

Carlomagno Araya Alpizar

ESTADÍSTICA



para
laboratorista químico

Editorial de la Universidad de Costa Rica



542.1
A663e

Araya Alpízar, Carlo Magno, 1961-
Estadística para laboratorista químico / Carlomagno
Araya A. - 1. ed. - San José, C.R. : Editorial de la
Universidad de Costa Rica, 2004.
1 disco óptico de computador : col. ; 4³/₄ plg.

ISBN: 9977-67-849-9

1. QUÍMICA - PROBLEMAS, EJERCICIOS -
DISCOS COMPACTOS. 2. ESTADÍSTICA
MATEMÁTICA. I. Título.

CIP/1298
CC/SIBDIUCR

Edición aprobada por la Comisión Editorial de la Universidad de Costa Rica.

Primera edición: 2004.

Diseño de portada: *Elisa Giacomini V.*

© Editorial de la Universidad de Costa Rica, Ciudad Universitaria "Rodrigo Facio".
Apdo. 75-2060. Fax: 207-5257, e-mail: editucr@cariari.ucr.ac.cr • San José, Costa Rica.

Prohibida la reproducción total o parcial. Todos los derechos reservados. Hecho el depósito de ley.

CONTENIDO

PRESENTACION

CAPÍTULO 1. INTRODUCCIÓN

1.1	Historia de la estadística	2
1.2	Poblaciones y muestras	4
1.3	Estadística descriptiva e inferencial	5
1.4	Tipos de variables aleatorias	6
1.5	Error experimental	7
1.6	Diseño estadístico de experimentos	8
1.7	Exactitud y precisión	9

CAPÍTULO 2. ESTADÍSTICA DESCRIPTIVA

2.1	Medidas de posición	
	2.1.1 La media aritmética	11
	2.1.2 La media geométrica	13
	2.1.3 La mediana	14
	2.1.4 La moda	15
	2.1.5 Proporciones	16
2.2	Medidas de variabilidad	
	2.2.1 El recorrido	18
	2.2.2 La variancia	19
	2.2.3 La desviación estándar	20
	2.2.4 El coeficiente de variación	21
2.3	Distribución de frecuencias	
	2.3.1 Selección del número de clases	23
	2.3.2 Cálculo del intervalo de clase	24
	2.3.3 Definición de los límites de clase	25
	2.3.4 Conteo del número de observaciones por clase	26
	2.3.5 Representación gráfica de distribuciones de frecuencias	28
2.4	Problemas	32

CAPÍTULO 3. PROBABILIDADES

3.1	Introducción a las probabilidades	45
3.2	Definición axiomática de probabilidad	47
3.3	Teoremas fundamentales de cálculos de probabilidades	
3.3.1	Teorema de la adición	49
3.3.2	Teorema de la multiplicación	50
3.4	Distribuciones de probabilidad	51
3.5	Distribuciones discretas	52
3.5.1	Distribución binomial	52
3.5.2	Distribución Poisson	55
3.6	Distribuciones continuas	
3.6.1	Distribución normal	57
3.6.2	Uso de la tabla normal estándar	59
3.7	Aplicaciones de la distribución normal	62
3.8	Aproximación de la distribución Binomial por la Normal	63
3.9	Distribución T de Student	65
3.10	Problemas	67

CAPÍTULO 4. ESTIMACIÓN

4.1	Propiedades deseables de un estimador	74
4.2	Distribuciones muestrales	76
4.3	Tipos de estimación: puntual o por intervalo	78
4.4	El concepto de estimación aplicado a la química	79
4.5	Explicación sobre intervalos de confianza	81
4.6	Intervalos de confianza para la distribución normal	81
4.7	Intervalo para el promedio si se conoce la varianza	82
4.7	Intervalo para el promedio si se conoce la varianza	82
4.8	Intervalos de confianza para el promedio con varianza desconocida	84
4.9	Intervalo de confianza para la varianza	85
4.10	Estimación de tamaño muestral	87

4.11	Muestras para la estimación de promedios	88
4.12	Intervalos de confianza para variables dicotómicas	90
4.13	Intervalos de confianza para una proporción	90
4.14	Elección del tamaño muestral para una proporción	92
4.15	Problemas	94

CAPÍTULO 5. PRUEBA DE HIPÓTESIS

5.1	Método de pruebas de hipótesis	
5.1.1	Formulación de las hipótesis	105
5.1.2	Errores posibles al tomar una decisión errónea	106
5.1.3	Determinación del valor observado del estadístico de prueba	108
5.1.4	Determinación del valor crítico del estadístico de prueba	108
5.1.5	Comparación del valor observado del estadístico de prueba con el valor crítico	109
5.2	Pruebas de hipótesis paramétricas en una población normal	
5.2.1	Prueba de hipótesis para un promedio con varianza conocida	109
5.2.2	Prueba de hipótesis para un promedio con varianza desconocida	112
5.3	Contrastes de dos distribuciones normales independientes	
5.3.1	Contrastes de promedios con varianzas conocidas	115
5.3.2	Contraste de promedios con varianzas desconocidas	116
5.4	Contrastes para la diferencia de medias pareadas	118
5.5	Contrastes de una proporción	122
5.6	Contrastes sobre la diferencia de proporciones	124
5.7	Problemas	127

CAPÍTULO 6. REGRESIÓN LINEAL SIMPLE Y CORRELACIÓN

6.1	Regresión	141
6.2	Regresión lineal	143
6.3	Ajuste del modelo de regresión lineal	144
6.4	Uso de la ecuación de regresión	147
6.5	Error estándar de la estimación	148

6.6	Prueba de hipótesis sobre el coeficiente de regresión	149
6.7	Bondad de un ajuste	150
6.8	Intervalo de predicción	152
6.9	Análisis de regresión: comentarios finales	152
6.10	Análisis de correlación lineal	153
6.11	Análisis colorimétrico	155
6.12	Problemas	157
APÉNDICE A. FÓRMULAS ESTADÍSTICAS		173
APÉNDICE B. TABLAS ESTADÍSTICAS		179
APÉNDICE C. GLOSARIO		191
RESPUESTAS A LOS EJERCICIOS IMPARES		203
BIBLIOGRAFÍA		218

Presentación

Siempre que se planea un nuevo libro de texto, se deben resolver los aspectos de cómo diferirá de los ya disponibles y qué contribuciones aportará al campo en estudio. Este libro fue escrito para usarse en un curso introductorio a la probabilidad y estadística, dando énfasis primordial a conceptos y aplicaciones de los métodos estadísticos básicos en el área de la química, con el fin de proporcionar al estudiante el enlace necesario entre los conceptos teóricos vistos en clase y las aplicaciones en la vida real. También pretende que los profesionales en química utilicen la ventaja que ofrece la Estadística, la cual les puede facilitar el diseño de experimentos, la recolección de los datos, mayor exactitud y precisión en los resultados.

Existe la creencia de que la estadística es muy complicada o consume demasiado tiempo cuando se la utiliza. Nuestro propósito es aportar técnicas que sean simples y rápidas, además de adecuadas para que el químico analice sus propios datos.

En esta primera edición, se incluyen los temas fundamentales de la estadística descriptiva e inferencial. Además, como una ayuda tanto para el instructor como para el estudiante, aparecen al final de la obra las soluciones a los problemas de número impar. El desarrollo de los temas está acompañado de ejemplos, que ilustran al lector sobre el alcance de las definiciones y sobre el desarrollo de las técnicas de cálculo. Asimismo, cada capítulo finaliza con una buena colección de problemas propuestos, muchos de ellos originales, en los que se combinan ejercicios prácticos con resultados complementarios de los obtenidos en el texto.

El capítulo primero ofrece una introducción a la estadística por medio de los conceptos básicos. El capítulo segundo introduce al lector en la estadística descriptiva. El tercer capítulo se centran en el desarrollo axiomático del concepto de probabilidad y en el estudio de sus propiedades más interesantes desde el punto de vista de los cálculos. En el primero de ellos se estudian las operaciones básicas con los sucesos y sus probabilidades, y a continuación las distribuciones de probabilidades como la normal.

La presentación es rigurosa y formal, lo que proporciona al estudio una cierta aridez, aunque los resultados vienen acompañados de ejemplos que ilustran su utilización práctica. En los capítulos cuarto al sexto se estudian los métodos básicos de la estadística inferencial; estimación, prueba de hipótesis, regresión y correlación. Cada capítulo finaliza con una amplia colección de ejercicios, a través de los cuales el lector podrá medir el grado de comprensión de los contenidos de cada tema.

Este texto ha sido utilizado, con carácter experimental, en el curso de Estadística para Laboratoristas Químicos del Recinto de Tacaes de la Sede de Occidente de la Universidad de Costa Rica. El libro está pensado no sólo para el curso anteriormente mencionado, sino que puede resultar también interesante para alumnos de disciplinas, como Ingeniería Industrial y Matemáticas, por ejemplo.

Se agregaron en los anexos fórmulas y tablas estadísticas de las principales distribuciones, tablas que permiten el cálculo aproximado y rápido de probabilidades. y rápido de probabilidades.

Para el estudio completo del libro se precisan unos conocimientos moderados de Cálculo diferencial e integral, algunas nociones del cálculo de límites y de las técnicas de sucesión de series numéricas sencillas, así como ciertos conceptos y desarrollos de Álgebra lineal. No obstante, las definiciones y resultados básicos (y muchos no tan básicos) van acompañados de observaciones que permiten que el lector poco avezado en las lides matemáticas pueda valorar su contenido, haciéndose, al menos, una idea intuitiva de ellos.

Esperamos que este texto, haga el estudio de la estadística más satisfactorio, comprensible y significativo para los estudiantes que lo utilicen.

Carlo Magno Araya Alpízar

Capítulo 1

Introducción

1.2 Poblaciones y muestras

1.3 Estadística descriptiva e inferencial

1.4 Tipos de variables aleatorias

1.5 Error experimental

1.6 Diseño estadístico de experimentos

1.7 Exactitud y precisión

Introducción

El análisis de los datos debe presentarse de manera que sea fácilmente comprensible. Con demasiada frecuencia, el lenguaje puramente estadístico significa poco o nada para los químicos; expresiones tales como “existe una diferencia significativa” puede ser verdadera, pero es demasiado general para ser útil.

En este capítulo se revisan algunos de los conceptos básicos que servirán de fundamento para gran parte del material del resto del texto.

En la sección 1.1 se presenta la historia y la definición de estadística. Describe la diferencia entre población y muestra en la sección 1.2. A continuación, en la sección 1.3 se estudia qué se entiende por estadística descriptiva e inferencia estadística. Seguidamente, se estudian los tipos de variables aleatorias, el error experimental, las medidas invalidadas y los diseños de experimentos. Por último, en la sección 1.7, se analiza la diferencia entre exactitud y precisión de las mediciones.

1.1 Historia de la estadística

Desde los comienzos de la civilización han existido formas sencillas de estadísticas, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o ciertas cosas.

Hacia el año 3000 A. C. los babilonios usaban ya pequeñas tablillas de arcilla para recopilar datos en tablas sobre la producción agrícola y de los géneros vendidos o cambiados mediante trueque. Los egipcios analizaban los datos de la población y la renta del país mucho antes de construir las pirámides en el siglo XXXI A. C.

Los libros bíblicos de Numéricos incluyen, en algunas partes, trabajos de estadística. El primero contiene dos censos de la población de Israel y el segundo describe el bienestar material de las diversas tribus judías. En China existían registros numéricos con anterioridad al año 2000 A. C. Los griegos clásicos

realizaban censos cuya información se utilizaba hacia el año 594 A. C. para cobrar impuestos.

El Imperio romano fue el primer gobierno que recopiló una gran cantidad de datos sobre la población, superficie y renta de todos los territorios bajo su control. Durante la Edad Media sólo se realizaron algunos censos exhaustivos en Europa. Los reyes carolingios Pipino el Breve y Carlomagno ordenaron hacer estudios minuciosos de las propiedades de la Iglesia en los años 758 y 762, respectivamente.

Después de la conquista normanda de Inglaterra en 1066, el rey Guillermo I de Inglaterra encargó un censo. La información obtenida con este censo, llevado a cabo en 1086, se recoge en el Domesdat Book. El registro de nacimientos y defunciones comenzó en Inglaterra a principios del siglo XVI, y en 1662 apareció el primer estudio estadístico notable de población, titulado "Observations on the London Bills of Mortality" (Comentarios sobre las partidas de defunción en Londres).

Un estudio similar sobre la tasa de mortalidad en la ciudad de Breslau, en Alemania, realizado en 1691, fue utilizado por el astrónomo inglés *Edmund Halley* como base para la primera tabla de mortalidad. En el siglo XIX, con la generalización del *método científico* para estudiar todos los fenómenos de las ciencias naturales y sociales, los investigadores aceptaron la necesidad de reducir la información a valores numéricos para evitar la ambigüedad de las descripciones verbales.

En nuestros días, la estadística se ha convertido en un método efectivo para describir con exactitud los valores de los datos económicos, políticos, sociales, psicológicos, biológicos y físicos, y sirve como herramienta para relacionar y analizar dichos datos.

El trabajo del experto estadístico no consiste ya sólo en reunir y tabular los datos, sino sobre todo el proceso de interpretación de esa información. El desarrollo de la teoría de la probabilidad ha aumentado el alcance de las aplicaciones de la estadística. Muchos conjuntos de datos se pueden aproximar,

con gran exactitud, utilizando determinadas distribuciones probabilísticas; los resultados de éstas se pueden utilizar para analizar datos estadísticas.

La probabilidad es útil para comprobar la fiabilidad de las inferencias estadísticas y para predecir el tipo y la cantidad de datos necesarios en un determinado estudio estadístico.

En Química, incluye tareas tan diversas como calcular el promedio aritmético de la cantidad de sólidos suspendidos en muestras de agua, determinar cuántos compuestos de tres elementos pueden elaborarse con cinco sustancias, comparar dos métodos de determinación del porcentaje de calcio, el crecimiento del número de bacterias según días de inoculación.

Por tanto, la ciencia de la estadística se puede considerar como la aplicación del método científico en el análisis de datos numéricos con el fin de tomar decisiones racionales.

1.2 Poblaciones y muestras

A fin de entender cómo se puede aplicar los métodos estadísticos, se debe distinguir entre población y muestra. Una **población** o universo es cualquier colección finita o infinita de individuos o elementos, para cada uno de los cuales se tiene que contar o medir una o varias características. Los **elementos** son los objetos que poseen la información que busca el investigador y acerca del cual deben hacerse las inferencias. Entre posibles elementos se tiene: una sustancia, un compuesto, un mineral, etc.

Aunque existe la libertad de llamar población a un grupo cualquiera de elementos definidos en el tiempo y el espacio, en la práctica depende del contexto en el cual se observarán los elementos. Suponga, por ejemplo, el número de total de compuestos que pueden elaborarse con 5 sustancias, puede considerarse como una población.

Una **muestra** es un subconjunto de elementos de la población que puede servir de base para generalizaciones válidas. En trabajos químicos, se toman

muestras de un material, se ensayan y se hacen deducciones para la totalidad de dicho material, a partir de los resultados obtenidos. La muestra no es exactamente la población, sino que se supone que la representa, y la validez de cualquier conclusión obtenida de ella depende de la verdadera representatividad que tenga.

La muestra debería ser seleccionada en forma aleatoria, es decir, cada elemento tiene una probabilidad conocida y no nula de selección. No siempre es posible obtener una muestra aleatoria. Esto es especialmente cierto cuando el estadístico puede confiar en su propio “juicio” o “conveniencia” al seleccionar los elementos de la muestra. Con la aleatoriedad se persigue que la muestra sea “representativa” de la población, para que concentre todas las características y particularidades de interés. Sin embargo, la aleatoriedad no garantiza necesariamente la representatividad.

Los factores que determinan la selección de muestras son: costo, tiempo y la imposibilidad práctica (o destrucción de la unidad de estudio). En tanto, el tamaño de la muestra queda determinado por la variabilidad de la población, el nivel de confianza y el error máximo permisible. Cuando los elementos de una población tienen cierto grado de heterogeneidad entre ellos, el tamaño de la muestra tiende a ser grande. Cabe destacar, que es completamente erróneo asegurar que la muestra depende del tamaño de la población.

Cuando se hace una investigación con todos los elementos de la población, se llama **censo o enumeración total**. En tanto, cuando se elabora con base en una muestra, se le denomina **estudio por muestreo**. En la mayoría de los casos los químicos no están interesados en los datos de enumeración total (o censo) sino en datos muestrales.

1.3 Estadística descriptiva e Inferencial

La **estadística descriptiva** se puede definir como los métodos estadísticos que pretenden describir las características más importantes de un conjunto de datos, sea que provenga de una muestra o de una población. Utiliza técnicas

estadísticas, como la representación gráfica, cuadros estadísticos, medidas de posición y de variabilidad.

Uno de los propósitos fundamentales de los métodos estadísticos es utilizar estadísticos muestrales para estimar los parámetros de la población. A este proceso de utilizar los estadísticos muestrales para llegar a conclusiones acerca de los verdaderos parámetros de la población, se le llama **inferencia estadística**.

La estadística inferencial generaliza los resultados observados en una muestra a toda la población bajo estudio, por medio del planteamiento y pruebas de hipótesis y cálculo de intervalos de confianza; se aplica a problemas como estimar, mediante pruebas, el rendimiento promedio de un proceso químico, verificar las especificaciones de producción a partir de mediciones efectuadas sobre muestras o predecir los residuos de cloro en una piscina basándose en una muestra de datos tomados en ciertos periodos de tiempo.

1.4 Tipos de variables aleatorias

Cada unidad que compone la población está caracterizada por uno o varios valores. Estas características, con los valores que toma, se denominan variables. Se distingue entre variables **cuantitativas** y **cualitativas**. Las primeras representan características con valor aritmético, mientras que las segundas hacen referencia a cualidades que sólo pueden ser contadas u ordenadas. Por ejemplo: color, sabor, textura, punto de fusión, solubilidad, diámetro y peso.

Si **S** es un espacio de muestra con una medida de probabilidad y **x** es una función con valor real definida con respecto a los elementos de **S**, entonces **x** se denomina **variable aleatoria**.

La diferencia entre ellas es que las variables aleatorias cuantitativas toman valores **numéricos**, mientras que las variables aleatorias cualitativas asumen valores categóricos, por ejemplo, el color de una sustancia.

Por otra parte, las variables aleatorias cuantitativas pueden dividirse en discretas y continuas. Las **variables aleatorias discretas** asumen valores aislados y surgen de un proceso de conteo. “El número de bacteria de una

muestra de agua” es un ejemplo de una variable cuantitativa discreta ya que toma un número finito de valores que se pueden contar

Las **variables aleatorias continuas** toman valores sobre un intervalo o una colección de intervalos, que surgen de un proceso de medición. La información importante para un químico no viene, generalmente del proceso de contar, sino de la medida de volúmenes y de los instrumentos de medidas. Todas las operaciones citadas entrañan medidas y éstas tienen un cierto grado de incertidumbre. La “conductividad térmica del cloruro de metilo” es un ejemplo de una variable aleatoria continua, ya que la respuesta puede tomar cualquier valor según la precisión del instrumento para la medición.

1.5 Error experimental

En Química Analítica se clasifican los errores en determinados e indeterminados. Los errores determinados se definen como aquellos que pueden ser evitados una vez que son conocidos. Este tipo de error es originado por factores como los siguientes:

1. Inadecuada calibración de los aparatos o inadecuada estandarización de los reactivos.
2. Error personal, tal como la dificultad de un analista para juzgar un cambio de color.
3. Avería del aparato.
4. Error de método.

Los errores determinados introducen un sesgo en las medidas. Por ejemplo, si el analista no está centrado en la escala del instrumento, su lectura será más alta o más baja de la debida a causa de paralaje.

Los errores indeterminados, por el contrario, no pueden ser eliminados, sino que existen por la propia naturaleza del dato medido. Por ejemplo, los ligeros errores obtenidos en la interpolación son indeterminados y el analista no conoce

su magnitud, ni si son positivos o negativos. Son llamados errores experimentales o estadísticamente errores de muestreo.

En lo relativo a la precisión, los resultados de estudios por muestreo siempre están sujetos a cierta incertidumbre debido a que sólo una parte de la población ha sido investigada y debido a los errores de medición. Esta incertidumbre puede ser reducida, tomando muestras más grandes y usando métodos de medición adecuados.

1.6 Diseño estadístico de experimentos

El hombre ha ido diseñando experimentos cada vez que ha formulado preguntas acerca del mundo que lo rodea, por lo cual el concepto “diseño de experimentos” no es nada nuevo. Ha desarrollado, también un sistema para la resolución de problemas, que ha llegado a ser conocido como “método científico”. Esencialmente, este plan consta de los siguientes elementos:

1. Situación y definición del problema
2. Formulación de una hipótesis para explicarla
3. Obtención de datos.
4. Confrontación de la hipótesis con los datos.
5. Aceptación o rechazo de la hipótesis según esté o no de acuerdo con los datos.

La confrontación de la hipótesis con los datos y la aceptación o rechazo de ella son materias de juicio crítico, y cualquier camino es bueno para ello, si el juicio puede ser hecho en forma más objetiva que subjetiva.

Un experimento diseñado estadísticamente proporciona una estimación del error que puede servir como norma por la que pueden ser medidos los resultados del experimento.

1.7 Exactitud y precisión

La diferencia entre el resultado obtenido con base en el análisis de muestras y el “verdadero”, se toma como medida de la **exactitud** de la técnica que se estudia. La **precisión** representa la probabilidad de que los resultados sucesivos caigan dentro un intervalo estrecho de valores centrado alrededor del valor medio. Este valor medio puede diferir apreciablemente del valor “verdadero”, esto es, puede existir un sesgo, positivo o negativo, en el método de estudio.

Un método satisfactorio de análisis debe poseer a la vez una exactitud y una precisión razonables. Es esencial que los químicos distinguan cuidadosamente estos dos conceptos. Todos los instrumentos de alta tecnología que se emplean actualmente están contruidos para funcionar con un alto grado de precisión. Así, si se somete repetidamente la misma muestra a un instrumento determinado, las lecturas tienden a ser iguales cada vez.

Estadísticamente, la precisión de un estimador* es un concepto que expresa la concentración de las estimaciones con respecto a su valor medio. Se mide por la raíz cuadrada de la varianza del estimador. Un estimador será más preciso cuanto menor sea su varianza.

* En el capítulo 3, se estudian las propiedades los estimadores.

Capítulo 2

Estadística Descriptiva

- 2.1 Medidas de posición
- 2.2 Medidas de variabilidad
- 2.3 Distribución de frecuencias
- 2.4 Problemas

Estadística Descriptiva

La finalidad de este capítulo, es mostrar cómo un método estadístico puede ser aplicado a problemas del analista químico que realiza trabajos de rutina. Estos problemas son resueltos algunas veces caprichosamente, otras mediante la intuición (la cual se encuentra, a menudo, altamente desarrollada en un analista de experiencia) y otras, también, permanecen sin resolver.

En la sección 2.1 se estudian las medidas de posición y seguidamente, en la sección 2.2, las medidas de variabilidad o dispersión de los datos obtenidos por experimentación o que son simplemente un conjunto de mediciones. A continuación, la sección 2.4 trata el problema relacionado con el agrupamiento de datos y su presentación gráfica y la finalmente, se analiza el problema de la detección de valores extremos en conjuntos de datos.

2.1 Medidas de Posición

La característica más importante que describe un grupo de datos es su posición en la recta numérica. La mayor parte de los conjuntos de datos muestran una tendencia definida a agruparse alrededor de cierto punto. Por ello, para cualquier conjunto particular de datos, suele ser posible seleccionar un valor típico para describir o resumir todo el conjunto de datos, dependiendo de la característica de estudio y de la muestra seleccionada.

2.1.1 La media aritmética

La media de la muestra, llamada también media aritmética o promedio aritmético, no es más que la suma de los valores observados divididos entre el número total de elementos de la misma. Es la medida de mayor uso de la centralidad de un conjunto de datos.

Si las observaciones en una muestra de tamaño n son x_1, x_2, \dots, x_n la media es el valor que se escribe de las siguientes formas equivalentes:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + \dots + x_n}{n} \quad (2.1)$$

La media de una población de N elementos se calcula de la misma forma, solamente que la suma de los valores se dividen entre el tamaño de la población. El estadístico muestral \bar{x} es el estimador¹ del parámetro² de la población, μ

La media aritmética de una serie de observaciones se expresa en las mismas unidades de medición que los datos; esto es, si la observación es en gramos, el valor medio resulta en gramos. Por ejemplo, supóngase que se determinó que a 22°C una pipeta de 5,00 mililitros, después de pesar por seis ocasiones, su volumen vertido, generó los siguientes pesos aparentes de agua en gramos:

4,995 4,993 4,994 4,996 4,998 4,992

La media para esta muestra se calcula como:

$$\begin{aligned} \bar{x} &= \frac{4,995 + 4,993 + 4,994 + 4,996 + 4,998 + 4,992}{6} \\ &= \frac{29,968}{6} \\ &= 4,995 \end{aligned}$$

Se diría que la pipeta escurre a una temperatura de 22°C en promedio 4,995 gramos de agua.

La principal debilidad de este estadístico como representante de un conjunto de datos es su susceptibilidad a valores extremos, ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema, hará que la media se desplace en esa dirección.

Al promediar estos datos, los resultados se vuelven dudosos como una descripción razonable respecto a la tendencia central de los datos. Se debe utilizar otras medidas de posición cuando se presenta este problema.

¹ Estimador: en términos generales es una función de la muestra, que proporciona un valor representativo de la característica poblacional.

² Parámetros: conjunto de valores poblacionales que definen la función de densidad de probabilidad de una variable aleatoria.

2.1.2 La media geométrica

La *media geométrica* de un grupo de mediciones es la raíz enésima de su producto. Expresado matemáticamente

$$\begin{aligned}\bar{x}_g &= \sqrt[n]{\prod_{i=1}^n x_i} \\ &= \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdots X_n}\end{aligned}\tag{2.2}$$

Cuando los valores de las mediciones están comprendidos dentro de un intervalo amplio y muchos de ellos se acumulan hacia un extremo, la media geométrica resulta ser una expresión de la tendencia central más adecuada que la media aritmética por la precisión de los resultados.

Tiene la ventaja que considera todos los valores de la distribución y es menos sensible en comparación al promedio aritmético a los valores extremos. Pero tiene la desventaja que tiene un significado estadístico menos intuitivo que el promedio aritmético y puede quedar indeterminada; por ejemplo, si un valor $x_i = 0$ entonces la media geométrica se anula.

Ejemplo. Supóngase, que los porcentajes de estudiantes reprobados en química general 1 durante el 2009, 2010 y 2011 por semestre fueron los siguientes: 25%, 15%, 35%, 32%, 22% y 27%. Calcular la media geométrica del porcentaje de reprobados por semestre.

Siendo en nuestro caso $n = 6$, $x_1 = 25$, $x_2 = 15$, $x_3 = 35$, $x_4 = 32$, $x_5 = 22$, $x_6 = 27$. Aplicando la fórmula (2.2) tenemos:

$$\begin{aligned}\bar{x}_g &= \sqrt[6]{25(15)(35)(32)(22)(27)} \\ &= 25.09\end{aligned}$$

También se puede utilizar para el cálculo de la media geométrica cualquier tipo de logaritmo. Usado logaritmos en base 10.

$$\begin{aligned} \bar{x}_g &= \text{anti log} \left[\frac{\sum_{i=1}^n \log(x_i)}{n} \right] \\ &= \text{anti log} \left[\frac{\log(25) + \log(15) + \log(35) + \log(32) + \log(22) + \log(27)}{6} \right] \\ &= \text{anti log}(1.3995) = 10^{1.3995} \\ &= 25.09 \end{aligned}$$

Entonces la media geométrica del porcentaje de estudiantes reprobados en química general 1 es 25.09%

2.1.2 La mediana

Dado que los valores extremos en un conjunto de datos distorsionan la media aritmética, no es una buena medida de tendencia central en esas circunstancias. Por ello, siempre que está presente un valor extremo, es más apropiado utilizar la mediana, al no ser afectada por la presencia de valores extremos en un conjunto de datos, tal que todos los valores tienen igual importancia en su cálculo.

La mediana de un conjunto de datos es el valor del centro cuando las observaciones están ordenadas en forma ascendente del mínimo hasta el máximo valor, es decir, dispuestos en orden de magnitud. Es decir, el 50% de las observaciones en conjunto de datos son menores o iguales y el otro 50% de las observaciones son mayores o iguales a la mediana.

Para calcular la mediana, primero hay que ordenar los datos en forma ascendente. Si el tamaño de la muestra es un número impar, la mediana se representa con el valor numérico de observación ordenado $(n+1)/2$. Por otra parte, si el tamaño de la muestra es un número par, la mediana se representa con la media de los dos valores centrales en el arreglo ordenado de datos.

Por ejemplo, suponga que las siguientes observaciones representan el punto de ebullición de un compuesto de silicio (en grados Celsius):

166 141 136 153 170 162 155 148 132 160 175 150

Para obtener el punto ebullición del compuesto del silicio de 12 observaciones realizadas, los datos ordenados serían como sigue:

132 136 141 148 150 153 155 160 162 166 170 175

Por tanto, dado que la mediana representa el punto medio en el arreglo ordenado y el número de observaciones en la muestra es par, la mediana se calcula como el promedio de los dos valores centrales de la siguiente manera:

$$M_e = \frac{153 + 155}{2} \\ = 154$$

Este resultado, indica que el 50% de los puntos de ebullición del compuesto de silicio observados son menores a 154 grados Celsius.

Entre las propiedades de la mediana, a destacar están las siguientes:

- Como medida descriptiva, tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello es adecuado su uso en distribuciones asimétricas.
- Es de cálculo rápido y de interpretación sencilla. A diferencia de la media, la mediana de una variable discreta es siempre un valor de la variable que estudiamos. Por ejemplo, la mediana del número de bacterias toma siempre valores enteros.
- Si una población está formada por 2 subpoblaciones de medianas mediana 1 y mediana 2, sólo se puede afirmar que la mediana de la población está comprendida entre mediana 1 y mediana 2.
- El mayor defecto de la mediana es que tiene unas propiedades matemáticas complicadas, lo que hace que sea muy difícil de utilizar en inferencia estadística.

2.1.4 La moda

La moda se define simplemente como el valor que se presenta con la más alta frecuencia absoluta en un conjunto de datos. No requiere hacer operaciones de

cálculo, solamente hay que contar y puede determinarse en relación con variables cualitativas y cuantitativas.

Un conjunto de datos puede tener más de una moda (o sea puede ser multimodal), esto un signo de la falta de homogeneidad en los datos. También es posible que en algunos conjuntos de datos no hay moda. Debe ser usada como medida de posición cuando el tamaño de muestra es grande.

Por ejemplo, los siguientes son resultados en porcentaje obtenidos de oxígeno disuelto al muestrear un río en 36 ocasiones diferentes:

7,1 6,7 6,2 6,1 3,4 5,9 8,7 7,1 6,5 3,9 7,2 8,1
3,3 7,1 8,0 5,4 7,4 7,0 4,0 5,5 5,1 7,1 6,4 7,1
7,7 4,5 7,6 5,8 7,1 7,1 6,3 6,1 6,3 5,8 5,8 7,5

En este caso, el valor más típico o valor modal es 7,1%. Como se mencionó anteriormente, la moda tiene una característica distintiva, pues es la única medida de tendencia central que se puede utilizar con datos cualitativos. Por ejemplo, puede ser usada para determinar el color más frecuente en un conjunto de compuestos químicos.

2.1.5 Proporciones

La proporción de un dato estadístico es el número de veces que se presenta ese dato respecto al total de datos. Por ejemplo, la fracción molar de un componente, es el número de moles del mismo dividido por el número total de moles de todos los componentes presentes en el sistema. Otro caso de investigación en un laboratorio, puede ser la proporción de alcohol metílico en muestras de un producto químico, mediante la técnica de cromatografía líquida de alta resolución (HPLC). Simbólicamente se representa como,

$$p = \frac{x}{n}$$

donde x = parte de la muestra que tiene la característica de interés (subconjunto) y n es el tamaño de la muestra (o conjunto).

Se conoce también como frecuencia relativa y es uno de los parámetros de cálculo más sencillo. Tiene la ventaja de que puede calcularse para variables cualitativas y cuantitativas.

Ejemplo. Si en un grupo de 35 estudiantes, 11 tienen problemas de sobrepeso. ¿Cuál es la proporción de estudiantes con sobrepeso?

$$p = \frac{x}{n}$$
$$= \frac{11}{35} = 0.31$$

Esto significa que, el 0.31 de los estudiantes del grupo tienen problemas de sobrepeso.

2.2 Medidas de variabilidad

La variabilidad en química analítica afecta la precisión y la exactitud de los resultados, por consiguiente, influye en nuestros juicios. Carece de sentido decir que un análisis está “dentro de los límites del error experimental”, si no tiene una idea de la magnitud del mismo.

Las medidas de variabilidad (o dispersión) describen el grado en que los valores no son iguales o en como varían entre sí. Los valores de éstas medidas serán grandes cuando los datos son muy heterogéneos y pequeños, cuando los datos son muy homogéneos.

Dos conjuntos de datos pueden diferir tanto en tendencia central como en dispersión o, pueden tener las mismas medidas de posición, pero diferir en términos de la variabilidad. Por tanto, al analizar un conjunto de datos es insuficiente resumir los datos únicamente presentando medidas de tendencia. Los datos deben ser estudiados en términos de su dispersión para evitar sacar conclusiones falsas de los mismos.

Hay cuatro maneras ordinarias de evaluar la variabilidad: el recorrido, la variancia, la desviación estándar y el coeficiente de variación.

2.2.1 El recorrido

La expresión más sencilla de la dispersión es el *recorrido*, esto es, la diferencia entre los valores máximo y mínimo.

$$R_x = \text{máximo}(x) - \text{mínimo}(x) \quad (2.3)$$

donde x representa la variable de estudio. Como el recorrido destaca los valores más extremos, no influye en él la dispersión típica de los datos.

Para muestras pequeñas, de pocas observaciones, el recorrido da aproximadamente tanta información acerca de la variabilidad de las observaciones como otras medidas de variabilidad más elaboradas. Dado que para su cálculo se utilizan los dos valores extremos, se desperdicia el resto de los datos; para muestras mayores no debería utilizarse el recorrido sino otras medidas de la dispersión. Otro inconveniente es que el rango aumenta con el número de observaciones, o bien se queda igual. En cualquier caso nunca disminuye.

Por ejemplo, considérese el estudio de J.J. Thompson (1856-1940), quien descubrió el electrón aislando partículas cargadas negativamente para las cuales midió la relación carga a masa. Esta relación parecía ser constante en un amplio margen de condiciones experimentales y, en consecuencia, podría ser una característica de esa nueva partícula. Sus observaciones de dos tubos distintos de rayos catódicos que usaban aire como gas dentro de ellos son:

Tubo 1	0,57	0,34	0,43	0,32	0,48	0,40	0,40
Tubo 2	0,53	0,47	0,47	0,51	0,63	0,61	0,48

Se encuentra que el recorrido es,

$$R_x = 0,63 - 0,32 = 0,31$$

En este ejemplo, el recorrido puede ser considerado como una buena medida de la variabilidad porque tanto 0,63 y 0,32 no son valores extremos comparados con las otras observaciones de ambos tubos.

Aunque no es tan eficiente como la desviación estándar, lo fácil de su cálculo ha hecho muy popular al recorrido. La relación existente entre él y la desviación

estándar ha sido estudiada por TIPPETT¹, quien ha demostrado que, para pequeños grupos de medidas (que es el caso usual en laboratorios químicos), el recorrido es, en la práctica, suficientemente cercano a la desviación estándar.

2.2.2 La varianza

La varianza es una medida de la dispersión de los valores de una distribución alrededor del valor medio. Se define como la media aritmética de los cuadrados de la diferencia entre los valores que toma la variable aleatoria y la media de la distribución. El conocimiento del investigador podrá ayudarle a analizar la varianza según el problema de estudio.

Si x_1, x_2, \dots, x_n es una muestra de n observaciones, entonces la varianza muestral está representada por,

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

$$= \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]$$

(2.4)

El divisor $n - 1$ representa los grados de libertad, indicando que solamente hay $n - 1$ desviaciones de la media que son independientes. Dividimos entre $n - 1$ porque se desea utilizar la varianza muestras² como un estimador insesgado de la varianza poblacional σ^2 . Las estimaciones obtenidas al dividir la suma de los cuadrados de las desviaciones entre n tienden a subestimar s^2 .

Por ejemplo, para encontrar el efecto de la carga de polvo en la salida de un sistema con un precipitante, se efectuaron las siguientes mediciones (en gramos por m^3 en el tubo del gas):

1,5	1,5	1,4	1,1	1,7	1,8	1,6	1,5	1,6	2,2
1,7	1,4	1,9	1,9	1,5	1,4	1,9	2,2	1,8	2,0

La variancia se calcula de la siguiente manera:

¹ TIPPETT, L. H. C.: "On the extreme individuals and range of samples taken from a normal population", *Biometrika*, 17, 364, (1925).

$$\begin{aligned}
 s^2 &= \frac{1}{19} \left[57,98 - \frac{(33,6)^2}{20} \right] \\
 &= \frac{1,502}{19} \\
 &= 0,08063
 \end{aligned}$$

El valor 1,502 representa las desviaciones con respecto a la media aritmética al cuadrado.

La varianza no tiene la misma magnitud que las observaciones (ejemplo, si las observaciones se miden en metros, la varianza lo hace en metros²). Si quiere que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada.

2.2.3 La desviación estándar

La desviación estándar de un conjunto de datos representa la variabilidad de los datos muestrales y constituye nuestra estimación de la desviación estándar de la población. Para obtener la desviación estándar simplemente se toma la raíz cuadrada de la variancia. Por tanto,

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \tag{2.5}$$

La desviación estándar del ejemplo presentado para cálculo de la variancia sobre la salida de polvo de un sistema precipitante, sería,

$$s = \sqrt{0,08063} = 0,2839$$

Las unidades de medición para **s** son las mismas de los datos originales. Por otra parte, la desviación estándar se utiliza para medir la variación promedio alrededor a la media aritmética.

Además de las propiedades citadas de la variancia y desviación estándar, será conveniente tener siempre en mente otras que serán enunciadas a continuación:

- Ambas son sensibles a la variación de cada una de las puntuaciones, es decir, si una puntuación cambia, cambia con ella la variancia. La razón es que

si miramos su definición, la varianza es función de cada una de las puntuaciones.

- Si se calculan a través de los datos agrupados en una tabla, dependen de los intervalos elegidos. Es decir, se comete cierto error en el cálculo de la varianza cuando los datos han sido resumidos en una tabla estadística mediante intervalos, en lugar de haber sido calculados directamente como datos no agrupados.
- No es recomendable el uso de ellas, cuando tampoco lo sea el de la media como medida de tendencia central.

2.2.4 Coeficiente de variación

El coeficiente de variación es una medida de variabilidad relativa, tal que expresa la magnitud de la desviación estándar como un porcentaje de la media. Se expresa como porcentaje en vez de las mismas unidades que los datos. La fórmula de cálculo es,

$$CV_x = \frac{s_x}{\bar{x}} * 100 \quad (2.6)$$

El coeficiente de variación se utiliza para comparar variabilidad entre dos o más variables que se miden en diferentes unidades o cuya media es muy diferente. Por ejemplo, las siguientes cifras son las horas de estudio de 10 alumnos para presentar un examen de Química General 1 y las calificaciones que obtuvieron:

Horas de estudio:	7	5	10	12	10	6	18	15	4	9
Calificación:	60	44	80	75	70	65	90	90	35	65

Por medio de las fórmulas 2.1, 2.4 y 2.5 se obtiene para las variables horas de estudio y calificación el promedio aritmético, la varianza y la desviación estándar:

Sustancia	\bar{x}	S^2	s	CV
Horas de estudio	9,6	19,8222	4,4522	46,4%
Calificación	67,4	323,156	17,9765	26,7%

Un error sería comparar la variabilidad absoluta usando la varianza o la desviación estándar entre las variables, tal que están medidas en diferentes unidades, con lo cual, podría concluirse que las calificaciones tienen mayor variabilidad.

Sin embargo, al calcular el coeficiente de variación, se observa que en realidad las horas de estudio presentan una mayor variabilidad relativa. Se podría decir que entre los alumnos existe una mayor variabilidad con respecto a las horas de estudio en relación a las calificaciones obtenidas en el examen de Química General 1.

Otras consideraciones que se deben tener en cuenta son:

- Sólo se debe calcular para variables con todos los valores positivos. Todo indicador de variabilidad es esencialmente no negativo. Las observaciones pueden ser positivas o nulas, pero su variabilidad debe ser siempre positiva.
- No es invariante ante cambios de origen. Es decir, si a los resultados de una medida le sumamos una cantidad positiva, $b > 0$, para tener $Y = X + b$, la desviación estándar no es sensible ante cambios de origen, pero la media sí lo es.
- Es invariante a cambios de escala. Si multiplicamos X por una constante a , para obtener $y = ax$, para ver que las siguientes consideraciones

$$CV_Y = \frac{S_Y}{\bar{y}} = \frac{S_{aX}}{a\bar{x}} = \frac{aS_X}{a\bar{x}} = CV_X$$

2.3 Distribución de frecuencia

Es un método para resumir grandes cantidades de datos en clases apropiadas, mostrando el número de observaciones en cada clase. En esta forma resumida, posibilita que los datos sean más manejables y comprensibles, se logra destacar las características sobresalientes de la variable de estudio.

Los resultados experimentales se representan muchas veces en forma de distribuciones de frecuencia, agrupados o no. En las distribuciones de frecuencia no agrupadas, los valores observados se disponen en orden ascendente de magnitud. Las distribuciones de frecuencia agrupadas se disponen de modo que se manifiesten las frecuencias con que se presentan los valores de la variable dentro de determinadas clases ordenadas.

Cuando los datos se agrupan según su magnitud numérica, la tabla resultante recibe el nombre de distribución numérica o cuantitativa. En tanto, si los datos se agrupan en categorías que difieren en tipo y no en grado, la tabla resultante se denomina distribución categórica o cualitativa.

La construcción de una distribución cuantitativa consta esencialmente de cuatro pasos:

1. Selección del número de clases
2. Cálculo del intervalo de clase
3. Definición de los límites de clase
4. Conteo del número de observaciones por clase

2.3.1 Selección del número de clases

El número de clases a utilizar, depende en forma primordial de tres factores: interés del investigador, costumbre y cálculo estadístico. Primeramente, podría ser de preferencia para el investigador agrupar los datos en número determinado de clases, porque su conocimiento del problema así lo indica o simplemente por interés personal. En tanto, la costumbre en las diferentes disciplinas científicas, guían al analista de datos en la selección del número apropiado de clases.

El número exacto que se utilice en una situación dada dependerá de la naturaleza, magnitud y orden de los datos. Con certeza, perderíamos más de lo que ganaríamos si agrupamos cinco observaciones en 12 clases, tal que la mayor parte de ellas quedarían vacías y por el contrario sería inconveniente agrupar 1000 mediciones en dos clases. No obstante, en general, la distribución de frecuencias debe tener el número de clases que permita una concentración o agrupamiento de los datos de modo conveniente para los propósitos del estudio.

Por otro lado, el número de clases puede ser definido al utilizar un cálculo estadístico, que relaciona la *amplitud general*, es decir, la diferencia entre el máximo y mínimo valor que tomó la característica y el intervalo de clase seleccionado o calculado previamente.

$$\text{Número de clases} = \frac{\text{Amplitud general}}{\text{Intervalo de clase}} \quad (2.7)$$

El intervalo de clase es simplemente la distancia (o “diferencia”) entre el límite superior e inferior de una clase.

2.3.2 Cálculo del intervalo de clase

Al construir una distribución de frecuencia, es deseable que cada clase tenga la misma longitud o intervalo de clase. Para determinar el ancho de cada clase, la amplitud general de los datos (la diferencia entre la observación más grande y la más pequeña), se divide entre el número de clases deseadas:

$$\text{Intervalo de clase} = \frac{\text{Amplitud general}}{\text{Número de clases}} \quad (2.8)$$

Por ejemplo, los siguientes datos son los tiempos de ignición de ciertos materiales expuestos al fuego, dados a la más cercana centésima de segundo:

2,58	5,50	6,75	2,65	7,60	6,25	3,78	4,90	5,21	2,51
6,20	5,92	5,84	7,86	8,79	4,79	3,90	3,75	3,49	4,04
3,87	6,90	4,72	9,45	7,41	2,45	3,24	5,15	3,81	2,50
1,52	4,56	8,80	4,71	5,92	5,33	3,10	6,77	9,20	6,43
1,38	2,46	7,40	6,25	9,64	8,64	6,43	5,62	1,20	1,58

Suponga, que se tiene interés de construir cinco clases. Con el arreglo ordenado de los tiempos se determina que la observación más grande es de 9,65 y la más pequeña, de 1,20. Por tanto, la amplitud se calcula como:

$$\begin{aligned} \textit{Amplitud} \textit{ general} &= \textit{Observación mayor} - \textit{observación menor} \\ &= 9,64 - 1,20 \\ &= 8,44 \end{aligned}$$

y se tiene la aproximación al intervalo de clase con

$$\textit{Intervalo de clase} = \frac{8,44}{5} = 1,688 \cong 1,69$$

Como regla general, el intervalo de clase tendrá igual número de decimales que los datos. Si el resultado del cálculo del intervalo tiene más dígitos, se debe redondear utilizando el criterio de redondeo hacia arriba.

2.3.3 Definición de los límites de clase

Para construir la tabla de distribución de frecuencia, es necesario establecer con claridad límites para cada clase, a fin de que las observaciones puedan ser ubicadas en una sola categoría. Las clases deben ser mutuamente excluyentes para evitar la superposición de ellas.

Como el intervalo de clase para los tiempos de ignición se ha establecido en 1.69, se deben establecer los “límites o fronteras” de las diversas clases con el propósito de incluir toda la amplitud de las observaciones. Siempre que sea posible, estos límites se deben seleccionar para facilitar la lectura e interpretación de los datos. Por tanto, el límite inferior y superior de la primera clase se establece desde 1,20 hasta 2,88, la segunda clase quedaría definida entre los límites de 2,89 hasta 4,57, etc.

En caso que tenga conjuntos de datos con valores extremos, puede ser necesario construir “clases abiertas”, estas no tienen límite inferior o límite superior. Por ejemplo: “menos de 2.89” y “más de 7.95”.

2.3.4 Conteo del número de observaciones por clase

Siempre debe asegurarse que cada elemento (medición u observación) esté solamente en una clase. Para ello, el valor menor y el mayor deben estar dentro de la clasificación definida y las clases deben ser mutuamente excluyentes. Los datos ordenados para cada clase y su frecuencia absoluta serían:

Tiempos De ignición	Conteo	Frecuencia Absoluta (f_i)
1,20 – 2,88	//// //// //	10
2,89 – 4,57	//// //// /	9
4,58 – 6,26	//// //// //// ////	16
6,27 – 7,95	//// //// /	9
7,96 – 9,64	//// //	6

Al establecer los límites de cada clase como se indica arriba, las 50 observaciones se han contado y separado en cinco clases, cada una de las cuales tiene un intervalo de 1,69 sin existir traslapamiento entre ellas una observación no podría ser contabilizada en dos clases a la vez.

