, Ed. Reverté. S.A., 7ª ed, México
- **Moore, D. S.**, 2005. Estadística Aplicada Básica, Antoni Bosch Editor S.A., Barcelona, España
- **Guerra Dávila, T; Marques Dos Santos, M J; López Reynoso, J.M.**; 2009, Cuaderno de Problemas Resueltos y Propuestos de Probabilidad y Estadística, Universidad Nacional Autónoma de México, FES Zaragoza, 2ªed, México.
- **Marques Dos Santos, M.J.** Probabilidad y Estadística para Ciencias Químico Biológicas; FES Zaragoza, UNAM; 2ª edición, 2004, México.
- **Wackerly, D. D., Mendenhall III, W.; Scheaffer, R. L.**; 2002, Estadística Matemática con Aplicaciones; Ed. Thompson, 6ª edición, México.
- **Salgado Ugarte , I. H.**, 2002, Suavización No Paramétrica para Análisis de Datos, UNAM, FES Zaragoza, México.
- **Hines, W.W y Montgomery D.C.**, 1998, Probabilidad y Estadística para Ingeniería, Ed. CECSA, 3ª ed., México
- **Parzen, E.** 1962 Modern Probability Theory and Its Applications, Ed, John Wiley. , New York

Bibliografía Complementaria:

- **Evans, M. J., Rosenthal, J. S.**, 2007. Probabilidad y Estadística, Ed Reverté. S.A., México.
- **Navidi, W.**, 2006. Estadística para Ingenieros y Científicos, Mc Graw-Hill–Interamericana de México.



Facultad de Estudios Superiores Zaragoza
Campus I. Av. Guelatao No. 66, Col. Ejército de Oriente,
Campus II. Batalla del 5 de Mayo s/n esq. Fuerte de Loreto, Col. Ejército de Oriente,
Iztapalapa, C.P. 09230, México, D.F.

<http://www.zaragoza.unam.mx>