La elaboración de esta tabla de resumen o distribución de frecuencia permite saber cómo se distribuyen las observaciones dentro de una clase en particular, sin tener necesidad de analizar los datos originales. Por tanto, para 10 materiales expuestos al fuego el tiempo de ignición se encuentra entre 1,20 y 2,88 centésima de segundo, sin embargo la tabla de resumen no indica con exactitud si las observaciones están distribuidas en forma uniforme en el intervalo, si todas son cercanas a 1,20 o todas están cercanas a 2,88.

Por ello, el *punto medio* del intervalo de clase, es el valor utilizado para representar todos los datos resumidos en una clase particular. El punto medio para el intervalo “1,2 – 2,88” es 2,04. Los otros puntos medios de clase, respectivamente, son:

3,73 5,42 7,11 8,8

Por otra parte, la ventaja principal de utilizar una distribución de frecuencias es que las características de los datos más sobresalientes son claramente identificadas por los usuarios de la información. Se observa que la mayoría de tiempos de ignición tienden a agruparse entre 4,58 y 6,26 centésima de segundo. La distribución de frecuencias es, en realidad, una tabla de resumen en la cual se agrupan los datos para facilitar el análisis.

Ahora bien, para facilitar todavía más el análisis de los datos, se pueden calcular las *frecuencias relativas* o *porcentajes*, según si el investigador prefiere proporciones o porcentajes. Las frecuencias relativas se calculan dividiendo la frecuencia absoluta de cada clase entre número total de observaciones o tamaño de la muestra. La distribución con las frecuencias absolutas y relativas se presentan a continuación:

Tiempos de ignición	f_i	Frecuencia relativa (f_i/n)	Porcentajes %
1,20 – 2,88	10	0,20	20
2,89 – 4,57	9	0,18	18
4,58 – 6,26	16	0,32	32
6,27 – 7,95	9	0,18	18
7,96 – 9,64	6	0,12	12

Con esta tabla, se pueden calcular los porcentajes por clase al multiplicar por 100 cada frecuencia relativa. Un 32% de los materiales fueron consumidos por el fuego entre 4,58 y 6,26 centésima de segundo. En realidad, el uso de las frecuencias relativas o distribución porcentual de las observaciones se vuelve importante siempre que un conjunto de datos se compara con otro conjunto, sobre todo si difieren el número de observaciones en cada conjunto.

Por tanto, para comparar los tiempos de ignición de 50 materiales con respecto a otro conjunto de 25 materiales, se hace necesario establecer, ya sea la distribución de las frecuencias relativas o de los porcentajes para este último conjunto.

2.3.5 Representación gráfica de distribuciones de frecuencias

Como anteriormente se ha citado, las distribuciones de frecuencias resume los datos que se dispone de una población, de forma que ésta se puede analizar de una manera más sistemática y resumida. Para darse cuenta de un sólo vistazo de las características de la población resulta aún más esclarecedor el uso de gráficos, cuya construcción abordamos en esta sección.

HISTOGRAMA

Un histograma se construye a partir de la distribución de frecuencias, representando sobre cada intervalo, un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de cada intervalo y su área. Los histogramas, son gráficos de barras verticales, contruidos sobre los límites reales de cada clase.

Los límites reales se les denominan también límites verdaderos, calculados para mantener la continuidad de las clases y se obtienen considerando que:

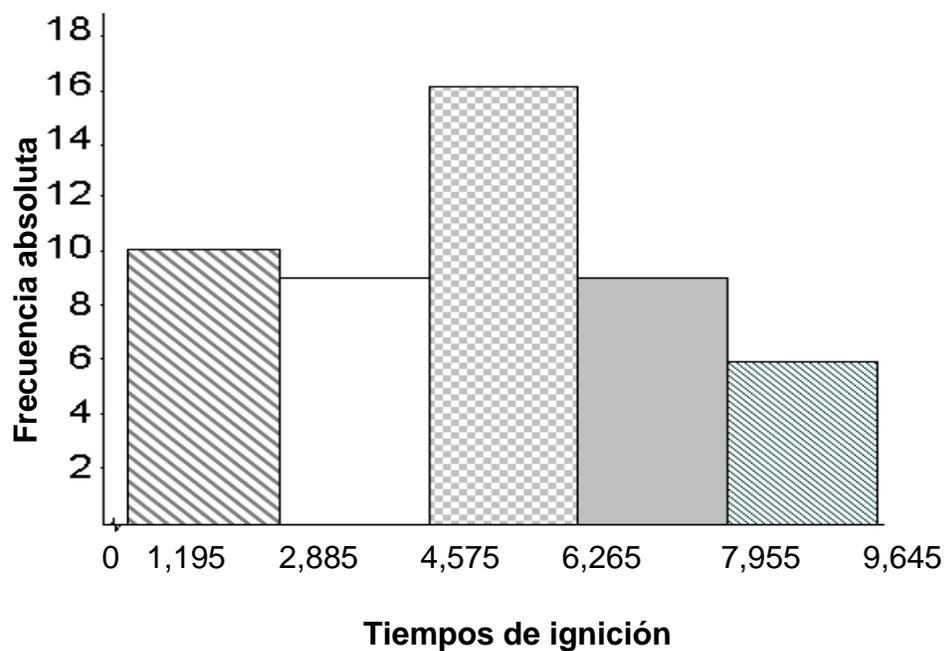
- Si los límites son números enteros, entonces, restar 0.5 al límite inferior ($L_{i-0.5}$) y sumar 0.5 al límite superior ($L_{s+0.5}$).
- Si los límites no son números enteros, se debe restar y sumar a los intervalos de clase 0.05 si tienen un solo decimal, 0.005 si tienen dos decimales, 0.0005 si tienen tres decimales, etc.

Para el caso de estudio, los límites reales se representan en la segunda columna de la siguiente tabla:

Tiempos de ignición	Limites reales	f_i
1,20 – 2,88	1,195 – 2,885	10
2,89 – 4,57	2,885 – 4,575	9
4,58 – 6,26	4,575 – 6,265	16
6,27 – 7,95	6,265 – 7,955	9
7,96 – 9,64	7,955 – 9,645	6

En términos generales, el límite real inferior de un intervalo es el valor que resulta de disminuir el valor del límite indicado (o dado) inferior en media unidad de medida. El límite real superior de un intervalo resulta de incrementar el límite superior indicado en media unidad de medida.

Al dibujar el histograma, la variable aleatoria de interés se representa a lo largo del eje horizontal; el eje vertical representa la frecuencia absoluta o la relativa por clase. A continuación se presenta el histograma para los tiempos de ignición de los materiales.



Sin embargo, cuando un analista de información quiera comparar dos o más conjuntos de datos, no se pueden construir dos histogramas en la misma gráfica, porque la superposición de las barras verticales ocasionaría dificultades en el análisis. Una forma de solucionar este problema, es construir un polígono de frecuencias absolutas o relativas.

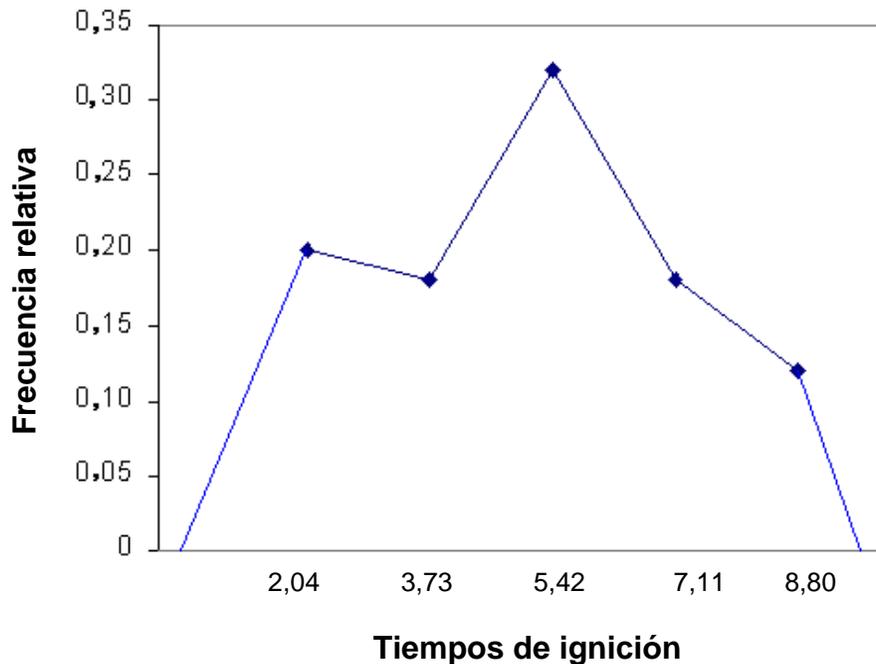
POLÍGONO DE FRECUENCIAS

Al igual que con los histogramas, al dibujar el polígono de frecuencias, la variable aleatoria de interés se representa a lo largo del eje horizontal, pero dada en puntos medios y en el eje vertical es representada la frecuencia absoluta o la

relativa. El polígono se construye al representar las frecuencias absolutas o relativas con sus respectivos puntos medios y, luego, al conectar la sucesión de puntos medios por medios líneas rectas.

Tiempos de ignición	Puntos medios x_i	Frecuencias relativas f_i/n
1,20 – 2,88	2,04	0,20
2,89 – 4,57	3,73	0,18
4,58 – 6,26	5,42	0,32
6,27 – 7,95	7,11	0,18
7,96 – 9,64	8,80	0,12

El siguiente gráfico muestra el polígono de frecuencias utilizando las frecuencias relativas para los tiempos de ignición de los materiales.



Se debe tener en cuenta que el polígono es una representación de la forma de la distribución de probabilidad de la variable aleatoria particular. Dado que el área debajo de la totalidad de la curva o distribución de las frecuencias relativas debe ser 1, es necesario conectar el primero y el último punto medio con el eje

horizontal y modificar el eje vertical dividiendo por el intervalo de clase que es 1,69 a fin de encerrar el área de la distribución observada.

El procedimiento consiste en conectar el primer punto medio observado con el punto medio de una clase anterior ficticia, que tenga 0 de frecuencia absoluta y al conectar el último punto medio observado con el punto medio de una clase posterior ficticia, que tenga 0 observaciones.

Se debe mencionar, que cuando construimos un polígono de frecuencia o histograma, el eje vertical debe mostrar el cero real (origen) a fin de representar correctamente las características de la distribución de los datos. Sin embargo, el eje horizontal no necesita especificar el punto cero para la variable de interés. El recorrido de la variable aleatoria debe constituir la mayor parte del gráfico y, cuando se tienen problemas de presentación de los valores es apropiado hacer cortes de escala y denotarlos simbólicamente con “ ζ ” en el eje.

2.4 PROBLEMAS

2.1 A. A. Michelson (1852-1931) efectuó muchas series de mediciones de la velocidad de la luz. Empleando una técnica de espejo giratorio, obtuvo

12 30 30 27 30 27 48 24 18 25 39 18

para las diferencias (velocidad de la luz en aire) – (299 700 km/s (Fuente: The Astrophysical Journal, 65 (1927): 11).

- Calcule el promedio aritmético y la mediana.
- Calcule la variancia y la desviación estándar.
- Calcule el coeficiente de variación.

2.2 Un analista desea determinar el número de moles de cobre II (Cu^{2+}) en un volumen dado de una solución, por electrólisis. Suponga que el químico empleó sólo 9 muestras de la solución para el experimento y que los resultados fueron, en moles de la sustancia en referencia, como sigue:

0,15 0,17 0,19 0,15 0,18 0,16 0,17 0,18 0,19

- Calcule e interprete la mediana.
- Calcule la variabilidad relativa de número de moles.

2.3 Para estimar la cantidad de alimentos contaminados en las fiestas de Zapote, un inspector del Ministerio de Salud, decide contar el número de muestras contaminadas con alguna bacteria en los lugares de venta. Para ello, decidió tomar 70 muestras de alimentos durante un día. Los resultados fueron son los siguientes:

7 10 4 8 6 9 6 4 9 10 9 8 3 9 5 9 9 8
2 7 4 8 5 10 9 6 8 8 8 7 8 6 11 9 11 7
10 8 8 5 9 9 8 8 9 10 7 7 7 5 8 7 9 9
9 5 8 8 7 9 3 8 7 8 7 10 7 11 6 8

- Construya una distribución de frecuencias con cinco clases.
- Calcule las frecuencias acumuladas, absolutas y relativas.
- Construya un histograma de las frecuencias absolutas.

2.4 La contaminación con petróleo de los mares estimula el crecimiento de ciertos tipos de bacterias. Un conteo de microorganismos oleolíticos (bacterias por 100 mililitros) en 10 muestras de agua de mar, dio los siguientes resultados:

49 70 54 67 59 40 61 69 71 52

- Calcule e interprete la mediana.
- Calcule el coeficiente de variación.

2.5 A continuación se dan 15 mediciones del punto de ebullición de un compuesto de silicio(Si) (en grados Celsius):

166 141 136 153 170 162 155 146 183 157 148 132 160 175 150

Calcule el promedio aritmético y la desviación estándar.

2.6 Los siguientes datos son los tiempos de ignición de ciertos materiales expuestos al fuego, dados a la más cercana centésima de segundo:

2,58 5,50 6,75 2,65 7,60 6,25 3,78 4,90 5,21 2,51
6,20 5,92 5,84 7,86 8,79 4,79 3,90 3,75 3,49 4,04
3,87 6,90 4,72 9,45 7,41 2,45 3,24 5,15 3,81 2,50
1,52 4,56 8,80 4,71 5,92 5,33 3,10 6,77 9,20 6,43
1,38 2,46 7,40 6,25 9,65 8,64 6,43 5,62 1,20 1,58

- Construya una distribución de frecuencias utilizando un intervalo de clase de 2 minutos.
- Represente la distribución de frecuencias por medio de un polígono de frecuencias.
- Calcule gráficamente la mediana.

2.7 Un ingeniero químico vigila la calidad del agua midiendo la cantidad de sólidos suspendidos en una muestra de agua pluvial. En 11 días distintos observó los sólidos suspendidos (partes por millón).

14 12 21 28 30 63 29 63 55 19 20

- a) Calcule e interprete la media aritmética y la mediana
- b) Calcule el coeficiente de variación

2.8 Los siguientes son los pesos en gramos (redondeados al gramo más próximo) de 30 ratones, de un ensayo realizado en un Laboratorio de Química usadas en un estudio de deficiencias vitamínicas:

136 92 115 118 121 137 132 120 104 125
119 115 101 129 85 108 110 133 135 126
127 103 110 126 118 82 104 137 120 148

- a) Construya una distribución de frecuencia con 6 clases que incluya límites dados y reales.
- b) Interprete la frecuencia absoluta de la clase 3.
- c) Construya el polígono de frecuencias.

2.9 Los siguientes son los resultados para la medida de oxígeno (O_2) disuelto en diferentes ocasiones en un punto de muestreo de un río, en gramos por litro.

7,1 6,7 6,2 6,1 3,4 5,9 8,7 6,8 6,5 3,9 7,2 8,1
3,3 7,1 8,0 5,4 7,4 7,0 4,0 5,5 5,1 8,9 6,0 7,2
7,7 4,5 7,6 5,8 7,1 7,1 6,3 6,1 6,3 5,8 5,8 7,5

- a) Construya una distribución de frecuencia que tenga 5 clases. (incluya los límites dados, las frecuencias absolutas y relativas).
- b) Calcule e interprete el valor de la tercera clase de la frecuencia absoluta acumulada a "más de".
- c) Represente gráficamente la distribución de frecuencias por medio de un polígono de frecuencia.

2.10 Ocho determinaciones de la densidad de un líquido orgánico dieron los resultados (en g/cm^3). Calcule e interprete la media aritmética y la mediana.

0,913 0,943 0,959 0,901 0,951 0,917 0,961 0,943

2.11 Se le pide a un analista experimentado la evaluación de dos métodos diferentes para la determinación de trazas de plomo en ácido acético glacial, y se le entrega una muestra que contiene precisamente 1.282 ppm. de Pb, por litro (dato desconocido por el analista). Se realizaron cinco determinaciones mediante cada método, que dieron los siguientes resultados para la concentración del plomo, en partes por millón (ppm):

Método A: 1,34 1,33 1,32 1,34 1,31

Método B: 1,30 1,26 1,30 1,33 1,24

Compare ambos métodos en cuanto a su exactitud y variabilidad.

2.12 Un estudiante de química preparó una serie de compuestos que contienen las siguientes masas de nitrógeno y oxígeno(en gramos):

Compuesto	N ₂	O ₂
A	16,8	19,2
B	17,1	39,0
C	33,6	57,3
D	25,4	28,2
E	27,9	26,1
F	14,3	45,8
G	35,2	53,5

- Calcule e interprete la masa de nitrógeno promedio de los compuestos.
- Calcule e interprete la mediana de la masa de oxígeno de los compuestos.
- ¿Cuál elemento de los compuestos presenta una menor variabilidad relativa? Justifique la respuesta

2.13 El espectro de masa de una muestra de óxido de plomo contiene iones de fórmula PbO. El óxido de plomo ha sido preparado a partir de ¹⁶O isotópicamente puro. Las masas de los iones en gramos por mol, se listan a continuación:

220,03 220,90 226,92 225,20 216,80 229,56

222,15 232,11 235,00 223,50 224,30 227,23

224,78	218,45	231,45	222,62	228,56	224,05
230,30	215,00	223,06	228,92	223,82	230,82
225,62	229,10	224,05	232,50	221,95	218,42

- Construya una distribución de frecuencias con 4 clases. Incluya en la tabla la frecuencia absoluta y la relativa.
- Calcule la frecuencia absoluta acumulada "a menos de" y la frecuencia relativa "a más de". Interprete el valor de la tercera clase de ambas frecuencias acumuladas.
- Represente gráficamente la distribución de frecuencias por medio de un polígono de frecuencias.

2.14 A continuación se presenta las medidas de pH y los porcentajes de Cl y SO₄ en muestra de agua del túnel Arenal 2, a diferentes distancias de la estación.

DISTANCIA	pH	% CL	% SO ₄
560	6,20	26,6	78,3
565	3,20	37,9	127
570	3,20	71,5	260
634	3,00	22,6	611
643	2,95	23,2	611
533	5,70	15,5	74,1
769	3,05	29,6	827
746	4,05	30,8	784
809	3,65	32,1	883
816	3,60	32,5	883

- Calcule e interprete el promedio y la mediana de los datos de SO₄
- ¿Qué variable -pH o CL- tiene menor variabilidad relativa?

2.15 Un ingeniero químico está investigando sobre el rendimiento del efecto de la temperatura en grados centígrados sobre un determinado proceso. El estudio da como resultado los siguientes datos:

Temperatura: 100 110 120 130 140 150 160 170 180 190
Rendimiento: 45 51 54 61 66 70 74 78 85 89

- a) Calcule la mediana del rendimiento.
- b) Determine cuál de las dos variables: temperatura o rendimiento, tiene mayor variabilidad relativa.

2.16 En una muestra de tabletas de aspirinas, de las cuales se cuantifica su peso expresado en gramos, se obtiene:

1,19 1,23 1,18 1,21 1,27 1,17 1,15 1,14 1,19

Calcule e interprete la mediana del peso de las tabletas de aspirinas.

2.17 Para retardar la fermentación de cierta cerveza se agrega un ingrediente en el proceso de elaboración. Para controlar el tiempo de fermentación, se llevó un seguimiento por días como sigue:

14 17 16 25 24 24 30 36 15 28 29 23 31 34 25 31 26
24 19 28 26 23 37 29 28 27 29 29 16 23 32 28 29 18

- a) Construya una distribución de frecuencias con 6 clases. Incluya los límites dados, la frecuencia absoluta y la relativa.
- b) Represente gráficamente la distribución de frecuencias utilizando un histograma.
- c) Calcule e interprete la moda.

2.18 Un estudiante obtiene los datos que siguen del volumen de aire como función de la temperatura a presión constante.

Temperatura (°C):	16	55	85	103	126	163
Volumen (cm ³)	31	35	38	40	43	47

¿Cuál variable tiene menor variabilidad relativa?

2.19 Se está investigando un método para producir gasolina a partir de desechos orgánicos. La planta ha proporcionado los siguientes datos referentes al rendimiento, en porcentaje del combustible:

24,1 21,0 26,6 26,0 25,7 21,8 20,9 20,4 20,0
 28,8 20,2 25,9 21,2 26,2 22,0 24,2 24,7 20,7
 25,9 26,7 30,0 24,0 21,3 26,6 21,0 22,1 21,8
 21,8 22,9 21,6 25,3 24,9 25,9 26,5 25,4 22,4

- Construya una distribución de frecuencia con 5 clases. Incluya los límites dados, la frecuencia absoluta y la frecuencia relativa acumulada "a menos de".
- Dibuje el polígono de frecuencias.

2.20 La distribución siguiente corresponde a las lecturas con un contador Geiger del número de partículas emitidas por una sustancia radiactiva en 100 intervalos sucesivos de 40 segundos:

Número de partículas	Frecuencia
5-9	1
10-14	10
15-19	37
20-24	36
25-29	13
30-34	2
35-39	1

- Calcule la frecuencia absoluta acumulada "a menos de" y la frecuencia relativa "a más de". Interprete el valor de la tercera clase de ambas frecuencias acumuladas.
- Represente gráficamente la distribución de frecuencias por medio de un histograma.
- Calcule e interprete: el promedio aritmético, la moda y la mediana.
- Calcule la desviación estándar y el coeficiente de variación.
- Calcule e interprete el percentil 75%.

2.21 De las siguientes determinaciones en porcentaje de humedad (H_2O):

0,48 0,37 0,47 0,40 0,44 0,46 0,43 %

Calcule e interprete:

- a) El promedio aritmético
- b) La desviación estándar
- c) El rango

2.22 A continuación se presenta los siguientes valores de pH de un conjunto de muestras de agua:

6,63	5,68	7,22	12,11	9,28
9,82	6,73	9,63	4,90	7,23
4,85	2,70	5,44	8,98	10,91
3,67	6,69	11,20	5,42	7,91
7,20	14,57	8,87	9,12	6,97

Construya una distribución de frecuencias con 5 clases y construya el histograma utilizando las frecuencias relativas.

2.23 El análisis de una mezcla indica que contiene 32,55% de Fe_2O_3 y pierde 1,25% de peso al secarse. Calcule el porcentaje de Fe_2O_3 en base seca.

2.24 El $CaCO_3$ (100,09) es el único componente de una mezcla que pierde peso por calcinación, al desprender CO_2 , peso molecular 44,01 g/mol. Al calcinar una muestra de 0,4532 g de esta mezcla, el peso del residuo resulta ser 0,4289 g.

- a) ¿Cuál es el porcentaje de pérdida por calcinación?
- b) ¿Cuál es el porcentaje de $CaCO_3$?

2.25 A continuación se presentan los resultados obtenidos en la determinación del % de nitrógeno en cinco muestras de heno tomadas al azar, como sigue:

1,24 1,37 1,29 1,28 1,33

- a) Calcule el promedio aritmético

- b) Calcule el recorrido y la desviación estándar. Compare los resultados
- c) Calcule la variabilidad relativa

2.26 Los siguientes datos representan los porcentajes de zinc de un conjunto de aleaciones:

% Zn.	Frecuencia
2,01-2,10	1
2,11-2,20	3
2,21-2,30	9
2,31-2,40	16
2,41-2,50	19
2,51-2,60	30
2,61-2,70	18
2,71-2,80	12
2,81-2,90	6
2,91-3,00	1

- a) Interprete en términos del problema de la frecuencia absoluta de la quinta clase.
- b) Calcule e interprete el valor correspondiente de la frecuencia relativa acumulada a “más de” de la cuarta clase.
- c) Representa gráficamente la distribución de frecuencias utilizando un histograma.

2.27 Varios análisis de una aleación dieron los siguientes porcentajes de cobre:

3,27 3,31 3,29 3,42 3,32

Calcular la media aritmética, la media geométrica y la mediana. ¿Cuál de estos valores se puede tomar como el mejor valor del contenido real en cobre de la aleación analizada?

2.28 Los datos siguientes representan el tanto por ciento de cenizas de diferentes muestras de un cargamento de carbón.

% de cenizas	Frecuencia
4,0-4,9	7
5,0-5,9	28
6,0-6,9	75
7,0-7,9	85
8,0-8,9	45
9,0-9,9	27
10,0-10,9	7
11,0-11,9	2

Calcular el promedio aritmético del contenido de cenizas de las muestras y representar la distribución utilizando un polígono de frecuencias.

2.29 Se analizó volumétricamente una cierta muestra para determinar el porcentaje de sulfatos, y se obtuvieron los resultados siguientes:

6,39 6,99 5,71 5,93 5,35 5,81 5,52 5,91 5,46

¿Se puede rechazar alguno de los resultados? Determinar la media, el recorrido y la desviación estándar del método.

2.30 Cinco determinaciones de mercurio sobre una muestra determinada dieron:

21,0 20,0 20,5 20,8 21,1 % Hg.

Calcular e interpretar la mediana del análisis.

2.31 Se hicieron cuatro determinaciones sobre una misma muestra para determinar el valor del contenido en el constituyente mayor. Los resultados en porcentajes fueron:

38,64 38,61 38,55 38,78

¿Qué valor debe darse como resultado?

2.32 Los siguientes resultados de sendas microdeterminaciones de C y H en ácido benzoico e hidrocloreuro de efedrina, fueron obtenidos por diversos analistas.

Acido benzoico		Hidrocloreuro de efedrina	
% C	% H	% C	% H
69,03	5,00	59,22	8,19
68,96	4,91	59,66	8,31
69,18	5,22	59,27	7,86
69,22	5,19	59,22	7,87
68,88	4,89	59,51	8,24
69,04	4,76	59,60	8,00

- Calcular la media de cada determinación.
- Calcular la mediana de cada determinación.
- Para cada muestra calcular el estimador de la desviación estándar de la media.

2.33 Se obtuvieron los siguientes resultados del porcentaje de nitrógeno en una muestra paralelamente con el porcentaje de nitrógeno en un blanco.

Análisis (% N)	Blanco (% N)
18,16	0,10
18,20	0,14
18,07	0,07
18,30	0,12

- Calcular la desviación estándar para el blanco y el propio análisis.
- Calcular la variabilidad relativa con ambos métodos.

2.34 Se obtuvieron los siguientes resultados en la valoración de una disolución de ácido clorhídrico en moles por litro, según el método de determinación

volumétrica, frente a carbonato de sodio puro, y otro frente a hidróxido de sodio, obteniéndose los siguientes resultados:

Molaridad con NaOH	Molaridad con Na ₂ CO ₃
0,1052	0,1060
0,1051	0,1055
0,1047	0,1049
0,1052	0,1062
0,1049	0,1059

- Calcular la media de cada método.
- Calcular la desviación estándar de la media de cada método.
- ¿Cuál método presenta mayor variabilidad relativa?

Capítulo 3

Probabilidades

3.1 Introducción a las probabilidades

3.2 Definición axiomática de probabilidad

3.3 Teoremas fundamentales del cálculo de probabilidades

3.4 Distribuciones de probabilidad

3.5 Distribuciones discretas

3.6 Distribuciones continuas

3.7 Aplicaciones de la distribución normal estándar

3.8 La distribución t de Student

3.9 Problemas

Probabilidades

En los dos capítulos anteriores se han examinado los conceptos básicos de estadística y los métodos de análisis descriptivo de datos. En este capítulo se empieza el estudio de diversas reglas de probabilidad básica y modelos de probabilidades, que son utilizados para hacer inferencias relacionadas con una población, basándose tan sólo en muestreo estadístico.

En la sección 3.1, se introduce los conceptos básicos de probabilidades y sus métodos de cálculo.

Seguidamente, en la sección 3.2 se presentan las distribuciones de probabilidades y sus aplicaciones en el análisis de datos. Se detalla en las siguientes secciones las características de las distribuciones Binomial, Poisson, Normal y t de Student.

Es importante estudiar el eslabón que une la teoría de la probabilidad y la estadística aplicada: la noción de variable aleatoria, mostrando de esta manera, como puede emplearse la teoría de la probabilidad para sacar conclusiones precisas acerca de una población con base en una muestra extraída de ella.

3.1 Introducción a las probabilidades

La teoría de probabilidades proporciona las bases matemáticas y el lenguaje de la estadística usada en este libro. En la mayoría de los problemas estadísticos de probabilidades, nos interesa saber el número de elementos en un espacio muestral finito, sin embargo resulta muy difícil determinar todo el conjunto de eventos factibles, por medio de la enumeración directa. A la vez, es importante conocer los resultados posibles de un experimento y sus probabilidades. Por ejemplo, podría interesar cuantificar el número de productos defectuosos producidos por una máquina, y las probabilidades de tener x artículos defectuosos

En estadística se acostumbra designar como un experimento a un proceso de observación o medición cualquiera. En ese sentido, un experimento consiste en determinar el número de elementos a combinar para encontrar un compuesto

anhelado. Lo que se obtiene de un experimento, ya sea lectura de instrumentos o valores obtenidos a través de operaciones, se denomina resultados del experimento.

Cuando se estudian los resultados de un experimento, por lo general se identifican los diversos resultados posibles con números, puntos u otras clases de símbolos, de manera que podamos tomar en cuenta todos los aspectos concernientes a los resultados desde un punto de vista matemático, sin tener que recorrer largas descripciones verbales de lo que ha sucedido, lo que está sucediendo o lo que ocurrirá.

Hay tres definiciones de las probabilidades de ocurrencia de un evento particular. La probabilidad estadística está basada en un conocimiento previo del proceso aleatorio, la probabilidad de un evento (resultado o acontecimiento), es la proporción de las veces en que ocurrirán eventos del mismo tipo.

La probabilidad clásica se define como el número de resultados favorables a A , dividido entre el número total de posibles resultados del experimento aleatorio, o sea:

$$P[A] = \frac{\text{Número de casos favorables a } A}{\text{Número de casos posibles}} \quad (3.1)$$

Una desventaja importante del concepto clásico de probabilidad es su limitada aplicación, ya que hay muchas situaciones en que las posibilidades que se presentan no pueden considerarse igualmente probables.

La probabilidad subjetiva son probabilidades que expresan la fuerza de nuestra creencia en relación con incertidumbres que están asociadas “suposiciones razonadas”, quizá intuición y otros factores subjetivos del fenómeno de estudio, está puede ser muy diferente a la probabilidad asignada por otra persona.

Los elementos básicos de la teoría de probabilidad son los resultados del proceso o fenómeno en estudio. Cada posible resultado de un experimento se llama evento.

Un evento simple esta formado solamente por un resultado posible. En tanto, un evento compuesto es la unión de conjunto de eventos simples. El conjunto de todos los eventos posibles se llama espacio probabilístico. La forma en que se divide

el espacio probabilístico, depende del tipo de probabilidades que se van a determinar.

La probabilidad simple, significa la probabilidad de ocurrencia de un evento simple.

La probabilidad compuesta resulta la ocurrencia que contiene dos o más eventos simples.

Esos términos se podrán comprender mejor mediante el siguiente ejemplo. Considere el experimento aleatorio del lanzamiento de un dado al aire, que tiene 6

$$E = \{1,2,3,4,5,6\}$$

resultados posibles (1 punto, 2 puntos, 3 puntos, 4 puntos, 5 puntos y 6 puntos):

El evento 3 puntos, es un evento simple porque está formado por solo un resultado. En tanto, el evento un número par es compuesto porque es la unión de los eventos simples: 2 puntos, 4 puntos y 6 puntos.

3.2 Definición axiomática de probabilidad

Para hacer una definición rigurosa de la probabilidad, necesitamos precisar ciertas leyes o axiomas que deba cumplir una función de probabilidad. Intuitivamente estos axiomas deberían implicar, entre otras, las siguientes cuestiones, que nos parecen lógicas en términos de lo que se puede esperar de una función de probabilidad:

- La probabilidad sólo puede tomar valores comprendidos entre 0 y 1 (no pueden haber sucesos cuya probabilidad de ocurrir sea 1,5 ni -0,50).
- La probabilidad del suceso seguro es 1.
- La probabilidad del suceso imposible debe ser 0.
- La probabilidad de la intersección de dos sucesos debe ser menor o igual que la probabilidad de cada uno de los sucesos por separado, es decir,

$$P[A \cap B] \leq P[A] \tag{3.2}$$

$$P[A \cap B] \leq P[B] \tag{3.3}$$

- La probabilidad de la unión de sucesos debe ser mayor que la de cada uno de los sucesos por separado:

$$P[A \cup B] \geq P[A] \tag{3.4}$$

$$P[A \cup B] \geq P[B] \tag{3.5}$$

- La probabilidad del suceso contrario de A, deber ser

$$P[\bar{A}] \leq 1 - P[A] \tag{3.6}$$

- La probabilidad de la unión numerable de sucesos disjuntos es la suma de sus probabilidades.

$$A_1, A_2, \dots, A_n \in \mathcal{A} \Rightarrow \text{Prob} \left[\bigcup_{i=1}^{\infty} A_i \right] = \sum_{i=1}^{\infty} \text{Prob}[A_i] \tag{3.7}$$

En las últimas líneas fueron mencionadas ciertas propiedades que deberían cumplir una función de probabilidad. Debe tenerse en cuenta entonces que se han de seguir los siguientes enunciados:

1. La función de probabilidad debe calcularse sobre subconjuntos del espacio probabilístico. No es estrictamente necesario que sean todos, pero sí es necesario que si se puede calcular sobre un conjunto, lo pueda ser también sobre su complementario, y que si se puede calcular sobre dos conjuntos A y B, que también se pueda calcular sobre su unión y su intersección.
2. Entre las leyes que cumplen una función de probabilidad, como ya fue citado, algunas son redundantes pues se pueden deducir de las demás.

Con la definición axiomática de la probabilidad se pretende dar el menor conjunto posible de estas reglas, para que las demás se deduzcan como una simple consecuencia de ellas.

3.3 Teoremas fundamentales del cálculo de probabilidades

Hay algunos resultados importantes del cálculo de probabilidades que son conocidos bajo los nombres de teorema de la adición y teorema de la multiplicación. Veamos cuales son estos teoremas.

3.3.1 Teorema de la adición

La probabilidad de la unión de sucesos indica que la probabilidad de ocurrencia de un evento “A” o “B”, es igual a la probabilidad de ocurrencia de “A”, más la probabilidad de ocurrencia de “B”, menos la probabilidad que los dos eventos se presenten al mismo tiempo. Esta regla se expresa simbólicamente en los siguientes términos:

$$P[A \cup B] = P[A] + P[B] - P[A \cap B] \quad (3.8)$$

Siempre que la ocurrencia de A y B a la vez sea imposible, la probabilidad $P(AB)$ es igual a cero; estos eventos se llaman mutuamente excluyentes, porque la ocurrencia de un evento impide la ocurrencia del otro.

Ejemplo 1. Un químico está interesado en analizar muestras de agua en una planta eléctrica de vapor. Las pruebas serán engañosas si se encuentra presente nitrógeno o silicio en las muestras. Los dispositivos automáticos indican que la probabilidad de que ambos elementos se encuentren presentes simultáneamente es despreciable. ¿Escriba una expresión general para la probabilidad de que las pruebas hayan sido engañosas?

$$P[N \text{ o } Si] = P[N] + P[Si]$$

Ejemplo 2. En una universidad el 65% de los alumnos practican el balompié, el 25% la natación y el 12% los dos deportes. ¿Cuál es la probabilidad de encontrar alumno que practique balompié o natación?

Solución:

Sea A el suceso practicar balompié:

$$Prob[A] = 0,65$$

Sea B el suceso practicar natación:

$$Prob[B] = 0,25$$

El suceso practicar balompié y natación es $A \cap B$:

$$Prob[A \cap B] = 0,12$$

Así:

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[A \cap B] \\ &= 0,65 + 0,25 - 0,12 \\ &= 0,78 \end{aligned}$$

3.3.2 Teorema de la multiplicación

Se refiere a dos o más eventos que podrían presentarse al mismo tiempo. Si la ocurrencia de uno no afecta la probabilidad que suceda otro evento cualquiera, se dice que son eventos estadísticamente independientes. En forma matemática, lo anterior puede expresarse como el producto de las probabilidades marginales así:

$$P[AB] = P[A].P[B] \tag{3.9}$$

La dependencia estadística existe cuando la probabilidad de un evento depende o está asociada a la ocurrencia de otros eventos. La probabilidad conjunta en condiciones de dependencia estadística:

$$P[AB] = P[A].P[B/A] \tag{3.10}$$

Entonces, la probabilidad conjunta de que los eventos A y B ocurran simultáneamente o en sucesión, es igual a la probabilidad que suceda el evento A, multiplicado por la probabilidad del evento B si ya tuvo lugar el evento A.

Ejemplo. En una planta productora de plástico, 3 tuberías vacían distintos productos químicos en un recipiente mezclador. Cada tubería tiene un control de 2 posiciones que mide la tasa de flujo que se hace llegar al recipiente. Cierta día, al experimentar con varias mezclas, se obtuvo una solución que emitió un gas venenoso. No se registraron las posiciones de los controles. ¿Cuál es la probabilidad de obtener de nuevo esta solución al experimentar aleatoriamente?

Suponiendo que las probabilidades de cada posición de las tuberías sean iguales ($1/2$) y usando la ecuación 3.9, se puede establecer la solución del problema en la siguiente forma:

$$\begin{aligned}P[T_1 T_2 T_3] &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\ &= \left(\frac{1}{2}\right)^3 \\ &= 0,125\end{aligned}$$

3.4 Distribuciones de probabilidad

En el capítulo 2, se trató de las distribuciones empíricas de frecuencia de muestras. Ahora, se estudia las distribuciones teóricas de frecuencia y sus propiedades. Cuando una distribución de frecuencia se basa en una muestra, se trata, en cierto sentido, de una estimación de la distribución de población correspondiente.

En la mayoría de los problemas estadísticos la muestra no es lo suficientemente grande para poder determinar la distribución de población con mucha precisión. Sin embargo, existe, por lo general, bastante información en la muestra, junto con la información obtenida de otras fuentes, para postular la naturaleza general de la distribución de población. Estos postulados dan lugar a las distribuciones de probabilidad.

Una distribución de probabilidad o teórica es un modelo matemático para la distribución de frecuencia real. En efecto, puede concebirse una distribución de probabilidad como una distribución teórica de frecuencia.

Una distribución teórica de frecuencia es una distribución de probabilidad que describe cómo se espera que varíen los resultados del experimento aleatorio. Dado que estas clases de distribuciones se ocupan de las expectativas, son modelos de gran utilidad para hacer inferencias y tomar decisiones en condiciones de incertidumbre.

Las distribuciones de probabilidades están formadas por los valores que toma la variable aleatoria y su respectiva probabilidad. Una variable aleatoria es una función con valor numérico sobre un espacio de muestreo.

Atendiendo a la clasificación de las variables aleatorias discretas y continuas se describen las principales leyes de probabilidad de cada una de ellas, las cuales constituirán el soporte subyacente de la inferencia estadística y a las que será necesario hacer referencia en el estudio de dicho bloque.

3.5 Distribuciones discretas

Se denomina distribución de variable discreta a aquella cuya función de probabilidad sólo toma valores positivos en un conjunto de valores de X finito. A dicha función se le llama función de masa de probabilidad. En este caso la distribución de probabilidad es la suma de la función de masa, por lo que tenemos entonces que:

$$F(x) = P(X \leq x) = \sum_{k=-\infty}^x f(k)$$

Y, tal como corresponde a la definición de distribución de probabilidad, esta expresión representa la suma de todas las probabilidades desde $-\infty$ hasta el valor x .

3.5.1 Distribución binomial

La distribución binomial es uno de los primeros ejemplos de las llamadas distribuciones discretas (que sólo pueden tomar un número finito de valores). Fue estudiada por Jakob Bernoulli (Suiza, 1654-1705), quién escribió el primer tratado importante sobre probabilidad, “Ars coniectandi” (El arte de pronosticar). Los Bernoulli formaron una de las leyendas de los matemáticos más importantes de la historia.

Existen muchos problemas aplicados en los que interesa la probabilidad de que un evento ocurra “ x veces en n ensayos”. En otras palabras, se está interesado en lograr x aciertos y $n - x$ fracasos en n ensayos.

Una variable discreta tiene distribución binomial cuando cumple con las siguientes condiciones:

1. El experimento consta de n ensayos o pruebas idénticas.

2. Cada ensayo puede tener uno de dos resultados. Un resultado se llama “éxito”, y al otro, “fracaso”.
3. La probabilidad de un éxito en un ensayo es igual a p y permanece constante de uno a otro ensayo. La probabilidad de un fracaso es $q(1 - p)$.
4. Los ensayos son estadísticamente independientes.
5. Interesa conocer x , el número de éxitos observados en n pruebas.

El resultado éxito no tiene que verse como un acierto o logro, en el sentido usual de la palabra. Además, los ensayos independientes significan que el resultado de uno de ellos no afecta la probabilidad de éxito de cualquier otro en el experimento.

Si realizamos n veces un experimento en el que podemos obtener “éxito” E , con probabilidad p y fracaso F , con probabilidad $q(1 - p)$, diremos que estamos ante una distribución binomial de parámetros n y p . La fórmula para obtener x “éxitos” viene dada por:

$$f(x) = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, \dots, n \tag{3.11}$$

Los parámetros de la distribución son: Media $\mu = np$ y desviación estándar $\sigma = \sqrt{npq}$.

Los cálculos necesarios para obtener las probabilidades de valores particulares de n y x , resultan bastante tediosos; razón por lo cual sean incluido la Tabla 1 en el apéndice, con la distribución binomial para $n \leq 20$ y valores disponibles de p que principian con 0,05 y proceden por pasos de 0,05 hasta el valor 0,95. Para utilizar la tabla 1 deben considerarse las reglas siguientes:

Regla	Descripción	Ejemplo
1	$P(x = k) = F(k) - F(k - 1)$	$P(x = 3) = F(3) - F(2)$
2	$P(x \leq k) = F(k)$	$P(x \leq 6) = F(6)$
3	$P(x \geq k) = 1 - F(k - 1)$	$P(x \geq 4) = 1 - F(3)$
4	$P(k_1 \leq x \leq k_2) = F(k_2) - F(k_1 - 1)$	$P(5 \leq x \leq 10) = F(10) - F(4)$

Estas reglas son necesarias ya que en la tabla se representan las probabilidades acumuladas, es decir,

$$P(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i q^{n-i}$$

Ejemplo. Un agente químico produce la decoloración del 5% de los rollos de papel producido por cierta compañía. ¿Cuál es probabilidad de que en una muestra aleatoria de 7 rollos de papel, 1 rollo presente decoloración debido al agente químico?

Este problema se puede considerar como un problema de distribución binomial para el cual $n = 7$ y $p = 0,05$. Por medio de la fórmula 3.11 la probabilidad de 1 rollo decolorado está dada por,

$$\begin{aligned} P(x = 1) &= \binom{7}{1} (0,05)^1 (0,95)^{7-1} \\ &= \frac{7!}{6! 1!} (0,05)^1 (0,95)^6 \\ &= 7(0,05)(0,7351) \\ &= 0,257 \end{aligned}$$

Por lo tanto, hay una probabilidad de 0,257 de que de los 7 rollos de papel ocurra que uno presente decoración debido al agente químico.

Utilizando la tabla 1, se tiene que encontrar primero $n = 7$ y la probabilidad de 0,05. Después se recorre hacia abajo esa columna hasta llegar al renglón $x = 1$. La respuesta allí es 0,956, cifra que representa la probabilidad de 0 o 1 rollo decolorado. Hay que restar 0,698, la probabilidad de ningún rollo decolorado; entonces resulta 0,258.

La elección de éxito o fracaso es subjetiva y queda a juicio del investigador. En general, la distribución binomial tiene poca aplicación en el análisis químico de laboratorio, su principal uso está en el control estadístico de calidad, en la detección de porcentajes de productos “defectuosos” en lotes de producción.

3.5.2 Distribución Poisson

La distribución de Poisson representa la probabilidad de que un evento aislado (o variable aleatoria discreta) ocurra un número específico de veces en un intervalo de tiempo, espacio o distancia, dado un promedio por unidad de medida.

Un rasgo característico de esta distribución es el hecho de que la ocurrencia de un evento puede ser cuantificada; no así su no-ocurrencia, dada su falta de significado en términos físicos. De este modo, no puede medirse el número total de eventos n , en consecuencia, no es posible aplicar con precisión la distribución binomial.

Los detectores de radiación como, por ejemplo, los contadores de Geiger, siguen generalmente una distribución de Poisson. En este tipo de distribución, la varianza es de igual magnitud que la media. Si x es el número de ocurrencias cuya probabilidad es necesario conocer, puede expresarse la distribución de Poisson como

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \rightarrow x = 0, 1, \dots \quad (3.12)$$

Los parámetros de la distribución tienen la característica, que tanto la media y desviación estándar son iguales a λ .

$$\mu = \sigma = \lambda$$

Lambda (λ) representa el número medio de ocurrencias por intervalo de tiempo. Los términos de tal distribución, se derivan de un proceso que constituye un mecanismo físico aleatorio, en el cual los eventos ocurren al azar en una escala de tiempo (o de espacio y distancia).

Ejemplo. Supóngase que partículas radiactivas dan en cierto blanco a una tasa promedio de 3 partículas por minuto. ¿Cuál es la probabilidad de que 2 partículas den en el blanco durante un minuto cualquiera? Aplicando la fórmula 3.12, se puede calcular la probabilidad de exactamente dos partículas:

$$\begin{aligned}
 P(x) &= \frac{3^2 e^{-3}}{2!} \\
 &= \frac{6(0,04979)}{2} \\
 &= 0,224
 \end{aligned}$$

Desde luego, puede resolverse este problema utilizando la tabla 2 del apéndice; la distribución de Poisson para diversos valores de lambda. Se utilizan las reglas de cálculo de la distribución binomial citadas en la página 60.

Primero, se localiza el valor de $\lambda = 3$, la probabilidad asociada con $x = 2$. Sin embargo, la tabla presenta la función de probabilidades acumuladas, por lo cual se debe restarle la probabilidad correspondiente a $x = 1$, de la siguiente forma,

$$\begin{aligned}
 P(x = 2) &= 0,4232 - 0,1991 \\
 &= 0,224
 \end{aligned}$$

3.6 Distribuciones continuas

Se denomina variable continua a aquella que puede tomar cualquiera de los infinitos valores existentes dentro de un intervalo. En el caso de variable continua la distribución de probabilidad es la integral de la función de densidad, por lo que tenemos entonces que:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

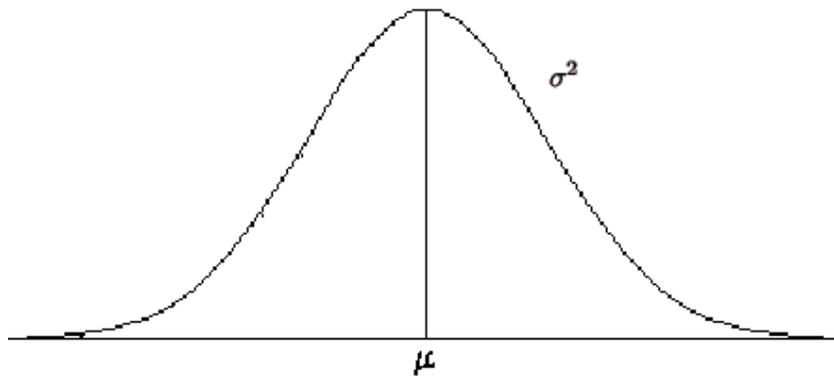
Cuando la variable aleatoria toma valores en el conjunto de los números reales, la distribución de probabilidad está completamente especificada por la función de distribución, cuyo valor en cada real x es la probabilidad de que la variable aleatoria sea menor o igual que x

3.6.1 Distribución normal

Esta distribución resulta útil no sólo porque un gran número de distribuciones de frecuencias presentan formas aproximadamente normales, sino también por su gran significado teórico en el campo de la estadística inferencial. La distribución normal es un ejemplo de las distribuciones continuas, y aparece en multitud de fenómenos de investigación. Fue estudiada, entre otros, por J.K.F.

Gauss (Alemania, 1777-1855), uno de los más famosos matemáticos de la historia. La gráfica de la distribución normal en forma de campana se denomina Campana de Gauss.

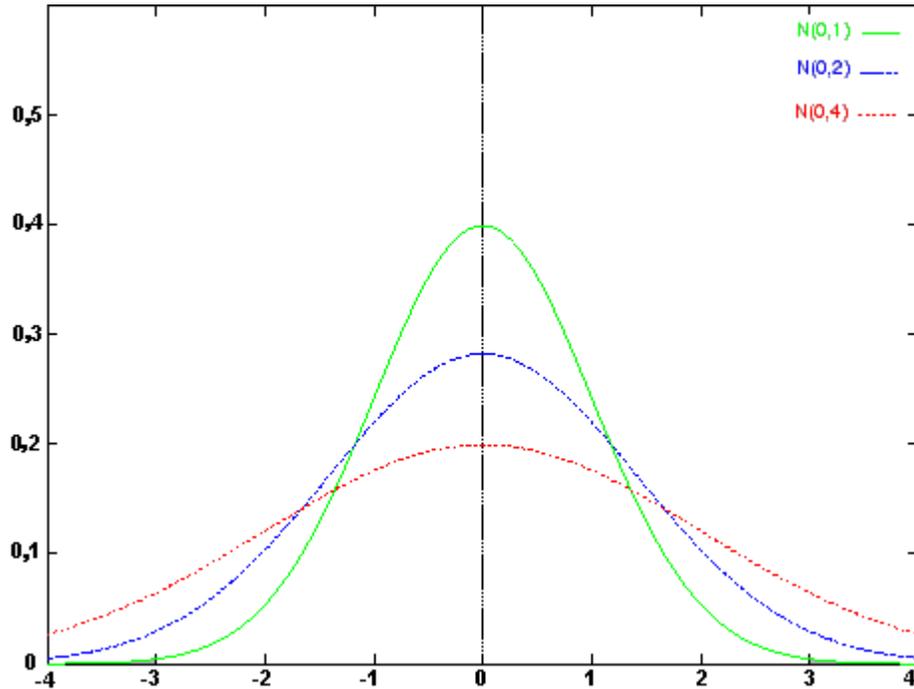
La distribución normal, o de Gauss, es una de las más útiles, y gran parte de la estadística matemática se basa en ella. La normal es en muchos aspectos la piedra angular de la estadística. La gráfica tiene una forma semejante al perfil de una campana.



El soporte de la distribución es todo \mathfrak{R} , de modo que la mayor parte de la masa de probabilidad (área comprendida entre la curva y el eje de abscisas) se encuentra concentrada alrededor de la media, y las ramas de la curva se extienden asintóticamente a los ejes, de modo que cualquier valor “muy alejado” de la media es posible (aunque poco probable). La forma de la campana de Gauss depende de los parámetros μ y σ .

Se ha encontrado experimentalmente que la función de distribución normal describe satisfactoriamente aquellos sistemas en los que las mediciones en estudio vienen afectadas por un número grande de errores que actúan todos independientemente.

Distribuciones gaussianas con igual media pero varianza diferente

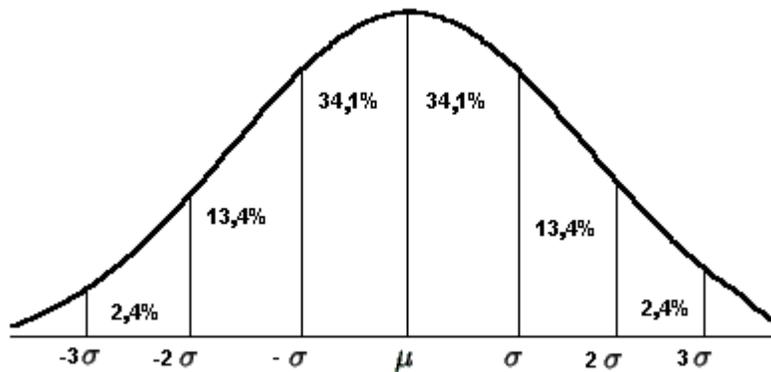


La distribución normal es una aproximación bastante buena a la distribución real de las fluctuaciones debidas al azar en los fenómenos físicos. La ecuación de la ley de distribución normal se puede escribir como sigue:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty \quad (3.13)$$

donde μ (*mu*) es la media aritmética y σ (*sigma*) es la desviación estándar (σ^2 es la varianza).

El área bajo la curva representa la probabilidad de que ocurra una observación dentro de los límites del área. El punto central en la distribución es la media, mientras que las distancias de la media se expresan en función de las desviaciones estándar. La siguiente figura ejemplifica la forma de la curva y las áreas correspondientes a los múltiplos de la desviación estándar a la izquierda y a la derecha de la media.



La distribución normal se define por la media de la población μ y la desviación estándar de la población σ . Puesto que existe infinito número de combinaciones de μ y σ , existe un número infinito de distribuciones normales. Un método de solucionar el problema es la estandarización los datos con el fin de obtener la distribución normal estándar.

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0,1) \quad (3.14)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty \quad (3.15)$$

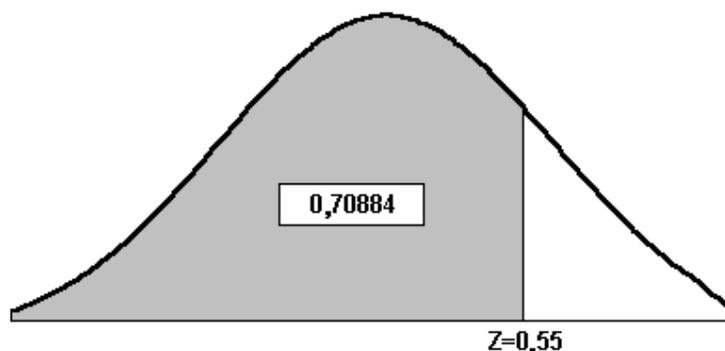
Para la utilización en problemas prácticos de la función de la distribución normal estándar, existen tablas donde se ofrecen (con varios decimales de precisión) los valores de la función de distribución de la normal para una serie limitada de valores x dados.

3.6.2 Uso de la tabla normal estándar

La Tabla 3 del apéndice ofrece las áreas bajo la curva normal estándar que corresponden a los diversos valores de Z . Puesto que la curva es simétrica (un lado es la imagen del otro), en la tabla se presentan primero las áreas con valores de Z negativos y en la siguiente página las áreas para valores de Z positivos.

Para usar la Tabla 3, se debe tener en cuenta que los valores Z se deben calcular con dos decimales.

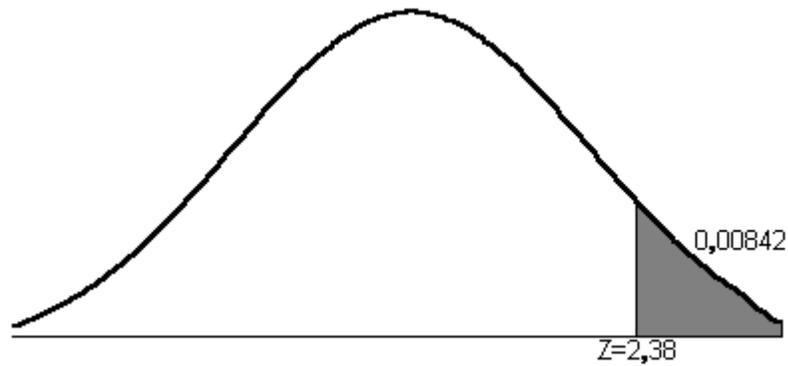
Ejemplo 1. Calcular la probabilidad de Z menor a 0,55, $P(Z \leq 0,55)$. Para determinar la probabilidad o área bajo la curva desde $-\infty$ a 0,55, se sigue hacia abajo la columna Z de la tabla normal estándar, hasta que se encuentra el valor de Z de interés (en décimas).



Por tanto, hay que detenerse en la fila $Z = 0,5$. Después se lee esta hilera en sentido horizontal hasta que se intercepta con la columna que contiene valor de Z en centésimas, 0,05. Entonces, en el cuerpo de la tabla, se encuentra que la probabilidad $P(Z \leq 0,55) = 0,70884$.

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54395	0,54776	0,55172	0,55567	0,55962	0,56356	0,56750	0,57124	0,57534
0,2	0,57926	0,58617	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61781	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80510	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82124	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891

Ejemplo 2. ¿Cuál es la probabilidad de que Z sea mayor o igual a 2,38, $P(Z \geq 2,38)$? Primero, para encontrar esta probabilidad hay que recordar la regla del complemento para el cálculo de probabilidades (fórmula 3.6). Gráficamente, la probabilidad se representa como,



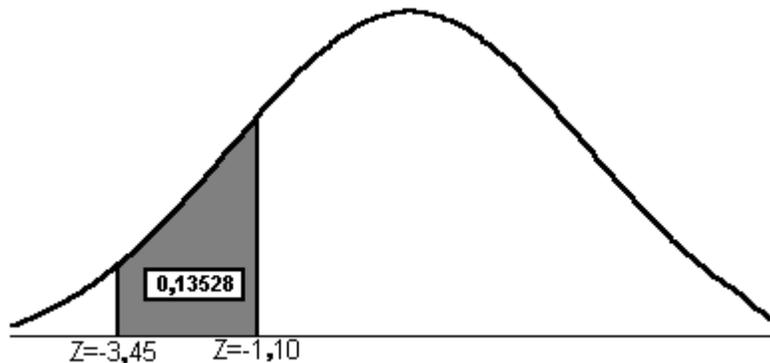
Entonces,

$$P(z \geq 2,38) = 1 - P(z \leq 2,38)$$

Se busca en la primera columna, la fila que corresponda a $Z = 2,3$. Seguidamente, en la columna que corresponda al segundo decimal, $0,08$. Entonces, la probabilidad es $0,99158$.

$$\begin{aligned} P(Z \geq 2,38) &= 1 - P(z \leq 2,38) \\ &= 1 - 0,99134 \\ &= 0,00866 \end{aligned}$$

Ejemplo 3. ¿Cuál es la probabilidad de que Z se encuentre entre $-3,45$ y $-1,10$, $P(-3,45 \leq Z \leq -1,10)$?



Como regla general para calcular probabilidades utilizando un intervalo, se tiene que,

$$P(Z_1 \leq Z \leq Z_2) = P(Z \leq Z_2) - P(Z \leq Z_1)$$

(3.16)

Seguidamente, se encuentra en la tabla las probabilidades asociadas a valores de Z que componen el intervalo. Como los valores de interés tienen signo negativo, es utilizada la parte negativa de la tabla normal estándar.

En la tabla se encuentre que la probabilidad asociada a $-3,45$ es $0,00039$, y para $-1,10$ es $0,13567$.

Entonces,

$$\begin{aligned}P(-3,45 \leq Z \leq -1,10) &= P(Z \leq -1,10) - P(Z \leq -3,45) \\ &= 0,13567 - 0,00039 \\ &= 0,13528\end{aligned}$$

3.7 Aplicaciones de la distribución normal estándar

Ahora que ya se ha aprendido a usar la tabla normal estándar, pueden resolver problemas relacionados con la distribución normal.

Ejemplo 1. Suponga que el tiempo medio de reacción de una sustancia es de 30 segundos y la desviación estándar de 4 segundos. ¿Cuál es la probabilidad de que en cierto experimento, la reacción se produzca en menos de 25 segundos?

Para resolver el problema, se utiliza la fórmula 3.8 para estandarizar el tiempo de reacción de la sustancia de la siguiente manera,

$$\begin{aligned}P(x \leq 25) &= P\left(Z \leq \frac{25-30}{4}\right) \\ &= P(Z \leq -1,25) \\ &= 0,10565\end{aligned}$$

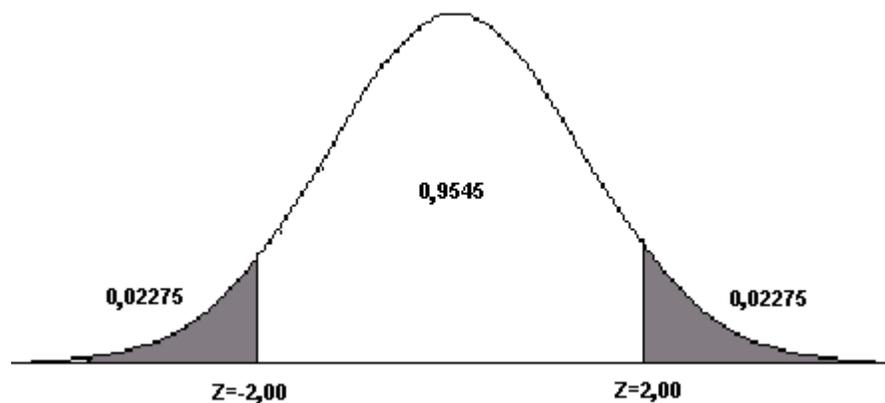
La probabilidad que la reacción se produzca en menos de 25 segundos es $0,10565$.

Ejemplo 2. Una analista químico realiza un gran número de medidas de una solución con absorbancia media de $0,435$ y una desviación estándar de $0,005$. ¿Qué porcentaje de las lecturas se estima fuera del intervalo $0,425$ y $0,445$?

Los cálculos dan

$$\begin{aligned}
P(0,425 \leq x \leq 0,445) &= P\left(\frac{0,425 - 0,435}{0,005} \leq Z \leq \frac{0,445 - 0,435}{0,005}\right) \\
&= P(-2,00 \leq Z \leq 2,00) \\
&= 0,97725 - 0,02275 \\
&= 0,9545
\end{aligned}$$

Así pues, el porcentaje de lecturas fuera de intervalo 0,425 y 0,445 es 4,55%. Simplemente, se calcula restando 1-0,9545. Gráficamente, el porcentaje fuera del intervalo es,



3.8 Aproximación de la distribución Binomial por la Normal

Cuando n es grande y p es constante el comportamiento de una distribución binomial es asemeja al de distribución normal con media $\mu = np$ y desviación estándar $\sigma = \sqrt{npq}$. Esto lo demostró el matemático Abraham de Moivre (1667-1750). La aproximación se puede aplicar (es buena aproximación) sólo si $np \geq 5$ y $nq \geq 5$. Si no se cumplen estas condiciones no podemos aproximar la binomial a través de la distribución normal.

Dado que por mucho que se parezca nunca es igual, una distribución binomial (discreta) que una distribución normal (continua), es necesario aplicar en el cálculo de probabilidades un ajuste que recibe el nombre de corrección por continuidad o Yates, que consiste en hacer determinados ajustes para que la aproximación realizada sea lo más precisa posible. El ajuste por continuidad se presenta en las siguientes reglas:

Regla	Descripción	Ejemplo
1	$P(x = k) = P(k - 0,5 \leq x \leq k + 0,5)$	$P(x = 5) = P(4,5 \leq x \leq 5,5)$
2	$P(x \leq k) = P(x \leq k + 0,5)$	$P(x \leq 7) = P(x \leq 7,5)$
3	$P(x \geq k) = P(k - 0,5)$	$P(x \geq 3) = P(x \geq 2,5)$
4	$P(k_1 \leq x \leq k_2) = P(k_1 - 0,5 \leq x \leq k_2 + 0,5)$	$P(3 \leq x \leq 6) = P(2,5 \leq x \leq 6,5)$

Así, usando la regla 1, si nos piden $P(x = k)$ en una distribución binomial, y aproximamos X por una distribución normal, no podemos calcular directamente $P(x = k)$ porque, como ya se ha mencionado anteriormente, en una distribución continua la probabilidad de un valor exacto (o entero) es 0. La corrección por continuidad consiste en tomar un pequeño intervalo de longitud 1 alrededor del punto k . De modo que, si nos piden $P(x = k)$ con X binomial, con la aproximación normal deberemos calcular $P(k - 0,5 \leq x \leq k + 0,5)$.

Ejemplos.

1. Suponga que la probabilidad de que los resultados de un experimento sean aceptables es 0.55. Si el experimento se repite 60 veces. ¿Cuál es la probabilidad de obtener al menos 28 resultados útiles?

Solución. Es una distribución binomial, los experimentos sólo pueden ser aceptables o no aceptables. La n es grande y por ello, podemos hacer la aproximación de la binomial a la normal.

Comprobamos que $np \geq 5 \Rightarrow 60(0,55) = 33$ y $nq \geq 5 \Rightarrow 60(0,45) = 27$. Seguidamente calculamos la media y la desviación estándar de la distribución normal.

$$\mu = np = 60(0,55) = 33$$

$$\sigma = \sqrt{npq} = \sqrt{60(0,55)(0,45)} = 3,85357$$

Entonces, la probabilidad se calcula como,

$$P(x \geq 28) = P(x \geq 27,5) = P\left(Z \geq \frac{27,5 - 33}{3,85357}\right)$$

$$\begin{aligned}
&= P(z \geq -1,43) = 1 - P(z \leq -1,43) \\
&= 1 - 0,07636 = 0,92364
\end{aligned}$$

2. Una compañía produce un compuesto químico y está preocupada por su contenido de impurezas. Con el fin de cumplir las normas establecidas, es importante que la proporción de productos con impurezas de un lote de 1000 unidades no sea superior al 2,5%. ¿Cuál es la probabilidad de encontrar como máximo 20 unidades con impurezas?

Solución. Es una distribución binomial, las unidades del producto pueden tener o no impurezas. La n es grande y por ello, podemos hacer la aproximación de la binomial a la normal.

Proamos que $np \geq 5 \Rightarrow 1000(0,025) = 25$ y $nq \geq 5 \Rightarrow 1000(0,975) = 975$. Seguidamente calculamos la media y la desviación estándar de la distribución normal.

$$\begin{aligned}
\mu &= np = 1000(0,025) = 25 \\
\sigma &= \sqrt{npq} = \sqrt{1000(0,025)(0,975)} = 4,9371
\end{aligned}$$

Entonces, la probabilidad se calcula como,

$$\begin{aligned}
P(x \leq 20) &= P(x \leq 20,5) = P\left(Z \leq \frac{20,5 - 25}{4,9371}\right) \\
&= P(z \leq -0,91) \\
&= 0,18141
\end{aligned}$$

3.9 La distribución t de Student

La teoría de la distribución normal se desarrolla a partir de tamaños de muestra suficientemente grandes, generalmente mayores a 30 observaciones y no aplicable a muestras pequeñas.

En el laboratorio no hay libertad de realizar un gran número de observaciones y, por ello, las pruebas de hipótesis estadísticas basadas en la distribución normal llevarían al químico a falsas conclusiones.

El hecho fue reconocido por W. S. Gosset, un químico irlandés que en 1908 publicó, bajo el pseudónimo de Student, un trabajo titulado “El error probable de una medida”. En parte por consideraciones teóricas y, en parte, por el uso de muestras aleatorias, obtuvo la distribución teórica del promedio de tamaños de muestra pequeñas ($n \leq 30$), ajustada a una distribución normal.

La distribución de Student tiene propiedades parecidas a $N(0,1)$:

- Es de media cero, y simétrica con respecto a la misma;
- Es algo más dispersa que la normal, pero la varianza decrece hasta 1 cuando el número de grados de libertad aumenta;
- Para un tamaño de muestra grande se puede aproximar la distribución de

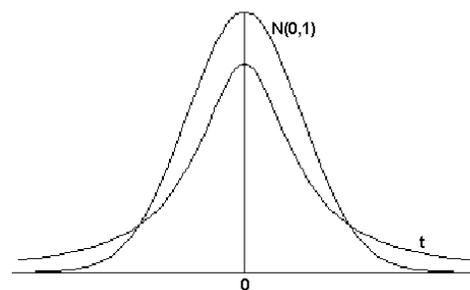
Student por la normal, es decir, $t_n \xrightarrow{n \rightarrow \infty} N(0,1)$

La curva t presenta menos dispersión que la curva normal, pero se aproxima a ella cuando crece el tamaño de la muestra y se hace igual cuando “ n ” tiende a infinito. Para fines prácticos es usada, corrientemente, la distribución normal con muestras de tamaño mayor a 30.

La teoría en la que está basada la distribución t , requiere presuponer que la variable básica x posee una distribución normal y, por consiguiente, el empleo de la distribución t sólo puede justificarse cuando es posible asegurar que x se distribuye por lo menos, en una forma aproximadamente normal.

La Tabla 4 del apéndice, da los valores de t asociados a los grados de libertad y a varias probabilidades o “valores críticos”. Los grados de libertad están dados por la fórmula $n - 1$. Esto equivale al empleo del divisor $n - 1$ en vez de n al definir la desviación estándar de una muestra, tal como se hizo en el capítulo 2.

Comparación entre las funciones de densidad de t y $N(0,1)$



3.10 PROBLEMAS

3.1 Suponga que el número de moles de sulfato de cobre II $[\text{Cu}(\text{SO}_4)]$, en un volumen dado de una solución, tiene una media de 0,17 moles de esa sustancia y la desviación estándar de 0,01 moles. ¿Cuál es la probabilidad de que el número de moles de cobre II sea mayor a 0,18?

3.2 Según una asociación de protección ambiental, el cloroformo, del cual se sospecha que en forma gaseosa sea un agente causante de cáncer, está presente en pequeñas cantidades en todas las fuentes de agua públicas. Si la media y la desviación estándar de las cantidades de cloroformo presentes en las fuentes de agua son de 34 y 53 microgramos por litro, respectivamente. Suponga que se toman 30 muestras de agua, ¿cuántas tendrán entre 33 y 37 microgramos por litro de cloroformo?

3.3 Dada la distribución de probabilidad para la variable aleatoria y:

Y	0	1	2	3	4
P(y)	1/9	2/9	3/9	2/9	1/9

Calcule las siguientes probabilidades: a) $P(y \leq 2)$, b) $P(y > 1)$ y c) $P(1 \leq y \leq 3)$.

3.4 Un dispositivo para la detección de incendios utiliza tres células sensibles a la temperatura, que actúan independientemente, de manera que una o varias de ellas pueden accionar la alarma. Cada célula tiene una probabilidad de 0,80 de accionar la alarma cuando la temperatura llega a los 60 grados o más. Sea x igual al número de células que accionan la alarma cuando la temperatura llega a los 60 grados.

a) Encuentre la distribución de probabilidad de x .

b) Calcule el valor esperado y la varianza.

3.5 Se ha encontrado que el número de cierto tipo de bacterias en el agua de servicio público tiene una distribución de Poisson. Supongamos que el número máximo permisible por muestra de agua para este tipo de bacteria

- es 5 y que el número medio de bacterias por muestra que reciben las viviendas son 2. Si fuera analizada una muestra, ¿cuál es probabilidad de que el número exceda el máximo tolerable?
- 3.6 Usando una técnica de trasplante nuclear, el núcleo de una célula de uno de los estados finales del desarrollo de un embrión se trasplanta a un huevo fertilizado para ver si ese puede desarrollarse normalmente. Si la probabilidad de que un trasplante tenga éxito es 0,683, ¿cuál es probabilidad de que como máximo 20 trasplantes de un total de 75 fracasen?
- 3.7 Se realiza un experimento para comparar la efectividad de un compuesto químico para producir resistencia en porcentaje de humedad en productos textiles. Suponga que las medidas de resistencia a la humedad tienen una distribución normal con media de 11,9 y variancia 1,21. ¿Qué porcentaje de las medidas de resistencia a la humedad es menor que 11,8?
- 3.8 Se realiza un experimento para investigar el efecto tóxico de una sustancia química sobre la piel de las ratas. La sustancia se aplicó en secciones de piel de una pulgada cuadrada y para cada sección se anotó una puntuación entre 0 y 10, dependiendo del grado de irritación. Por experiencia anterior se sabe que la puntuación sigue una distribución de Poisson con media de 3,5 por pulgada cuadrada. ¿Cuál es la probabilidad de encontrar una puntuación de 4 en 2 pulgadas cuadrada?
- 3.9 Una compañía de productos químicos elimina sus residuos en un río situado en la vecindad de su ubicación. Para verificar el grado de contaminación creado por estos residuos, se desea obtener algunas estimaciones. Suponga que los residuos de contaminante tienen una distribución normal con media de 1,7 gramos por litro (g/l) y desviación estándar de 0,223 gramos por litro (g/l) ¿Cuál es la probabilidad de que los

- residuos del contaminante: a) Sean mayores de 1,72 b) Entre 1,46 y 1,88
c) Inferiores de 2,05
- 3.10 Sí la probabilidad de que el vapor se condense en un tubo delgado de aluminio a 10 atm de presión es 0,30. ¿Cuál es probabilidad de que en las condiciones establecidas el vapor se condense?
a) En 5 de 9 de tales tubos
b) Entre 20 y 30 (inclusive) de 40 de tales tubos
- 3.11 Al inspeccionar la aplicación de estaño (**Sn**) por un proceso electrolito continuo, se descubren en promedio 1,4 imperfecciones por minuto. Calcule las probabilidades de encontrar:
a) Una imperfección en 2 minutos.
b) Al menos dos imperfecciones en 4 minutos.
- 3.12 El número de rayos gamma que emite por segundo cierta sustancia radiactiva es una variable aleatoria que tiene distribución de Poisson con lambda igual a 5,5. Si un detector deja de operar cuando hay más de 10 rayos por segundo, ¿cuál es la probabilidad de que este instrumento deje de funcionar durante un segundo cualquiera?
- 3.13 El 73% de todas las nubes impregnadas con yoduro de plata muestran un significativo crecimiento. Entre 60 nubes impregnadas con yoduro de plata, ¿cuál es la probabilidad que de 20 a 40 (inclusive) muestren un crecimiento?
- 3.14 Como parte de un análisis de riesgo con respecto a una planta (**AgIO₃**) de energía nuclear, los ingenieros deben modelar la resistencia de los soportes de un generador de vapor en función de su capacidad de resistir la aceleración máxima ocasionada por temblores. La opinión de los expertos sugiere que la aceleración se distribuye normalmente con promedio 4,0 y

variancia 0,09. ¿Cuál es la probabilidad de que los soportes resistan una aceleración máxima de 3,3?

- 3.15 En ciertos experimentos el error cometido al determinar la densidad en g/ml de una sustancia es una variable aleatoria cuya distribución es normal con promedio 0,015 y desviación estándar de 0,045. Cuáles son las probabilidades de que tal error:
- Este entre 0,011 y 0,013
 - Sea como máximo 0,017
- 3.16 Una compañía que fabrica purificadores de agua, desea estimar el número medio de gramos por litro de impurezas que elimina su máquina. La cantidad que elimina tiene una distribución normal con $\mu=6,6$ y $\sigma^2=3,2$ gramos. ¿Cuál es la probabilidad de que el purificador elimine entre 5,5 y 7,3 gramos de impurezas del agua por litro?
- 3.17 El 90% de los químicos determinan en forma correcta el porcentaje de calcio en cierto compuesto químico. Se pide a cinco químicos seleccionados aleatoriamente que determinen el porcentaje de calcio del compuesto. ¿Cuál es la probabilidad de que tres químicos calculen el verdadero contenido de calcio?
- 3.18 Una muestra aleatoria de tamaño 70 se extrae de una población con media 76,1 y variancia 256,34. ¿Cuál es la probabilidad de obtener una media muestral entre 75,0 y 78,5?
- 3.19 Se sabe que un componente electrónico tiene una tasa media de 0,0024 fallas por hora. ¿Cuál es probabilidad de que fallen 3 componentes en 1000 horas de funcionamiento del sistema?

- 3.20 El punto de vaporización de algunas sustancias se distribuye normalmente con promedio de 64,5 y desviación estándar de 25,8. ¿Cuál es la probabilidad de que el punto de vaporización de las sustancias fuera al menos 65,7?
- 3.21 La probabilidad de que un experimento de laboratorio tenga un resultado exitoso es 0,65. ¿Cuál es la probabilidad de que en 15 ensayos del experimento 5 resultados sean positivos?
- 3.22 Si las mediciones del peso específico de un metal pueden considerarse como una muestra de una población normal cuya desviación estándar es de 0,04, ¿cuál es probabilidad de que la media de una muestra aleatoria de tamaño 60 difiera a lo sumo por 0,003 del promedio de la población?
- 3.23 Se está estudiando el rendimiento de un proceso químico. Se sabe, debido a experiencias anteriores, que la variancia y el promedio respectivamente del rendimiento con este proceso son 2,32 y 90,48 (unidades porcentuales). ¿Cuál es la probabilidad de encontrar un rendimiento durante un ensayo entre 89,95 y 91,30?
- 3.24 La tensión superficial de un líquido está representada por T (dinas/cm), y bajo ciertas condiciones, $T \cong 2(1 - 0,005x)^{1,2}$, en donde x es la temperatura del líquido en grados centígrados. Si X tiene función de densidad de probabilidad, $f(x) = 3000x^{-4}$ para $x \geq 10$, calcule el promedio teórico y su variancia.
- 3.25 Una empresa envasa un cierto producto en frascos cuyos contenidos cuantificados por su peso en gramos, tienen distribución aproximadamente normal, con desviación estándar 5,50 g. Si el 10% de los frascos tienen un peso menor de 139 gramos, ¿cuál es el peso promedio de ellos?

- 3.26 La concentración del reactivo en un proceso químico es una variable aleatoria que tiene distribución de probabilidad, $f(x) = 6r(1 - r)$ para $0 \leq r \leq 1$. Calcule la probabilidad de que la concentración sea menor a 0,50.
- 3.27 Un agente químico produce la decoloración en 7% de los rollos de papel producido por cierta compañía. ¿Cuál es probabilidad de que en una muestra aleatoria de 15 rollos de papel, 9 no presenten decoloración debido al agente químico?
- 3.28 El contenido del porcentaje de magnesio en una aleación es una variable aleatoria, dada por la siguiente función de densidad de probabilidad $f(x) = \frac{x}{18}$ $0 \leq x \leq 6$. Calcule el contenido de magnesio esperado y su variancia.
- 3.29 En promedio se encuentran 3 glóbulos rojos por unidad cuadrada. ¿Cuál es la probabilidad de que más de 4 glóbulos rojos sean visibles para un observador con un microscopio en dos unidades cuadradas?
- 3.30 Para estudiar la regulación hormonal de una línea metabólica se inyectan ratas albinas con un fármaco que inhibe la síntesis de proteínas del organismo. En general, 4 de cada 20 ratas mueren a causa del fármaco antes de que el experimento haya concluido. Si se trata a 10 animales con el fármaco, ¿cuál es la probabilidad de que al menos 8 lleguen vivas al final del experimento?
- 3.31 Un fabricante afirma que mediante el uso de un aditivo en la gasolina, el 0,75 de todas las marcas de automóviles podrían aumentar el recorrido por litro. Se usa una muestra aleatoria de 300 automóviles para evaluar este producto. ¿Cuál es la probabilidad que 250 automóviles logren alcanzar la meta de aumentar el recorrido por litro de gasolina?

Capítulo 4

Estadística Inferencia: Estimación

- 4.1 Propiedades deseables de un estimador
- 4.2 Distribuciones muestrales
- 4.3 Tipos de estimación: puntual o por intervalo
- 4.4 El concepto de estimación aplicado a la química
- 4.5 Explicación sobre intervalos de confianza
- 4.6 Intervalos de confianza para la distribución normal
- 4.7 Intervalo para la media si se conoce la varianza
- 4.8 Intervalos de confianza para la media con varianza desconocida
- 4.9 Intervalo de confianza para la varianza
- 4.10 Estimación de tamaño muestral
- 4.11 Muestras para la estimación de medias
- 4.12 Intervalos de confianza para variables dicotómicas
- 4.13 Intervalos de confianza para una proporción
- 4.14 Elección del tamaño muestral para una proporción
- 4.15 Problemas

Estimación

El propósito de un estudio estadístico suele ser extraer conclusiones acerca de la naturaleza de una población. Al ser la población grande y no poder ser estudiada en su integridad en la mayoría de los casos, las conclusiones obtenidas deben basarse en el examen de solamente una parte de ésta, lo que nos lleva a la necesidad de estudiar diferentes métodos de la estadística inferencial.

En este capítulo, se examinarán los procedimientos estadísticos básicos para la estimación que pretende medir cierta característica particular de una población, como la media aritmética o la proporción.

4.1 Propiedades deseables de un estimador

Los primeros términos obligados a los que se debe hacer referencia, serán los de **estadístico** y **estimador**. Un estimador es una variable aleatoria con una determinada distribución, y que será la pieza clave en las dos amplias categorías de la inferencia estadística: la estimación y el contraste de hipótesis.

Estimador. En términos generales es una función de la muestra, que proporciona un valor representativo de la característica poblacional que se quiere conocer.

El concepto de estimador, como herramienta fundamental, lo caracterizamos mediante una serie de propiedades que nos servirán para elegir el "mejor" para un determinado parámetro de una población, así como algunos métodos para la obtención de ellos, tanto en la estimación puntual como por intervalos.

Un estimador no es un valor concreto sino una variable aleatoria, ya que aunque depende unívocamente de los valores de la muestra observados, la elección de la muestra es un proceso aleatorio. Una vez que la muestra ha sido elegida, se denomina estimación el valor numérico que toma el estimador sobre esa muestra.

Intuitivamente, las características que serían deseables para esta nueva variable aleatoria (que se usa para estimar el parámetro desconocido) deben ser:

Consistencia

Cuando el tamaño de la muestra crece, el valor estimado se aproxima al parámetro desconocido. Decimos que $\hat{\theta}$ es un estimador consistente con el parámetro θ si:

$$\forall \epsilon > 0, \quad \lim_{n \rightarrow \infty} P[|\hat{\theta} - \theta| < \epsilon] = 1 \quad (4.1)$$

Cuando el número de observaciones n tiende a infinito, es lo que se denomina propiedades asintóticas.

Insesgado

Un estimador es insesgado cuando su esperanza matemática coincide con el valor del parámetro poblacional. A efectos de esta aplicación un estimador será insesgado cuando al repetirse el proceso de muestreo, mediante reiteraciones*, un número suficiente de veces la diferencia entre el valor medio de estas reiteraciones y el valor objetivo poblacional converja a cero. Se dice que un estimador $\hat{\theta}$ de un parámetro θ es insesgado si

$$E(\hat{\theta}) = \theta \quad (4.2)$$

La carencia de sesgo puede interpretarse del siguiente modo. Suponga que se tiene un número indefinido de muestras de una población, todas ellas del mismo tamaño n . Sobre cada muestra el estimador nos ofrece una estimación concreta del parámetro. Pues bien, el estimador es insesgado, si sobre dicha cantidad indefinida de estimaciones, el valor medio obtenido en las estimaciones es θ (el valor que se desea conocer).

***Reiteraciones:** o número de muestras, se definen como repeticiones del proceso de selección de una muestra. Esto es, con esta expresión se identifica el número de veces que se va a repetir el proceso de selección de muestras. Evidentemente en el trabajo de campo de los estudios por muestreo, únicamente se investiga una muestra.

Eficiencia

Al estimador, al ser variable aleatoria., no puede exigírsele que para una muestra cualquiera se obtenga como estimación el valor exacto del parámetro. Sin embargo, se puede esperar que su dispersión con respecto al valor central (varianza) sea tan pequeña como sea posible.

Entonces, dados dos estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ de un mismo parámetro θ , se dice que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2) \quad (4.3)$$

Suficiencia

El estimador debería aprovechar toda la información existente en la muestra.

Más adelante, en este capítulo, se presentarán los tipos de estimación. A continuación, se estudia las distribuciones muestrales.

4.2 Distribuciones muestrales

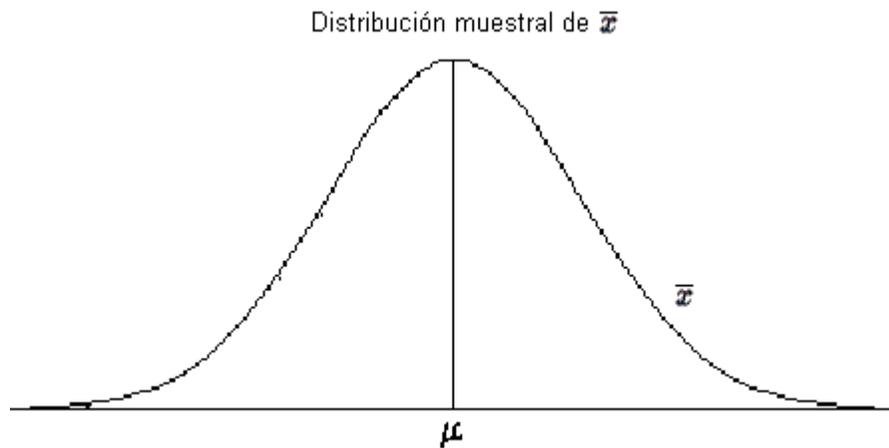
Una **distribución muestral** es la distribución de todos los posibles valores del estadístico de la muestra, que se pueden obtener de la población para un determinado tamaño de muestra. Por ejemplo, se podría tomar, de una población de desodorantes antitranspirante producidos por una empresa farmacéutica, una muestra aleatoria de 10 unidades y pesarlas para calcular después la media de su peso neto.

Puede pensarse en esta media muestral \bar{X} , como si se hubieran extraído de la distribución de todas las medias muestrales posibles, para muestras de tamaño 10 que pudieran tomarse de la población. En forma similar, cada estadístico muestral a ser calculado, a partir de datos muestrales, puede considerarse como que ha sido extraído de una distribución muestral.

El **teorema del límite central** establece que, al hacerse más grande la muestra, la distribución de muestreo de las medias de la muestra tiende hacia la distribución normal, con media aritmética μ y desviación estándar es $\frac{\sigma}{\sqrt{n}}$ (a este valor se le conoce como *error estándar de la distribución muestral*). Esta distribución

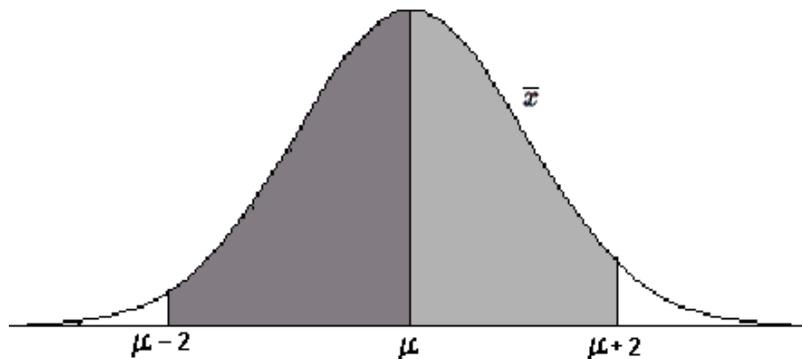
muestral tenderá hacia la normalidad, sin importar la forma de la distribución de la población de la que se extrajeron los elementos de la muestra. La siguiente figura muestra cómo lucirá dicha distribución muestral.

El *teorema del límite central* es de particular importancia en las estadísticas, ya que permite al analista calcular la probabilidad de diversos resultados de muestras a través del conocimiento de las posibilidades de las probabilidades de la curva normal.



Ejemplo. ¿Cuál es la probabilidad de que la media de una muestra aleatoria de 10 unidades, extraídas de una población, se encuentre en un margen de 2 gramos de la media real del peso de la población, si la desviación estándar de la población se estima en 3,5 gramos?

La figura muestra la distribución apropiada:



El error estándar es:

$$\frac{\sigma}{\sqrt{n}} = \frac{3.5}{\sqrt{10}} = 1,107$$

Luego

$$z = \frac{2}{1.107} = 1,806$$

con lo que se obtiene en la tabla de la normal estándar un área 0,93124.

La probabilidad de que la media muestral esté en un margen de 2 gramos de la media real, es de alrededor del 93%, reflejando un tamaño de muestra de 10 unidades y la variabilidad estimada de la población de 3,5 gramos. Como se verá, esta capacidad para calcular probabilidades de resultados de muestras permitirá al analista llegar a útiles resultados en la estimación y en la prueba de hipótesis.

4.3 Tipos de estimación: puntual o por intervalo

Hay dos tipos de estimación; **puntual** y por **intervalo**. Una *estimación puntual* utiliza un solo valor de la muestra para estimar el parámetro de la población de estudio. Por ejemplo, la media de la muestra \bar{x} es una estimación puntual de la media μ de la población. La variancia s^2 de la muestra es una estimación puntual de la variancia σ^2 de la población.

La *estimación por intervalo* de confianza consiste en determinar un posible rango de valores o intervalo, en los que pueda precisarse --con una determinada probabilidad-- que el valor de un parámetro se encuentra dentro de esos límites.

Evidentemente esta técnica no tiene porqué dar siempre un resultado correcto. A la probabilidad de acertar al decir que el parámetro estaba contenido en dicho intervalo se la denomina **nivel de confianza**. También se denomina **nivel de significación** a la probabilidad de no atinar con la estimación.

Es común entre químicos en la práctica, realizar el análisis por duplicado o triplicado. Se promedian entonces los resultados y se juzga el error experimental por la cercanía de lo hallado en los dos o tres análisis. Si el rango de los resultados es pequeño, se piensa en forma intuitiva que el análisis es exacto y preciso, lo que no es necesariamente cierto, porque la precisión de resultados, no necesariamente conduce a una buena exactitud.

Por ejemplo, si el análisis consistiera en pesar una muestra, disolverla y analizar porciones alícuotas de la solución, los resultados quizá se repetirían a la perfección, a pesar de lo cual el análisis podría haber enmascarado algún error en la pesada, así como otros circunstanciales. Si se hacen dos pesadas y, simultáneamente, dos análisis, se habrá aumentado la posibilidad de desenmascarar los errores de pesada, pero pueden persistir los ambientales y de manipulación.

POWERS¹ encontró, por otra parte, que un químico que hace varios análisis obtiene mejores resultados análogos que varios químicos analizando la misma muestra. Ambos estudios demuestran con claridad la importancia de efectuar de forma independiente la repetición de análisis. El término “*repetición*” se refiere a la repetición de observaciones hechas de forma independiente.

Si el analista realiza un segundo análisis verdaderamente independiente, aumentará sus conocimientos en dos sentidos:

- El promedio de los dos análisis dará una mejor estimación del verdadero valor, μ , que si se hiciera uno sólo.
- La diferencia entre las dos medidas encontradas dará una estimación del error. El analista puede entonces hacer dos estimaciones:
- Una estimación puntual de μ utilizando \bar{x} , el promedio de las replicaciones. Es la mejor estimación insesgada que puede tener de μ .
- Una estimación por intervalo en la que puede afirmar que se halla el verdadero valor μ .

¹ POWERS, R. F.: “Determination of the precision of analytical control methods”. Ind. Eng. Chem., Anal, Ed., 15, 361 (1943)

4.5 Explicación sobre intervalos de confianza

Se presenta el siguiente ejemplo para explicar el cálculo de los límites de confianza. Suponga que un químico analizando una sustancia mediante un método de variabilidad desconocida, obtiene el resultado de 49,69, el cual es una estimación del verdadero valor, pero como no tiene idea de la magnitud del error experimental tampoco puede tener idea de la confianza que puede poner en esta estimación.

Repite el ensayo y obtiene como resultado 50,90, con lo cual ahora tiene otra estimación de μ (50,3) y un rango (50,9-49,69=1,21), a partir del cual estima la región en la que se encuentra el verdadero valor desconocido. Puede tener el 50 de 100 de seguridad de que este valor verdadero se halla entre 50,90 y 49,69 y alrededor del 80 por 100 de confianza en que está 52,11 y 48,48.

Si continua repitiendo los análisis y obteniendo resultados como 48,99, 51,23, 51,47 y 48,80, mejora su estimación de μ y la confianza que puede depositar en esta estimación.

Algunos químicos rechazan como “discrepantes” los resultados que caen por fuera de los límites de probabilidad del 95%; otros rechazan los que caen fuera de los límites de $\pm 3s$. Cualquiera que sea el criterio escogido. El cálculo del intervalo de confianza proporciona una cierta justificación estadística para el rechazo de resultados.

4.6 Intervalos de confianza para la distribución normal

Dada una variable aleatoria de distribución gaussiana, $X \sim N(\mu, \sigma^2)$, interesa en primer lugar, en calcular intervalos de confianza para sus dos parámetros, μ y σ^2 . He aquí un resumen de las situaciones a considerar:

- **Intervalo para el promedio si se conoce la varianza.** Este no es un caso práctico (no se puede conocer σ^2 sin conocer previamente μ), pero sirve para introducir en el problema de la estimación confidencial de la media.
- **Intervalos de confianza para el promedio** (caso general). Este se trata del caso con verdadero interés práctico. Por ejemplo sirve para estimar intervalos

que contenga la media del colesterol en sangre en una población, la altura, el peso, etc., cuando se dispone de una muestra de la variable.

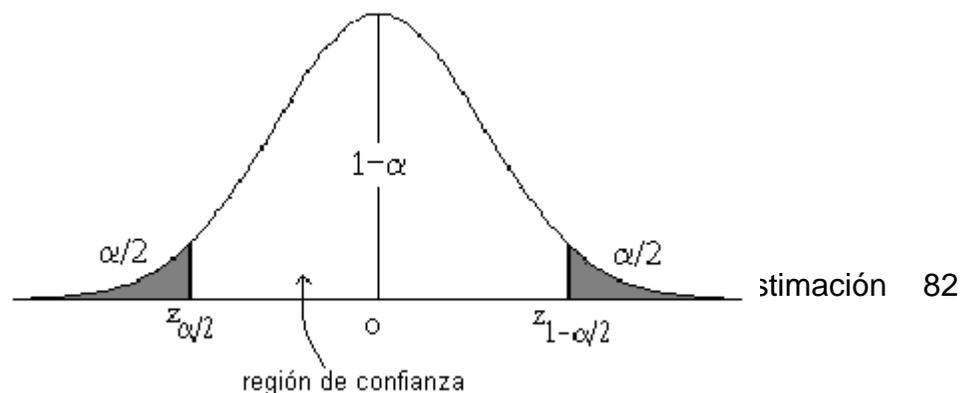
- **Intervalo de confianza para la varianza.** Este es otro caso de interés en las aplicaciones. El objetivo es calcular un intervalo de confianza para σ^2 , cuando sólo se dispone de una muestra.
- **Estimación de tamaño muestral.** La utilidad consiste en decidir cuál deberá ser el tamaño necesario de una muestra para obtener intervalos de confianza para una media, con precisión y significación dadas de antemano. Para que esto sea posible es necesario poseer cierta información previa, que se obtiene a partir de las denominadas muestras piloto.

4.7 Intervalo para el promedio si se conoce la varianza

Este caso que se plantea es más a nivel teórico que práctico: difícilmente vamos a poder conocer con exactitud σ^2 mientras que μ es desconocido. Sin embargo nos aproxima del modo más simple a la estimación por intervalo de las medias.

En una situación real, la media de la población es desconocida y es el parámetro que se debe estimar con base en la información de la muestra. La desviación estándar de la población rara vez se conoce, pero por medio del **teorema del límite central**, se podría estimar si el tamaño de la muestra es mayor a 30.

Al trabajar con un nivel de confianza del 95%, esto implica que se tendría 95% de confianza de que la muestra seleccionada es una en la cual la media real se encuentra dentro del intervalo encontrado. El 95% **no significa** que exactamente 95 de cada 100 muestras incluirán la media de la población dentro del intervalo de confianza, sino sólo que en el promedio de todas las posibles muestras, 95% incluirá la media real.



Ya que μ es desconocida, no se sabe en realidad si el intervalo específico incluye o no la media de la población. Para una muestra particular, se puede establecer el nivel de confianza $(1 - \alpha)$ de que el intervalo incluya la media de la población. De este modo, fijado $\alpha \in (0,1)$, considere la v.a. normal estándar y un intervalo que contenga una de probabilidad de $1 - \alpha$

Este intervalo es tan pequeño como sea posible. Por ello lo mejor es tomarlo simétrico con respecto a la media (0) (véase la figura). Así las dos colas de la distribución (zonas más alejadas de la media) se repartirán a partes iguales el resto de la masa de probabilidad α .

La fórmula de cálculo para obtener la estimación del intervalo de confianza de la media cuando se conoce σ , se define como

$$L_i = \bar{x} \pm Z_{\alpha/2} \sqrt{\frac{N-n}{N-1} \frac{\sigma_x}{\sqrt{n}}} \quad (4.1)$$

En el caso, de estar muestreando con remplazo o si el factor de corrección,

$$fc = \frac{N-n}{N-1}$$

es menor al 5%, se puede ignorar este término de la fórmula de cálculo.

Ejemplo. La conductividad térmica del grafito varía con la temperatura según la ecuación $K=K_0-\alpha T$. Experimentalmente, solo es posible obtener una conductividad media en un intervalo de temperatura. Se requiere estimar la conductividad promedio con un intervalo de confianza del 95%, dado que una muestra aleatoria de 35 ensayos, resultó una media aritmética 1,2825 y una desviación estándar de 0,1315 K_m , determinado T y 25°C en kilo-ergio/cm² seg.°C cm⁻¹.

Al suponer una población infinita (muestreo con reemplazo) y, por medio de la fórmula 4.1, se obtiene:

$$1,2825 \pm 1,96 \frac{0,1315}{\sqrt{35}}$$

$$1,2825 \pm 0,04357$$

Entonces,

$$L_1 = 1,2825 - 0,04357 = 1,2389$$

$$L_2 = 1,2825 + 0,04357 = 1,3260$$

El valor 1,96 se obtiene de la distribución normal estándar. Por lo cual, se puede estimar la conductividad promedio entre el 1,2389 y 1,3260 K_m, con un nivel de confianza del 95%.

4.8 Intervalos de confianza para el promedio con varianza desconocida

Como sea ha mencionado, los casos anteriores se presentarán poco en la práctica, ya que lo usual es que sobre una población quizás se puede conocer si se distribuye normalmente, pero el valor exacto de los parámetros μ y σ^2 no son conocidos. De ahí nuestro interés en buscar intervalos de confianza para ellos.

En tales situaciones, es necesario utilizar \bar{x} y s^2 para obtener una estimación del intervalo de confianza de μ . La distribución que se ha desarrollado para aplicarla a esta situación es la **distribución t de Student**.

El intervalo de confianza para la media con σ desconocida, se expresa como

$$L_i = \bar{x} \pm t_{\alpha/2(n-1)gl} \sqrt{\frac{N-n}{N-1}} \frac{s_x}{\sqrt{n}} \quad (4.2)$$

Los valores críticos de la **distribución t**, para algunos grados de libertad, se pueden obtener haciendo uso de la tabla 4 del apéndice. El valor en la parte superior de cada columna de la tabla de **t** indica el nivel de significancia utilizado, mientras la primera columna representa el valor particular de los grados de libertad. Por ejemplo, con 5 grados de libertad, si se desea 99% de confianza, el valor crítico de **t** sería 3,3649.

Con el fin de estudiar cómo se calcula el intervalo de confianza para una media, cuando se desconoce la desviación estándar de la población, se examinará el siguiente problema.

Ejemplo. Se analizó gravimétricamente una cierta muestra para determinar sulfatos y se obtuvieron los siguientes resultados en partes por millón:

3,390 6,990 5,71 5,930 5,350 5,810 5,520 5,910 5,460 5,810

¿Cuáles son los límites de confianza al 99% sobre el contenido promedio de sulfatos?.

$$5,588 \pm 2,8214 \frac{0,89496}{\sqrt{10}}$$

$$5,588 \pm 0,7985$$

Finalmente,

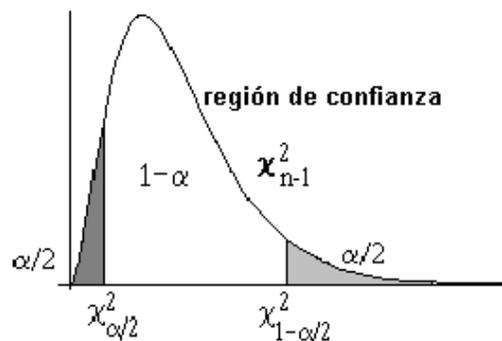
$$L_1 = 5,588 - 0,7985 = 4,7895$$

$$L_2 = 5,588 + 0,7985 = 6,3865$$

Con resultados anteriores, se estima que el contenido promedio de sulfatos se encuentra comprendido entre 4,7895 y 6,3865, a un nivel de confianza del 99%.

4.9 Intervalo de confianza para la varianza

Para compensar cualquier incertidumbre que pueda existir en la estimación de la varianza, se pueden calcular sus límites de confianza empleando las tablas de valores de la distribución de probabilidad Ji-cuadrada. En estas tablas se dan los valores numéricos del término χ^2 para diferentes grados de libertad y probabilidades.



Considere dos cuantiles de esta distribución que dejen una probabilidad $1-\alpha$ en la "zona central" de la distribución.

En vista de la imposibilidad práctica de tener la varianza de la población, en muchos procedimientos estadísticos como los límites de confianza para el promedio

con muestras pequeñas, se tiene que sustituir σ^2 por s^2 . Se puede asegurar con una probabilidad de $1-\alpha$ que la variancia de poblacional, σ^2 , se encuentre los límites;

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2} \quad (4.3)$$

Al extraer la raíz cuadrada a cada miembro de esta desigualdad se obtiene un intervalo correspondiente con un nivel de confianza de $(1-\alpha)$ 100% para σ .

Nótese que los intervalos de confianza para σ o σ^2 considerando “colas iguales”, como en la fórmula anterior, en realidad no dan los intervalos de confianza más reducidos, debido a que la distribución Ji-cuadrada no es simétrica. Sin embargo, son utilizados en la mayoría de las aplicaciones a fin de evitar cálculos complicados.

Ejemplo. Cinco determinaciones de la densidad de un líquido orgánico dieron los resultados:

$$0,9132 \quad 0,9138 \quad 0,9129 \quad 0,9131 \quad 0,9133 \text{ g/cm}^3$$

Construya un intervalo con un nivel de confianza del 95% para σ , o sea la desviación estándar de la población.

Para $5-1=4$ grados de libertad,

$$\chi_{0,975}^2 = 0,4844$$

$$\chi_{0,025}^2 = 11,143$$

de acuerdo con la tabla 5; así la sustitución en la fórmula produce,

$$\frac{4(1,13 \times 10^{-7})}{11,852} \leq \sigma^2 \leq \frac{4(1,13 \times 10^{-7})}{0,4844}$$

$$3,8137 \times 10^{-8} \leq \sigma^2 \leq 9,3311 \times 10^{-7}$$

y en consecuencia

$$1,95287 \times 10^{-4} \leq \sigma \leq 9,65978 \times 10^{-4}$$

Esto significa que se tiene un 95% de confianza que el intervalo comprendido entre $1,95287 \times 10^{-4}$ y $9,65978 \times 10^{-4}$ contiene a la desviación estándar verdadera de la densidad de un líquido orgánico.

El método que expuesto se aplica sólo a muestras aleatorias de poblaciones normales o al menos a las poblaciones que pueden aproximarse a la distribución normal.

4.10 Estimación de tamaño muestral

El químico, especialmente el químico analista, se pregunta con frecuencia qué tamaño de muestra le conviene usar, lo que no es tanto una pregunta como una exclamación, a menudo desesperada. Lo que el químico en realidad quiere decir es: “¿Cuál es el tamaño de muestra más económico que puedo tomar para estar razonablemente seguro de que las deducciones que se obtienen de la muestra son válidas para la población?”.

Siempre que se trabaje con muestreo, es necesario de contar con un plan preciso para delimitar el tamaño de la muestra a extraer de la población para cumplir con los objetivos de la investigación. Un error común consiste en pensar que una muestra debe ser grande para que realmente sea representativa de la población.

Hay que puntualizar que sólo las muestras probabilísticas pueden abordarse mediante las técnicas de delimitación del tamaño de la muestra que serán expuestas aquí. Recuérdese que las muestras no probabilísticas hacen imposible al cálculo de error muestral.

El conocimiento previo de la población es una ayuda para decidir el tamaño de la muestra. Si sabe que se muestrea una población homogénea (poca variabilidad), tal como una solución bien mezclada, se entiende que una muestra tomada de cualquier parte da tanta información como diez muestras tomadas de diez diferentes partes de la solución. Por el contrario, si se sospecha o hay certeza de que el material que se muestrea no es homogéneo (tiende a tener mucha variabilidad), se tiene que tomar muestras de varias partes del material. El

conocimiento previo respecto al error del muestreo permite al analista trabajar con muestras más pequeñas.

El método que hay que seguir para tomar muestras para el análisis, viene determinado por el estado de la sustancia a estudiar. Los productos gaseosos y líquidos en general son esencialmente homogéneos, siendo suficiente en tal caso recoger de ellos una sola muestra discreta. Una forma de suprimir la posibilidad de toma de muestras subjetivas, es empleando un procedimiento de muestreo aleatorio.

4.11 Tamaño de muestras para la estimación de promedios

En cada uno de los ejemplos relacionados con la estimación del intervalo de confianza, el tamaño de muestra se determinó en forma arbitraria. La determinación del tamaño de muestra es un procedimiento más riguroso, se tendría que fijar una magnitud de error de muestreo permisible dispuestos a aceptar al estimar el parámetro de la población.

El **error de muestreo** es igual a la diferencia entre la estimación de la muestra y el parámetro. También hay que determinar por anticipado un nivel de confianza deseada para la estimación correcta del parámetro de la población.

La mayor dificultad al determinar el tamaño de la muestra necesaria para estimar la media de la población consiste en calcular la desviación estándar de la población. Después de todo, si tiene un conocimiento completo sobre la población, no habría necesidad de realizar una investigación sobre sus parámetros.

Al dudar de trabajos anteriores para calcular la desviación estándar de la población, las alternativas incluyen el juicio del investigador o el empleo de estudios exploratorios con muestras pequeñas y emplear los resultados para obtener una estimación de la desviación estándar.

De manera alternativa, el rango de una variable con distribución normal es aproximadamente igual a más o menos tres desviaciones estándar y, así, se puede estimar la desviación estándar al dividir el rango entre seis. Con la frecuencia, el investigador puede estimar el rango con base en el conocimiento del fenómeno.

Luego el tamaño de muestra para la estimación de la media de la población se obtiene como,

$$n = \left(\frac{Z_{\alpha/2} \sigma}{d} \right)^2 \quad (4.4)$$

donde:

$Z_{\alpha/2}$ = es valor de la distribución normal estándar asociado al nivel de confianza.

d = error de muestreo permitido o diferencia máxima entre la media muestra y la media de la población que se está dispuestos a aceptar en el nivel de confianza que es indicado.

σ = desviación estándar de la población (conocida o estimada a partir de estudios anteriores)

Ejemplo. Un investigador desea probar la hipótesis de que el punto de fusión del estaño es 232,0 grados Celsius. Se considera un margen de error de muestreo de 0,10 grados y un nivel de confianza del 98%. Además, se puede asumir que el punto de fundición es una variable aleatoria distribuida normalmente, con una desviación estándar de 0,18 grados. ¿Cuántas determinaciones deberán realizarse?

Sustituyendo $d=0,10$, $\sigma=0,18$ y $Z_{0,99}=2,33$ en la fórmula 5.6 para n , se tiene

$$n = \left[\frac{2,33(0,18)}{0,10} \right]^2 \cong 18$$

Por tanto, el químico si desea realizar un estudio con toda la precisión requerida en el enunciado se debería tomar una muestra de 18 determinaciones, para estimar la verdadera media del punto de fundición con una confianza del 98%, con un margen de error a lo sumo de 0,10 grados Celsius.

Esto es una indicación de gran utilidad antes de comenzar el estudio. Una vez que el muestreo haya sido realizado, se debe confirmar que el error para el nivel de significación dado es inferior o igual a 1 cm, utilizando la muestra obtenida.

Como regla empírica a utilizar en la determinación del tamaño de la muestra, se considera conveniente redondear al siguiente entero a fin de sobrestimar ligeramente el tamaño y así satisfacer los criterios deseados.

4.12 Intervalos de confianza para variables dicotómicas

Cuando se tiene una variable dicotómica (o de Bernoulli) a menudo interesa saber en qué proporción de casos, p , ocurre el éxito en la realización de un experimento. También es de interés calcular para un nivel de significación dado, el tamaño muestral necesario para calcular un intervalo de confianza cuya amplitud sea menor que cierta cantidad.

4.13 Intervalos de confianza para una proporción

El concepto de estimación se puede extender a los datos dicotómicos para estimar la proporción de la población con base en los datos de la muestra. La distribución del número de éxitos es binomial, y puede ser aproximada a la normal cuando el tamaño de la muestra n es grande, y p no es una cantidad muy cercana a cero o uno.

$$X \sim B(n, p) \Rightarrow X \sim N(np, npq) \quad (4.5)$$

Es decir, al considerar como estimación de p la proporción de éxitos obtenidos en las n pruebas, \hat{p}

$$\hat{p} = \frac{X}{n} \quad (4.6)$$

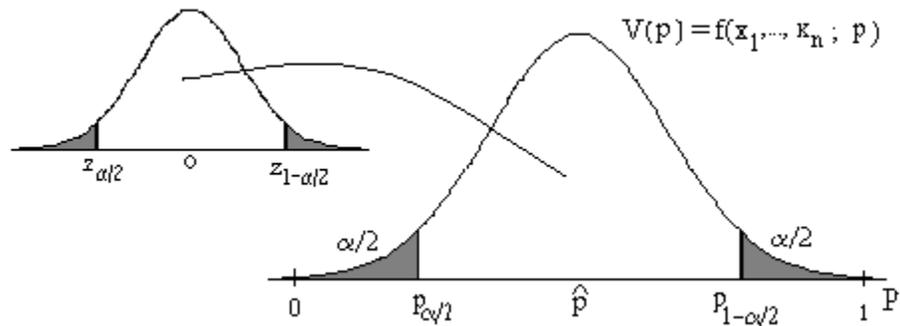
El estimador \hat{p} no es más que un cambio de escala de X , por tanto

$$\hat{p} = \frac{X}{n} \sim N\left(p, \frac{pq}{n}\right) \Rightarrow \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \approx Z \sim N(0,1) \quad (4.7)$$

Para encontrar el intervalo de confianza al nivel de significación α para p se considera el intervalo que hace que la distribución normal estándar deje α fuera del mismo. Esto se resume en la siguiente expresión.

$$L_i = \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{N-n}{N-1} \frac{\hat{p}\hat{q}}{n}} \quad (4.8)$$

Intervalo de confianza para una proporción



En caso de tamaños de muestras pequeñas o si la proporción es muy pequeña, se debe utilizar la distribución binomial en lugar de la distribución normal. A fin de ver la forma en que se puede utilizar esta estimación del intervalo de confianza para la proporción de la población, se examinará el siguiente problema.

Ejemplo. En un estudio sobre seguridad de explosivos usada en ciertas minas, los explosivos que contenían nitrato de potasio se utilizaron en 95 de 250 casos estudiados. Construya un intervalo de confianza del 90% sobre la proporción de empresas que utilizan nitrato de potasio en los explosivos.

$$0,38 \pm 1,64 \sqrt{\frac{0,38(0,62)}{250}}$$

$$0,38 \pm 0,05$$

$$L_1 = 0,38 - 0,05 = 0,33$$

$$L_2 = 0,38 + 0,05 = 0,43$$

La proporción de empresas que utilizan nitrato de potasio en los explosivos se encuentra entre el 33% y el 43%, con un nivel de confianza del 90%.

4.14 Elección del tamaño muestral para una proporción

Los métodos para determinar el tamaño de muestra que se utilizan en la estimación de una proporción real son similares a los empleados en la estimación del promedio aritmético. La fórmula apropiada es:

$$n = \left(\frac{Z_{\alpha/2} \sqrt{PQ}}{d} \right)^2 \quad (4.9)$$

donde:

$Z_{\alpha/2}$ = es valor de la distribución normal estándar asociado al nivel de confianza

d = error de muestreo permitido o diferencia máxima entre la proporción muestral y la proporción de la población que se está dispuestos a aceptar en el nivel de confianza escogido

P = proporción de la población que posee la característica de interés (sí puede estimar la proporción, hágalo y utilícela como P ; en caso contrario, sea conservador y use $P = 0,5$ en la fórmula).

Al aplicar está fórmula, primero hay que decidir si es posible estimar en forma aproximada el valor de la proporción de la población. En caso, aceptar con seguridad que esa proporción difiere mucho de 0,5 en una u otra dirección, se está en condiciones de obtener la precisión deseada con un tamaño de muestra más pequeño (y de menos costo) de la muestra. Por ello, este camino de estimar P debería utilizarse con precaución y no como solución simplista.

Ejemplo. Se está estudiando el contenido de cobre en una aleación con la esperanza de reducir finalmente el costo de manufactura. Por estudios anteriores se estima que la proporción es 0,12. ¿Cuántas mezclas deben estudiarse para estimar la proporción de cobre en una aleación con un error de muestreo de 10%, a un nivel de confianza de 95%?

Sustituyendo $d=0,10$, $P=0,12$ y $Z_{0,975} = 1,96$ en la fórmula 4.9, se obtiene

$$n = \left[\frac{1,96\sqrt{0,12(0,88)}}{0,10} \right]^2 \cong 40$$

El investigador tendrá que observar 40 mezclas para estimar la proporción de cobre en una aleación, con un nivel de confianza del 95% y margen de error de muestreo del 10%.

4.15 PROBLEMAS

- 4.1 Cinco determinaciones de la densidad de un líquido orgánico dieron los resultados 0,9132, 0,9138, 0,9129, 0,9131 y 0,9133 g/cm³. Calcule un intervalo de confianza del 95% para el promedio real de la densidad del líquido orgánico.
- 4.2 De acuerdo con un método de análisis del contenido de plomo en el agua, el promedio aritmético y desviación estándar del agua son 20,2 y 10,8 µg/L, respectivamente. Un estudio de 40 muestras de agua produce una media de 24,75 µg/L. Halle el intervalo de confianza del 93% para la media.
- 4.3 Una investigadora quiere determinar el tiempo promedio que un mecánico tarda en cambiar las llantas en un automóvil. Esto con una confianza del 95% que la media de su muestra difiera a lo sumo de 1,2 minutos de la poblacional. Se puede presumir por experiencia que la desviación estándar es 3,4 minutos. ¿Qué tamaño debe tener la muestra?
- 4.4 Nueve repeticiones de la determinación de la solubilidad molar de un ácido orgánico en agua a 25°C, dieron un valor medio de 0,00673 mol/L, con una desviación estándar de 0,00008 mol/L. Calcule un intervalo de confianza del 90% para el verdadero valor medio de la solubilidad molar del ácido.
- 4.5 Observando la cantidad de contaminantes atmosféricos en una ciudad durante 500 días, se encontró un exceso de 200 microgramos en sólo 15 días. Construya un intervalo para la probabilidad de que la contaminación exceda 200 microgramos por metro cúbico en un día cualquiera. $\alpha=0,01$
- 4.6 Se determinó que a 22°C una pipeta volumétrica presenta los siguientes pesos aparentes de agua:

4,995 4,993 4,997 4,991 4,994 4,992 gramos

Calcule un intervalo de confianza del 95% para el verdadero valor del volumen promedio de agua escurrida.

- 4.7 La validación de un método para la determinación de cobre II, mostró una desviación estándar de 0,08%. Los resultados para un análisis de 6 réplicas de una muestra se presentan a continuación:

Réplica	1	2	3	4	5	6
Contenido Cu (%)	3,25	3,68	3,27	3,30	3,98	3,46

Calcule e interprete un intervalo de confianza del 90% para la verdadera media del contenido de cobre.

- 4.8 En un estudio sobre seguridad de explosivos usada en ciertas minas, los explosivos que contenían nitrato de potasio se utilizaron en 95 de 250 casos estudiados. Construya un intervalo de confianza del 97,5% sobre la proporción de empresas que utilizan nitrato de potasio en los explosivos.
- 4.9 Se desea estimar el número de horas de uso continuo antes de que cierto equipo de mezcla de pintura produzca una inconsistencia. Se puede suponer que la desviación estándar es 5,3 horas, ¿cuántas observaciones del equipo se necesitan, con una confianza del 90% y un error de muestreo de 2,2 horas?
- 4.10 Un método para la determinación de cobre tiene una desviación estándar de 0,08 (unidades porcentuales). Se toma una muestra de 50 réplicas y por cada una de ellas, se le mide el contenido de cobre, resultando una media de 3,528 (unidades porcentuales). Calcule un intervalo de confianza del 97,5% para la media verdadera del contenido de cobre.
- 4.11 Los siguientes son los resultados obtenidos en la determinación de plomo en partículas suspendidas en el aire en $\mu\text{g}/\text{m}^3$:

2,4 2,1 1,6 2,2 1,9 2,5 2,8 3,7 3,5 3,9

Encuentre el intervalo del 99% de confianza para el promedio real del plomo.

4.12 Se quiere estimar la proporción de productos defectuosos dentro de una cadena de producción; por experiencia se sabe que aproximadamente el 10% de los productos presenta algún defecto. De qué tamaño debe ser la muestra para estimar la proporción de artículos que presentan algún defecto con un error de muestreo de 5% y un nivel de confianza del 98%.

4.13 En una muestra al azar de 60 secciones de tubo en una planta química, 8 de ellos mostraron señales de corrosión seria. Construya un intervalo de confianza del 93% para la parte de los tramos de tubo con corrosión seria.

4.14 En 6 determinaciones del punto de fusión del estaño, un químico obtuvo una media de 232,26 grados Celsius con una desviación estándar de 0,14 grados. Si utiliza esta media como punto de fusión real del estaño, ¿qué puede asegurar el químico con una confianza del 90% acerca del error máximo?

4.15 En un estudio sobre la contaminación del aire efectuado por una estación experimental, se recolectaron cantidades de benceno soluble suspendido en materia orgánica (en microgramos por metro cúbico) en 12 diferentes muestras de aire:

2,2 2,6 1,8 1,9 3,1 2,0 3,3 2,4 2,0 2,6 1,2 2,0

Suponiendo que la población de estudio sea normal, construya un intervalo con un nivel de confianza del 95% para la media real correspondiente.

- 4.16 La reacción de un individuo al estímulo durante un experimento psicológico puede tomar una de dos formas, A o B. Si un experimentador desea estimar la probabilidad p de que una persona reaccione a favor de A, ¿cuántas personas deben incluirse en el experimento?. Suponga que el investigador quedará satisfecho si el error de estimación es menor que 0,10 con una confianza del 90%. Además, también que él espera que p tenga un valor cercano de 0,6
- 4.17 Debido tanto a las variaciones en las técnicas de laboratorio, como a las impurezas de los materiales y a otros factores desconocidos, los resultados de un experimento en un laboratorio de química no siempre dan las mismas respuestas numéricas. En un experimento de electrólisis, una clase mide la cantidad de cobre precipitado de una solución saturada de sulfato de cobre, durante un período de 30 minutos. Los 40 estudiantes consiguieron una media de 0,145 moles y una desviación estándar de 0,051 moles. Encuentre un intervalo de confianza del 90% para la cantidad media de cobre precipitado.
- 4.18 Un químico ha preparado un producto diseñado para matar el 98% de un tipo particular de insectos. ¿De qué tamaño debe ser la muestra si él desea tener un 95% de confianza de que su estimación se aleja menos de 0,01 de la verdadera fracción de insectos eliminados?
- 4.19 En un experimento realizado para probar el efecto de un ambiente presurizado de oxígeno sobre los pacientes que padecen de osteomielitis se encontró que de 70 pacientes sometidos al tratamiento, todos mejoraron y 63% se liberó de la enfermedad. Si las personas tratadas representan una muestra aleatoria de 600 pacientes con osteomielitis, estime la proporción de los pacientes tratados que se liberarán de la enfermedad. $\alpha=0,05$

- 4.20 Un experimentador ha preparado una droga para inducir el sueño a la población de pacientes con problemas mentales ($N=2500$). ¿Cuántas personas deben ser tratadas si se desea estimar la verdadera fracción con un error menor que 0,05 a un nivel de significación del 10%?
- 4.21 Un informe del Ministerio de Salud indica que cierta marca de cigarrillo con filtro tiene un contenido de alquitrán y nicotina de 5 y 0,4 mg por cigarrillo., respectivamente. Sin embargo, no indica cuántos cigarrillos se analizaron para obtener estos resultados ni da una medida de la variación de las observaciones de cigarrillo a cigarrillo. Considere que la desviación estándar del contenido de alquitrán es aproximadamente igual a 1 mg por cigarrillo, ¿cuántos cigarrillos habría que analizar? Supóngase que se desea un error de estimación menor que 0,1 mg con un nivel de confianza del 99%.
- 4.22 Se desea estimar la producción media por hora de un proceso para la fabricación de un antibiótico. El proceso se observa durante 100 períodos de una hora escogidos al azar, resultando una media muestral de 34 gramos/hora y una variancia de 400 gramos/hora. Estime la producción por hora usando un intervalo de confianza. $\alpha=0,025$
- 4.23 ¿De qué tamaño debe ser una muestra para estimar el parámetro binomial, p , con un error de muestreo de 0,03, con un nivel de confianza del 90%? Supóngase que el valor de p es aproximadamente 0,15 y el tamaño de la población de 500 elementos.
- 4.24 Al intentar determinar las calorías por ración de lasaña en una asignación de laboratorio de nutrición, 13 estudiantes obtuvieron:
- 239 335 347 318 316 322 330 351 362 315
342 353 327

Calcule e interprete un intervalo de confianza del 95% para la media de calorías por ración de lasaña.

- 4.25 Como parte de una investigación acerca de la contaminación de las aguas de un río, se tomó una muestra de mariscos y se encontró que los niveles de cierta sustancia contaminante en ellos eran (ppm):

15 16 10 13 16 25 12 15 18

- a) Calcule e interprete la media real del contaminante en los mariscos de este río, con un intervalo de confianza del 95%.
- b) ¿Qué suposiciones deben hacerse para que las inferencias sean válidas?
- 4.26 Los químicos orgánicos a menudo purifican los compuestos orgánicos por un método conocido como cristalización fraccionada. Un experimentador deseaba preparar y purificar 4,85 gramos de anilina. Se prepararon individualmente diez cantidades de anilina de 4,85 gramos y se purificaron convirtiéndose en acetanilida. Las cantidades obtenidas fueron las siguientes:
- 3,85 3,88 3,90 3,62 3,72 3,80 3,85 3,36 4,01 3,83
- Estime el número medio de gramos de acetanilida que se pueden recuperar de una cantidad inicial de 4,85 gramos de anilina. Use un intervalo de confianza del 90%.

- 4.27 Un investigador desea probar la hipótesis de que el punto de fusión de una aleación es de 1000°C. Si el verdadero punto difiere de éste por más de 5°C deberá cambiarse la composición de la aleación. Si se considera que el punto de fusión es una variable aleatoria distribuida normalmente, con una desviación estándar de 10°C. ¿Cuántas observaciones deberán realizarse?. Utilice $\alpha=0,01$

- 4.28 Las mediciones de la cantidad de cloroformo (microgramos por litro) de 40 muestras de agua potable de una ciudad produjeron los siguientes resultados: $\bar{x} = 34.8$ y $s^2 = 24.01$. Calcule e interprete un intervalo de confianza de 97,5%, para la cantidad promedio de cloroformo del agua potable de esta ciudad.
- 4.29 En un experimento de Química General, se desea determinar la cantidad (en mililitros) de una solución de hidróxido de sodio (NaOH) que se necesita para neutralizar 1 ml de cierto ácido. Al realizar el experimento en el laboratorio habrá variación debida al error experimental, por lo que llevan a cabo 35 análisis volumétricos, usando fenolftaleína como indicador, produciendo $\bar{x} = 77.97$ ml. y $s^2 = 3.6618$. Calcule e interprete un intervalo de confianza del 90% para estimar el número medio de mililitros que se requieren para neutralizar 1 ml de ácido.
- 4.30 El punto de ebullición del azufre tiene una variancia de 0,6889 grados Celsius. ¿Qué tamaño debe tener la muestra para asegurar con una confianza del 99% que el error para estimar el punto promedio de ebullición del azufre sea a lo más de 0,15 grados Celsius?
- 4.31 Los siguientes datos representan el contenido de colesterol (en miligramos por mililitro, mg/ml) de un conjunto de alimentos dietéticos:
- $$n = 32 \quad \bar{x} = 2,57 \quad s = 0,5466$$
- Calcule un intervalo de confianza del 95% para el contenido promedio de colesterol del alimento.
- 4.32 Se sabe que la concentración de un componente en una mezcla líquida se distribuye aproximadamente en forma normal, con una variancia de 2,3095. Se desea obtener una confianza del 97,5% de que el error al estimar la concentración media sea de 0,7. ¿Qué tamaño de muestra debería emplearse?

- 4.33 Se está estudiando el contenido de cobre en una aleación con la esperanza de reducir finalmente el costo de manufactura. Por estudios anteriores se estima que la proporción es 0,12. ¿Cuánto mezclas deben estudiarse para estimar la proporción de cobre en una aleación con un error de muestreo de 0,04, a un nivel de confianza de 90%?
- 4.34 Un ingrediente para retardar la fermentación de cierta cerveza se agrega en el proceso de elaboración. El tiempo de fermentación en semanas fueron los siguientes:
- 2 4 6 2 4 6 2 4 6 2 4 6
- Suponiendo que el tiempo de fermentación de la cerveza es normal, ¿calcule e interpretar un intervalo de confianza del 95% para el promedio en semanas del tiempo de fermentación de la cerveza después de agregarle el ingrediente?
- 4.35 Se sabe por un estudio elaborado anteriormente, que el porcentaje de calcio en cierto compuesto tiene una desviación estándar de 8,047 (puntos porcentuales). Se pide a 20 de los 60 estudiantes de un curso de Química General, que determinen el porcentaje de calcio de igual número de mezclas del mismo compuesto. Los resultados mostraron que el promedio de calcio encontrado por los estudiantes era de 84,715. ¿Calcule e interprete un intervalo de confianza del 90% para el verdadero promedio del porcentaje de calcio del compuesto?
- 4.36 El contenido de cobre en una aleación tiene una variancia de 43,81 (puntos porcentuales). Con la esperanza de reducir el costo de manufactura, se quiere analizar una mezcla reciente que tenga un error de estimación de 2 puntos al estimar el promedio poblacional, a un nivel de confianza del 97%. Calcule el número de mezclas necesarias que deben analizarse para cumplir con las condiciones antes citadas.

4.37 Según la experiencia del gerente de producción de una empresa, el 12,5% de los artículos presentan algún defecto de elaboración o de empaque. ¿Calcule un tamaño de muestra necesario para estimar el porcentaje real de artículos defectuosos producidos por la empresa, con un error de muestreo del 2% y nivel de significación del 10%?

4.38 La conductividad térmica del grafito varía con la temperatura según la ecuación $K=K_0-\alpha T$. Experimentalmente, solo es posible obtener una conductividad media en un intervalo de temperatura. Se requiere estimar la conductividad promedio con un intervalo de confianza del 93%, dado que de una muestra aleatoria de 35 ensayos, resultó una media aritmética 1,2825 y una desviación estándar de 0,1315 K_m , determinado T y 25°C en kilo-ergio/cm² seg.°C cm⁻¹.

4.39 Considere los puntos de ebullición en grados centígrados de algunas sustancias como sigue:

80,2 78,3 34,6 65,2 48,9 78,5

Calcule e interprete un intervalo de confianza del 95% para el punto promedio de ebullición de las sustancias.

4.40 Considere la carga iónica para algunos iones metálicos, como se establece a continuación:

1,0 1,5 2,1 3,1 2,7 2,8 6,7 4,8 4,7

Calcule e interprete un intervalo de confianza del 95% para la media de la carga iónica.

Capítulo 5

Estadística Inferencia: Prueba de Hipótesis

- 5.1 Método de pruebas de hipótesis
- 5.2 Pruebas de hipótesis paramétricas en una población normal
- 5.3 Contrastes de dos distribuciones normales independientes
- 5.4 Contrastes para la diferencia de medias pareadas
- 5.5 Contrastes de una proporción
- 5.6 Contrastes sobre la diferencia de proporciones
- 5.7 Problemas

Prueba de Hipótesis

Pueden presentarse en la práctica, situaciones en las que exista una teoría preconcebida relacionada con la característica de la población sometida a estudio. Tal sería el caso, por ejemplo, se piensa que un método de producción nuevo puede tener un porcentaje de rendimiento mayor que otro estándar, o cuando se plantea si los contenidos de alcohol de distintas cervezas vendidas en el mercado nacional son similares.

Este tipo de circunstancias es el que nos lleva al estudio de la Estadística Inferencial y, específicamente, a la prueba de Hipótesis. Implica, en cualquier investigación, la existencia de dos teorías o hipótesis implícitas, que se denomina hipótesis nula e hipótesis alternativa, que de alguna manera reflejarán esa idea a priori que se tiene y que es contrastada con la "realidad".

De la misma manera, aparecen, implícitamente, diferentes tipos de errores que es posible cometer durante el procedimiento. No puede olvidarse que, habitualmente, el estudio y las conclusiones obtenidas para una población cualquiera, se habrán apoyado exclusivamente en el análisis de sólo una parte de ésta. La probabilidad con la que se está dispuesto a asumir estos errores, dependerá, por ejemplo, el tamaño de la muestra requerida.

Son estudiadas en este capítulo las pruebas de hipótesis para los parámetros más usuales que fueron citadas en los capítulos anteriores: medias, varianzas y proporciones, para una o dos poblaciones. Los contrastes desarrollados se apoyan en que los datos de partida siguen una distribución normal.

5.1 Método de pruebas de hipótesis

La prueba de hipótesis es un procedimiento mediante el cual se compara lo propuesto, contra la evidencia empírica que proporciona la observación de datos, provenientes de la población sobre la cual se hace la hipótesis. Este procedimiento contempla etapas que describiremos en los siguientes apartados.

5.1.1 Formulación de las hipótesis

Hipótesis Nula (H_0)

Tiene un carácter negativo y escéptico, reta al investigador a demostrar que el enunciado es erróneo. Cuando no es rechazada la hipótesis nula implica que no existe suficiente información estadísticamente abrumadora que la refute. En tanto, el rechazo de la hipótesis nula implicará la aceptación de la hipótesis alternativa. En general, la hipótesis nula se formula de manera que su rechazo lleva a la aceptación de la conclusión deseada.

Hipótesis Alternativa (H_1)

Es la afirmación que plantea el investigador. Establece que el parámetro poblacional tiene un valor diferente, de alguna manera, al establecido en la hipótesis nula.

Ejemplo. Suponga que se debe realizar un estudio sobre la altura media de los estudiantes universitarios. Antes de tomar una muestra, lo lógico es hacer la siguiente suposición a priori,

H_0 : La altura media no difiere del resto de habitantes.

Al obtener una muestra de tamaño $n=8$, se podría encontrar los siguientes casos:

Muestra 1	1,50	1,52	1,48	1,55	1,57	1,60	1,53	1,45
Muestra 2	1,65	1,74	1,68	1,75	1,85	1,78	1,65	1,70

Intuitivamente, en el caso de la muestra 1 sería lógico suponer que salvo que la muestra obtenida sobre los habitantes del pueblo sea muy poco representativa, la

hipótesis H_0 debe ser rechazada. En el caso de la muestra 2 tal vez no pueda afirmarse con rotundidad que la hipótesis H_0 sea cierta; sin embargo, no se podría descartar y la aceptarla por una cuestión de simplicidad.

Este ejemplo sirve como introducción de los siguientes conceptos: En una prueba de hipótesis (también denominado Contraste de Hipótesis) se decide si cierta hipótesis H_0 que es denominada *hipótesis nula* puede ser rechazada o no a la vista de los datos suministrados por una muestra de la población. Para realizar el contraste es necesario establecer previamente una hipótesis alternativa (H_1) que será admitida cuando H_0 sea rechazada. Normalmente H_1 es la negación de H_0 , aunque esto no es necesariamente así.

5.1.2. Errores posibles al tomar una decisión errónea

La decisión de rechazar o no la hipótesis nula está al fin y al cabo basada en la elección de una muestra tomada al azar y, por tanto, es posible cometer decisiones erróneas. Los errores que se pueden cometer se clasifican como sigue:

Error Tipo I

Es el error que consiste en rechazar H_0 cuando es cierta. La probabilidad de cometer este error es lo que recibe el nombre de nivel de significancia. Es una costumbre establecida el denotarlo siempre con la letra griega alfa (α). En la práctica de análisis de datos, se presenta cuando los resultados de la muestra llevan al rechazo de la hipótesis nula que en realidad es verdadera.

$$\alpha = P[\text{Rechazar } H_0 / H_0 \text{ es verdadera}] \quad (5.1)$$

Error Tipo II

Es el error que consiste en no rechazar H_0 cuando es falsa. La probabilidad de cometer este error es representada con la letra β . A diferencia de α , que especifica el investigador, la magnitud de β depende del valor real del parámetro de la población.

$$\beta = P[\text{No rechazar } H_0 / H_0 \text{ es falsa}] \quad (5.2)$$

Observaciones importantes sobre los errores:

1. Los *errores de tipo I y II* no están relacionados más que del siguiente modo: Cuando α decrece β crece. Por tanto, no es posible encontrar un contraste que hagan tan pequeños ambos errores simultáneamente. De este modo, es siempre necesario privilegiar a una de las hipótesis, de manera que no será rechazada, a menos que su falsedad se haga muy evidente. En los contrastes, la hipótesis privilegiada es H_0 que sólo será rechazada cuando la evidencia de su falsedad supere el umbral del 100 $(1-\alpha)$ %.
2. Al tomar α muy pequeño el valor de β tiende aproximarse a uno. Lo ideal a la hora de definir un prueba es encontrar un compromiso satisfactorio entre α y β (aunque siempre a favor de H_0). Se denomina *potencia de un contraste* a la cantidad $1 - \beta$, es decir

$$\begin{aligned} \text{Potencia del contraste} &= 1 - \beta \\ &= P[\text{Rechazar } H_0 / H_0 \text{ es falsa}] \end{aligned} \tag{5.3}$$

3. En el momento de elegir una hipótesis privilegiada es posible titubear entre si elegir una dada o bien su contraria. Criterios a tener en cuenta en estos casos son los siguientes:

- **Simplicidad científica:** A la hora de elegir entre dos hipótesis científicamente razonables, se toma como H_0 aquella que sea más simple.
- **Las consecuencias de equivocarse:** Por ejemplo, al juzgar el efecto que puede causar cierto tratamiento médico que está en fase de experimentación, en principio, se ha de tomar como hipótesis nula aquella cuyas consecuencias por no rechazarla siendo falsa son menos graves, y como hipótesis alternativa aquella en la que el aceptarla siendo falsa trae peores consecuencias. Es decir,

H_0 : El paciente empeora o queda igual ante el tratamiento

H_1 : El paciente mejora con el tratamiento

Otro ejemplo claro es cuando acaban de instalar un nuevo ascensor en un edificio y se quiere saber si el ascensor caerá o no al vacío con las personas dentro.

Una persona prudente es la que espera a que un número suficiente de personas hayan usado el ascensor (muestra aleatoria) y realiza un contraste del tipo:

H_0 : El ascensor se caerá

H_1 : El ascensor no se caerá

y sólo aceptará la hipótesis alternativa para $\alpha \approx 0$ aunque para ello tenga que ocurrir que $\beta \approx 1$, ya que las consecuencias del *error de tipo I* (ir al hospital) son mucho más graves que las del *error del tipo II* (subir a pie varios pisos).

Es decir, a la hora de decidirse por una de las dos hipótesis no basta con elegir la más probable (nadie diría "voy a tomar el ascensor pues la probabilidad de que no se caiga es del 0,65"). Hay que elegir siempre la hipótesis H_0 a menos que la evidencia a favor de H_1 sea muy significativa.

5.1.3. Determinación del valor observado del estadístico de prueba

En la tercera etapa se establece el valor real del estadístico que se utilizará en la prueba. El estadístico de prueba es una medida de cuánto se acercan los resultados obtenidos en la muestra a la hipótesis nula. Con frecuencia, el estadístico sigue una distribución conocida, como la Normal, t de Student, Ji Cuadrada o Fisher. Según la finalidad de ésta, el estadístico puede ser uno de los siguientes:

- Una media muestral
- Una proporción de la muestra
- La diferencia entre dos medias muestrales
- La diferencia entre dos proporciones de muestras
- El estadístico calculado de Ji cuadrada
- La razón F calculada

5.1.4. Determinación del valor crítico del estadístico de prueba

En esta etapa es calculado el valor crítico que el estadístico deberá rebasar a fin de rechazar la hipótesis nula. Este valor se obtiene consultando la tabla correspondiente del estadístico.

Para las pruebas de hipótesis bilaterales destinadas a determinar diferencias entre dos medias aritméticas o proporciones muestrales, habrá dos valores críticos, uno en cada dirección de la distribución de probabilidad.

5.1.5. Comparación del valor observado del estadístico de prueba con el valor crítico

El valor crítico definirá las *regiones de aceptación y rechazo* de la hipótesis nula. Si el valor observado del estadístico supera al valor crítico, la hipótesis nula será rechazada. En caso contrario, no se rechaza. Cuanto más pequeño es el nivel de significancia elegido, más difícil será que un estadístico observado logre alcanzar la *región de "rechazo"*.

5.2 Pruebas de hipótesis paramétricas en una población normal

Suponga que la variable aleatoria X estudiada sobre la población sigue una distribución normal y es seleccionada una muestra de tamaño n , mediante muestreo aleatorio simple. A continuación, se estudian las técnicas para contrastar hipótesis sobre los parámetros que rigen X .

5.2.1 Prueba de hipótesis para un promedio con varianza conocida

Suponiendo que $X \rightarrow N(\mu, \sigma^2)$ donde σ^2 es conocido y se quiere contrastar si es posible que μ (desconocida) sea en realidad cierto valor μ_0 fijado. Esto es un supuesto teórico que nunca se dará en la realidad pero servirá para introducir la teoría sobre contrastes. La prueba se escribe entonces como:

$$H_0: \mu = \mu_0 \qquad H_1: \begin{cases} \mu < \mu_0 \\ \mu > \mu_0 \\ \mu \neq \mu_0 \end{cases}$$

Como fue mencionado anteriormente, la técnica para hacer el contraste consiste en suponer que H_0 es cierta. Si H_0 es cierta, el valor Z_{obs} obtenido sobre la muestra es,

$$Z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{5.4}$$

esté con una gran probabilidad cercana a cero. Esto se expresa fijando un nivel de significación α , y tomando como región teórica(C) a los valores que son extremos, con probabilidad total α , según sea el tipo de contraste: una cola o dos colas. Para contraste de dos colas la región teórica está definida por,

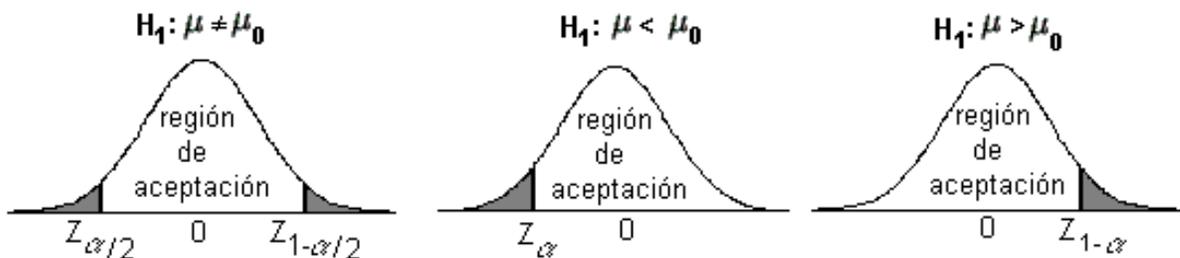
$$C = \{Z_{obs} : |Z_{obs}| \leq Z_{1-\alpha/2}\} \quad (5.5)$$

Luego es rechazada la hipótesis nula si

$$|Z_{obs}| > Z_{1-\alpha/2}$$

Consideré un contraste donde ahora la hipótesis alternativa sea compuesta: $H_1: \mu < \mu_0$ y $H_0: \mu > \mu_0$. La región crítica está formada por los valores extremadamente bajos o altos respectivamente de Z_{obs} , con probabilidad α . Entonces la región de aceptación, o de modo más correcto, de no rechazo de la hipótesis nula es $Z_{obs} > Z_\alpha$ para los contrastes unilaterales con la cola izquierda de la distribución y $Z_{obs} < Z_{1-\alpha}$ para la prueba con la cola derecha.

Regiones críticas para la prueba bilateral y unilateral de una media con varianza conocida



Ejemplo. Un investigador desea contrastar la hipótesis que el límite de concentración del isótopo radiactivo estroncio 90 en la leche es de 5 por litro. Para tal efecto, selecciona aleatoriamente una muestra de lecherías. La hipótesis nula considera el límite de referencia $\mu = 5$, contra la hipótesis alternativa que es mayor a 5.

Para verificar si las especificaciones se cumplían en cierta región del país, fueron seleccionadas aleatoriamente 40 lecherías, encontrándose una media aritmética de 5,4 y una desviación estándar de la muestra de 0,493 por litro. ¿Existe

una diferencia en la concentración promedio con el límite tolerable? Realizar la prueba con un nivel de confianza del 90%.

Simbólicamente, se puede escribir las hipótesis como;

$$H_0: \mu = 5$$

$$H_a: \mu > 5$$

y para los datos en uso,

$$\begin{aligned} Z_{obs} &= \frac{5,4 - 5}{0,493/\sqrt{40}} \\ &= \frac{0,4}{0,07795} \\ &= 5,13 \end{aligned}$$

El valor teórico de la distribución normal al nivel significancia establecido de 0,10, resulta aproximadamente 1,64. Dado que 5,13 es mayor a 1,64, entonces la decisión sería rechazar la hipótesis nula H_0 .



La conclusión a que se llegaría al nivel de confianza del 90% es que la concentración del isótopo radiactivo estroncio 90 en todas las lecherías es significativamente mayor al límite tolerable.

5.2.2 Prueba de hipótesis para un promedio con varianza desconocida

En la sección anterior, el valor de la varianza estaba a nuestra disposición por experiencia pasada o con base en el teorema del límite central y bajo el supuesto de normalidad de la distribución del promedio muestral. Esto es satisfactorio para muestras grandes ($n > 30$), pero el error que resulte puede ser muy considerable si n es pequeña.

Sea $X \rightarrow N(\mu, \sigma^2)$ donde ni μ ni σ^2 son conocidos y se quiere realizar el contraste

$$H_0: \mu = \mu_0 \qquad H_1: \begin{cases} \mu < \mu_0 \\ \mu > \mu_0 \\ \mu \neq \mu_0 \end{cases}$$

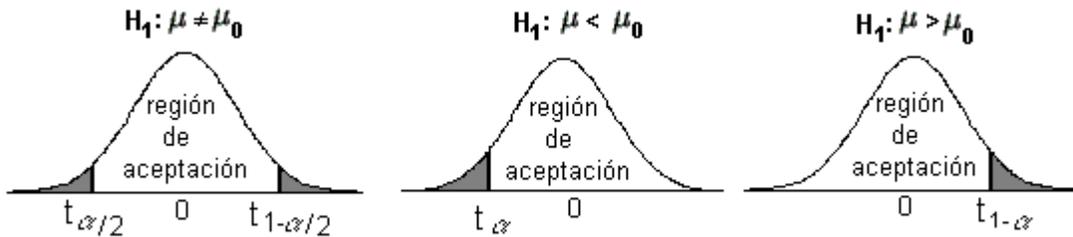
Al no conocer σ^2 va a ser necesario estimarlo a partir de su estimador insesgado: la varianza muestral s^2 . Por ello la distribución del estimador del contraste será una t de Student.

Para dar una forma homogénea a todas las pruebas de hipótesis es costumbre denominar al valor del estadístico del contraste calculado sobre la muestra como *valor observado* y a los extremos de la región crítica, como *valores teóricos*. Definiendo entonces,

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{5.6}$$

Para la prueba bilateral, $H_0: \mu = \mu_0$ el valor teórico está definido como $t_{teorica} = t_{(n-1)(1-\alpha/2)}$. En tanto, cuando el contraste es unilateral de la forma $H_1: \mu < \mu_0$ o $H_1: \mu > \mu_0$, el valor teórico es $t_{teorica} = t_{(n-1)(1-\alpha)}$.

Regiones críticas para la prueba bilateral y unilateral
de una media con varianza desconocida



Ejemplo. Cinco determinaciones de mercurio metálico sobre una muestra dieron como resultados:

21,0 20,0 20,5 20,8 21,2 % Hg⁰

El verdadero valor del porcentaje en Hg⁰ es 20,0%. *¿Difiere la determinación media de las muestras del verdadero valor, al nivel de confianza del 95%?*

La hipótesis nula y alternativa para este problema se expresarían como sigue:

$$H_0: \mu = 20$$

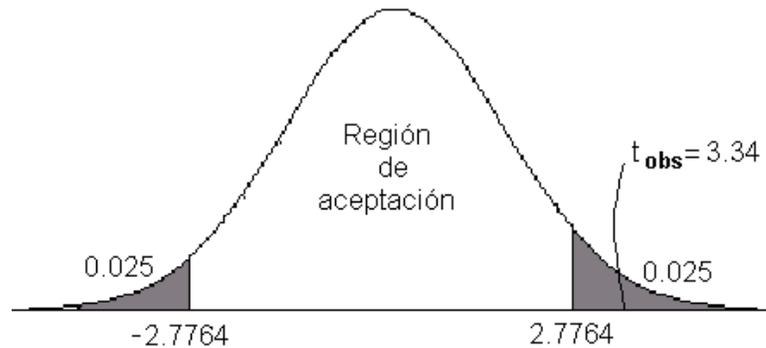
$$H_1: \mu \neq 20$$

La región de rechazo estaría dividida en las dos colas de la distribución, esto implica que el nivel de significancia $(1-0,95)100=5\%$ se divide en dos partes iguales de 2,5%. La media aritmética y la desviación estándar de las cinco determinaciones resulta 20,7 y 0,469, respectivamente.

Para estos datos se tiene,

$$\begin{aligned} t_{obs} &= \frac{20,7 - 20}{0,469/\sqrt{5}} \\ &= \frac{0,7}{0,2097} \\ &= 3,34 \end{aligned}$$

Dado que el valor teórico determinado con la tabla **t** de Student para un área de 0,025 en cada cola de la distribución con 4 grados de libertad es 2,7764, entonces se rechaza H_0 .



La conclusión es que las determinaciones medias de las muestras difieren del verdadero valor, a un nivel de confianza del 95%.

Es importante mencionar que, conforme aumenta el número de grados de libertad, la distribución t se aproxima a la normal. Para muestras de tamaño grande, los valores críticos de la distribución t serán virtualmente los mismos que la distribución normal.

El valor crítico al nivel de significancia de 0,01 (dos colas) de la distribución normal es 2,58. El valor correspondiente de la distribución t es de 2,75 para 30 grados de libertad, 2,6603 para 60 grados de libertad, 2,6259 para 100 grados de libertad y 2,609 para 150 grados de libertad. Por tanto, se podría decir que cuando hay, por lo menos, 150 grados de libertad, la distribución normal se puede intercambiar con la distribución t para obtener los valores críticos correctos.

5.3 Contrastes de dos distribuciones normales independientes

Se considera a lo largo de toda esta sección a dos poblaciones normales que representadas mediante $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$. De las que, de manera *independiente*, se extraen muestras de tamaño respectivo n_1 y n_2 . La prueba a realizar está relacionada con las diferencias existentes entre ambas medias.

5.3.1 Contraste de promedios con varianzas conocidas

De manera similar al caso del contraste para una media, se quiere en esta ocasión contrastar la hipótesis de que las dos poblaciones (cuyas varianzas se suponen conocidas) sólo difieren en una cantidad δ , esto es $H_0: \mu_1 - \mu_2 = \delta$. Las hipótesis alternativas darán lugar a contrastes unilaterales o bilaterales como se estudiará más tarde. Para ello se tiene la distribución del siguiente estadístico de prueba calculado bajo el supuesto que las poblaciones (o variables consideradas en cada población) tienen aproximadamente una distribución normal y las varianzas poblacionales σ_1^2 y σ_2^2 son conocidas.

$$Z_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (5.7)$$

Consideraré en primer lugar el contraste de dos colas

$$H_0: \mu_1 - \mu_2 = \delta$$

$$H_0: \mu_1 - \mu_2 \neq \delta$$

Se define entonces el

$$Z_{teórica} = Z_{1-\alpha/2}$$

y la prueba consiste en

No rechazar H_0 si $|Z_{obs}| < |Z_{teórica}|$

Rechazar H_0 si $|Z_{obs}| > |Z_{teórica}|$

Ejemplo. Se obtuvieron los resultados siguientes de la normalidad de una disolución de ácido clorhídrico según el método de determinación gravimétrica de cloruros y el método de valoraciones frente a carbono sódico puro.

Método	Números de muestras	Media aritmética	Desviación estándar
Gravimétrica	35	0,10505	$2,38 \times 10^{-4}$
Volumétrica	40	0,1055	$6,04 \times 10^{-4}$

El estudio deseaba saber si había alguna diferencia significativa entre los valores medios de las normalidades obtenidas por los dos métodos, con un nivel de significación del 95%. La hipótesis nula y alternativa para este ejemplo serían,

$$H_0: \mu_{gra} - \mu_{vol} = 0$$

$$H_1: \mu_{gra} - \mu_{vol} \neq 0$$

Para los datos con que se trabaja se tiene,

$$\begin{aligned} Z_{obs} &= \frac{0,10505 - 0,1055}{\sqrt{\frac{(2,38 \times 10^{-4})^2}{35} + \frac{(6,04 \times 10^{-4})^2}{40}}} \\ &= \frac{-4,5 \times 10^{-4}}{\sqrt{1,6184 \times 10^{-9} + 9,1204 \times 10^{-9}}} \\ &= \frac{-4,5 \times 10^{-4}}{1,03626 \times 10^{-4}} \\ &= -4,34 \end{aligned}$$

Al nivel de confianza del 95%, el valor tabular en la distribución normal estándar corresponde a $\pm 1,96$. Como se observa, al comparar el valor observado y el teórico, $|-4,34| > |1,96|$, se toma la decisión de rechazar H_0 . Se concluye que se tiene evidencia estadística para afirmar que existe una diferencia significativa entre los valores medios de las normalidades obtenidas por los dos métodos.

5.3.2 Contraste de promedios con varianzas desconocidas

Ahora considere el problema de contrastar $H_0: \mu_1 - \mu_2 = \delta$, cuando sólo es conocida que las varianzas de ambas poblaciones son iguales ($\sigma_1^2 = \sigma_2^2$ o sea, homocedasticidad), pero desconocidas. El estadístico que a usar para el contraste es

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \quad (5.8)$$

Para el contraste de significación

$$H_0 : \mu_1 - \mu_2 = \delta$$

$$H_1 : \mu_1 - \mu_2 \neq \delta$$

el valor teórico se calcula como, $t_{teórica} = t_{(n_1+n_2-2), (1-\alpha/2)}$ y tiene que rechazarse o admitir la hipótesis nula siguiendo el criterio

$$\text{No rechazar } H_0 \text{ si } |t_{obs}| < |t_{teórica}|$$

$$\text{Rechazar } H_0 \text{ si } |t_{obs}| > |t_{teórica}|$$

El caso más problemático para el contraste de medias, es cuando sólo es conocido de las dos poblaciones que su distribución es normal, y que sus varianzas no son conocidas y significativamente diferentes (no homocedasticidad).

En este caso el estadístico de contraste tendrá una ley de distribución muy particular. Consistirá en una distribución *t* de Student, con un número de *grados de libertad* que en lugar de depender de modo determinista de la muestra (a través de su tamaño), depende de un modo aleatorio mediante las varianzas muestrales. Concretamente, el estadístico que interesa es

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (5.9)$$

El número de grados de libertad que se calcula mediante la *fórmula de Welch*,

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 + 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 + 1} \left(\frac{s_2^2}{n_2}\right)^2} - 2 \quad (5.10)$$

No es desarrollado en detalle los cálculos a realizar, pues la técnica para efectuar los contrastes es análoga a las vistas anteriormente cuando las varianzas son desconocidas e iguales.

Finalmente, si las medias poblacionales de dos muestras independientes obtenidas de poblaciones normales son idénticas, esto se reduce a los casos anteriores tomando $\delta = 0$, es decir, realizando el contraste:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Ejemplo. Diez muestras de piedra caliza, tomadas del suministro de materia prima almacenada en un tajo, dieron los siguientes resultados analíticos en % de CaO:

50,6 50,9 48,5 50,0 46,5 50,5 49,2 49,7 43,0 50,5

Otras diez muestras se tomaron después de haber sometido a molienda la materia prima, dieron los siguientes resultados en % de CaO:

50,5 50,3 50,9 50,7 50,3 51,4 49,2 49,9 50,4 50,6

El problema consiste en medir estadísticamente si hay evidencia de una diferencia significativa antes y después de someter la materia prima a molienda en los porcentajes de CaO, a un nivel de confianza del 99%. Los resultados se pueden resumir como sigue:

Tratamiento	Tamaño de Muestra	Promedio Aritmético	Variación	Desviación Estándar
Antes	10	48,94	6,0738	2,4645
Después	10	50,42	0,344	0,5865

La prueba es de dos colas para determinar si hay alguna diferencia entre las medias, entonces la hipótesis nula y alternativa serán,

$$H_0: \mu_{antes} = \mu_{después}$$

$$H_1: \mu_{antes} \neq \mu_{después}$$

Con los datos de la tabla anterior se tiene,

$$\begin{aligned} t_c &= \frac{48,94 - 50,42}{\sqrt{9(6,0738) + 9(0,344)}} \sqrt{\frac{10(10)(18)}{20}} \\ &= \frac{-1,48}{\sqrt{57,7602}} \sqrt{90} \\ &= -1,8474 \end{aligned}$$

El valor teórico de la tabla de *t-Student*, se encuentra calculando los grados de libertad como la suma de los tamaños de muestra menos dos ($10+10-2=18$) y tal que, es una prueba de dos colas con un nivel de significancia del 1% este se divide entre 2. Se obtiene el valor tabular de 2,8784 y puesto que el estadístico de prueba es inferior no se rechaza H_0 . Esto es, no se tiene fuerte evidencia para concluir que existe una diferencia significativa promedio antes y después de someter la materia prima a molienda, para determinar los porcentajes de CaO.

5.4 Contrastes para la diferencia de medias pareadas

El procedimiento consiste en tomar dos observaciones de un mismo sujeto, en condiciones homogéneas. Un ejemplo de observaciones pareadas consiste en considerar a un conjunto de n personas a las que se le aplica un tratamiento médico y se mide por ejemplo el nivel de insulina en la sangre antes (X) y después del mismo (Y).

Paciente	x_i	y_i	d_i
1	140	130	20
2	190	150	40
...	...		
n	155	120	35

No es posible considerar a X e Y como variables independientes ya que va a existir una dependencia clara entre las dos variables. Cuando se quiere contrastar el que los pacientes han experimentado o no una mejoría con el tratamiento, simbólicamente d_i es la diferencia entre las observaciones antes y después del tratamiento.

$$d_i = x_i - y_i$$

Suponga que la *variable aleatoria*, que define la diferencia entre el antes y después del tratamiento es una v.a. que se distribuye normalmente, pero cuyas media y varianza son desconocidas.

$$d \sim N(\mu_d, \sigma_d^2)$$

Luego se contrasta la hipótesis de que el tratamiento ha producido cierto efecto θ , utilizando la media observada de las diferencias pareadas resultantes.

$$H_0 : \mu_d = \theta$$

El estadístico de contraste está definido por,

$$t_{obs} = \frac{\bar{d} - \Delta}{S_d / \sqrt{n}} \quad (5.11)$$

donde \bar{d} es la media muestral de las diferencias d_i y s_d representa la desviación estándar muestral de las mismas. Sin embargo, generalmente se calcula la varianza de las diferencias y luego su raíz cuadrada con propósito de obtener la desviación estándar.

$$\begin{aligned} s_d^2 &= \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n d_i^2 - \frac{\left(\sum_{i=1}^n d_i \right)^2}{n} \right] \end{aligned} \quad (5.12)$$

Ahora, considere el contraste bilateral de tipo,

$$H_0 : \mu_d = \theta$$

$$H_1 : \mu_d \neq \theta$$

Entonces se define $t_{teo} = \pm t_{(n-1)(1-\alpha/2)}$, y se rechaza la hipótesis nula cuando $|t_{obs}| > |t_{teórica}|$.

No supone ninguna dificultad para la varianza conocida (σ_d^2), ya que el estadístico del contraste es,

$$Z_{obs} = \frac{\bar{d} - \theta}{\sigma_d / \sqrt{n}} \quad (5.13)$$

Ejemplo. Considere un análisis realizado por dos diferentes laboratorios, donde se estudian nueve muestras de peróxido de hidrógeno (H_2O_2), como sigue:

Muestra	Lab. A	Lab. B	$d_i=A - B$
1	93,08	92,97	0,11
2	92,59	92,85	-0,26
3	91,36	91,86	-0,50
4	91,60	92,17	-0,57
5	91,91	92,33	-0,42
6	93,49	93,28	0,21
7	92,03	92,30	-0,27
8	92,80	92,70	0,10
9	91,03	91,50	-0,47

Se desea determinar si existe diferencia en los análisis de los laboratorios. En este caso $\theta = 0$, donde las hipótesis nula y alternativa pueden expresarse como sigue:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

Para estos datos, el promedio de las diferencias es,

$$\begin{aligned} \bar{d} &= \frac{\sum_{i=1}^n d_i}{n} \\ \bar{d} &= \frac{(0,11 - 0,26 - 0,50 - 0,57 + \dots + - 0,47)}{9} \\ &= \frac{- 2,04}{9} = -0,23 \end{aligned}$$

La variancia de la diferencia es,

$$\begin{aligned} s_d^2 &= \frac{1}{8} \left[1,1789 - \frac{(- 2,07)^2}{9} \right] \\ &= \frac{1}{8} [1,1789 - 0,4761] \\ &= \frac{1}{8} 0,7028 \\ &= 0,08785 \end{aligned}$$

de modo que,

$$s_d = \sqrt{0,08785} = 0,296$$

Utilizando la fórmula 5.11, el estadístico de prueba resultaría,

$$\begin{aligned}t &= \frac{-0,23}{0,296/\sqrt{9}} \\ &= \frac{-0,23}{0,0987} \\ &= -2,33\end{aligned}$$

El valor teórico de la distribución t, con 8 grados de libertad y un nivel de significancia del 5% es $\pm 2,306$. Dado que $-2,33 < -2,306$, se rechaza H_0 . Entonces, se concluye que existe una diferencia en las mediciones entre los laboratorios, a un nivel de confianza del 95%.

5.5 Contrastes de una proporción

En las secciones anteriores se utilizaron procedimientos de contraste de hipótesis para datos cuantitativos. El concepto de la prueba de hipótesis también se puede utilizar en relación con datos cualitativos, esto permite decidir con base en los datos muestrales, si la proporción poblacional \mathbf{P} , es igual, mayor o menor que una constante determinada.

Suponga que se tiene una sucesión de observaciones independientes, de modo que cada una de ellas se comporta como una distribución de Bernoulli de parámetro \mathbf{P} : La variable aleatoria X , definida como el número de éxitos obtenidos en una muestra de tamaño n es por definición una v.a. de distribución binomial,

$$X = \sum_{i=1}^n X_i \tag{5.14}$$

La proporción muestral (estimador del verdadero parámetro \mathbf{P} a partir de la muestra) es,

$$\hat{p} = \frac{X}{n} \tag{5.15}$$

Interesa el contraste de significación de,

$$H_0: P = P_0$$

donde P_0 es un valor prefijado, frente a otras hipótesis alternativas. El estadístico de contraste sigue una distribución aproximadamente normal para tamaños muestrales suficientemente grandes, si:

$$[np \text{ y } (1 - p) \geq 5]$$

Cuando n es pequeña, las pruebas referentes a proporciones se pueden hacer directamente con la distribución binomial.

$$Z_{obs} = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \quad (5.16)$$

Para el contraste bilateral,

$$H_0: P = P_0$$

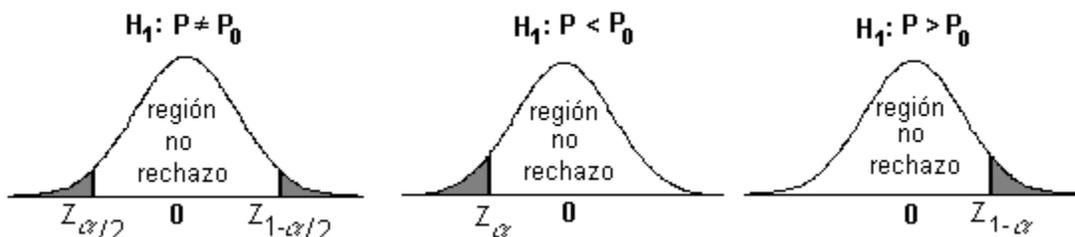
$$H_1: P \neq P_0$$

Entonces se define Z_{obs} como la ecuación (5.16) y el valor teórico $Z_{teórica} = Z_{1-\alpha/2}$. Siendo el criterio de no rechazo o rechazo de la hipótesis nula,

$$\text{No rechazar } H_0 \text{ si } |Z_{obs}| \leq |Z_{teórica}|$$

$$\text{Rechazar } H_0 \text{ si } |Z_{obs}| > |Z_{teórica}|$$

Regiones críticas para la prueba bilateral y unilateral de una proporción



Ejemplo. Una compañía farmacéutica afirma que un antibiótico tiene 80% de efectividad. Se somete a prueba una muestra aleatoria de 50 casos de estudio y se obtiene una proporción de 75,7%. ¿Presentan los datos suficiente evidencia

para refutar la afirmación de la compañía? Para este problema, si se selecciona un nivel de significación α de 0,05.

La hipótesis nula y alternativa para este problema se pueden expresar como sigue:

$$H_0 : P = 0,80$$

$$H_1 : P \neq 0,80$$

Con los datos que se tienen,

$$\begin{aligned} Z_{obs} &= \frac{0,757 - 0,80}{\sqrt{\frac{0,8(0,2)}{50}}} \\ &= \frac{-0,043}{0,0566} \\ &= -0,76 \end{aligned}$$

Al 95% de confianza el valor teórico Z es -1,645; por tanto, la hipótesis nula no se rechazaría porque el estadístico de prueba (-0,76) no ha caído en la región de rechazo, o simplemente comparando los valores absolutos, se observa que 1,645 es mayor 0,76. Se llegaría a la conclusión de que no hay suficiente evidencia para refutar la afirmación de la compañía.

5.6 Contrastes sobre la diferencia de proporciones

En lugar de preocuparse por las diferencias entre dos poblaciones en términos de una variable cuantitativa, el interés podría estar, centrarse en las diferencias en alguna característica cualitativa. Entonces, el problema sería decidir si la diferencia observada entre dos proporciones, proveniente de dos muestras independientes es significativa, o pueden ser atribuidas al azar y también concluir que las proporciones son iguales.

Suponga dos muestras independientes tomadas sobre dos poblaciones, en la que se estudia una variable de tipo dicotómico (Bernoulli). Se contabilizan en cada caso el número de éxitos en cada muestra se tiene que cada una de ellas se distribuye como una variable aleatoria binomial, de modo que los estimadores de las proporciones en cada población tienen distribuciones aproximadamente normales (cuando n_1 y n_2 son bastante grandes).

$$\hat{p}_1 = \frac{X_1}{n_1} \approx N\left(P_1, \frac{P_1 Q_1}{n_1}\right)$$

$$\hat{p}_2 = \frac{X_2}{n_2} \approx N\left(P_2, \frac{P_2 Q_2}{n_2}\right)$$

El contraste que interesa realizar es si la diferencia entre las proporciones en cada población es una cantidad conocida δ .

$$H_0 : P_1 - P_2 = \delta$$

Si H_0 fuese cierta se tendría que,

$$\hat{p}_1 - \hat{p}_2 \approx N\left(P_1 - P_2, \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}\right)$$

Desafortunadamente ni P_1 ni P_2 son conocidos de antemano y son utilizados sus estimadores, lo que da lugar a un error que es pequeño cuando los tamaños muestrales son grandes. El estadístico se define como,

$$Z_{obs} = \frac{(\hat{p}_1 - \hat{p}_2) - \delta}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

(5.17)

donde:

$\hat{p}_1 = \frac{x_1}{n_1}$	Proporción muestral en la población 1
$\hat{p}_2 = \frac{x_2}{n_2}$	Proporción muestral en la población 2
x_1	Números de éxitos en la muestra 1
x_2	Números de éxitos en la muestra 2
n_1	Tamaño de la muestra en la población 1
n_2	Tamaño de la muestra en la población 2
$\hat{q}_1 = 1 - \hat{p}_1$	Proporción muestral de fracasos en la muestra 1
$\hat{q}_2 = 1 - \hat{p}_2$	Proporción muestral de fracasos en la muestra 2

El contraste bilateral sobre la diferencia de proporciones es,

$$H_0 : P_1 - P_2 = \delta$$

$$H_1 : P_1 - P_2 \neq \delta$$

Se define el estadístico de prueba como la ecuación (5.17) y se rechaza la hipótesis nula si $|Z_{obs}| > |Z_{\alpha/2}|$.

Ejemplo. Una planta de energía eléctrica operada con carbón ha considerado dos sistemas diferentes para reducir la contaminación del aire. El primer sistema redujo la emisión de contaminantes a niveles de 68% del tiempo, según se determinó de 200 muestras de aire. El segundo sistema, a niveles de 76% de las veces, según quedó determinado en 250 muestras de aire. ¿Pruebe la hipótesis de que no hay diferencia entre los sistemas, a un nivel de confianza de 97,5%?

La hipótesis nula y alternativa para este problema son:

$$H_0 : P_1 - P_2 = 0$$

$$H_1 : P_1 - P_2 \neq 0$$

Para los datos en uso:

$$\begin{aligned} Z_{obs} &= \frac{0,68 - 0,76}{\sqrt{\frac{0,68(0,32)}{200} + \frac{0,76(0,24)}{250}}} \\ &= \frac{-0,08}{\sqrt{\frac{0,2176}{200} + \frac{0,1824}{250}}} \\ &= \frac{-0,08}{\sqrt{0,0018176}} \\ &= \frac{-0,08}{0,04263} \\ &= -1,876 \end{aligned}$$

El valor teórico de la prueba con dos colas según la distribución normal estándar para una confianza del 97,5% es $\pm 2,24$. Resulta que $-1,876 > -2,24$, por tanto, se toma la decisión de no rechazar H_0 . La conclusión es que no hay pruebas de diferencias entre los sistemas en la reducción de la contaminación del aire, a un nivel de confianza del 95%.

5.7 PROBLEMAS

- 5.1 Se le pide a una analista experimentada la evaluación de dos métodos diferentes para la determinación de trazas de plomo en ácido acético glacial. Se realizaron ocho determinaciones mediante cada método, que dieron los siguientes resultados para la concentración del plomo (en ppm):

Método A 1,34 1,33 1,32 1,35 1,32 1,43 1,34 1,31

Método B 1,30 1,26 1,30 1,33 1,20 1,24 1,24 1,33

¿Existe una diferencia significativa en el promedio de trazas de plomo determinado por ambos métodos? $\alpha=0,05$

- 5.2 Un biólogo sostiene que las concentraciones altas de actinomisina D inhiben la síntesis de ARN en las células y, por consiguiente, también la producción de proteínas. En un experimento realizado para probar esta teoría, se compara la síntesis del ARN en células tratadas con dos concentraciones de actinomisina D, 0,6 y 0,7 microgramos por milímetro, respectivamente.

De las 70 células tratadas con la concentración más baja, 55 se desarrollaron normalmente, mientras que sólo 28 de las 50 tratadas con la concentración alta se desarrollaron normalmente. ¿Proporcionan estos datos suficiente evidencia para concluir que hay diferencia entre las tasas de síntesis del ARN normales para las células expuestas a las dos concentraciones de actinomisina D? Use $\alpha=0,10$.

- 5.3 De acuerdo con las normas establecidas para un examen de actitud mecánica, las personas de 18 años deberían promediar 73,2 puntos con una desviación estándar de 8,6 puntos. Si 45 personas escogidas en forma aleatoriamente promediaron 76,7 puntos. Pruebe la hipótesis de que la media de la población en realidad es mayor a un nivel de significación de 0,05.

- 5.4 Una solución que contiene 102,2 mg/L de plomo, es el estándar de referencia certificado. Si 17 muestras dieron como resultado $\bar{x} = 98,6$ mg/L y $s = 3,5$ mg/L. ¿Existe diferencia entre la media obtenida y la media poblacional? Use $\alpha = 0,01$.
- 5.5 Se desea comparar un nuevo método enzimático para determinar colesterol con el método estándar de Lieberman (colorimétrico). Para ello, se observó una muestra para ambos métodos con los resultados siguientes: (en mg/dl)
- | | | | | | | | |
|---------------|-----|-----|-----|-----|-----|-----|-----|
| Enzimático | 305 | 385 | 193 | 162 | 478 | 455 | 238 |
| Colorimétrico | 300 | 392 | 185 | 152 | 480 | | |
- ¿Existe diferencia significativa al nivel de confianza del 99% entre los métodos?
- 5.6 Un organismo de control farmacéutico investiga una muestra de 20 frascos de cierto medicamento para controlar el contenido de cierta droga que afecta el ritmo cardíaco. Se pretende determinar si se están cumpliendo las especificaciones del caso, que establecen que ese contenido no debe diferir de 0,12 gr/100 ml. Al evaluar la muestra, se encontró que el contenido medio es de 0,10 gr /100 ml, con una desviación estándar de 0,02 gr/100 ml. ¿Se estarán o no infringiendo las especificaciones? Use $\alpha = 0,01$.
- 5.7 Las pruebas efectuadas en una muestra de 40 motores diesel producidos por un fabricante mostraron que tenían una deficiencia térmica promedio de 31,4% con una desviación estándar 1,6%. Dado un nivel de significación de 0,05, pruebe la hipótesis nula de que $\mu = 32,3\%$ contra la hipótesis alternativa de que es diferente.
- 5.8 Se utilizaron dos métodos diferentes para determinar la concentración (ppm) de Mg en agua natural. Se hallan los siguientes resultados:

Método A 5,65 5,75 5,64 5,73 5,69 5,71

Método B 5,83 5,78 5,79 5,92 5,95 5,82

¿Existe diferencia significativa al nivel de confianza del 90% entre los métodos?

5.9 Una fábrica de productos químicos ha producido en promedio 800 toneladas por semana. Las producciones en las últimas semanas fueron 785, 805, 790, 793 y 802 toneladas. ¿Indican estos datos que la producción promedio fue menor que 800 toneladas y que, por lo tanto, el rendimiento de la planta ha disminuido? Realice una prueba con un nivel de significación del 5%.

5.10 Cinco mediciones del contenido de alquitrán de cierta marca de cigarrillos producen los resultados:

14,5 14,2 14,4 14,3 14,6 mg por cigarrillo

Pruebe que la diferencia entre promedio de esta muestra y la media del contenido de alquitrán que indica el fabricante, $\mu=14$, es significativa en $\alpha=0,10$.

5.11 Considere los datos del ejercicio 4.39, una vez seleccionada la muestra aleatoria de 195 artículos, 30 tenían algún defecto. Probar la hipótesis, que la proporción de defectuosos es diferente al 10%? Use $\alpha=0,05$.

5.12 Las observaciones de la cantidad de agua tomada por una muestra de 17 ratas que habían sido inyectadas con una solución de cloruro de sodio, tuvieron una media y una desviación estándar de 31,7 y 6,4 centímetros cúbicos, respectivamente. Dado que la cantidad promedio de agua ingerida por las ratas que no han sido inyectadas, en un período de tiempo comparable, es de 27,0 centímetros cúbicos, ¿indican los datos que las ratas inyectadas beben más que las no inyectadas? Haga la prueba usando un nivel de significación del 1%.

- 5.13 Los desechos industriales y la basura que se descargan en los ríos absorben oxígeno, y por lo tanto, reducen la cantidad de oxígeno disuelto disponible para los peces y otras formas de vida acuática. Según investigadores de la Escuela de Biología de una universidad estatal se requiere un mínimo de 5 ppm de oxígeno disuelto para que sea suficiente para la vida acuática. Durante los meses de diciembre y enero se tomaron 45 muestras de agua, dando como resultado una media aritmética de 4,352 ppm y una desviación estándar de 1,261 ppm. ¿Hay suficiente evidencia en los datos que indique que el contenido de oxígeno disuelto es menor que 5 ppm? Use $\alpha=0,025$.
- 5.14 Un fabricante puede tolerar una pequeña cantidad (0,05 miligramos por litro) de impurezas en una materia prima que necesita para fabricar su producto. Debido a que los análisis de laboratorio para detectar las impurezas están sujetos a errores experimentales, el fabricante analiza cada lote 10 veces. Suponga que el valor medio del error experimental es cero y que, por lo tanto, el valor medio de los 10 análisis constituye una estimación insesgada de la verdadera cantidad de impurezas en el lote. Para un lote de la materia prima, la media de los 10 análisis es 0,058 mg/l, con una desviación estándar de 0,012 mg/l. ¿Proporcionan los datos suficiente evidencia que indique que la cantidad de impurezas en el lote excede los 0,05 mg/l? Use $\alpha=0,10$.
- 5.15 Un método de impregnar nubes tuvo éxito en 57 de 150 intentos, mientras que otro tuvo éxito en 33 de 100 intentos. Con un nivel significación de 0,05, ¿se puede concluir que el primer método es mejor que el segundo?
- 5.16 La variabilidad de la cantidad de impurezas presentes en un compuesto químico usado para un proceso particular, depende del tiempo en que el proceso esté en operación. Un fabricante que usa las líneas de producción 1 y 2, ha introducido un ligero ajuste al proceso 2, con la esperanza de

reducir tanto la variabilidad como el promedio de la cantidad de impurezas en el compuesto químico. Muestras aleatorias de 35 observaciones de los dos procesos tuvieron las medias y varianzas siguientes:

$$\begin{aligned}\bar{y}_1 &= 3.2 & s_1^2 &= 1.14 \\ \bar{y}_2 &= 3.0 & s_2^2 &= 1.51\end{aligned}$$

¿Presentan los datos suficiente evidencia que indique que el promedio del proceso 2 es menor que la del proceso 1? Use $\alpha=0,01$.

- 5.17 Cierta estudio sugiere que la aspirina puede proteger a los pacientes de cirugía de la formación de coágulos postoperatorios en las venas. De 43 individuos que recibieron 4 tabletas de aspirina por día, sólo 5 tuvieron coágulos de sangre en comparación con 8 de 50 individuos que tomaron placebos. ¿Proporcionan estos datos suficiente evidencia para concluir que hay relación entre el uso de la aspirina y la frecuencia de la formación de coágulos postoperatorios? Use $\alpha=0,01$.
- 5.18 Los límites de emisión de hidrocarburos para automóviles el año pasado eran de 1,5 gramos por milla. Los análisis de la emisión de hidrocarburos de una muestra aleatoria de 12 autos, produjeron una media aritmética de 1,394 y una desviación estándar de 0,175 gramos por milla. ¿Proporcionan los datos suficiente evidencia para concluir que la emisión media de hidrocarburos para este modelo de automóvil es menor que el límite de 1,5 gramos por milla? Use $\alpha=0,05$.
- 5.19 Cierta producto lleva escrito en su envase: “contenido 16 onzas netas”. Un agente del Departamento de Protección del Consumidor del Ministerio de Economía, selecciona 70 unidades del producto y pesa su contenido, encontrando que el promedio aritmético es 14,5 onzas y una desviación estándar de 1,04 onzas. ¿Presentan los datos suficiente evidencia que indique que el peso neto es menor que lo indicado en el envase? Use $\alpha=0,02$.

- 5.20 Se realiza un experimento para comparar la resistencia (en onzas) de dos tipos de papel. Se seleccionó una muestra aleatoria de tiras papel de cada tipo. El tipo 1, papel estándar, una muestra de tamaño 60 presentó una resistencia media aritmética de 1,364 y una desviación estándar de 0,112. El tipo 2, tratado con una sustancia química, el promedio aritmético y desviación estándar de una muestra de tamaño 50 resultaron, 1,442 y 0,123, respectivamente. Pruebe la hipótesis de que no hay diferencia entre la resistencia de los dos tipos de papel. Use $\alpha=0,10$.
- 5.21 El fabricante de un producto para remover manchas afirma que su producto remueve cuando menos el 90% de todas las manchas. ¿Qué puede concluir acerca de esta afirmación, si el producto sólo eliminó 34 de 50 manchas elegidas al azar de la ropa llevada a una lavandería? Use $\alpha=0,025$.
- 5.22 Los datos que se muestran a continuación son los grados de dureza Brinell obtenidos para muestras de dos aleaciones de magnesio:
Aleación 1: 66,3 63,5 64,9 61,8 64,3 64,7 65,1 64,5
Aleación 2: 71,3 60,4 62,6 63,9 68,8 70,1
¿Existe una diferencia significativa en los grados dureza de las dos aleaciones? Use $\alpha=0,10$.
- 5.23 La dureza del agua en la salida de cierta planta de energía eléctrica tiene una distribución normal con $\mu=55$ ppm. Doce muestras de agua dieron como resultado una media aritmética de 53 y una variancia de 97,4 ppm. ¿Se puede concluir que la dureza media del agua es inferior a 55 ppm? Use $\alpha=0,10$.
- 5.24 El gerente de producción de una empresa de productos derivados de la leche afirma que cuando mucho el 2% de las bolsas de leche se descomponen en los lugares de venta. Pruebe esta aseveración con

$\alpha=0,01$, si una revisión hecha en 10 establecimientos, 2 de 50 bolsas estaban en mal estado.

5.25 Según las especificaciones de producción el peso de las bolsas de jabón en polvo llenadas por una máquina se distribuye normalmente con promedio aritmético de 2 lbs. El encargado de control de calidad seleccionó aleatoriamente 25 bolsas, encontrando que tenían un peso promedio de 1,95 y una desviación estándar de 0,145 lbs. ¿Qué se puede concluir acerca del peso promedio de las cajas? Use $\alpha=0,10$.

5.26 Se instala un nuevo dispositivo de filtrado en una unidad química. Antes y después de su instalación, una muestra aleatoria proporcionó la siguiente resultados acerca del porcentaje de impurezas:

Tiempo	Tamaño de muestra	Promedio aritmético	Variancia
Antes	40	12,5	101,17
Después	35	10,2	94,73

¿Ha reducido significativamente el porcentaje de impurezas el dispositivo de filtrado? Use $\alpha=0,05$.

5.27 El límite de concentración del isótopo radiactivo estroncio 90 es 5 por litro. Para verificar si las especificaciones se cumplían en cierta región del país, fueron seleccionadas aleatoriamente 60 lecherías, encontrándose una media aritmética de 6,6 y una desviación estándar de 0,7958 por litro. ¿Existe alguna diferencia significativa en la concentración promedio con el límite tolerable? Use $\alpha=0,10$.

5.28 Una compañía dedicada a pintar casas encontró que de 83 casas pintadas con pintura de la marca A, 20 mostraron indicios de resquebraaduras y de 58 casas pintadas con pintura de la marca B, 15 presentaron el mismo

problema transcurrir un año. ¿Determine con un nivel de significación de 0,03 si la marca A tiene más durabilidad?

- 5.29 En un laboratorio se ha estado utilizando el método de absorción atómica para determinar el contenido de calcio en sangre. En una publicación reciente, es presentado un método colorimétrico que promete ser rápido y barato. Se toman 8 muestras de sangre, siendo analizadas por ambos métodos. A continuación se presentan los resultados:

Absorción	10,9	10,1	10,6	11,2	9,7	10,0	10,6	10,7
Colorimétrico	9,2	10,5	9,7	11,5	11,6	9,3	10,1	11,2

¿Existe una diferencia promedio en la determinación del contenido de calcio entre los métodos? Use $\alpha=0,01$.

- 5.30 Un ingeniero químico sospecha que el porcentaje de sulfato de sodio de una reacción es superior al 93,4%. Para poner a prueba su hipótesis toma una muestra aleatoria 35 reacciones, determinando que el porcentaje promedio del Na_2SO_4 es 95,5%. ¿Qué puede concluirse acerca del porcentaje de sulfato de sodio de la reacción? Use $\alpha=0,04$.

- 5.31 Se encontró que la molécula de piridina, $\text{C}_5\text{H}_5\text{N}$, se absorbe en la superficie de ciertos óxidos metálicos. En una muestra aleatoria 12 porciones de 5 g de ZnO , la absorción promedio fue 0,068 g, con una desviación estándar 0,0272 g. Pruebe la hipótesis nula de que $\mu=0,067$ g contra la hipótesis alternativa de que $\mu<0,067$ g con un nivel de significancia de 0,05.

- 5.32 Se pide a dos estudiantes de la carrera de Laboratorista Químico de la Universidad de Costa Rica, que determinen el porcentaje de calcio de cierto compuesto. Cada estudiante realiza 6 determinaciones. Utilizando los datos que se presentan a continuación:

Estudiante 1:	84,99	84,04	84,38	84,24	85,13	84,75
Estudiante 2:	84,72	84,48	83,78	85,15	84,33	87,92

¿Difieren significativamente los resultados encontrados por los dos estudiantes? Use $\alpha=0,01$.

5.33 Una empresa de embutidos ubicada en Alajuela elabora tortas de pollo. La compañía asegura que, a lo sumo, el 5% de las tortas presentan algún defecto de peso, empaque o de forma. Una muestra aleatoria de 80 tortas seleccionadas en los establecimientos de venta, indicó que 10 tenían alguno de los problemas anteriormente citados. ¿Pruebe la hipótesis de la compañía? Use $\alpha=0,02$

5.34 Los siguientes datos son los niveles del residuo de plaguicida (ppb) en la sangre de dos muestras aleatorias de personas de cierta comunidad:

Muestra A	27	35	5	11	9	33	8	4	14	16
Muestra B	12	32	30	20	44	18	25	17	13	21

Con un 0,05 de probabilidad máxima de rechazar la hipótesis nula siendo verdadera, pruebe la hipótesis de que no hay diferencia entre los dos promedios.

5.35 Al comparar la resistencia a la tensión de dos tipos de acero para estructuras, un experimento dio los siguientes resultados.

Tipo	Tamaño de muestra	Variación	Promedio
Acero A:	43	19,3	14,64
Acero B:	36	13,5	11,48

Suponiendo que las mediciones constituyen muestras tomadas al azar de dos poblaciones normales, ¿existe una diferencia significativa en el promedio de resistencia a la tensión de los dos tipos de acero? Use $\alpha=0,10$.

5.36 Se ensayan dos métodos para la determinación del calor de fusión del hielo en calorías por gramo. Se desea conocer si existe diferencia entre los métodos a un nivel de significación del 5%.

Método A	79,9	80,0	79,9	80,2	80,5	80,4	80,3	80,1
Método B	80,0	79,9	80,3	79,9	79,9	79,9		

- 5.37 Para comparar las cualidades de dos procesos de fabricación se escoge 100 unidades de la producción obtenida por cada proceso, resultando:

PROCESO	DEFECTUOSOS
A	6
B	11

¿Qué conclusión puede extraerse de estos datos? Use $\alpha=0,01$.

- 5.38 Los propietarios de una empresa agrícola de exportación, están estudiando el efecto de la temperatura de almacenamiento en la conservación de tomates. Interesa cuantificar el número de tomates que se pudren después de una semana de almacenamiento. Deciden usar dos niveles de temperatura a 50°C y 60°C, divididos en 7 lotes de 10 unidades. Los resultados que se presentan a continuación son el número de tomates podridos por lote.

50°C	8	7	4	6	5	6	7
60°C	7	3	3	6	5	4	3

¿Son significativos los efectos de la temperatura? Use $\alpha=0,05$.

- 5.39 Un fabricante sospecha que el contenido de nitrógeno en un producto varía de un lote a otro. Selecciona una muestra aleatoria de cuatro lotes y realiza cinco determinaciones del contenido de nitrógeno en cada lote. ¿Existe una diferencia significativa en el contenido de nitrógeno de un lote a otro? Use $\alpha=0,05$.

Lote	Observaciones				
1	26,15	26,25	26,39	26,18	26,20
2	24,95	25,01	24,89	24,85	25,13
3	25,00	25,36	25,20	25,09	25,12
4	26,81	26,75	26,15	26,50	26,70

5.40 Una compañía química produce cierta droga cuyo peso tiene una desviación estándar de 4 mg. Se ha propuesto un nuevo método de producción aun cuando está involucrado un cierto costo adicional. La gerencia autorizará un cambio en la técnica de producción solamente si el promedio aritmético del peso del nuevo proceso es menor de 16,625 mg.

Supóngase que los investigadores seleccionan $n=10$ y obtienen los datos que a continuación se presentan. Asumiendo que el peso se distribuye normalmente y que $\alpha=0,05$, ¿Cuál debería ser su decisión?

16,628 16,623 16,615 16,634 16,622
 16,622 16,617 16,630 16,624 16,636

5.41 Los siguientes son los datos de un estudio de la seguridad y exactitud de las determinaciones microanalíticas de carbono e hidrógeno, hechas por dos analistas distintos:

Pedro	59,09	59,17	59,27	59,13	59,10	59,14
Ana	59,51	59,75	59,61	59,60		

¿Existe una diferencia significativa promedio al 95% de confianza entre las determinaciones de los dos analistas?

5.42 El límite superior especificado de impurezas de una sustancia es 0,01 por 100. El analista repite el análisis cinco veces y encuentra los siguientes resultados:

0,009 0,012 0,0113 0,0821 0,0915

Teniendo en cuenta el error experimental, ¿es significativamente superior el promedio de impurezas al especificado del 0,01%? Use $\alpha=0,01$.

5.43 Los resultados de varias determinaciones de cobre en un mineral son los siguientes (en % Cu). El primer grupo de determinaciones fue realizado por el analista A, el segundo por el analista B. Comparar las precisiones de los dos analistas. Use $\alpha=0,10$.

A	6,2	5,7	6,5	6,0	6,3	5,8	5,7	6,0	6,0	5,8
B	5,6	5,9	5,6	5,8	6,0	5,5	5,7	5,5		

5.44 Considere dos yacimientos de mineral de cobre. Los yacimientos se dividieron en secciones, y se analizaron muestras procedentes de secciones correspondientes, determinando su contenido en cobre como muestras en paralelo. Los datos obtenidos fueron los siguientes:

Yacimiento 1	25,5	23,1	28,2	24,4	31,5	25,7	27,9	28,7
Yacimiento 2	24,8	21,6	21,6	17,5	25,2	19,2	15,1	27,8

¿Existe una diferencia significativa entre los dos yacimientos en lo que refiere a su contenido en metal?. Use $\alpha=0,05$.

5.45 Muestras de tamaño 60 y 70 produjeron proporciones de 0,25 y 0,28, respectivamente. ¿Pruebe la igualdad de las proporciones de la población al nivel de significación del 0,05?

5.46 Los resultados que se dan a continuación son varios análisis de proteínas en una misma variedad de trigo cultivado en dos lugares diferentes. ¿Hay una diferencia significativa entre las dos medias? Use $\alpha=0,01$.

Lugar A	12,5	11,9	12,7	13,3		
Lugar B	13,2	12,9	13,5	12,7	12,3	13,3

5.47 Un estudio tenía el objetivo de comparar las proporciones de dos poblaciones. Muestras de tamaño 160 y 110 dieron como resultado proporciones de 0,425 y 0,345. ¿Pruebe la igualdad entre las verdaderas proporciones de la población? Use $\alpha=0,10$.

5.48 Los siguientes resultados de dos determinaciones de C e H en ácido benzoico e hidrocloreuro de efedrina, fueron obtenidos por diversos analistas.

Acido benzoico		Hidrocloruro de efedrina	
% C	% H	% C	% H
69,03	5,00	59,22	8,19
68,96	4,91	59,66	8,31
69,18	5,22	59,27	7,86
69,22	5,19	59,22	7,87
68,88	4,89	59,51	8,24
69,04	4,76	59,60	8,00

¿Cómo podría compararse las determinaciones de carbono y las de hidrógeno entre sí para ver si unas son significativamente más precisas que las otras?

Capítulo 6

Regresión y Correlación Lineal Simple

6.1 Regresión

6.2 Regresión lineal

6.3 Ajuste del modelo de regresión lineal

6.4 Uso de la ecuación de regresión

6.5 Error estándar de estimación

6.6 Prueba de hipótesis

6.7 Bondad de un ajuste

6.8 Intervalo de predicción

6.9 Análisis de regresión: comentarios finales

6.10 Análisis de correlación lineal

6.11 Análisis colorimétrico

6.12 Problemas

Regresión y Correlación Lineal Simple

En trabajos científicos es necesario, a menudo, determinar el efecto que una variable ejerce sobre otra. Así, por ejemplo, se desea comprobar si una reacción colorimétrica sigue la ley de Beer-Lambert, medir la velocidad de una reacción química o conocer la validez de nuevo método en relación con una serie de normas conocidas. Un experimento de laboratorio es una medida del efecto de una variable sobre la otra; con base a una cantidad de muestra, se efectúa la reacción y posteriormente, se cuantifica la respuesta.

En el presente capítulo se ocupará exclusivamente de la regresión y correlación lineal simple que utiliza dos variables. Al discutir sobre estas variables se examinarán varias medidas como el coeficiente de regresión y determinación, pruebas de hipótesis e intervalos de confianza, con el propósito de sacar conclusiones acerca de ellas.

6.1 Regresión

La regresión es una técnica estadística para estudiar la naturaleza de la relación entre dos o más variables. Aunque puede utilizarse en esto las ecuaciones no lineales, el capítulo se limitará a la explicación de las ecuaciones de regresión del tipo lineal (línea recta).

Un caso práctico de análisis es construir las curvas de calibración utilizadas en los métodos fotométricos: hay que proceder a la medición de la respuesta de un aparato en relación con cantidades variables del constituyente a estudiar. En análisis gravimétrico, puede presentarse la necesidad de relacionar el peso de precipitado con la temperatura, el pH, el contenido en electrolito inerte o con otras variables.

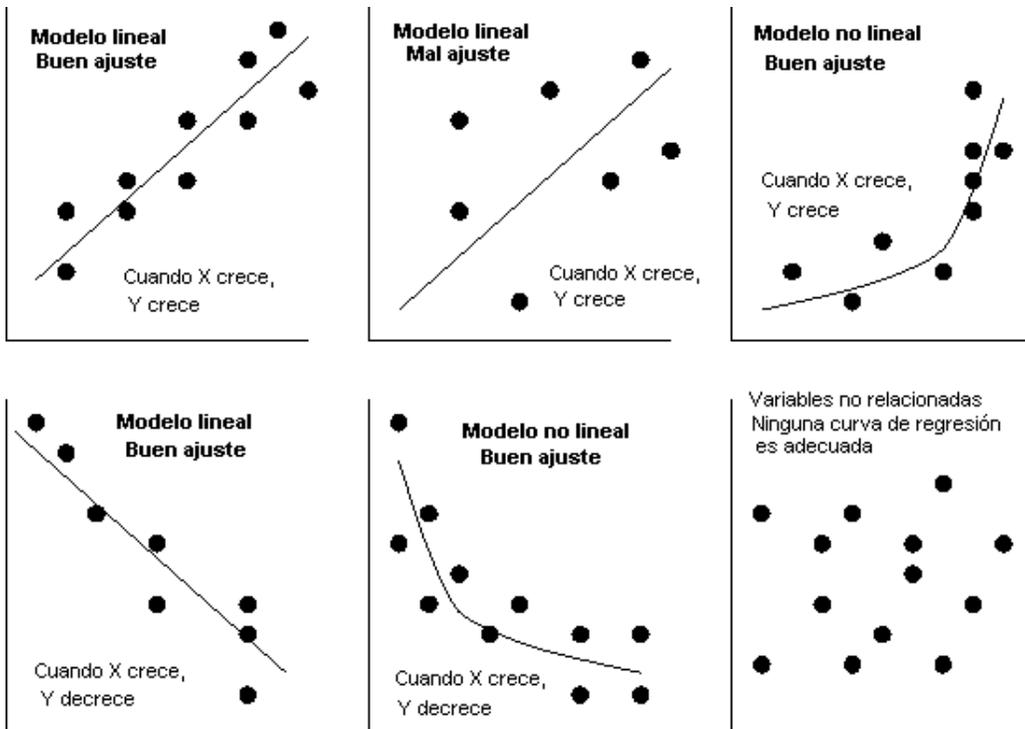
Los objetivos de la regresión son mostrar *la forma como* la variable independiente (X) se relaciona con la variable dependiente (Y), hacer pronósticos sobre los valores de la variable dependiente, con base en el conocimiento de los valores de la variable independiente. Matemáticamente se expresa como $\hat{Y} = f(X)$.

Esto es conocido como la *relación funcional*. El criterio para construir \hat{Y} , tal que la diferencia entre Y e \hat{Y} sea pequeña, esto es que $e_i = Y_i - \hat{Y}_i$ tienda a hacer mínima.

El término que se denomina *error* debe ser tan pequeño como sea posible. El objetivo será buscar la función (también llamada *modelo de regresión*) $\hat{Y} = f(X)$ que lo minimice. Véase las figuras de la siguiente página (o *diagramas de dispersión*).

Como se observa en las figuras, el diagrama de dispersión es la representación gráfica de dos variables, que tiene el objetivo de examinar la dispersión de los datos observados, para determinar visualmente el patrón de comportamiento de los puntos (pares de datos x, y). Con ayuda de esta técnica, podemos establecer la clase de relación que existe entre las variables, ésta puede ser una línea recta, una parábola, una exponencial, etc.

Diferentes nubes de puntos y modelos de regresión para ellas.



6.2 Regresión lineal

La forma de la función de regresión en principio podría ser arbitraria, y tal vez se tenga que la relación más exacta entre las variables X e Y sea la siguiente,

$$\hat{Y} = f(x) = 3,2746e^{x+\text{sen } x}$$

Por el momento no se pretende encontrar relaciones tan complicadas entre variables, pues la discusión se limitará al caso de la *regresión lineal*. Con este tipo de regresiones será suficiente hallar relaciones funcionales de tipo lineal, es decir, se busca cantidades a y b tales que se pueda escribir,

$$\hat{Y} = a + bx \tag{6.1}$$

con el menor error posible entre \hat{Y} e Y , de forma que $Y_i - \hat{Y}_i$ sea una variable que toma valores próximos a cero. El problema que se plantea es entonces el de cómo calcular las cantidades a y b a partir de un conjunto de n observaciones,

$$\begin{cases} (x_1, y_1) \\ (x_2, y_2) \\ \dots \\ (x_n, y_n) \end{cases}$$

de forma que se minimice el error. Las etapas en que se divide el proceso que vamos a desarrollar son, de forma esquemática, las que siguen:

1. Dadas dos variables X, Y , sobre las que definimos $\hat{Y} = a + bx$, se cuantifica el error que se comete al aproximar Y mediante \hat{Y} calculando la suma de las diferencias entre los valores reales y los aproximados al cuadrado (para que sean positivas y no se compensen los errores):

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 \tag{6.2}$$

2. Una aproximación $\hat{Y} = a + bx$ de Y , se define a partir de dos cantidades a y b . Se calcula aquellas que minimizan la función,

$$\text{Error}(a, b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2 \tag{6.3}$$

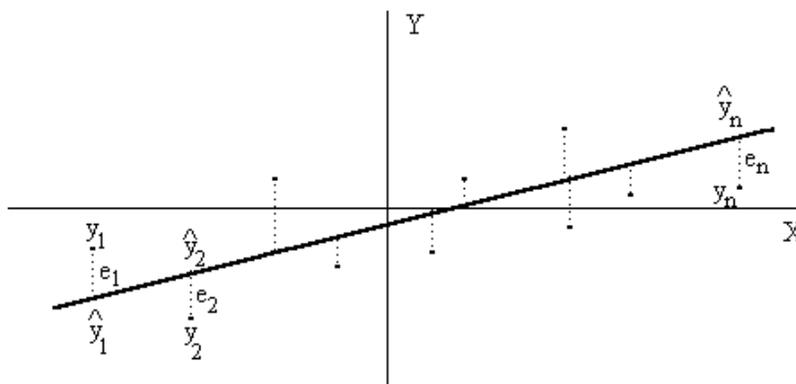
3. Posteriormente se encuentran las fórmulas para el cálculo directo de **a** y **b** que sirvan para cualquier problema.

6.3 Ajuste del modelo de regresión lineal

Para calcular la recta de regresión de Y sobre X se utiliza el método de *mínimos cuadrados*, el que garantiza que la ecuación escogida minimizará la suma de cuadrados de las desviaciones entre cada valor de Y y su valor correspondiente predicho por la ecuación (\hat{Y})¹.

Los errores a minimizar son las cantidades

$$e_i^2 = (y_i - \hat{y}_i)^2$$



Una vez definido el error de aproximación mediante la relación mostrada en la figura, las cantidades que lo minimizan se calculan derivando con respecto a ambas e igualando a cero (procedimiento de los *mínimos cuadrados*). Obteniéndose dos *ecuaciones normales*. La primera se escribe como,

$$\begin{aligned} \sum_{i=1}^n y_i - an - b \sum_{i=1}^n x_i &= 0 \\ \Rightarrow a &= \bar{y} - b\bar{x} \end{aligned}$$

(6.4)

Mediante la segunda ecuación normal,

$$\sum_{i=1}^n y_i x_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = 0$$

¹ Estas desviaciones son llamados también errores o residuos de estimación; y se parte del supuesto que los errores se hallan distribuidos normalmente y tienen una variancia constante.

(6.5)

y sustituyendo, se obtiene que

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad (6.6)$$

La constante a de regresión indica el valor pronosticado de Y cuando X es cero. El coeficiente de regresión b representa la *cantidad de cambio* (bien, puede ser aumento o disminución) *que se pronostica en Y para un aumento de una unidad en X* . Un signo negativo en el coeficiente b indica que medida que aumenta X , disminuye Y . Gráficamente, b es la pendiente de la línea de regresión.

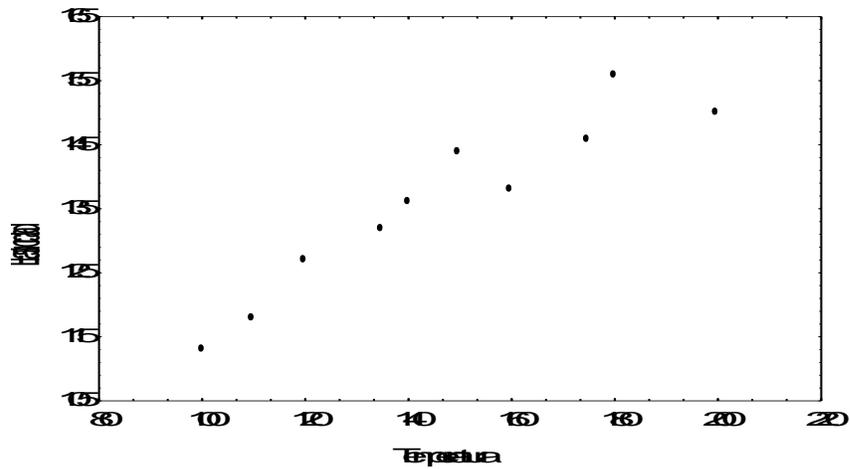
Ejemplo. Un investigador tiene interés de estudiar la elasticidad de cierto plástico (en grados) como una función de la temperatura ($^{\circ}\text{F}$) a la que se produce. Se preparan diez piezas de plástico utilizando distintas temperaturas y los valores observados de la elasticidad fueron:

Temperatura	100	110	120	135	140	150	160	175	180	200
Elasticidad	113	118	127	132	136	144	138	146	156	150

Es evidente, al observar el siguiente gráfico donde se han representado los valores, que es razonable suponer que la relación entre temperatura y elasticidad del plástico es lineal; esto es, una línea recta da una buena aproximación sobre el rango¹ de los datos disponibles.

¹ Temperaturas entre el valor mínimo (100) y la máxima (200).

Diagrama de dispersión entre la temperatura
y la elasticidad de cierto plástico



Las cantidades que necesitamos para sustituir en las fórmulas 6.4 y 6.6 son:

$$\begin{aligned} \sum_{i=1}^{10} x_i y_i &= 203.690 & \sum_{i=1}^{10} x_i &= 1.470 & \sum_{i=1}^{10} y_i &= 1.360 \\ \sum_{i=1}^{10} x_i^2 &= 225.450 & \sum_{i=1}^{10} y_i^2 &= 186.674 & n &= 10 \end{aligned}$$

Al utilizar estos resultados, se calcula **b** como,

$$\begin{aligned} b &= \frac{10(203.690) - 1.470(1.360)}{10(225.450) - (1.470)^2} \\ &= \frac{2.036.900 - 1.999.200}{2.254.500 - 2.160.900} \\ &= \frac{37.700}{93.600} \\ &= 0,40278 \end{aligned}$$

y la constante de regresión, **a**, sería:

$$\begin{aligned} a &= \frac{1.360}{10} - 0,40278 \frac{1.470}{10} \\ &= 136 - 0,40278(147) \\ &= 76,79167 \end{aligned}$$

Por tanto, la ecuación de la línea recta que mejor ajusta los datos en el sentido de mínimos cuadrados es,

$$\hat{y} = 76,79167 + 0,40278X$$

El coeficiente de regresión resultó 0,40278. Este valor significa que, para aumento de un grado en la temperatura, hay un cambio en la elasticidad del plástico de 0,40 grados, es decir un aumento. Se puede considerar que esta pendiente representa la parte variable de la elasticidad, la cual varía con la temperatura de fabricación del plástico.

La intersección con el eje Y se calculó que era 76,79167. Esta se puede considerar que representa la parte fija de la elasticidad del plástico, o sea, la elasticidad que no varía en función de la temperatura.

6.4 Uso de la ecuación de regresión

La ecuación de regresión que se ha ajustado a los datos, se puede utilizar ahora para predecir el valor de la elasticidad para un valor dado de la temperatura. Por ejemplo, para $x=115$ °F predecimos que la elasticidad será 123,1 grados, que se calcula de la siguiente forma,

$$\hat{y} = 76,79167 + 0,40278(115) = 123,1$$

Cuando se utiliza el análisis de regresión para predicciones, es importante considerar sólo el rango de variación de la variable independiente al hacer las predicciones. Este rango abarca todos los valores de X, desde el mínimo hasta el máximo utilizados para encontrar la ecuación de regresión.

Predecir Y para un valor dado de X, se llama *interpolar* dentro de este rango de los valores de X, pero no se puede *extrapolar más allá del rango de los valores de X*. Cualquier predicción de Y fuera del rango de X, supone que la relación ajustada es válida fuera de este recorrido de 113 y 200 grados. Este supuesto necesariamente no se cumple, existen muchos problemas de investigación donde después de ciertos valores de X, la forma de la relación es curvilínea.

6.5 Error estándar de estimación

El error estándar de la estimación se parece al utilizado para determinar un intervalo de confianza para una media o proporción de la población, salvo que en este caso el intervalo está en la dirección vertical. Por ejemplo, si sustituyéramos $X=10$ en la ecuación de regresión $\hat{y} = 76.79167 + 0.40278X$, se calcula que el valor de Y es de 80,82. No obstante, se trata de una mera estimación puntual y necesariamente hay cierta incertidumbre respecto al nivel probable de Y cuando X sea igual a 10. Es imprescindible conocer el error estándar de estimación para calcular un intervalo más “apropiado” en sentido estadístico.

El error de la estimación puede calcularse por,

$$S_e = \sqrt{\frac{\sum_{i=1}^n y_i^2 - a \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i y_i}{n-2}} \quad (6.7)$$

donde: a = la intersección de la recta

b = el coeficiente de regresión

n = tamaño de la muestra

Con el uso de la ecuación (6.7) y el problema de la elasticidad del plástico, el error estándar de la estimación se calcula como,

$$\begin{aligned} S_e &= \sqrt{\frac{186.674 - 76,79167(1.360) - 0,40278(203,690)}{8}} \\ &= \sqrt{\frac{186.674 - 104.436,6712 - 82.042,2582}{8}} \\ &= \sqrt{\frac{195,0706}{8}} \\ &= \sqrt{24,383825} \\ &= 4.9378 \end{aligned}$$

Este error estándar de la estimación, igual a 4,94378 (grados), representa una medida de la variación en torno a la recta ajustada de regresión. Se mide en la unidad de medición de la variable Y dependiente. La interpretación del error estándar de estimación, es semejante al de la desviación estándar. Así como la desviación

estándar mide la variabilidad alrededor del promedio, el error estándar de estimación mide la variabilidad en torno a la recta ajustada de regresión.

6.6 Prueba de hipótesis sobre el coeficiente de regresión

Una vez estimada la ecuación de regresión es conveniente determinar el aporte de la variable independiente al modelo ajustado. Hay que tomar la decisión de eliminarla e incluir otras que podrían contribuir con más información sobre la variabilidad de **Y**.

La hipótesis nula plantea que la pendiente de la recta es cero contra la hipótesis alternativa que es diferente. Simbólicamente se denota como,

$$H_0: \beta = 0 \quad (\text{no hay relación lineal})$$

$$H_1: \beta \neq 0 \quad (\text{si hay relación lineal})$$

La hipótesis nula implica que no existe ninguna relación lineal entre las variables. La hipótesis alternativa es que existe una relación lineal, positiva o negativa. Podríamos decir, que cuando no se rechaza, entonces indica que la variable independiente no aporta información para predecir la variabilidad de **Y**. Para realizar la prueba se utiliza una estadística *t* con **n-2** grados de libertad, donde

$$t_{obs} = \frac{b}{S_e / \sqrt{SC_x}} \quad (6.8)$$

Donde SC_x son la suma de cuadrados de la variable independiente *X* y se tiene que,

$$SC_x = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} \quad (6.9)$$

Para los datos de temperatura y elasticidad del plástico, según las ecuaciones 6.8 y 6.9, se obtiene

$$\begin{aligned} sc_x &= 225450 - \frac{(1470)^2}{10} \\ &= 225450 - 216090 \\ &= 9360 \end{aligned}$$

$$\begin{aligned}
 t_{obs} &= \frac{0.40278}{4.94378 / \sqrt{9360}} \\
 &= \frac{0.40278}{0.0511} \\
 &= 7.882
 \end{aligned}$$

Si se deseó un nivel de confianza del 99%, se tendría un valor teórico con 8 grados de libertad de 3,3554. Ahora, como el valor observado es mayor al teórico, ($3,3554 < 7,882$), se rechaza H_0 . Entonces, la conclusión a que se llegó es rechazar la hipótesis nula y decir que hay alguna relación entre la temperatura de fabricación y la elasticidad del plástico.

6.7 Bondad de un ajuste

Si el ajuste de Y mediante la ecuación de regresión $\hat{Y} = f(x)$ es bueno, cabe esperar que la varianza del error de estimación debe ser pequeña (en comparación con la de Y). Ello se puede expresar definiendo una nueva variable E que mida las diferencias entre los auténticos valores de Y y los teóricos suministrados por la regresión,

$$E = Y - \hat{Y} \rightarrow \begin{cases} e_1 = y_1 - \hat{y}_1 \\ e_2 = y_2 - \hat{y}_2 \\ \dots \\ e_n = y_n - \hat{y}_n \end{cases}$$

Dicho de otro modo, E debe ser una variable cuya media debe ser 0, y la varianza S_E^2 pequeña en comparación con S_Y^2 . Por ello, se define el coeficiente de determinación de la regresión de Y sobre X , como

$$\begin{aligned}
 r^2 &= 1 - \frac{S_E^2}{S_Y^2} \\
 &= \frac{\text{Variancia de } Y \text{ explicada por } X}{\text{Variancia total de } Y}
 \end{aligned} \tag{6.10}$$

donde,

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \cdot \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}} \quad (6.11)$$

Si el ajuste es bueno se debe tener que r^2 tome un valor próximo a 1. El coeficiente de determinación sirve entonces para medir de qué modo las diferencias entre los verdaderos valores de una variable y los de su aproximación mediante una ecuación de regresión son pequeños en relación con los de la variabilidad de la variable que intentamos aproximar. Por esta razón, estas cantidades miden el *grado de bondad del ajuste*.

En el problema tratado en las secciones anteriores y utilizando la ecuación 6.11, se calcula:

$$\begin{aligned} r &= \frac{10(203.690) - 1.470(1.360)}{\sqrt{[10(225.450 - (1.470)^2)] - [10(186.674 - (1.360)^2)]}} \\ &= \frac{2.036.900 - 1.999.200}{\sqrt{[2.254.500 - 2.160.900][1.866.740 - 1.849.600]}} \\ &= \frac{37.700}{\sqrt{93.600(17.140)}} \\ &= \frac{37.700}{40.053,76} \\ &= 0,94 \end{aligned}$$

entonces $r^2 = 0,8836$. El 88,36% de la variabilidad de la elasticidad del plástico se encuentra explicada por relación lineal con la temperatura de fabricación. Esto implica, que el restante 11,64% está explicado por otras variables que afectan la elasticidad, no consideradas en el modelo de regresión lineal.

6.8 Intervalo de predicción

El intervalo de confianza alrededor de la recta de regresión se refiere a la verdadera respuesta media en $\mathbf{x} = \mathbf{x}_0$ (esto es un parámetro de la población), y no a observaciones futuras. La fórmula de cálculo del intervalo de predicción sería,

$$\hat{y} \pm t_{\alpha/2(n-2)gl} \cdot S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SC_x}}$$

(6.10)

donde: $t_{\alpha/2(n-2)gl}$ = valor de la distribución t para el nivel de confianza deseado y (n-2) grados de libertad

S_e = error estándar de estimación

\hat{y} = valor de Y calculado a partir de X a partir de la línea de regresión

\bar{x} = la media de los valores de X en los datos de la muestra

SC_x = suma de cuadrados de X (véase ecuación 6.9).

Una vez presentada la ecuación 6.10, se puede desarrollar estimaciones del intervalo de confianza para el problema que se ha venido comentando. Por ejemplo, si se desea una estimación del intervalo con 95% de confianza de la elasticidad promedio real para una temperatura de fabricación de 145 grados, se calcula lo siguiente:

$$\hat{y} = 76,79167 + 0,40278(145) = 135,2$$

Además, $t_{\alpha/2(n-2)gl} = 3,3554$

$$\bar{x} = 147$$

$$x_0 = 145$$

$$SC_x = 9.360$$

Por tanto,

$$135,2 \pm 3,3554(4,94378) \sqrt{\frac{1}{10} + \frac{(145 - 147)^2}{9,360}}$$

$$135,2 \pm 16,588\sqrt{0,10}$$

$$135,2 \pm 16,588(0,316)$$

$$135,2 \pm 5,24$$

De modo que:

$$130,0 < \mu_{y/x_0} < 140,4$$

Por consiguiente, se estima que la elasticidad promedio real del plástico fabricado a 145 grados es entre 130,0 y 140,4. Se tiene 95% de confianza de que el intervalo estima correctamente la verdadera elasticidad promedio del plástico.

6.9 Análisis de regresión: comentarios finales

Un problema potencial es tratar de hacer predicciones con relación a los valores de X que se encuentran *fuera de los límites de los datos sobre los que se calcula la línea de regresión*. La linealidad de la relación no siempre se extiende a esta región, haciendo que una predicción realizada a partir de esta extensión muestre un grave error. Es muy posible que un modelo que ajusta bien en la región de los datos originales ya no se ajuste igual fuera de esa región.

Por último, cuando un diagrama de dispersión indica que una relación posiblemente no sea lineal, puede incluir un término X^2 en la ecuación, por ejemplo:

$$\hat{y}_1 = b_0 + b_1x + b_2x^2$$

quizá dividirlo en varios intervalos de X y luego calcular una ecuación de mínimos cuadrados para cada uno de los intervalos. La alternativa consiste en el análisis de regresión no lineal, la cual rebasa el ámbito de este libro. Con todo, una alternativa práctica es emplear varias líneas de regresión para aproximar una relación no lineal.

6.10 Análisis de correlación lineal

En las anteriores secciones fueron estudiados los aspectos básicos de la identificación de la naturaleza de una relación lineal ajustada a un conjunto de datos para dos variables. Así pues, es necesario definir una medida de la relación entre ellas, y que no esté afectada por los cambios de unidad de medida. Una forma posible de conseguir este objetivo es dividir la **covarianza**¹ por el producto de las desviaciones estándares de cada variable, ya que así se obtiene un coeficiente adimensional, r , que se denomina coeficiente de correlación lineal de Pearson (véase la fórmula de cálculo en la ecuación 6.11).

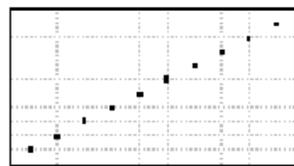
¹ La covarianza es una medida de la variación conjunta (covariación) de las variables X e Y . Esta medida no debe ser utilizada de modo exclusivo para medir la relación entre las dos variables, ya que es sensible al cambio de unidad de medida.

$$r = \frac{S_{xy}}{S_x S_y}$$

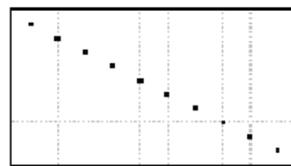
El coeficiente de correlación lineal posee las siguientes propiedades:

- Es invariante para transformaciones lineales (cambio de origen y escala) de las variables.
- Sólo toma valores comprendidos entre -1 y 1 , $-1 \leq r \leq 1$
- Carece de unidades de medida (adimensional).
- Cuando $|r|$ esté próximo a uno, existe una relación lineal muy fuerte entre las variables.
- Cuando $r \approx 0$, puede afirmarse que no existe relación lineal entre ambas variables.

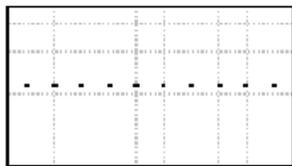
La figura 1 ilustra una correlación lineal positiva, perfecta, entre X e Y. Por tanto, hay una asociación perfecta uno a uno entre X e Y, de modo que Y aumentará en una forma perfectamente predecible según aumenta X.



Correlación positiva perfecta
1



Correlación negativa perfecta
2



No hay correlación
3



No hay correlación
4

La gráfica 2 ilustra una correlación negativa perfecta entre X e Y. En este caso, Y disminuirá en una forma perfectamente predecible según aumenta X.

La gráfica 3 y 4 ilustran el caso donde no hay relación entre X e Y. Según aumenta X no hay cambio en Y o bien, al mantenerse constante X hay un incremento en Y, de modo que no hay asociación entre el valor de X y el valor de Y.

Para los datos de temperatura y elasticidad del plástico, se obtiene una $r=0,94$ (ver el procedimiento de cálculo en la página 170). El coeficiente de correlación indica una asociación positiva alta, es decir, el aumento en la temperatura de fabricación del plástico tiene una fuerte asociación con el aumento de la elasticidad del plástico.

Hay que tener presente que la fuerza de asociación no significa necesariamente **causalidad**. El simple hecho de que los cambios en X parezca explicar gran parte de la variación en Y no significa automáticamente que X sea la causa de Y. En efecto, lo inverso puede suceder o bien ambas variables pueden ser causadas por una tercera que no ha sido identificada.

Algunas veces los altos valores de correlación se registran para variables que, según el sentido común, no parecerían tener relación directa entre sí. Se refieren a correlaciones *espurias*. Un estudio de correlación probablemente descubre que el tiempo del almacenamiento del químico, guarda estrecha correlación con el porcentaje de impurezas de ciertas sustancias. Pero la causalidad en una u otra dirección parece poco probable.

Una segunda consideración es que un alto valor del coeficiente de correlación puede no indicar realmente una relación muy significativa; por ejemplo, bastan dos puntos para determinar una recta. Por consiguiente, con sólo dos puntos de datos, automáticamente se tendría un coeficiente de correlación igual a $+1$ o -1 . Una recta sería entonces un ajuste perfecto de los “datos”.

6.11 Análisis colorimétrico

Los ensayos colorimétricos implican la preparación de una serie de disoluciones estándar y la comparación de la intensidad del color con la producida en la solución de concentración desconocida. Esta concentración es deducida de su absorbancia en comparación con las de las soluciones estándar, medidas en un colorímetro o espectrofotómetro. La relación de absorbancia a concentración es lineal (ley de Beer) y, como tal, sigue las leyes que se han discutido antes para la regresión lineal simple.

En colorimetría, la concentración de la sustancia que se está midiendo es la variable X, y la absorbancia la variable Y. La pendiente es la medida del cambio de absorbancia por unidad de cambio en la concentración, y la ordenada (a) es la medida del ensayo en blanco.

Hay dos técnicas de uso general en colorimetría. Una implica la interpretación de la concentración desconocida en una curva estándar. La otra hace uso de la relación entre la solución problema y la estándar, aproximadamente, a la misma concentración. La hipótesis es que la pendiente de la curva problema es igual a la pendiente de la estándar.

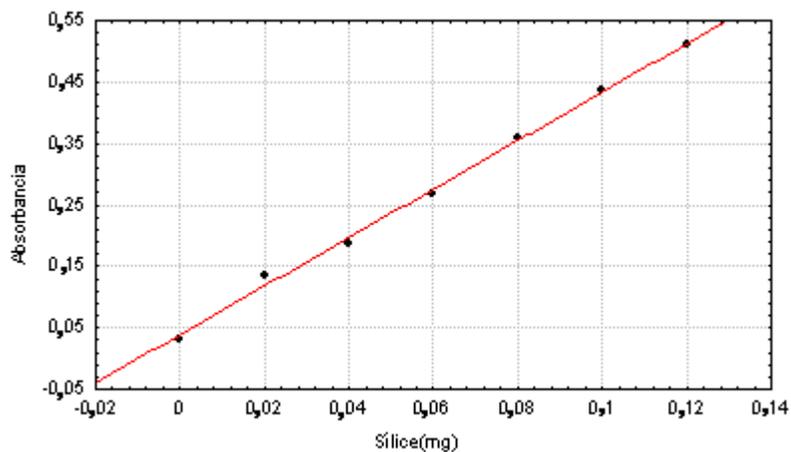
El siguiente ejemplo se desarrolla usando los datos de un análisis citado en métodos calorimétricos de análisis, de Snell y Snell¹.

Sílice (mg)	Absorbancia
X	Y
0	0,032
0,02	0,135
0,04	0,187
0,06	0,268
0,08	0,359
0,10	0,435
0,12	0,511

El diagrama de dispersión presenta una tendencia marcada en la relación entre sílice y la absorbancia, indicando que un modelo de regresión lineal es adecuado para el análisis de la relación entre las dos variables.

Los autores encuentran un contenido de 0,053 mg de silicio en la muestra cuando usan el método gráfico y de 0,0243 mg cuando utilizan el método aritmético. La pendiente de la recta de regresión es 0,971 mg; al aumentar la cantidad sílice en un miligramo, la absorbancia se incrementa en 0,971.

¹ Snell, F.D. y Snell, C.: Colorimetric Methods of Analysis, 3ª. Edición, vol. 1, pág. 151. Van Nostranda-Reinhold, Pricenton, New Jersey (1.951).



El modelo de regresión lineal permite describir el 94,3% de la variabilidad de la absorbancia utilizando como variable independiente el contenido de sílice. La correlación entre las variables de estudio, como puede observarse en el diagrama de dispersión es positiva o directa y bastante alta (0,97).

6.12 PROBLEMAS

- 6.1 Los siguientes datos corresponden al tiempo de secado (en horas) de cierto barniz y la cantidad de un aditivo (en gramos) con el que se intenta reducir el tiempo de secado:

Aditivo	1	1,5	2	2,5	3	3,5
Tiempo	2	4	6	8	9	11

- Construya el diagrama de dispersión.
 - Calcule el coeficiente de correlación.
 - Estime el tiempo de secado del barniz cuando se han utilizado 4 gramos del aditivo.
- 6.2 Los siguientes datos son las mediciones de la velocidad del aire, X , (cm/seg.) y el coeficiente de evaporación (mm^2/seg), Y , de las gotitas de combustible en una turbina de propulsión.

X	20	60	100	140	180	220	260	300	340
Y	0,18	0,37	0,35	0,78	0,56	0,75	1,18	1,36	1,17

- Ajuste la recta de regresión por el método de mínimos cuadrado.
 - Calcule e interprete el coeficiente de determinación.
 - Con un nivel de significación de 0,01, ¿hay una relación lineal entre la velocidad del aire y el coeficiente de evaporación?
- 6.3 Las materias primas empleadas en la producción de una fibra sintética son almacenadas en un lugar en donde no se tiene control sobre la humedad. Las mediciones de la humedad relativa en el lugar de almacenamiento y su contenido en una muestra de las materias primas (ambas en porcentajes) en 12 días dieron los siguientes resultados:

Humedad	42	35	50	43	48	62	31	36	44	39	55	48
Contenido	12	8	14	9	11	16	7	9	12	10	13	11

- Encuentre la ecuación de regresión lineal.
- Interprete la pendiente de la recta en términos del problema.

- c) Encuentre un intervalo con un nivel de confianza del 95% para el contenido de humedad de las materias primas cuando la humedad del lugar de almacenamiento es del 40%.
- d) Encuentre una estimación de intervalo con 95% de confianza para la pendiente real.

6.4 Los datos siguientes se refieren a los días desde la inoculación(**X**) y al crecimiento de una colonia de bacterias (**Y**) en un cultivo.

X	3	6	9	12	15	18
Y	115	147	239	356	579	864

- a) Calcule el coeficiente de correlación
- b) Calcule e interprete el coeficiente de determinación
- c) Interprete la constante de regresión en términos del problema
- d) Calcule el error estándar de estimación

6.5 La mayoría de los estudiantes universitarios de física realizan experimentos para verificar la ley de Hooke. La ley de Hooke establece que cuando se aplica una fuerza a un cuerpo que es largo en comparación con su área transversal, el cambio de su longitud (mm) es proporcional a la fuerza(Kg). En la siguiente tabla se muestran los resultados de un experimento de laboratorio. Se usaron seis piezas de alambre de acero de 0,34 mm de diámetro y 3 m de largo para obtener las mediciones de fuerza y cambio de longitud.

Fuerza	29,4	39,2	49,0	58,8	68,6	78,4
Cambio	4,25	5,25	6,50	7,85	8,75	10,0

- a) Dibuje el diagrama de dispersión.
- b) Ajuste el modelo de regresión lineal.
- c) Encuentre un intervalo de confianza del 95% para la pendiente de la recta.

- d) De acuerdo a la ley de Hooke, la recta debe pasar por el origen, es decir, la constante debe ser igual a 0. Pruebe esta hipótesis, use un nivel de significación del 1%.
- e) Encuentre una predicción del aumento de longitud de un alambre de 2 m de largo cuando se le aplica una fuerza de 55 kg. Use un intervalo de predicción del 95%.

6.6 Se considera que el rendimiento de un proceso químico es una función de la cantidad de catalizador agregada a la reacción. Se realiza un experimento y se obtienen los siguientes datos:

X= catalizador (lb) Y= rendimiento (%)

X	0,9	1,4	1,6	1,7	1,8	2,0	2,1	2,3
Y	60,54	63,86	63,76	60,15	66,66	71,66	70,81	65,72

- a) Calcule e interprete el coeficiente de regresión lineal.
- b) Pruebe la significación de la regresión. Use $\alpha=0,10$.
- c) Calcule e interprete el coeficiente de correlación.

6.7 Los siguientes datos pertenecen al residuo de cloro en partes por millón, que hay en una piscina en diferentes momentos después de ser tratada con sustancias químicas:

Número de horas	2	4	6	8	10	12	14	16
Residuo (ppm) Cl ₂	1,8	1,5	1,4	1,1	1,0	0,9	0,7	0,4

- a) Determine la bondad de ajuste del modelo de regresión lineal.
- b) Encuentre una estimación de intervalo del 99% de confianza del residuo de cloro promedio para cinco horas después del tratamiento químico.
- c) Pruebe la significación de la regresión. Use $\alpha=0,05$.

6.8 El objetivo de un experimento consistió en determinar la relación entre los porcentajes de manganeso en la resistencia a la ruptura (pies/libra), obteniéndose los siguientes resultados:

Manganeso	0,1	0,2	1,0	1,2	1,4	3,0	3,2	3,4
Resistencia	26,7	27,5	30,6	34,1	47,8	33,2	20,5	15,4

- Construya el diagrama de dispersión para verificar la utilidad del modelo de regresión lineal.
- Calcule e interprete el coeficiente de determinación.
- Calcule un intervalo de confianza del 90% para el promedio de la resistencia cuando el porcentaje de manganeso es 1,35.

6.9 Reversiones inducidas a la independencia por 10^7 células sobrevivientes (ergs/bacterias), X , 10^{-5} de Escherichia coli estreptomycin, Y , dependiente sometida a radiación ultravioleta monocromática de $2,967 = A$ longitud de onda.

X	Y	X	Y
13,6	52	39,8	130
13,9	78	40,1	139
21,1	72	43,9	173
25,6	89	51,9	208
26,4	80	53,2	225
30,1	112	65,2	259

- Construya el diagrama de dispersión.
- Ajuste la recta de regresión lineal de mínimos cuadrados.
- Calcule el coeficiente de correlación.

6.10 Un investigador tiene interés de estudiar la elasticidad de cierto plástico(Y) como una función de la temperatura(X) a la que se produce. Se preparan diez piezas de plástico utilizando distintas temperaturas y los valores observados de la elasticidad fueron:

X	100	110	120	135	140	150	160	175	180	200
Y	113	118	127	132	136	144	138	146	156	150

- Ajuste un modelo de regresión lineal.
- Pruebe la significación de la temperatura a un nivel de confianza del 95%.

- c) Analice la bondad de ajuste de la regresión lineal.
- d) Calcule e interprete un intervalo de confianza del 99% para el promedio de elasticidad a una temperatura de 130.

6.11 De una muestra hipotética de 10 pacientes en los que se ha recogido los siguientes datos: edad (en años), nivel de colesterol en plasma sanguíneo (en mg/100 ml) y consumo de grasas saturadas (en gr/semana).

Edad	Colesterol	Grasas
80	350	35
30	190	40
42	263	15
50	320	20
45	280	35
35	198	50
18	232	70
32	320	40
49	303	45
35	220	35

- a) Calcule el coeficiente de correlación lineal entre el nivel de colesterol y consumo de grasas saturadas.
- b) Estime el nivel de colesterol para una persona de 60 años. Utilice el modelo de regresión lineal.
- c) Calcule el coeficiente de determinación para el modelo anterior.

6.12 En una compañía de helados se sospecha que el almacenamiento del helado a temperaturas bajas durante largos períodos tiene un efecto en la pérdida de peso del producto. En la planta de almacenamiento de la compañía se obtuvieron los siguientes datos:

Pérdida (onzas):	1,01	1,32	1,30	1,06	1,00	1,30	1,26
Tiempo(días):	10	15	25	30	35	40	50

- a) Dibuje y comente el diagrama de dispersión.

- b) Ajuste un modelo de regresión lineal simple.
- c) Interprete el coeficiente de regresión.
- d) Estime la pérdida de peso a 20 días de almacenamiento.

6.13 Una muestra de agua contaminada se oxidó a 25°C. El porcentaje de materia orgánica en la muestra que se oxidó varió con el tiempo como sigue:

X = tiempo (en días)

Y= materia oxidada (%)

X	1	2	3	4	5	6	7	10	12
Y	21	37	50	60	68	75	80	90	99

- a) Encuentre la ecuación de regresión lineal.
- b) Calcule e interprete el coeficiente de correlación.
- c) Estime el porcentaje de materia oxidada a 8 días.
- d) Calcule el error estándar del coeficiente de regresión.

6.14 Los siguientes datos representan la conductividad térmica del cloruro de metilo a las temperaturas (°C).

X= temperaturas

Y= conductividad térmica

X	100	37,7	65,5	93,3	121,1	148,8	176,6	204,4
Y	0,085	0,103	0,121	0,137	0,155	0,73	0,189	0,207

Estime la conductividad térmica a 211,6 °C.

6.15 En la fabricación de cierto compuesto se sabe que el porcentaje de impurezas está relacionado con la temperatura a que se realiza la mezcla de sus componentes. Se llevó a cabo un experimento y se encontró la

ecuación de regresión: $\hat{y} = 2,2 + 0,5x$

donde x= temperatura en grados centígrados

y= porcentaje de impurezas

- a) Interprete el coeficiente de regresión
- b) ¿Qué significa en términos del problema la constante de regresión?
- c) ¿Cuál sería el porcentaje de impurezas para una temperatura de 50 °C?

- 6.16 Un estudiante de química preparó una serie de compuestos que contienen nitrógeno y oxígeno.

Compuesto	Masa de nitrógeno	Masa de oxígeno
A	16,8	19,2
B	17,1	39,0
C	33,6	57,3
D	22,5	45,1
E	30,4	50,0
F	35,7	51,8
G	22,6	34,9

Calcule e interprete el coeficiente de correlación entre la masa de nitrógeno y la masa de oxígeno.

- 6.17 A continuación se presenta algunas propiedades físicas y químicas de los metales alcalinotérreos.

Elemento	Número atómico	Masa Atómica	Punto de congelación	Densidad
Berilio	4	9,01	1283	1,85
Magnesio	12	24,30	650	1,74
Calcio	20	40,08	851	1,54
Estroncio	38	87,62	757	2,58
Bario	56	137,33	704	3,65

- Calcule e interprete el coeficiente de determinación entre el punto de congelación y la densidad de los metales.
 - Ajuste el modelo de regresión lineal entre el número y su masa atómica.
 - Interprete el coeficiente de regresión de la recta.
 - Pruebe la significancia del modelo. $\alpha=0,05$
- 6.18 Un estudiante obtiene los datos que siguen del volumen de aire como función de la temperatura a presión constante:

Temperatura (°C):	16	55	85	103	126	163
-------------------	----	----	----	-----	-----	-----

Volumen (cm ₃):	31	35	38	40	43	47
-----------------------------	----	----	----	----	----	----

- Dibuje y comente el diagrama de dispersión.
- Ajuste un modelo de regresión lineal simple.
- Interprete el coeficiente de regresión.
- Estime el volumen del aire a 70 °C de temperatura.

6.19 Los siguientes datos representan a la viscosidad y tensión superficial de algunas sustancias (10^{-3}).

Viscosidad	0,65	1,20	0,23	1,55	0,652	0,466	0,356
Tensión	2,89	2,23	1,70	7,29	6,99	6,70	6,40

- Ajuste el modelo de regresión lineal simple.
- Calcule e interprete el coeficiente de correlación.
- ¿Qué significa en términos del problema la constante de regresión?
- ¿Cuál sería la tensión superficial para una viscosidad de 1,0?

6.20 Se determinaron valores para una magnitud experimental Y, mediante el empleo de varios valores controlados de una variable X. Se esperan que exista una relación lineal. Los datos obtenidos fueron los siguientes:

X	1	2	6	10	4	5	3	9	8
Y	0,16	0,21	0,26	0,30	0,25	0,23	0,22	0,29	0,27

- Dibuje el diagrama de dispersión de Y en función de X. Trace “a ojo” la mejor recta a través de los puntos. Lea el intercepto y calcule la pendiente de su línea.
- Use el método de mínimos cuadrados para hallar el ajuste lineal de los datos. Compare los resultados para la pendiente e intercepto con sus estimaciones en a).
- ¿Qué valor pediría para Y, cuando X=5? ¿Qué incertidumbre habría de agregarle?

6.21 La estandarización del nitrito de sodio se realiza por una valoración de una cantidad pesada de novocaína en solución ácida fría. El ácido nitroso

realiza la diazotación² del grupo amino de la novocaína y un exceso del mismo forma un anillo azul cuando se ensaya en papel de yoduro de almidón. Para manifestar el punto final se requiere una cantidad en blanco, de 0,1 – 0,2 ml. Supongamos que un analista pesa 6 muestra de novocaína y valora esta muestra con la solución, defectuosamente estandarizada, de nitrito de sodio, obteniendo los resultados que se muestran a continuación:

Novocaína (mg)	Nitrito sódico (ml)
200	7,50
300	11,3
400	14,9
500	17,6
600	22,25
700	26,1
800	29,65

- Calcule la ecuación lineal que mejor ajusta la relación entre mililitros de nitrito de sodio y miligramos de novocaína, bajo las condiciones del experimento.
- Interprete el coeficiente de regresión.
- Para conocer la precisión de estos resultados, calcule el error estándar de estimación.
- Pruebe si el factor (b) es significativamente diferente de 1,00. Use $\alpha=0,05$
- Pruebe si la ordenada (a) es una medida de la valoración en blanco. Use $\alpha=0,01$

6.22 Suponga que un analista efectúa un análisis calorimétrico de sílice y el ensayo en blanco, le da 0,032 nanómetros de absorbancia, y los estándares de 0,02, 0,04 y 0,06 mg, generan los resultados de absorbancia que a continuación se presentan en el cuadro siguiente:

² La reacción de diazotación fue puesta a punto por Peter Griess en 1860. Esta reacción tiene lugar entre una amina primaria aromática y NaNO_2 en presencia de HCl o H_2SO_4 , para formar una sal de diazonio

Sílice Mg (X)	Absorbancia (Y)
0,00	0,032
0,02	0,135
0,04	0,187
0,06	0,268

Calcule la ecuación de regresión lineal. Interprete la constante y el coeficiente de regresión.

- 6.23 La hidrólisis de cierto compuesto orgánico da lugar a un producto que es relativamente insoluble. Cuando el 1 por 100 de los productos de la hidrólisis está presente en la solución, se forman cristales que impiden su comercialización. Se diseñó un experimento para predecir la durabilidad del producto. La hidrólisis era una reacción monomolecular y, por tanto, se estudió en función del tiempo (meses).

Tiempo (meses) X	Hidrólisis (%) Y
3	0,220
5	0,255
7	0,304
9	0,402

Analice la conveniencia de utilizar el modelo de regresión lineal para describir la hidrólisis en función del tiempo. Posteriormente, ajuste el modelo no lineal, $\log Y = a + bX$, análisis los resultados.

- 6.24 Los pesos de hidrogenoftalato de potasio, $C_6H_5KO_4$, (Peso Molecular 204,23), que se dan a continuación, requieren los volúmenes de solución de NaOH que se señalan para titularse al punto final de la fenolftaleína. Analice la bondad de ajuste de modelo de regresión lineal.

$C_6H_5K_0_4$ gramos	Solución NaOH, ml
1,0000	45,64
1,1263	39,25
0,8567	28,75
0,8169	40,00
0,0679	8,96
1,1212	35,62

6.25 Algunas reacciones irreversibles siguen la ecuación de Nernst razonablemente bien, y como ejemplo se puede citar la reducción del permanganato. El valor del factor en la ecuación de Nernst, varía con la temperatura en la siguiente forma:

Temperatura °C	Valor del factor
10	0,05618
15	0,05717
20	0,05816
25	0,05916
30	0,06014
35	0,06114
40	0,06213

Suponiendo un modelo de regresión lineal, estime el valor del factor a una temperatura de 28°C, utilizando un nivel significancia de 1%.

6.26 La calibración de un método espectroquímico dio los siguientes resultados:

Mg plomo	1	2	3	4	5
Absorbancia	0,33	0,55	0,96	1,21	1,47

Calcular la línea recta correspondiente según el método de los mínimos cuadrados. Una muestra problema presentó una absorbancia de 0,78. ¿Cuál era su contenido en plomo?

6.27 Una muestra de 1,0255 g de un ácido orgánico soluble en agua, se disuelve y se diluye con agua hasta un volumen de exactamente 250 ml. Una alícuota de 25,00 ml se transfiere a una celda de conductividad y sin ninguna otra dilución, se titula con NaOH 0,1000 Molar. Se obtiene la siguiente información, donde R es la resistencia medida en ohms y V es el volumen de solución de NaOH en mililitros.

R	V	R	V	R	V
29,4	0,00	32,5	1,80	23,8	3,60
32,6	0,20	31,7	2,00	23,0	3,80
34,1	0,40	31,1	2,20	22,3	4,00
38,0	0,60	30,1	2,40	21,7	4,20
36,9	0,80	29,1	2,60	21,0	4,40
35,8	1,00	27,7	2,80	20,4	4,60
34,9	1,20	26,7	3,00	19,6	4,80
34,1	1,40	25,7	3,20	18,5	5,00
33,3	1,60	24,7	3,40	17,3	5,20

Calcule e interprete el coeficiente de correlación la resistencia(R) y el volumen de solución (V).

6.28 Los datos tabulados representan la variación del voltaje de un sistema en función de un agente acomplejante, concretamente la variación del potencial de semi-onda de los iones cadmio en función de la concentración de ion cloruro. Como los potenciales de reducción en los datos polarográficos se toman negativos, en todos los cálculos arrastramos este signo. Además, el gráfico que nos interesa es el del potencial de semi-onda en función del logaritmo de la concentración del ion acomplejante, por lo cual en el cálculo de la constante de regresión (a) y del coeficiente de regresión (b) tenemos que tomar el logaritmo de esta concentración.

$X = \text{Log}[Cl^-]$	$Y = E_{1/2}$
-1,1938	-0,710
-0,7959	-0,750
-0,5918	-0,772
-0,4949	-0,782
-0,4157	-0,794
-0,3188	-0,808
-0,2396	-0,814
-0,1938	-0,824

- Calcule la ecuación de regresión lineal.
- Calcule el error de estimación.
- Calcular e interpretar el coeficiente de determinación

6.29 Bennett, Brooks y Glasstone (J. Chem. Soc. 1935, 1821) obtuvieron los siguientes resultados en la valoración del fluorofenol en alcohol al 30% a 25° C. Al agregar x ml de NaOH 0.01 Molar a 50,00 ml de una disolución 0,01 Molar del fluorofenol se obtuvieron los valores de pH siguientes:

ml NaOH 0,01 N	10	15	20	30	40
pH	8,73	9,01	9,20	9,56	10,00

Calcular la constante de disociación del o-fluorofenol. Suponer que el coeficiente de actividad del ácido no disociado es la unidad.

6.30 En disoluciones acuosas diluidas, el valor de la constante H_0 tiende a aumentar el pH cuando la fuerza iónica tiende a cero. Para todos los otros disolventes H_0 es una constante empírica que disminuye cuando la acidez de la disolución aumente y que mide, en una escala logarítmica, el poder de una disolución o de un disolvente para dar protones a una base no cargada (neutra). A continuación se presenta los valores de H_0 para varias mezclas $H_2SO_4:H_2O$:

% H ₂ SO ₄ en peso	Ho
5,96	+0,44
10,59	0,08
15,92	-0,33
23,9	-0,87
31,9	-1,34
41,4	-2,18
50,8	-3,04
75	-6,16
80	-6,85
85	-7,60
90	-8,19
95	-8,75
100	-1,60

- Ajustar el modelo de regresión lineal a los datos.
- Calcular la bondad de ajuste del modelo.
- Calcular e interpretar el coeficiente de correlación.

6.31 La determinación de níquel en aceites vegetales hidrogenados se puede realizar calcinando la muestra a cenizas, añadiendo molibdeno como patrón interno y analizando por espectroscopia de emisión.

Ni (ppm)	0,2	0,4	0,6	0,8	1,0
Razón	0,43	0,70	0,93	1,15	1,34

- Dibujar el diagrama de dispersión.
- Se encontró que un aceite hidrogenado problema daba una razón de intensidad de 0,85. ¿determinar la cantidad de níquel (en ppm.) en dicha muestra, suponga un modelo de regresión lineal?

6.32 Se analizaron muestras de mayonesas para determinar su contenido graso; los análisis se efectuaron por un método rápido (x) y también por el método estándar de la Asociación de Químicos Agrícolas (y).

x	80,5	30,3	25,2	77,4	48,1	38,7	18,6
y	79,3	30,4	26,0	77,9	47,5	35,5	16,7

- Marcar estos puntos en un diagrama de dispersión y ajustar al conjunto una línea de regresión utilizando el método de mínimos cuadrados.
- Calcule e interprete la bondad de ajuste del modelo.
- Calcular un intervalo de confianza del 99% para el método estándar, cuando el método rápido es 50%.

6.33 Curva patrón para una determinación cuantitativa de la espectrofotometría de absorción. Se hace la reacción coloreada con concentraciones conocidas de un compuesto y se mide la densidad óptica obtenida con distintas concentraciones.

Concentración g/l	0.8	0.9	1.0	1.5	2.8	3.0	4.0	4.5
Densidad óptica	0.09	0.15	0.29	0.30	0.45	0.53	0.60	0.59

Se unen los puntos gráficamente y la figura resultante es una recta si la coloración obedece a la ley de Beer-Lambert. Dibuje el diagrama de dispersión y represente aproximadamente la recta de regresión lineal.

APÉNDICE

A

FÓRMULAS ESTADÍSTICAS

MEDIDAS DE POSICIÓN	174
MEDIDAS DE VARIABILIDAD	175
MUESTREO ALEATORIO SIMPLE	176
ESTADÍSTICO DE PRUEBA DE HIPÓTESIS	177
ANÁLISIS DE REGRESIÓN LINEAL SIMPLE	178
ANÁLISIS DE CORRELACIÓN	178

MEDIDAS DE POSICIÓN

Datos sin agrupar	Datos agrupados
Promedio aritmético de muestras	
$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$	$\bar{x} = \frac{\sum_{i=1}^k x_i f_i}{\sum_{i=1}^k f_i}$
Promedio ponderado $\bar{x} = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}$	Mediana $M_e = L_i + \left(\frac{\frac{n}{2} - F_{i-1}}{f_i} \right) * c$
Mediana para n impar $M_e = X_{\left(\frac{n+1}{2}\right)}$	Moda $M_o = L_i + \left(\frac{d_1}{d_1 + d_2} \right) * c$ $d_1 = f_i - f_{i-1}$ $d_2 = f_i - f_{i+1}$
Mediana para n par $M_e = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$	Percentiles $P_m = L_i + \left(\frac{\frac{m \cdot n}{100} - F_{i-1}}{f_i} \right) * c$
Percentiles $P_m = X_{\left[\frac{m}{100}(n+1) \right]}$	
Media geométrica $\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$	Media armónica $\bar{x}_a = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$

MEDIDAS DE VARIABILIDAD

Datos sin agrupar	Datos agrupados
Variancia de una muestra	
$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ $s_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right]$	$s_x^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i - \bar{x})^2 \cdot f_i$ $s_x^2 = \frac{1}{n-1} \left[\sum_{i=1}^k x_i^2 f_i - \frac{\left(\sum_{i=1}^k x_i f_i \right)^2}{n} \right]$
Variancia de la población	
$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$ $\sigma_x^2 = \frac{1}{N} \left[\sum_{i=1}^N x_i^2 - \frac{\left(\sum_{i=1}^N x_i \right)^2}{N} \right]$	$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^k (x_i - \mu)^2 \cdot f_i$ $\sigma_x^2 = \frac{1}{N} \left[\sum_{i=1}^k x_i^2 f_i - \frac{\left(\sum_{i=1}^k x_i f_i \right)^2}{N} \right]$
Coeficiente de variación de una población $CV_x = \frac{\sigma_x}{\mu} * 100$	Coeficiente de variación de una muestra $CV_x = \frac{s_x}{\bar{x}} * 100$
MEDIDAS DE VARIABILIDAD ENTRE Y DENTRO DE GRUPOS	
Variancia entre grupos $\sigma_e^2 = \frac{\sum_{i=1}^k N_i (\mu_i - \mu)^2}{N}$	Variancia dentro de grupos $\sigma_d^2 = \frac{\sum_{i=1}^k N_i \sigma_i^2}{N}$
Coeficiente de determinación $R^2 = \frac{\sigma_e^2}{\sigma^2} * 100$	Promedio total $\mu = \frac{N_1 \mu_1 + N_2 \mu_2 + \dots + N_k \mu_k}{N}$
Medida de variabilidad para muestras pareadas $s_d^2 = \frac{1}{n-1} \left[\sum_{i=1}^n d_i^2 \right] \quad d_i = X_{1i} - X_{2i}$	Variancia para variables dicotómicas $\sigma^2 = PQ \quad s^2 = \hat{p}\hat{q}$

MUESTREO ALEATORIO SIMPLE

Población finita	Población infinita
Variancia del promedio	
$s_x^2 = \frac{N-n}{N-1} \cdot \frac{s_x^2}{n}$	$s_x^2 = \frac{s_x^2}{n}$
$\sigma_x^2 = \frac{N-n}{N-1} \cdot \frac{\sigma_x^2}{n}$	$\sigma_x^2 = \frac{\sigma_x^2}{n}$
Variancia de una proporción	
$s_{\hat{p}}^2 = \frac{N-n}{N-1} \cdot \frac{\hat{p}\hat{q}}{n}$	$s_{\hat{p}}^2 = \frac{\hat{p}\hat{q}}{n}$
$\sigma_{\hat{p}}^2 = \frac{N-n}{N-1} \cdot \frac{PQ}{n}$	$\sigma_{\hat{p}}^2 = \frac{PQ}{n}$
Tamaño de muestra para la estimación De un promedio y una proporción poblacional	
$n = \frac{n_1}{1 + \frac{n_1}{N}}$ donde $n_1 = \left(\frac{Z_{\alpha/2} \sigma}{d} \right)^2$	$n = \left(\frac{Z_{\alpha/2} \sigma}{d} \right)^2$
$n = \frac{n_1}{1 + \frac{n_1}{N}}$ donde $n_1 = \left(\frac{Z_{\alpha/2} \sqrt{PQ}}{d} \right)^2$	$n = \left(\frac{Z_{\alpha/2} \sqrt{PQ}}{d} \right)^2$
Intervalos de confianza para el promedio cuando la variancia de la población es conocida	
$\bar{x} \pm Z_{\alpha/2} * \sqrt{\frac{N-n}{N-1}} * \frac{\sigma_x}{\sqrt{n}}$	$\bar{x} \pm Z_{\alpha/2} * \frac{\sigma_x}{\sqrt{n}}$
Intervalos de confianza para el promedio cuando la variancia de la población es desconocida y $n \leq 30$	
$\bar{x} \pm t_{\alpha/2(n-1)gl} * \sqrt{\frac{N-n}{N-1}} * \frac{s_x}{\sqrt{n}}$	$\bar{x} \pm t_{\alpha/2(n-1)gl} * \frac{s_x}{\sqrt{n}}$
Intervalos de confianza para una proporción si $np > 5$ y $nq > 5$	
$\hat{p} \pm Z_{\alpha/2} * \sqrt{\frac{N-n}{N-1}} * \sqrt{\frac{\hat{p}\hat{q}}{n}}$	$\hat{p} \pm Z_{\alpha/2} * \sqrt{\frac{\hat{p}\hat{q}}{n}}$

ESTADÍSTICO PARA PRUEBA DE HIPÓTESIS

Promedios	Proporciones
Para un promedio: variancia conocida $Z_{obs} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$	Para una proporción $Z_{obs} = \frac{\hat{p} - P}{\sqrt{\frac{PQ}{n}}}$
Para un promedio: variancia desconocida $t_{obs} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$	Diferencia de proporciones $Z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$
Diferencia de dos promedios: variancia conocida $Z_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Otra alternativa de cálculo: $Z_{obs} = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ $p = \frac{x_1 + x_2}{n_1 + n_2}$
Diferencia de dos promedios: variancia desconocida $t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} * k$	
	donde $k = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$
Estadístico de prueba de independencia y de homogeneidad Ji-Cuadrada $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ $E_{ij} = \frac{N_i N_j}{N}$	Estadístico de prueba para muestras pareadas $t_{obs} = \frac{\bar{d}}{S_d / \sqrt{n}}$

REGRESIÓN LINEAL SIMPLE

<p>Constante de regresión</p> $a = \bar{y} - b\bar{x}$	<p>Coefficiente regresión lineal</p> $b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$
<p>Intervalos de confianza para el promedio de y dado un x_0</p> $\hat{y} \pm t_{\alpha/2(n-2)gl} * S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SC_x}}$	<p>Intervalos de confianza para una observación de y dado un x_0</p> $\hat{y} \pm t_{\alpha/2(n-2)gl} * S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SC_x}}$
<p>Error estándar de estimación</p> $S_e = \sqrt{\frac{\sum_{i=1}^n y_i^2 - a \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i y_i}{n - 2}}$	<p>Suma de cuadrados de x</p> $SC_x = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n}$
<p>Inferencia sobre la constante y coeficiente de regresión</p>	
<p>Intervalos de confianza</p> $a \pm t_{(n-2)gl} S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SC_x}}$ $b \pm t_{(n-2)gl} \frac{S_e}{\sqrt{SC_x}}$	<p>Estadístico de prueba de hipótesis</p> $t_c = \frac{a}{S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SC_x}}}$ $t_c = \frac{b}{S_e / \sqrt{SC_x}}$

ANÁLISIS DE CORRELACIÓN

<p>Coefficiente de correlación lineal</p> $r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] * \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}}$	
<p>Estadístico para prueba de hipótesis sobre el coeficiente de correlación</p> $t_{obs} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$	<p>Coefficiente de correlación parcial</p> $r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$

APÉNDICE

B

TABLAS ESTADÍSTICAS

TABLA 1. DISTRIBUCIÓN BINOMIAL	180
TABLA 2. DISTRIBUCIÓN DE POISSON	185
TABLA 3. DISTRIBUCIÓN NORMAL ESTANDAR	187
TABLA 4. DISTRIBUCIÓN T DE STUDENT	189
TABLA 5. DISTRIBUCIÓN JI CUADRADA	190

TABLA 1. DISTRIBUCIÓN BINOMIAL

n	x	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95
1	0	0,950	0,900	0,850	0,800	0,750	0,700	0,650	0,600	0,550	0,500	0,450	0,400	0,350	0,300	0,250	0,200	0,150	0,100	0,050
2	0	0,903	0,810	0,723	0,640	0,563	0,490	0,423	0,360	0,303	0,250	0,203	0,160	0,123	0,090	0,063	0,040	0,023	0,010	0,003
	1	0,998	0,990	0,978	0,960	0,938	0,910	0,878	0,840	0,798	0,750	0,698	0,640	0,578	0,510	0,438	0,360	0,278	0,190	0,098
3	0	0,857	0,729	0,614	0,512	0,422	0,343	0,275	0,216	0,166	0,125	0,091	0,064	0,043	0,027	0,016	0,008	0,003	0,001	0,000
	1	0,993	0,972	0,939	0,896	0,844	0,784	0,718	0,648	0,575	0,500	0,425	0,352	0,282	0,216	0,156	0,104	0,061	0,028	0,007
	2	1,000	0,999	0,997	0,992	0,984	0,973	0,957	0,936	0,909	0,875	0,834	0,784	0,725	0,657	0,578	0,488	0,386	0,271	0,143
4	0	0,815	0,656	0,522	0,410	0,316	0,240	0,179	0,130	0,092	0,063	0,041	0,026	0,015	0,008	0,004	0,002	0,001	0,000	0,000
	1	0,986	0,948	0,891	0,819	0,738	0,652	0,563	0,475	0,391	0,313	0,242	0,179	0,127	0,084	0,051	0,027	0,012	0,004	0,001
	2	1,000	0,996	0,988	0,973	0,949	0,916	0,874	0,821	0,759	0,688	0,609	0,525	0,437	0,348	0,262	0,181	0,110	0,052	0,014
	3	1,000	1,000	1,000	0,998	0,996	0,992	0,985	0,974	0,959	0,938	0,909	0,870	0,822	0,760	0,684	0,590	0,478	0,344	0,186
5	0	0,774	0,591	0,444	0,328	0,237	0,168	0,116	0,078	0,050	0,031	0,019	0,010	0,005	0,002	0,001	0,000	0,000	0,000	0,000
	1	0,977	0,919	0,835	0,737	0,633	0,528	0,428	0,337	0,256	0,188	0,131	0,087	0,054	0,031	0,016	0,007	0,002	0,001	0,000
	2	0,999	0,991	0,973	0,942	0,897	0,837	0,765	0,683	0,593	0,500	0,407	0,317	0,235	0,163	0,104	0,058	0,027	0,009	0,001
	3	1,000	1,000	0,998	0,993	0,984	0,969	0,946	0,913	0,869	0,813	0,744	0,663	0,572	0,472	0,367	0,263	0,165	0,082	0,023
	4	1,000	1,000	1,000	1,000	0,999	0,998	0,995	0,990	0,982	0,969	0,950	0,922	0,884	0,832	0,763	0,672	0,556	0,410	0,226
6	0	0,735	0,531	0,377	0,262	0,178	0,118	0,075	0,047	0,028	0,016	0,008	0,004	0,002	0,001	0,000	0,000	0,000	0,000	0,000
	1	0,967	0,886	0,777	0,655	0,534	0,420	0,319	0,233	0,164	0,109	0,069	0,041	0,022	0,011	0,005	0,002	0,000	0,000	0,000
	2	0,998	0,984	0,953	0,901	0,831	0,744	0,647	0,544	0,442	0,344	0,255	0,179	0,117	0,071	0,038	0,017	0,006	0,001	0,000
	3	1,000	0,999	0,994	0,983	0,962	0,930	0,883	0,821	0,745	0,656	0,559	0,456	0,353	0,256	0,169	0,099	0,047	0,016	0,002
	4	1,000	1,000	1,000	0,998	0,995	0,989	0,978	0,959	0,931	0,891	0,836	0,767	0,681	0,580	0,466	0,345	0,224	0,114	0,033
	5	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,996	0,992	0,984	0,972	0,953	0,925	0,882	0,822	0,738	0,623	0,469	0,265
7	0	0,698	0,478	0,321	0,210	0,134	0,082	0,049	0,028	0,015	0,008	0,004	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,956	0,850	0,717	0,577	0,445	0,329	0,234	0,159	0,102	0,063	0,036	0,019	0,009	0,004	0,001	0,000	0,000	0,000	0,000
	2	0,996	0,974	0,926	0,852	0,756	0,647	0,532	0,420	0,316	0,227	0,153	0,096	0,056	0,029	0,013	0,005	0,001	0,000	0,000
	3	1,000	0,997	0,988	0,967	0,929	0,874	0,800	0,710	0,608	0,500	0,392	0,290	0,200	0,126	0,071	0,033	0,012	0,003	0,000
	4	1,000	1,000	0,999	0,995	0,987	0,971	0,944	0,904	0,847	0,773	0,684	0,580	0,468	0,353	0,244	0,148	0,074	0,026	0,004
	5	1,000	1,000	1,000	1,000	0,999	0,996	0,991	0,981	0,964	0,938	0,898	0,841	0,766	0,671	0,555	0,423	0,283	0,150	0,044
	6	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,996	0,992	0,985	0,972	0,951	0,918	0,867	0,790	0,679	0,522	0,302
8	0	0,663	0,431	0,273	0,168	0,100	0,058	0,032	0,017	0,008	0,004	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,943	0,813	0,657	0,503	0,367	0,255	0,169	0,106	0,063	0,035	0,018	0,009	0,004	0,001	0,000	0,000	0,000	0,000	0,000
	2	0,994	0,962	0,895	0,797	0,679	0,552	0,428	0,315	0,220	0,145	0,089	0,050	0,025	0,011	0,004	0,001	0,000	0,000	0,000
	3	1,000	0,995	0,979	0,944	0,886	0,806	0,706	0,594	0,477	0,363	0,260	0,174	0,106	0,058	0,027	0,010	0,003	0,000	0,000
	4	1,000	1,000	0,997	0,990	0,973	0,942	0,894	0,826	0,740	0,637	0,523	0,406	0,294	0,194	0,114	0,056	0,021	0,005	0,000
	5	1,000	1,000	1,000	0,999	0,996	0,989	0,975	0,950	0,912	0,856	0,780	0,685	0,572	0,448	0,322	0,203	0,105	0,038	0,006
	6	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,992	0,982	0,965	0,937	0,894	0,831	0,745	0,633	0,497	0,343	0,187	0,057
	7	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,996	0,992	0,983	0,968	0,942	0,900	0,832	0,728	0,570	0,337

<i>n</i>	<i>x</i>	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95
9	0	0,630	0,387	0,232	0,134	0,075	0,040	0,021	0,010	0,005	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,929	0,775	0,600	0,436	0,300	0,196	0,121	0,071	0,039	0,020	0,009	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,992	0,947	0,859	0,738	0,601	0,463	0,337	0,232	0,150	0,090	0,050	0,025	0,011	0,004	0,001	0,000	0,000	0,000	0,000
	3	0,999	0,992	0,966	0,914	0,834	0,730	0,609	0,483	0,361	0,254	0,166	0,099	0,054	0,025	0,010	0,003	0,001	0,000	0,000
	4	1,000	0,999	0,994	0,980	0,951	0,901	0,828	0,733	0,621	0,500	0,379	0,267	0,172	0,099	0,049	0,020	0,006	0,001	0,000
	5	1,000	1,000	0,999	0,997	0,990	0,975	0,946	0,901	0,834	0,746	0,639	0,517	0,391	0,270	0,166	0,086	0,034	0,008	0,001
	6	1,000	1,000	1,000	1,000	0,999	0,996	0,989	0,975	0,950	0,910	0,851	0,768	0,663	0,537	0,399	0,262	0,141	0,053	0,008
	7	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,991	0,981	0,962	0,930	0,879	0,804	0,700	0,564	0,401	0,225	0,071
	8	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,995	0,990	0,979	0,960	0,925	0,866	0,768	0,613	0,370	
10	0	0,599	0,349	0,197	0,107	0,056	0,028	0,014	0,006	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,914	0,736	0,544	0,376	0,244	0,149	0,086	0,046	0,023	0,011	0,005	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,989	0,930	0,820	0,678	0,526	0,383	0,262	0,167	0,100	0,055	0,027	0,012	0,005	0,002	0,000	0,000	0,000	0,000	0,000
	3	0,999	0,987	0,950	0,879	0,776	0,650	0,514	0,382	0,266	0,172	0,102	0,055	0,026	0,011	0,004	0,001	0,000	0,000	0,000
	4	1,000	0,998	0,990	0,967	0,922	0,850	0,752	0,633	0,504	0,377	0,262	0,166	0,095	0,047	0,020	0,006	0,001	0,000	0,000
	5	1,000	1,000	0,999	0,994	0,980	0,953	0,905	0,834	0,738	0,623	0,496	0,367	0,249	0,150	0,078	0,033	0,010	0,002	0,000
	6	1,000	1,000	1,000	0,999	0,997	0,989	0,974	0,945	0,898	0,828	0,734	0,618	0,486	0,350	0,224	0,121	0,050	0,013	0,001
	7	1,000	1,000	1,000	1,000	1,000	0,998	0,995	0,988	0,973	0,945	0,900	0,833	0,738	0,617	0,474	0,322	0,180	0,070	0,012
	8	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,996	0,989	0,977	0,954	0,914	0,851	0,756	0,624	0,456	0,264	0,086	
	9	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,994	0,987	0,972	0,944	0,893	0,803	0,651	0,401	
11	0	0,569	0,314	0,167	0,086	0,042	0,020	0,009	0,004	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,898	0,697	0,492	0,322	0,197	0,113	0,061	0,030	0,014	0,006	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,985	0,910	0,779	0,617	0,455	0,313	0,200	0,119	0,065	0,033	0,015	0,006	0,002	0,001	0,000	0,000	0,000	0,000	0,000
	3	0,998	0,982	0,931	0,839	0,713	0,570	0,426	0,296	0,191	0,113	0,061	0,029	0,012	0,004	0,001	0,000	0,000	0,000	0,000
	4	1,000	0,997	0,984	0,950	0,885	0,790	0,668	0,533	0,397	0,274	0,174	0,099	0,050	0,022	0,008	0,002	0,000	0,000	0,000
	5	1,000	1,000	0,997	0,988	0,966	0,922	0,851	0,754	0,633	0,500	0,367	0,247	0,149	0,078	0,034	0,012	0,003	0,000	0,000
	6	1,000	1,000	1,000	0,998	0,992	0,978	0,950	0,901	0,826	0,726	0,603	0,467	0,332	0,210	0,115	0,050	0,016	0,003	0,000
	7	1,000	1,000	1,000	1,000	0,999	0,996	0,988	0,971	0,939	0,887	0,809	0,704	0,574	0,430	0,287	0,161	0,069	0,019	0,002
	8	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,994	0,985	0,967	0,935	0,881	0,800	0,687	0,545	0,383	0,221	0,090	0,015
	9	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,994	0,986	0,970	0,939	0,887	0,803	0,678	0,508	0,303	0,102	
	10	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,991	0,980	0,958	0,914	0,833	0,686	0,431	
12	0	0,540	0,282	0,142	0,069	0,032	0,014	0,006	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,882	0,659	0,444	0,275	0,158	0,085	0,042	0,020	0,008	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,980	0,889	0,736	0,558	0,391	0,253	0,151	0,083	0,042	0,019	0,008	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,998	0,974	0,908	0,795	0,649	0,493	0,347	0,225	0,135	0,073	0,036	0,015	0,006	0,002	0,000	0,000	0,000	0,000	0,000
	4	1,000	0,996	0,976	0,927	0,842	0,724	0,583	0,438	0,304	0,194	0,112	0,057	0,026	0,010	0,003	0,001	0,000	0,000	0,000
	5	1,000	1,000	0,995	0,981	0,946	0,882	0,787	0,665	0,527	0,387	0,261	0,158	0,085	0,039	0,014	0,004	0,001	0,000	0,000
	6	1,000	1,000	0,999	0,996	0,986	0,961	0,915	0,842	0,739	0,613	0,473	0,335	0,213	0,118	0,054	0,019	0,005	0,001	0,000
	7	1,000	1,000	1,000	0,999	0,997	0,991	0,975	0,943	0,888	0,806	0,696	0,562	0,417	0,276	0,158	0,073	0,024	0,004	0,000
	8	1,000	1,000	1,000	1,000	1,000	0,998	0,994	0,985	0,964	0,927	0,866	0,775	0,653	0,508	0,351	0,205	0,092	0,026	0,002
	9	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,992	0,981	0,958	0,917	0,849	0,747	0,609	0,442	0,264	0,111	0,020
	10	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,992	0,980	0,958	0,915	0,842	0,725	0,557	0,341	0,118
	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,994	0,986	0,968	0,931	0,858	0,718	0,460	

<i>n</i>	<i>x</i>	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95
13	0	0,513	0,254	0,121	0,055	0,024	0,010	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,865	0,621	0,398	0,234	0,127	0,064	0,030	0,013	0,005	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,976	0,866	0,692	0,502	0,333	0,203	0,113	0,058	0,027	0,011	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,997	0,966	0,882	0,747	0,584	0,421	0,278	0,169	0,093	0,046	0,020	0,008	0,003	0,001	0,000	0,000	0,000	0,000	0,000
	4	1,000	0,994	0,966	0,901	0,794	0,654	0,501	0,353	0,228	0,133	0,070	0,032	0,013	0,004	0,001	0,000	0,000	0,000	0,000
	5	1,000	0,999	0,993	0,970	0,920	0,835	0,716	0,574	0,427	0,291	0,179	0,098	0,046	0,018	0,006	0,001	0,000	0,000	0,000
	6	1,000	1,000	0,999	0,993	0,976	0,938	0,871	0,771	0,644	0,500	0,356	0,229	0,130	0,062	0,024	0,007	0,001	0,000	0,000
	7	1,000	1,000	1,000	0,999	0,994	0,982	0,954	0,902	0,821	0,710	0,573	0,426	0,284	0,165	0,080	0,030	0,008	0,001	0,000
	8	1,000	1,000	1,000	1,000	0,999	0,996	0,987	0,968	0,930	0,867	0,772	0,647	0,500	0,346	0,206	0,099	0,034	0,007	0,000
	9	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,992	0,980	0,954	0,907	0,831	0,722	0,579	0,416	0,253	0,118	0,034	0,003
	10	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,989	0,973	0,942	0,887	0,798	0,667	0,498	0,308	0,134	0,025	0,000
	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,995	0,987	0,970	0,936	0,873	0,766	0,602	0,379	0,135	0,000
	12	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,990	0,976	0,945	0,879	0,746	0,487	0,000
14	0	0,488	0,229	0,103	0,044	0,018	0,007	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,847	0,585	0,357	0,198	0,101	0,048	0,021	0,008	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,970	0,842	0,648	0,448	0,281	0,161	0,084	0,040	0,017	0,007	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,996	0,956	0,854	0,698	0,521	0,355	0,221	0,124	0,063	0,029	0,011	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	4	1,000	0,991	0,953	0,870	0,742	0,584	0,423	0,279	0,167	0,090	0,043	0,018	0,006	0,002	0,000	0,000	0,000	0,000	0,000
	5	1,000	0,999	0,989	0,956	0,888	0,781	0,641	0,486	0,337	0,212	0,119	0,058	0,024	0,008	0,002	0,000	0,000	0,000	0,000
	6	1,000	1,000	0,998	0,988	0,962	0,907	0,816	0,693	0,546	0,395	0,259	0,150	0,075	0,032	0,010	0,002	0,000	0,000	0,000
	7	1,000	1,000	1,000	0,998	0,990	0,969	0,925	0,850	0,741	0,605	0,454	0,308	0,184	0,093	0,038	0,012	0,002	0,000	0,000
	8	1,000	1,000	1,000	1,000	0,998	0,992	0,976	0,942	0,881	0,788	0,663	0,514	0,360	0,220	0,112	0,044	0,012	0,002	0,000
	9	1,000	1,000	1,000	1,000	1,000	0,998	0,994	0,983	0,957	0,910	0,833	0,721	0,577	0,416	0,259	0,130	0,047	0,009	0,000
	10	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,989	0,971	0,937	0,876	0,780	0,645	0,479	0,302	0,147	0,044	0,004
	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,994	0,983	0,960	0,916	0,839	0,719	0,552	0,352	0,158	0,030
	12	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,992	0,980	0,953	0,899	0,802	0,643	0,415	0,153	0,000
	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,993	0,982	0,956	0,897	0,771	0,512	0,000
15	0	0,463	0,206	0,087	0,035	0,013	0,005	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,829	0,549	0,319	0,167	0,080	0,035	0,014	0,005	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,964	0,816	0,604	0,398	0,236	0,127	0,062	0,027	0,011	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,995	0,944	0,823	0,648	0,461	0,297	0,173	0,091	0,042	0,018	0,006	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	4	0,999	0,987	0,938	0,836	0,687	0,516	0,352	0,217	0,120	0,059	0,026	0,009	0,003	0,001	0,000	0,000	0,000	0,000	0,000
	5	1,000	0,998	0,983	0,939	0,852	0,722	0,564	0,403	0,261	0,151	0,077	0,034	0,012	0,004	0,001	0,000	0,000	0,000	0,000
	6	1,000	1,000	0,996	0,982	0,943	0,869	0,755	0,610	0,452	0,304	0,182	0,095	0,042	0,015	0,004	0,001	0,000	0,000	0,000
	7	1,000	1,000	0,999	0,996	0,983	0,950	0,887	0,787	0,654	0,500	0,347	0,213	0,113	0,050	0,017	0,004	0,001	0,000	0,000
	8	1,000	1,000	1,000	0,999	0,996	0,985	0,958	0,905	0,818	0,696	0,548	0,390	0,245	0,131	0,057	0,018	0,004	0,000	0,000
	9	1,000	1,000	1,000	1,000	0,999	0,996	0,988	0,966	0,923	0,849	0,739	0,597	0,436	0,278	0,148	0,061	0,017	0,002	0,000
	10	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,991	0,975	0,941	0,880	0,783	0,648	0,485	0,314	0,164	0,062	0,013	0,001
	11	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,994	0,982	0,958	0,910	0,827	0,703	0,539	0,352	0,177	0,056	0,006	0,000
	12	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,989	0,973	0,938	0,873	0,764	0,602	0,396	0,184	0,036	0,000
	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,995	0,986	0,965	0,920	0,833	0,681	0,451	0,171	0,000
	14	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,995	0,987	0,965	0,913	0,794	0,537	0,000

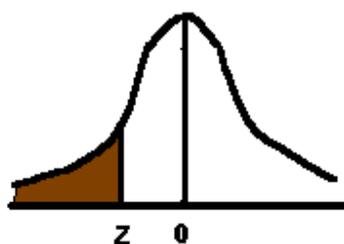
<i>n</i>	<i>x</i>	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	
16	0	0,440	0,185	0,074	0,028	0,010	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,811	0,515	0,284	0,141	0,064	0,026	0,010	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,957	0,789	0,561	0,352	0,197	0,099	0,045	0,018	0,007	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,993	0,932	0,790	0,598	0,405	0,246	0,134	0,065	0,028	0,011	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	4	0,999	0,983	0,921	0,798	0,630	0,450	0,289	0,167	0,085	0,038	0,015	0,005	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	5	1,000	0,997	0,977	0,918	0,810	0,660	0,490	0,329	0,198	0,105	0,049	0,019	0,006	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	6	1,000	1,000	0,994	0,973	0,920	0,825	0,688	0,527	0,366	0,227	0,124	0,058	0,023	0,007	0,002	0,000	0,000	0,000	0,000	0,000
	7	1,000	1,000	0,999	0,993	0,973	0,926	0,841	0,716	0,563	0,402	0,256	0,142	0,067	0,026	0,008	0,002	0,000	0,000	0,000	0,000
	8	1,000	1,000	1,000	0,999	0,993	0,974	0,933	0,858	0,744	0,598	0,437	0,284	0,159	0,074	0,027	0,007	0,001	0,000	0,000	0,000
	9	1,000	1,000	1,000	1,000	0,998	0,993	0,977	0,942	0,876	0,773	0,634	0,473	0,312	0,175	0,080	0,027	0,006	0,001	0,000	0,000
	10	1,000	1,000	1,000	1,000	1,000	0,998	0,994	0,981	0,951	0,895	0,802	0,671	0,510	0,340	0,190	0,082	0,024	0,003	0,000	0,000
	11	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,995	0,985	0,962	0,915	0,833	0,711	0,550	0,370	0,202	0,079	0,017	0,001	0,000
	12	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,989	0,972	0,935	0,866	0,754	0,595	0,402	0,210	0,068	0,007	0,000
	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,993	0,982	0,955	0,901	0,803	0,648	0,439	0,211	0,043	0,000
	14	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,990	0,974	0,937	0,859	0,716	0,485	0,189	0,000	0,000
	15	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,990	0,972	0,926	0,815	0,560	0,000
17	0	0,418	0,167	0,063	0,023	0,008	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,792	0,482	0,253	0,118	0,050	0,019	0,007	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,950	0,762	0,520	0,310	0,164	0,077	0,033	0,012	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,991	0,917	0,756	0,549	0,353	0,202	0,103	0,046	0,018	0,006	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	4	0,999	0,978	0,901	0,758	0,574	0,389	0,235	0,126	0,060	0,025	0,009	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	5	1,000	0,995	0,968	0,894	0,765	0,597	0,420	0,264	0,147	0,072	0,030	0,011	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	6	1,000	0,999	0,992	0,962	0,893	0,775	0,619	0,448	0,290	0,166	0,083	0,035	0,012	0,003	0,001	0,000	0,000	0,000	0,000	0,000
	7	1,000	1,000	0,998	0,989	0,960	0,895	0,787	0,641	0,474	0,315	0,183	0,092	0,038	0,013	0,003	0,001	0,000	0,000	0,000	0,000
	8	1,000	1,000	1,000	0,997	0,988	0,960	0,901	0,801	0,663	0,500	0,337	0,199	0,099	0,040	0,012	0,003	0,000	0,000	0,000	0,000
	9	1,000	1,000	1,000	1,000	0,997	0,987	0,962	0,908	0,817	0,686	0,526	0,360	0,213	0,105	0,040	0,011	0,002	0,000	0,000	0,000
	10	1,000	1,000	1,000	1,000	0,999	0,997	0,988	0,965	0,917	0,834	0,710	0,552	0,381	0,225	0,107	0,038	0,008	0,001	0,000	0,000
	11	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,989	0,970	0,928	0,853	0,736	0,580	0,403	0,235	0,106	0,032	0,005	0,000	0,000
	12	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,991	0,976	0,940	0,874	0,765	0,611	0,426	0,242	0,099	0,022	0,001	0,000
	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,994	0,982	0,954	0,897	0,798	0,647	0,451	0,244	0,083	0,009	0,000	0,000
	14	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,988	0,967	0,923	0,836	0,690	0,480	0,238	0,050	0,000	0,000
	15	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,993	0,981	0,950	0,882	0,748	0,518	0,208	0,000	0,000
	16	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,993	0,978	0,937	0,833	0,582	0,000
18	0	0,397	0,150	0,054	0,018	0,006	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,774	0,450	0,224	0,099	0,040	0,014	0,005	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,942	0,734	0,480	0,271	0,135	0,060	0,024	0,008	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,989	0,902	0,720	0,501	0,306	0,165	0,078	0,033	0,012	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	4	0,999	0,972	0,879	0,716	0,519	0,333	0,189	0,094	0,041	0,015	0,005	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	5	1,000	0,994	0,958	0,867	0,718	0,534	0,355	0,209	0,108	0,048	0,018	0,006	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	6	1,000	0,999	0,988	0,949	0,861	0,722	0,549	0,374	0,226	0,119	0,054	0,020	0,006	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	7	1,000	1,000	0,997	0,984	0,943	0,859	0,728	0,563	0,392	0,240	0,128	0,058	0,021	0,006	0,001	0,000	0,000	0,000	0,000	0,000
	8	1,000	1,000	1,000	0,996	0,981	0,940	0,861	0,737	0,578	0,407	0,253	0,135	0,060	0,021	0,005	0,001	0,000	0,000	0,000	0,000
	9	1,000	1,000	1,000	0,999	0,995	0,979	0,940	0,865	0,747	0,593	0,422	0,263	0,139	0,060	0,019	0,004	0,001	0,000	0,000	0,000
	10	1,000	1,000	1,000	1,000	0,999	0,994	0,979	0,942	0,872	0,760	0,609	0,437	0,272	0,141	0,057	0,016	0,003	0,000	0,000	0,000
	11	1,000	1,000	1,000	1,000	1,000	0,999	0,994	0,980	0,946	0,881	0,774	0,626	0,451	0,278	0,139	0,051	0,012	0,001	0,000	0,000
	12	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,994	0,982	0,952	0,892	0,791	0,645	0,466	0,283	0,133	0,042	0,006	0,000	0,000
	13	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,995	0,985	0,959	0,906	0,811	0,667	0,481	0,284	0,121	0,028	0,002	0,000	0,000
	14	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,988	0,967	0,922	0,835	0,694	0,499	0,280	0,098	0,011	0,000	0,000
	15	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,998	0,992	0,976	0,940	0,865	0,729	0,520	0,266	0,058	0,000	0,000
	16	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,995	0,986	0,961	0,901	0,776	0,550	0,227	0,000
	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,994	0,982	0,946	0,850	0,603	0,000	0,000

<i>n</i>	<i>x</i>	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95
19	0	0,377	0,135	0,046	0,014	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,755	0,420	0,199	0,083	0,031	0,010	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,934	0,705	0,441	0,237	0,111	0,046	0,017	0,006	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,987	0,885	0,684	0,455	0,263	0,133	0,059	0,023	0,008	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	4	0,998	0,965	0,856	0,673	0,465	0,282	0,150	0,070	0,028	0,010	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	5	1,000	0,991	0,946	0,837	0,668	0,474	0,297	0,163	0,078	0,032	0,011	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	6	1,000	0,998	0,984	0,932	0,825	0,666	0,481	0,308	0,173	0,084	0,034	0,012	0,003	0,001	0,000	0,000	0,000	0,000	0,000
	7	1,000	1,000	0,996	0,977	0,923	0,818	0,666	0,488	0,317	0,180	0,087	0,035	0,011	0,003	0,001	0,000	0,000	0,000	0,000
	8	1,000	1,000	0,999	0,993	0,971	0,916	0,815	0,668	0,494	0,324	0,184	0,089	0,035	0,011	0,002	0,000	0,000	0,000	0,000
	9	1,000	1,000	1,000	0,998	0,991	0,967	0,913	0,814	0,671	0,500	0,329	0,186	0,088	0,033	0,009	0,002	0,000	0,000	0,000
	10	1,000	1,000	1,000	1,000	0,998	0,990	0,965	0,912	0,816	0,676	0,506	0,333	0,186	0,084	0,029	0,007	0,001	0,000	0,000
	11	1,000	1,000	1,000	1,000	1,000	0,997	0,989	0,965	0,913	0,820	0,683	0,512	0,334	0,182	0,078	0,023	0,004	0,000	0,000
	12	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,988	0,966	0,917	0,827	0,692	0,519	0,335	0,175	0,068	0,016	0,002	0,000
	13	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,989	0,968	0,922	0,837	0,703	0,526	0,332	0,163	0,054	0,009	0,000
	14	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,990	0,972	0,930	0,850	0,718	0,535	0,327	0,144	0,035	0,002	0,000
	15	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,992	0,977	0,941	0,867	0,737	0,545	0,316	0,115	0,013
	16	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,995	0,983	0,954	0,889	0,763	0,559	0,295	0,067
	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,990	0,969	0,917	0,802	0,580	0,245
	18	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,986	0,954	0,865	0,623
20	0	0,359	0,122	0,039	0,012	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,736	0,392	0,176	0,069	0,024	0,008	0,002	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,925	0,677	0,405	0,206	0,091	0,036	0,012	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,984	0,867	0,648	0,411	0,225	0,107	0,044	0,016	0,005	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	4	0,997	0,957	0,830	0,630	0,415	0,238	0,118	0,051	0,019	0,006	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	5	1,000	0,989	0,933	0,804	0,617	0,416	0,245	0,126	0,055	0,021	0,006	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	6	1,000	0,998	0,978	0,913	0,786	0,608	0,417	0,250	0,130	0,058	0,021	0,007	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	7	1,000	1,000	0,994	0,968	0,898	0,772	0,601	0,416	0,252	0,132	0,058	0,021	0,006	0,001	0,000	0,000	0,000	0,000	0,000
	8	1,000	1,000	0,999	0,990	0,959	0,887	0,762	0,596	0,414	0,252	0,131	0,057	0,020	0,005	0,001	0,000	0,000	0,000	0,000
	9	1,000	1,000	1,000	0,997	0,986	0,952	0,878	0,755	0,591	0,412	0,249	0,128	0,053	0,017	0,004	0,001	0,000	0,000	0,000
	10	1,000	1,000	1,000	0,999	0,996	0,983	0,947	0,873	0,751	0,588	0,409	0,245	0,122	0,048	0,014	0,003	0,000	0,000	0,000
	11	1,000	1,000	1,000	1,000	0,999	0,995	0,980	0,944	0,869	0,748	0,586	0,404	0,238	0,113	0,041	0,010	0,001	0,000	0,000
	12	1,000	1,000	1,000	1,000	1,000	0,999	0,994	0,979	0,942	0,868	0,748	0,584	0,399	0,228	0,102	0,032	0,006	0,000	0,000
	13	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,994	0,979	0,942	0,870	0,750	0,583	0,392	0,214	0,087	0,022	0,002	0,000
	14	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,994	0,979	0,945	0,874	0,755	0,584	0,383	0,196	0,067	0,011	0,000	0,000
	15	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,994	0,981	0,949	0,882	0,763	0,585	0,370	0,170	0,043	0,003	0,000
	16	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,995	0,984	0,956	0,893	0,775	0,589	0,352	0,133	0,016	0,000
	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,996	0,988	0,965	0,909	0,794	0,595	0,323	0,076
	18	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,992	0,976	0,931	0,824	0,608	0,264
	19	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,997	0,989	0,961	0,878	0,642

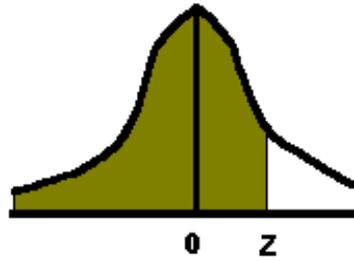
X	3,50	4,00	4,50	5,00	5,50	6,00	6,50	7,00
0	0,0302	0,0183	0,0111	0,0067	0,0041	0,0025	0,0015	0,0009
1	0,1359	0,0916	0,0611	0,0404	0,0266	0,0174	0,0113	0,0073
2	0,3208	0,2381	0,1736	0,1247	0,0884	0,0620	0,0430	0,0296
3	0,5366	0,4335	0,3423	0,2650	0,2017	0,1512	0,1118	0,0818
4	0,7254	0,6288	0,5321	0,4405	0,3575	0,2851	0,2237	0,1730
5	0,8576	0,7851	0,7029	0,6160	0,5289	0,4457	0,3690	0,3007
6	0,9347	0,8893	0,8311	0,7622	0,6860	0,6063	0,5265	0,4497
7	0,9733	0,9489	0,9134	0,8666	0,8095	0,7440	0,6728	0,5987
8	0,9901	0,9786	0,9597	0,9319	0,8944	0,8472	0,7916	0,7291
9	0,9967	0,9919	0,9829	0,9682	0,9462	0,9161	0,8774	0,8305
10	0,9990	0,9972	0,9933	0,9863	0,9747	0,9574	0,9332	0,9015
11	0,9997	0,9991	0,9976	0,9945	0,9890	0,9799	0,9661	0,9467
12	0,9999	0,9997	0,9992	0,9980	0,9955	0,9912	0,9840	0,9730
13	1,0000	0,9999	0,9997	0,9993	0,9983	0,9964	0,9929	0,9872
14	1,0000	1,0000	0,9999	0,9998	0,9994	0,9986	0,9970	0,9943
15	1,0000	1,0000	1,0000	0,9999	0,9998	0,9995	0,9988	0,9976
16	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9996	0,9990
17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9996
18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999
19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

X	7,50	8,00	8,50	9,00	9,50	10,0	15,0	20,0
0	0,0006	0,0003	0,0002	0,0001	0,0001	0,0000	0,0000	0,0000
1	0,0047	0,0030	0,0019	0,0012	0,0008	0,0005	0,0000	0,0000
2	0,0203	0,0138	0,0093	0,0062	0,0042	0,0028	0,0000	0,0000
3	0,0591	0,0424	0,0301	0,0212	0,0149	0,0103	0,0002	0,0000
4	0,1321	0,0996	0,0744	0,0550	0,0403	0,0293	0,0009	0,0000
5	0,2414	0,1912	0,1496	0,1157	0,0885	0,0671	0,0028	0,0001
6	0,3782	0,3134	0,2562	0,2068	0,1649	0,1301	0,0076	0,0003
7	0,5246	0,4530	0,3856	0,3239	0,2687	0,2202	0,0180	0,0008
8	0,6620	0,5925	0,5231	0,4557	0,3918	0,3328	0,0374	0,0021
9	0,7764	0,7166	0,6530	0,5874	0,5218	0,4579	0,0699	0,0050
10	0,8622	0,8159	0,7634	0,7060	0,6453	0,5830	0,1185	0,0108
11	0,9208	0,8881	0,8487	0,8030	0,7520	0,6968	0,1848	0,0214
12	0,9573	0,9362	0,9091	0,8758	0,8364	0,7916	0,2676	0,0390
13	0,9784	0,9658	0,9486	0,9261	0,8981	0,8645	0,3632	0,0661
14	0,9897	0,9827	0,9726	0,9585	0,9400	0,9165	0,4657	0,1049
15	0,9954	0,9918	0,9862	0,9780	0,9665	0,9513	0,5681	0,1565
16	0,9980	0,9963	0,9934	0,9889	0,9823	0,9730	0,6641	0,2211
17	0,9992	0,9984	0,9970	0,9947	0,9911	0,9857	0,7489	0,2970
18	0,9997	0,9993	0,9987	0,9976	0,9957	0,9928	0,8195	0,3814
19	0,9999	0,9997	0,9995	0,9989	0,9980	0,9965	0,8752	0,4703
20	1,0000	0,9999	0,9998	0,9996	0,9991	0,9984	0,9170	0,5591

TABLA 3. DISTRIBUCIÓN NORMAL ESTÁNDAR

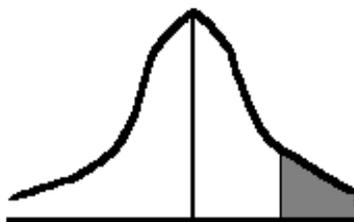


Z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0,00
-3,9	0,00005	0,00005	0,00005	0,00005	0,00006	0,00006	0,00006	0,00006	0,00007	0,00007
-3,8	0,00007	0,00007	0,00008	0,00008	0,00008	0,00009	0,00009	0,00010	0,00010	0,00010
-3,7	0,00011	0,00011	0,00012	0,00012	0,00013	0,00013	0,00014	0,00014	0,00015	0,00015
-3,6	0,00016	0,00016	0,00017	0,00018	0,00018	0,00019	0,00020	0,00021	0,00021	0,00022
-3,5	0,00023	0,00024	0,00025	0,00026	0,00027	0,00028	0,00029	0,00030	0,00031	0,00032
-3,4	0,00034	0,00034	0,00036	0,00038	0,00039	0,00040	0,00042	0,00043	0,00045	0,00047
-3,3	0,00048	0,00050	0,00052	0,00054	0,00056	0,00058	0,00060	0,00062	0,00064	0,00066
-3,2	0,00069	0,00071	0,00074	0,00076	0,00079	0,00082	0,00084	0,00087	0,00090	0,00093
-3,1	0,00097	0,00100	0,00103	0,00107	0,00111	0,00114	0,00118	0,00122	0,00126	0,00131
-3,0	0,00100	0,00103	0,00107	0,00111	0,00114	0,00118	0,00122	0,00126	0,00131	0,00135
-2,9	0,00139	0,00144	0,00149	0,00154	0,00159	0,00164	0,00169	0,00175	0,00181	0,00187
-2,8	0,00193	0,00199	0,00205	0,00212	0,00219	0,00226	0,00233	0,00240	0,00248	0,00255
-2,7	0,00263	0,00272	0,00280	0,00289	0,00298	0,00307	0,00317	0,00326	0,00336	0,00347
-2,6	0,00357	0,00368	0,00379	0,00391	0,00402	0,00414	0,00427	0,00440	0,00453	0,00466
-2,5	0,00489	0,00494	0,00508	0,00523	0,00539	0,00554	0,00570	0,00587	0,00604	0,00621
-2,4	0,00639	0,00657	0,00676	0,00695	0,00714	0,00734	0,00755	0,00776	0,00798	0,00820
-2,3	0,00842	0,00866	0,00889	0,00914	0,00939	0,00964	0,00990	0,01017	0,01044	0,01072
-2,2	0,01101	0,01130	0,01160	0,01191	0,01222	0,01254	0,01287	0,01321	0,01355	0,01390
-2,1	0,01426	0,01463	0,01500	0,01539	0,01578	0,01618	0,01659	0,01700	0,01743	0,01786
-2,0	0,01831	0,01876	0,01923	0,01970	0,02018	0,02067	0,02118	0,02169	0,02222	0,02275
-1,9	0,02329	0,02385	0,02442	0,02500	0,02559	0,02619	0,02680	0,02743	0,02807	0,02872
-1,8	0,02938	0,03005	0,03074	0,03144	0,03216	0,03288	0,03362	0,03438	0,03515	0,03593
-1,7	0,03673	0,03754	0,03837	0,03920	0,04006	0,04093	0,04181	0,04272	0,04363	0,04456
-1,6	0,04551	0,04648	0,04746	0,04846	0,04947	0,05050	0,05155	0,05262	0,05370	0,05480
-1,5	0,05592	0,05705	0,05821	0,05938	0,06057	0,06178	0,06301	0,06425	0,06552	0,06681
-1,4	0,06811	0,06944	0,07078	0,07214	0,07353	0,07493	0,07636	0,07780	0,07927	0,08076
-1,3	0,08226	0,08379	0,08534	0,08691	0,08851	0,09012	0,09176	0,09342	0,09510	0,09680
-1,2	0,09852	0,10027	0,10204	0,10383	0,10565	0,10749	0,10935	0,11123	0,11314	0,11507
-1,1	0,11702	0,11900	0,12100	0,12302	0,12507	0,12714	0,12924	0,13136	0,13350	0,13567
-1,0	0,13786	0,14007	0,14231	0,14457	0,14686	0,14917	0,15150	0,15386	0,15625	0,15865
-0,9	0,16109	0,16354	0,16602	0,16853	0,17105	0,17361	0,17619	0,17879	0,18141	0,18406
-0,8	0,18673	0,18943	0,19215	0,19489	0,19766	0,20045	0,20327	0,20611	0,20897	0,21185
-0,7	0,21476	0,21769	0,22065	0,22363	0,22663	0,22965	0,23269	0,23576	0,23885	0,24196
-0,6	0,24510	0,24825	0,25143	0,25463	0,25785	0,26109	0,26435	0,26763	0,27093	0,27425
-0,5	0,27759	0,28096	0,28434	0,28774	0,29116	0,29460	0,29806	0,30153	0,30503	0,30854
-0,4	0,31207	0,31561	0,31918	0,32276	0,32635	0,32997	0,33360	0,33724	0,34090	0,34459
-0,3	0,34827	0,35197	0,35569	0,35942	0,36317	0,36693	0,37070	0,37448	0,37828	0,38209
-0,2	0,38591	0,38974	0,39358	0,39743	0,40129	0,40516	0,40905	0,41294	0,41683	0,42074
-0,1	0,42465	0,42858	0,43250	0,43644	0,44038	0,44433	0,44828	0,45224	0,45620	0,46017
-0,0	0,46414	0,46812	0,47210	0,47609	0,48006	0,48405	0,48803	0,49202	0,49601	0,50000



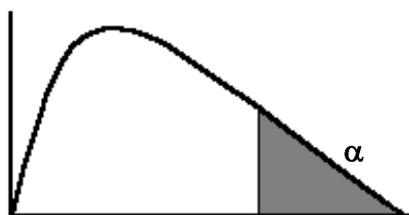
Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54395	0,54776	0,55172	0,55567	0,55962	0,56356	0,56750	0,57124	0,57534
0,2	0,57926	0,58617	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61781	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79373	0,79955	0,80234	0,80510	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82124	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87492	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89616	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91308	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96079	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96637	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98299	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99001	0,99036	0,99061	0,99086	0,99110	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99491	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99597	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99830	0,99836	0,99841	0,99846	0,99851	0,99856	0,99860
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99897	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99956	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

TABLA 4. DISTRIBUCIÓN T-STUDENT



Grados de libertad	$\alpha=0,25$	$\alpha=0,10$	$\alpha=0,075$	$\alpha=0,05$	$\alpha=0,025$	$\alpha=0,01$	$\alpha=0,005$	$\alpha=0,0001$
1	1,0000	3,0777	4,1653	6,3138	12,7062	31,8217	63,657	318,31
2	0,8165	1,8856	2,2819	2,9200	4,3027	6,9646	9,9248	22,327
3	0,7649	1,6377	1,9243	2,3534	3,1824	4,5407	5,8409	10,215
4	0,7407	1,5332	1,7782	2,1318	2,7764	3,7469	4,6041	7,1732
5	0,7267	1,4759	1,6994	2,0150	2,5706	3,3649	4,0321	5,8934
6	0,7176	1,4398	1,6502	1,9432	2,4469	3,1427	3,7074	5,2076
7	0,7111	1,4149	1,6166	1,8946	2,3646	2,9980	3,4995	4,7853
8	0,7064	1,3968	1,5922	1,8595	2,3060	2,8965	3,3554	4,5008
9	0,7027	1,3830	1,5737	1,8331	2,2622	2,8214	3,2498	4,2968
10	0,6998	1,3722	1,5592	1,8125	2,2281	2,7638	3,1693	4,1437
11	0,6974	1,3634	1,5476	1,7959	2,2010	2,7181	3,1058	4,0247
12	0,6955	1,3562	1,5380	1,7823	2,1788	2,6810	3,0545	3,9296
13	0,6938	1,3502	1,5299	1,7709	2,1604	2,6503	3,0123	3,8520
14	0,6924	1,3450	1,5231	1,7613	2,1448	2,6245	2,9768	3,7874
15	0,6912	1,3406	1,5172	1,7531	2,1314	2,6025	2,9467	3,7328
16	0,6901	1,3368	1,5121	1,7459	2,1199	2,5835	2,9208	3,6862
17	0,6892	1,3334	1,5077	1,7396	2,1098	2,5669	2,8982	3,6458
18	0,6884	1,3304	1,5037	1,7341	2,1009	2,5524	2,8784	3,6105
19	0,6876	1,3277	1,5002	1,7291	2,0930	2,5395	2,8609	3,5794
20	0,6870	1,3253	1,4970	1,7247	2,0860	2,5280	2,8453	3,5518
21	0,6864	1,3232	1,4942	1,7207	2,0796	2,5176	2,8314	3,5272
22	0,6858	1,3212	1,4916	1,7171	2,0739	2,5083	2,8188	3,5050
23	0,6853	1,3195	1,4893	1,7139	2,0687	2,4999	2,8073	3,4850
24	0,6848	1,3178	1,4871	1,7109	2,0639	2,4922	2,7969	3,4668
25	0,6844	1,3163	1,4852	1,7081	2,0595	2,4851	2,7874	3,4502
26	,6840	1,3150	1,4834	1,7056	2,0555	2,4786	2,7787	3,4350
27	0,6837	1,3137	1,4817	1,7033	2,0518	2,4727	2,7707	3,4210
28	0,6834	1,3125	1,4801	1,7011	2,0484	2,4671	2,7633	3,4082
29	0,6830	1,3114	1,4787	1,6991	2,0452	2,4620	2,7564	3,3962
30	0,6828	1,3104	1,4774	1,6973	2,0423	2,4573	2,7500	3,3852
40	0,6807	1,3031	1,4677	1,6839	2,0211	2,4233	2,7045	3,3069
50	0,6794	1,2987	1,4620	1,6759	2,0086	2,4033	2,6778	3,2614
60	0,6786	1,2958	1,4582	1,6706	2,0003	2,3901	2,6603	3,2317
120	0,6765	1,2886	1,4488	1,6577	1,9799	2,3578	2,6174	3,1595
∞	0,6745	1,2842	1,4430	1,6496	1,9674	2,3381	2,5913	3,1159

TABLA 5. DISTRIBUCIÓN JI-CUADRADA



Grados de libertad	$\alpha=0,995$	$\alpha=0,99$	$\alpha=0,975$	$\alpha=0,95$	$\alpha=0,90$	$\alpha=0,10$	$\alpha=0,05$	$\alpha=0,025$	$\alpha=0,01$	$\alpha=0,005$
1	0,0000	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2103	10,597
3	0,0717	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,345	12,838
4	0,2070	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,143	13,277	14,860
5	0,4117	0,5543	0,8312	1,1455	1,6103	9,2364	11,070	12,833	15,086	16,750
6	0,6757	0,8721	1,2373	1,6354	2,2041	10,645	12,592	14,449	16,812	18,548
7	0,9893	1,2390	1,6899	2,1673	2,8331	12,017	14,067	16,013	18,475	20,278
8	1,3444	1,6465	2,1797	2,7326	3,4895	13,362	15,507	17,535	20,090	21,955
9	1,7349	2,0879	2,7004	3,3251	4,1682	14,684	16,919	19,023	21,666	23,589
10	2,1559	2,5582	3,2470	3,9403	4,8652	15,987	18,307	20,483	23,209	25,188
11	2,6032	3,0535	3,8157	4,5748	5,5778	17,275	19,675	21,920	24,725	26,757
12	3,0738	3,5706	4,4038	5,2260	6,3038	18,549	21,026	23,337	26,217	28,300
13	3,5650	4,1069	5,0088	5,8919	7,0415	19,812	22,362	24,736	27,688	29,819
14	4,0747	4,6604	5,6287	6,5706	7,7895	21,064	23,685	26,119	29,141	31,319
15	4,6009	5,2293	6,2621	7,2609	8,5468	22,307	24,996	27,488	30,578	32,801
16	5,1422	5,8122	6,9077	7,9616	9,3122	23,542	26,296	28,845	32,000	34,267
17	5,6972	6,4078	7,5642	8,6718	10,085	24,769	27,587	30,191	33,409	35,718
18	6,2648	7,0149	8,2307	9,3905	10,865	25,989	28,869	31,526	34,805	37,156
19	6,8440	7,6327	8,9065	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,4338	8,2604	9,5908	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,0337	8,8972	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,6427	9,5425	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,2604	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,8862	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,43	104,21
80	51,172	53,540	57,153	60,391	64,278	96,578	101,88	106,63	112,33	116,32
90	59,196	61,754	65,647	69,126	73,291	107,57	113,15	118,14	124,12	128,30
100	67,328	70,065	74,222	77,929	82,358	118,50	124,34	129,56	135,81	140,17

APÉNDICE

C

GLOSARIO

Al Azar o Aleatorio. Son todos aquellos eventos fortuitos o productos de la suerte.

Aleatoriamente. Actividades o métodos producidos o llevados a cabo simulando un comportamiento al azar.

Aleatoriedad. Método para controlar las variables extrínsecas que comprende la asignación aleatoria de las unidades de prueba a los grupos experimentales, con el uso de números aleatorios. Las condiciones de tratamiento se asignan al azar a los grupos experimentales.

Alfa (α). Probabilidad de cometer un error de tipo I.

Análisis de correlación lineal. Técnica para determinar el grado hasta el cual las variables están linealmente asociadas.

Análisis de regresión. Procedimiento estadístico para analizar las relaciones de relación entre una variable dependiente métrica y una o más variables independientes.

Análisis de variancia (ANOVA). Técnica estadística utilizada para probar la igualdad de tres o más medias de población y, de este modo, hacer inferencias sobre si las muestras provienen de poblaciones que tienen la misma media.

Análisis exploratorio de datos. Métodos para analizar datos que requieren de muy pocas suposiciones principales.

Arreglo de datos. Organización de los datos sin procesar por observación, tomados en orden descendente o ascendente.

Asimetría. Característica de una distribución que evalúa su simetría con respecto de la media.

Atributos. Variables cualitativas que sólo poseen categorías.

Beta (β). Probabilidad de cometer un error de tipo II.

Censo. Cuento completo de los elementos de una población u objetos de estudio.

Clase de extremo abierto. Clase que permite que el extremo superior o inferior de un esquema de clasificación cuantitativo no tenga límite.

Clase mediana. Clase de una distribución de frecuencias que contiene el valor mediano de un conjunto de datos.

Confiabilidad. Grado hasta el cual el proceso de medición está libre de errores aleatorios.

Correlación. Es una medida de la relación entre dos o más variables. La correlación puede tomar valores entre -1 y $+1$. El valor de -1 representa una correlación negativa perfecta mientras un valor de $+1$ representa una correlación perfecta positiva. Un valor de 0 representa una falta de correlación.

Curtosis. El grado de agudeza de una distribución de puntos.

Datos continuos. Datos que pueden pasar de una clase a la siguiente sin interrumpirse y que pueden expresarse mediante números enteros o fraccionarios.

Datos discretos. Datos que no pasan de una clase a la siguiente sin que haya una interrupción; esto es, en donde las categorías representan valores o cuentas distintas que pueden representarse mediante números enteros.

Datos métricos. Datos que tienen naturaleza de intervalo o razón.

Datos no métricos. Datos que se derivan de una escala nominal u ordinal.

Datos. Son los valores cualitativos o cuantitativos mediante los cuales se miden las características de los objetos, sucesos o fenómenos a estudiar.

Dependencia estadística. Condición en la que la probabilidad de presentación de un evento depende de la presentación de algún otro evento, o se ve afectada por ésta.

Desviación estándar. Raíz cuadrada positiva de la varianza; medida de dispersión con las mismas unidades que los datos originales, más bien que en las unidades al cuadrado en que está la varianza.

Diagrama de dispersión. Gráfica de puntos en una red rectangular; las coordenadas X y Y de cada punto corresponden a las dos mediciones hechas sobre un elemento particular de muestra, y el patrón de puntos ilustra la relación entre las dos variables.

Distribución asimétrica. Se presenta cuando la distribución de un conjunto de datos resulta con un promedio, una mediana y una moda con valores diferentes.

Distribución bimodal. Distribución de puntos de datos en la que dos valores se presentan con más frecuencias que los demás elementos del conjunto de datos.

Distribución binomial. Distribución que describe los resultados de un experimento conocido como proceso de Bernoulli.

Distribución de frecuencias. Distribución matemática cuyo objetivo es obtener un conteo del número de respuestas asociadas con los distintos valores de una variable y expresar estos conteos en términos de porcentajes.

Distribución de ji cuadrada. Distribución asimétrica cuya forma depende únicamente del número de grados de libertad. Conforme se incrementa el número de grados de libertad, la distribución de ji cuadrada se hace más simétrica.

Distribución de la muestra. La distribución de los valores de la estadística de una muestra (calculada para cada muestra posible), que pueda tomarse de la población meta de acuerdo con un plan de muestreo específico.

Distribución de Poisson. Distribución discreta en la que la probabilidad de presentación de un evento en un intervalo muy pequeño es un número también

pequeño, la probabilidad de que dos o más de estos eventos se presenten dentro del mismo intervalo es efectivamente igual a cero, y la probabilidad de presentación del evento dentro del periodo dado es independiente de cuándo se presenta dicho periodo.

Distribución de probabilidad. Lista de los resultados de un experimento con las probabilidades que se esperarían ver asociadas con cada resultado.

Distribución normal estándar. Distribución normal de probabilidad con media cero y una desviación estándar de 1.

Distribución normal. Distribución de una variable aleatoria continua que una curva de un solo pico y con forma de campana. La media cae en el centro de la distribución y la curva es simétrica con respecto a una línea vertical que pase por la media. Los dos extremos se extienden indefinidamente, sin tocar nunca el eje horizontal. Base para la inferencia estadística clásica que tiene forma de campana y apariencia simétrica. Todas sus medias de tendencias central son idénticas.

Distribución t de Student. Familia de distribución de probabilidad que se distinguen por sus grados de libertad individuales; es parecida, en forma, a la distribución normal; y se utiliza cuando se desconoce la desviación estándar de la población y el tamaño de la muestra es relativamente pequeño ($n \leq 30$).

Ecuación de estimación. Fórmula matemática que relaciona la variable desconocida con las variables conocidas en el análisis de regresión.

Eficiencia estadística. Comparación de los errores estándar generados por los diferentes procedimientos de muestreo.

Error aleatorio. Error que surge de diferencias o cambios aleatorios en los entrevistados o las situaciones de medición.

Error de medición. La variación en la información que el investigador y la información que genera el proceso de medición empleado.

Error de muestreo. Error o variación entre estadísticas de muestra debido al azar; es decir, diferencias entre cada muestra y la población, y entre varias muestras que se deben únicamente a los elementos que elegimos para la muestra.

Error estándar de la estimación. Medida de la confiabilidad de la ecuación de estimación, que indica la variabilidad de los puntos observados alrededor de la línea de regresión, esto es, hasta qué punto los valores observados difieren de sus valores predichos sobre la línea de regresión.

Error estándar de la media. La desviación estándar de la distribución de muestreo de la media; una medida del grado en que se espera que varíen las medias de las diferentes muestras de la media de la población, debido al error aleatorio en el proceso de muestreo.

Error muestral. Diferencia entre el estadístico observado de la muestra probabilística y el parámetro de la población.

Error sistemático. Error que afecta la medición de manera constante y representa los factores estables que afectan, de la misma forma, la clasificación observada cada vez que se realiza la medición.

Error tipo 1. Error que se presenta cuando los resultados de la muestra llevan al rechazo de la hipótesis nula que en realidad es verdadera. También se conoce como error alfa.

Error tipo 2. Error que se presenta cuando los resultados de la muestra llevan a la aceptación de la hipótesis nula que en realidad es falsa. También se conoce como error beta.

Escala de razón. Escala de medición que tiene todas las propiedades de una escala de intervalo más un punto cero absoluto.

Escala nominal. Escala cuyos números sirven sólo como etiquetas o señales para identificar y clasificar los objetos, con una correspondencia estricta de uno a uno entre los números y los objetos.

Escala ordinal. Escala de clasificación en la que los números se asignan a los objetos para indicar el grado relativo con que se posee cierta característica. De modo, que es posible determinar si un objeto tiene mayor o menor cantidad de la característica que algún otro.

Espacio muestral. Conjunto de todos los resultados posibles de un experimento.

Esperanza. La esperanza (valor esperado o media) de una variable aleatoria discreta es la suma de los productos de sus valores por sus probabilidades asociadas.

Estadística descriptiva. Rama de la estadística que proporciona a los investigadores las mediciones resumidas para los datos en las muestras.

Estadística inferencial. Rama de la estadística que permite a los investigadores hacer juicios de la población con base en los resultados generados por las muestras.

Estadística. Ciencia que trata del desarrollo y aplicación de métodos eficientes de recolección, elaboración, presentación, análisis e interpretación de datos numéricos.

Estadísticas. Mediciones que describen las características de una muestra.

Estadístico. Descripción resumida de una medida en la muestra seleccionada.

Estimación por intervalo. Estimación del parámetro de la población utilizando un conjunto de valores comprendidos dentro de un intervalo.

Estimación puntual. Estimación del parámetro de la población con valor un solo valor calculado con la información de la muestra.

Estimación. Valor específico observado de un estimador.

Estimador coherente. Estimador que produce valores que se acercan más al parámetro de la población conforme aumenta el tamaño de la muestra.

Estimador consistente. Estadístico que se aproxima al parámetro de la población a medida que aumenta el tamaño de la muestra.

Estimador eficiente. Estimador con un menor error estándar que algún otro estimador del parámetro de la población, esto es, cuando más pequeño sea el error estándar de un estimador, más eficiente será ese estimador.

Estimador insesgado. Estimador cuyo valor esperado es el parámetro o valor de la población.

Estimador suficiente. Estimador que utiliza toda la información disponible en los datos correspondientes a un parámetro.

Estimador. Estadística de muestra utilizada para estimar un parámetro de población.

Evento. Uno o más de los resultados posibles de hacer algo, o uno de los resultados posibles de realizar un experimento.

Eventos independientes. Dos eventos evento son independientes si el conocimiento de que uno ocurrirá o ya ha ocurrido no afecta la probabilidad del otro; más precisamente, si la probabilidad condicional de cada uno dada por el otro es la misma que la probabilidad incondicional.

Eventos mutuamente excluyentes. Eventos que no se pueden presentar juntos.

Fracción de muestreo. La fracción o porción de la población contenida en la muestra.

Frecuencia absoluta. Número total de elementos que aparecen en una determinada categoría.

Frecuencia relativa. Porcentaje de elementos totales que aparecen en una determinada categoría.

Grados de libertad. Número de valores de una muestra que podemos especificar libremente, después de que ya sabemos algo sobre dicha muestra.

Gráfica lineal. Presentación gráfica de magnitud en el conjunto de datos mostrado por la pendiente de una línea (o líneas) que ha sido situada con respecto a una escala horizontal o vertical.

Gráfico circular. Círculo que divide en secciones de tal manera que el tamaño de cada una de éstas corresponde a una proporción del total.

Gráfico de barras. Presentación gráfica de magnitud en el conjunto de datos, representada por la longitud de diferentes barras trazadas con referencia a una escala horizontal o vertical.

Hipótesis alternativa. Afirmación de que se espera alguna diferencia o efecto. La aceptación de la hipótesis alternativa dará lugar a cambios en las opiniones o acciones.

Hipótesis nula. Afirmación en la cual no se espera ninguna diferencia ni efecto. Si la hipótesis nula no se rechaza, no se hará ningún cambio.

Hipótesis. Enunciado o proposición no probados acerca de un factor o fenómeno de interés para el investigador. Una hipótesis estadística a un enunciado

respecto a una población y usualmente es un enunciado respecto a uno a más parámetros de la población.

Histograma. Gráfica de un conjunto de datos compuesta de una serie de rectángulos cada uno con un ancho proporcional al alcance de los valores de cada clase y altura proporcional al número de elementos que entran en la clase, o altura proporcional a la fracción de elementos de la clase.

Incertidumbre. Falta de un conocimiento completo acerca de los posibles resultados de las acciones, con desconocimiento de las probabilidades de los posibles resultados.

Independencia estadística. Condición en la que presentación de algún evento no tiene efecto sobre la probabilidad de presentación de otro evento.

Inferencia estadística. Proceso de generalizar los resultados de la muestra a los resultados de la población.

Intersección. Constante para cualquier línea recta dada cuyo valor representa el valor de la variable Y cuando la variable X tiene un valor de 0.

Intervalo de confianza. Intervalo de valores que tiene designada una probabilidad de que incluya el valor real del parámetro de la población.

Línea de regresión. Una línea ajustada a un grupo de puntos para estimar la relación entre dos variables.

Media. El promedio; valor que se obtiene al sumar todos los elementos en un conjunto y dividirlos entre el número de elementos.

Mediana. Medida de tendencia central que se da como el valor arriba del cual caen la mitad de los valores y abajo del cuál cae la otra mitad.

Medibilidad. Es una característica de los diseños que permite calcular, a partir de la propia muestra, estimaciones válidas o aproximaciones de su variabilidad de muestreo.

Medición. Asignación de números u otros signos a las características de los objetos, de acuerdo con ciertas reglas especificadas con anterioridad.

Medidas de tendencia. Estadística que describe una ubicación dentro de un conjunto de datos. Las medidas de la tendencia describen el centro de la distribución.

Método de mínimos cuadrados. Técnicas para ajustar una línea recta a través de un conjunto de puntos de tal manera que la suma de las distancias verticales cuadradas desde los n puntos a la línea se minimiza.

Moda. Medida de tendencia central que se da como el valor que ocurre con mayor frecuencia en la distribución de una muestra.

Muestra representativa. Muestra que contiene las características importantes de la población en las mismas proporciones en que están contenidas en la población.

Muestra. Colección de algunos elementos, pero no de todos, de la población bajo estudio, utilizada para describir poblaciones.

Muestras pareadas. En la prueba de hipótesis seorean las observaciones de modo que los dos conjuntos de observación se relacionan con los mismos sujetos.

Muestreo aleatorio simple. Métodos de selección de muestras que permiten a cada muestra posible una probabilidad igual de ser elegida y a cada elemento de la población una oportunidad igual de ser incluidos en la muestra.

Muestreo aleatorio. Las técnicas de muestreo aleatorio aseguran que cada elemento en la población de interés tenga una probabilidad (no nula) de ser incluido en la muestra.

Muestreo con reemplazo. Procedimiento de muestreo en el que los elementos se regresan a la población después de ser elegidos, de tal forma que algunos elementos de la población pueden aparecer en la muestra más de una vez.

Muestreo de aleatorio. Método para seleccionar una muestra de una población en el que todos los elementos de la población tienen igual oportunidad de ser elegidos en la muestra.

Muestreo no probabilístico. Técnicas de muestreo que no utilizan procedimientos de selección por casualidad, sino que más bien dependen del juicio personal del investigador.

Muestreo probabilístico. Procedimiento de muestreo en el cual cada elemento de la población tiene una oportunidad probabilística fija de ser seleccionado para la muestra.

Muestreo sistemático. Un método de muestreo aleatorio usado en estadística en el que los elementos que se muestran seleccionando de la población en un intervalo uniforme que se mide con respecto al tiempo, al orden o el espacio.

Nivel de confianza. Probabilidad que los estadísticos asocian con una estimación de intervalo de un parámetro de población. Ésta indica qué tan seguros están de que la estimación de intervalo incluirá al parámetro de la población.

Nivel de significancia. Valor que indica el porcentaje de valores de muestra que están fuera de ciertos límites, suponiendo que la hipótesis nula es correcta, es decir, se trata de la probabilidad de rechazar la hipótesis nula cuando es cierta.

Ojiva. Gráfica de una distribución de frecuencias acumuladas.

Parámetros. Valores que describen las características de una población.

Pendiente. Constante para cualquier recta dada cuyo valor representa qué tanto el cambio de unidad de la variable independiente cambia la variable dependiente.

Población finita. Población que tiene un tamaño establecido o limitado.

Población infinita. Población en el que es teóricamente imposible observar todos los elementos.

Población. Conjunto de todos los elementos que comparten un grupo común de características, y forman el universo para el propósito del problema de investigación. Esta debe definirse en términos de: el contenido, las unidades, la extensión y el tiempo.

Polígono de frecuencias. Línea que une los puntos medios de cada clase de un conjunto de datos, trazada a la altura correspondiente a la frecuencia de los datos.

Porcentaje. Cociente de un valor actual entre un valor base cuyo resultado es multiplicado por cien.

Precisión. El grado de exactitud con el que la media de la muestra puede estimar la media de la población, según revela el error estándar de la media.

Probabilidad clásica. Número de resultados favorables a la presentación de un evento dividido entre el número total de resultados posibles.

Probabilidad condicional. Probabilidad de que se presente un evento, dado que otro evento ya se ha presentado.

Probabilidad conjunta. Probabilidad de que se presentan dos o más eventos simultáneamente o en sucesión.

Probabilidad marginal. Probabilidad incondicional de que se presente un evento; probabilidad de que se presente un sólo evento.

Probabilidad subjetiva. Probabilidad basada en las creencias personales de quien hace la estimación de probabilidad.

Probabilidad. La posibilidad de que algo suceda.

Prueba t. Prueba de hipótesis univariada que utiliza la distribución t y que se utiliza cuando se desconoce la desviación estándar y el tamaño de la muestra es pequeño.

Prueba z. Prueba de hipótesis univariada que utiliza la distribución normal estándar.

Pruebas de dos muestras. Pruebas de hipótesis basadas en muestras tomadas de dos poblaciones con el fin de comparar sus medias o proporciones.

Rango. Diferencia entre los valores más bajo y más alto de una distribución.

Razón. Es una frecuencia relativa que relaciona la frecuencia absoluta de una categoría con la frecuencia absoluta de otra categoría.

Regresión múltiple. Técnica estadística que desarrolla simultáneamente una relación matemática entre dos o más variables independientes y una variable dependiente con escala de intervalo.

Relación directa. Relación entre dos variables en las que, al incrementares el valor de la variable independiente, se incrementa el valor de la variable dependiente.

Relación espuria. Relación observada entre variables cuando una existe; la relación desaparece cuando se presenta una variable.

Relación inversa. Relación entre dos variables en la que, al incrementares la variable independiente, decrece la variable dependiente.

Relación lineal. Tipo particular de asociación entre dos variables que puede describirse matemáticamente mediante una línea recta.

Sesgo. Es el error humano, intencional o no intencional que se comete al ejecutar el muestreo y que generalmente es sistemático. Este error se minimiza a través de programas de entrenamiento, capacitación y motivación de inspectores y recolectores de información estadística.

Teorema del límite central. Resultado que asegura que la distribución de muestreo de la media se acerca a la normalidad cuando el tamaño de la muestra se incrementa, sin importar la forma de la distribución de la población que se selecciona la muestra.

Valor esperado. Es el valor promedio de una variable aleatoria en muchas pruebas u observaciones.

Valor z. Número de errores estándar en que un punto se encuentra alejado de la media.

Variable aleatoria continua. Variable aleatoria que puede tomar cualquier valor dentro de un intervalo dado de valores.

Variable aleatoria discreta. Variable aleatoria que puede tomar sólo un número limitado de valores.

Variable aleatoria. Es una función real en un espacio probabilístico: hace corresponder a cada evento elemental con un número real, el valor de la variable aleatoria en ese evento elemental.

Variable aleatoria. Variable que toma diferentes valores como resultado de un experimento aleatorio.

Variable confusa. Variable externa sin control, cuyo efecto es invalidar las conclusiones de un experimento.

Variable dependiente. La variable que tratamos de predecir en el análisis de regresión.

Variabes cuantitativas. Variables con valores numéricos que resultan de medir o de contar.

Variabes independientes. Variables (s) conocida(s) en el análisis de regresión.

Variancia. Desviación cuadrada media de todos los valores de la media.

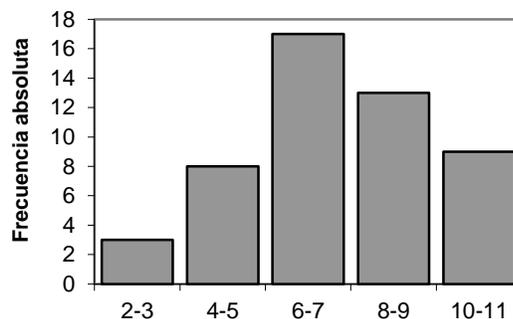
RESPUESTAS A LOS PROBLEMAS IMPARES

CAPÍTULO 2

- 2.1 a) $\bar{x} = 28.18$ $M_e = 27$
b) $s^2 = 91,56$ $s = 9,57$
c) $cv = 34\%$

- 2.3 Respuestas al punto a) y b).

Muestras	f_i	f_i/n	F_i	$F_i \uparrow$	F_i/n	$F_i/n \uparrow$
2-3	3	0,043	3	70	0,043	1,000
4-5	8	0,114	11	67	0,157	0,957
6-7	17	0,243	28	59	0,400	0,843
8-9	33	0,471	61	42	0,871	0,600
10-11	9	0,129	70	9	1,000	0,129



2.5 $\bar{x} = 155,6$ $s=14,26184$

2.7 a) $\bar{x} = 32,18182$ $M_e = 28$ b) $cv = \frac{19,07259}{32,18182} * 100 = 59,265\%$

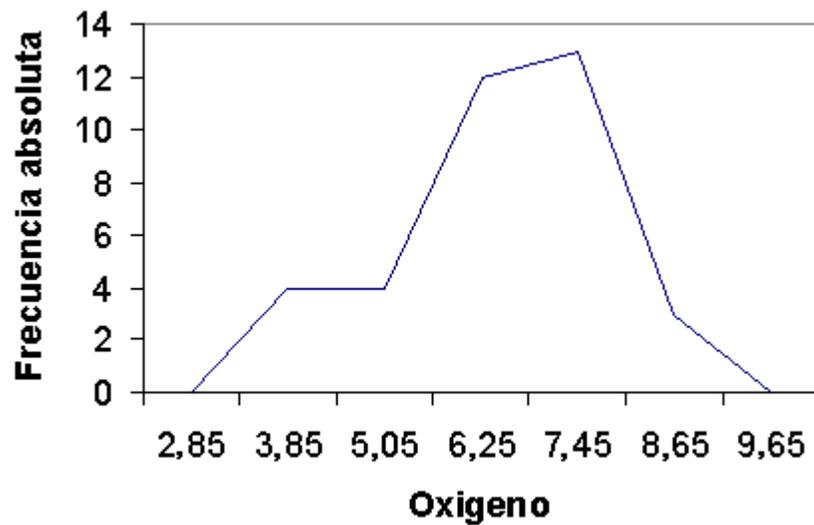
2.9 a)

Oxígeno	f_i	f_i/n
3,3-4,4	4	0,111
4,5-5,6	4	0,111
5,7-6,8	12	0,334
6,9-8,0	13	0,361
8,1-9,2	3	0,083

b) $F_3=28$

Interpretación: 28 muestras presentaron una medida de oxígeno superior a 5,6

Respuesta 2.9 c), polígono de frecuencias



2.11 $\bar{x}_A = 1,328$ $\bar{x}_B = 1,286$

Presenta mayor exactitud el Método B, el promedio aritmético de las trazas de plomo es prácticamente igual al verdadero valor 1,282 mg Pb por litro.

$s_A=0,013$

$s_B=0,0358$

El Método A tiene menor variabilidad, su desviación estándar es menor.

2.13 Respuesta al punto a) y b).

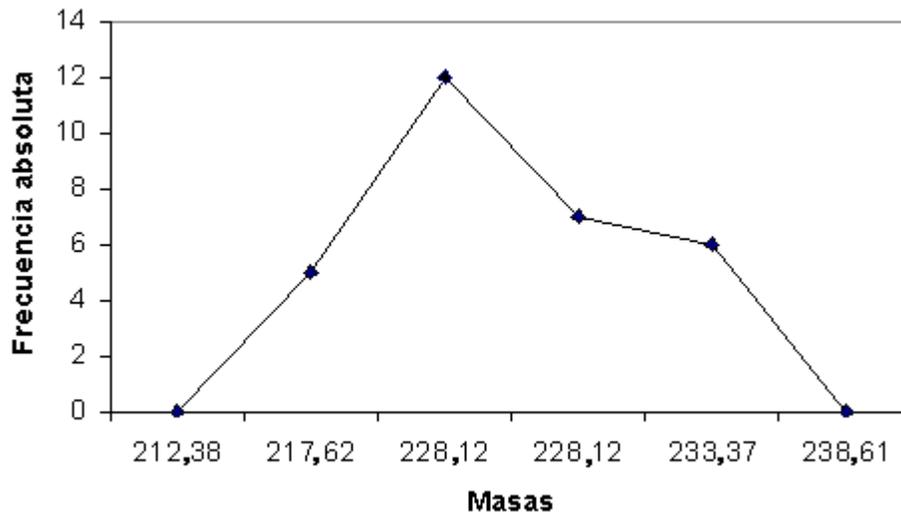
Masas	f_i	f_i/n	F_i	$F_i/n \uparrow$
215,00 – 220,24	5	0,167	5	1,000
220,25 – 225,49	12	0,400	17	0,833
225,50 – 230,74	7	0,233	24	0,433
230,75 - 235,99	6	0,200	30	0,200

Interpretación:

24 iones presentaron una masa inferior a 230,75.

Un 0,433 de los iones tienen una masa superior a 227.

Respuesta al punto c), polígono de frecuencias



2.15 a) $M_e=68$

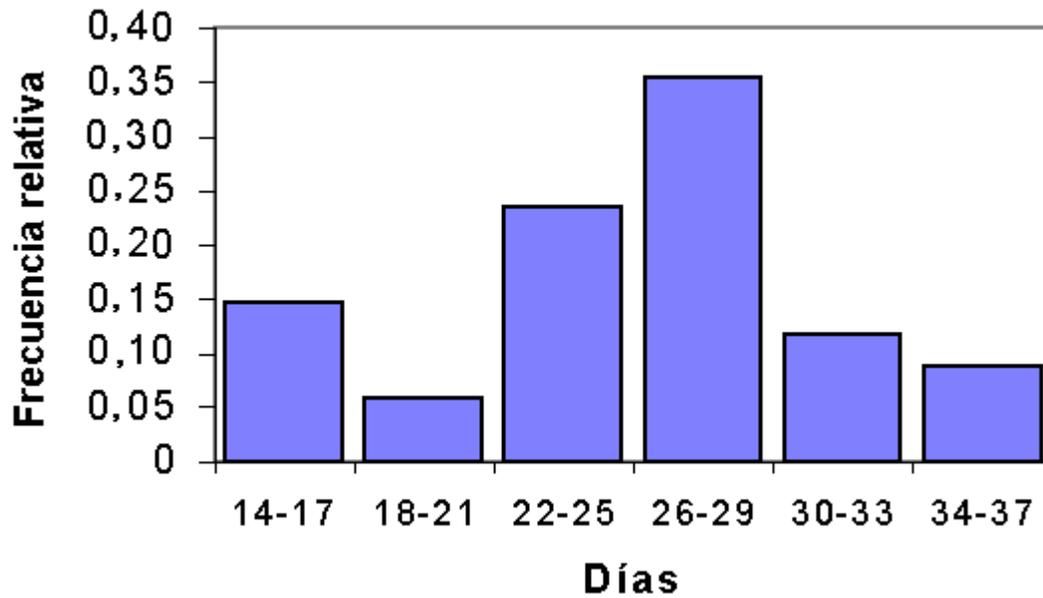
b) $cv_T = \frac{30,28}{145} * 100 = 20,88\%$

$cv_R = \frac{14,652}{67,3} * 100 = 21,77\%$

El rendimiento(B) presenta mayor variabilidad relativa que la temperatura(A).

2.17 Respuesta al punto a)

Días	f_i	f_i/n
14-17	5	0,147
18-21	2	0,059
22-25	8	0,235
26-29	12	0,353
30-33	4	0,118
34-37	3	0,088

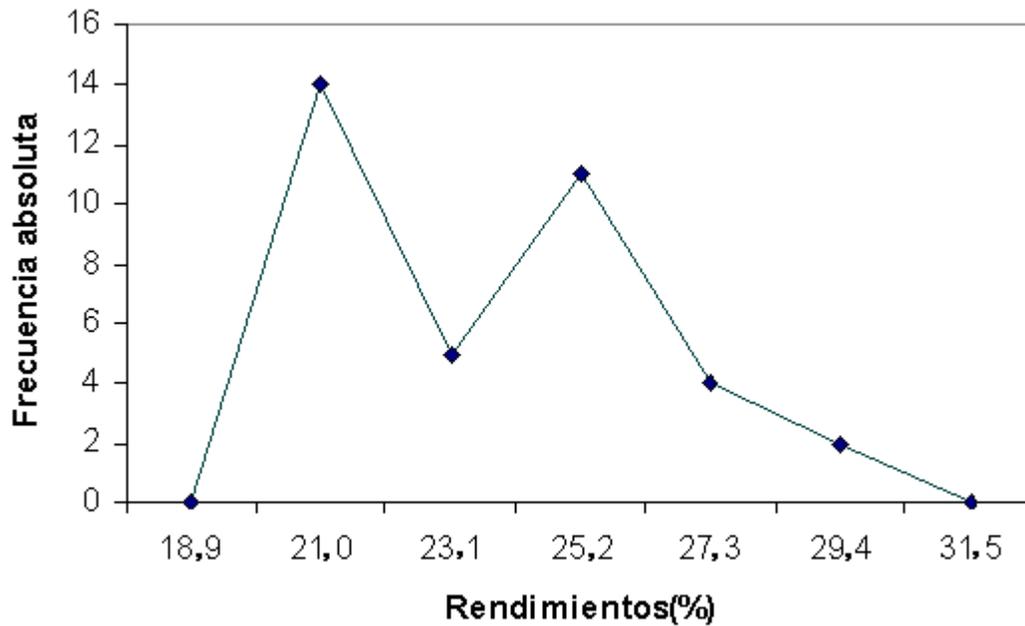


c) $M_0=29$. La mayoría de muestras de cerveza se fermentaron a los 5 días.

2.19 Respuesta al punto a).

Días	f_i	F_i/n
20,0-22,0	14	0,389
22,1-24,1	5	0,528
24,2-26,2	11	0,834
26,3-28,3	4	0,945
28,4-30,4	2	1,000

Respuesta al punto 2.19 b), polígono de frecuencias.



2.21 a) 0,436 b) 0,0395 c) 0,11

2.23 32,96%

2.25 a) 1,302 b) $R= 0,13$ $S=0,0497$ c) 3,82%

2.27 $\bar{x} = 3,3216$ $\bar{x}_g = 3,3218$ $M_e=3,31$ $R/ 3,32$

2.29 Rechazar el resultado 6,99; puede ser considerado como un valor extremo (o anormal)
 $\bar{x} = 5,897$ $S= 0,514$ $R= 1,64$

2.31 38,625

2.33 a) 0,0957 0,0299
 b) 52,4% 27,8%

CAPÍTULO 3

3.1 0,15866

3.3 a) $\frac{2}{3}$

b) $\frac{2}{3}$

c) $\frac{7}{9}$

3.5 0,0166

3.7 0,46414

3.9 a) 0,46414

b) 0,65096

c) 0,94179

3.11 a) 0,1703

b) 0,9756

3.13 0,16853

3.15 a) 0,01991

b) 0,51595

3.17 0,073

3.19 0,209

3.21 0,141

3.23 0,34223

3.25 146,04

3.27 3,0643E-04

3.29 0,7149

3.31 0.0001

CAPÍTULO 4

4.1 $\bar{x} = 0,91326$ $s=3,36155e-4$ $t=2,7764$ $d=0,00042$
 $L_1=0,91284$ $L_2=0,91368$

4.3 $n=31$

4.5 $\hat{p} = 0,03$ $z=2,57$ $d=0,0196$ $L_1=0,0104$ $L_2=0,0496$

4.7 $\bar{x} = 3,49$ $s=0,289$ $d=0,2377$ $L_1=3,25$ $L_2=3,73$
Se estima que la verdadera media del contenido de cobre se encuentra entre 3,25% y 3,73%, con un nivel de confianza del 90%.

4.9 $n=16$

4.11 $\bar{x} = 2,66$ $s=0,7933$ $d=0,815$ $L_1=1,845$ $L_2=3,475$

4.13 $\hat{p} = 0,133$ $z=1,81$ $d=0,079$ $L_1=0,054$ $L_2=0,212$

4.15 $\bar{x} = 2,258$ $s=0,581$ $d=0,369$ $L_1=1,889$ $L_2=2,627$

4.17 $\bar{x} = 0,145$ $s=0,051$ $d=0,013$ $L_1=0,132$ $L_2=0,158$

4.19 $\hat{p} = 0,63$ $z=1,96$ $d=0,106$ $L_1=0,524$ $L_2=0,736$

4.21 $n=660$

4.23 $n=216$

4.25 a) $\bar{x} = 15,556$ $s=4,275$ $t=2,306$ $d=3,286$
 $L_1=12,27$ $L_2=18,84$

b) Los niveles de contaminación por la sustancia de estudio tienen aproximadamente una distribución normal.

4.27 $n=26$

4.29 $d=0,5305$ $L_1=77,44$ $L_2=78,50$
Se requieren en promedio entre 77,44 y 78,50 mililitros para neutralizar 1 gramo de ácido, con un nivel de confianza del 90%.

4.31 $d=0,189$ $L_1=2,38$ $L_2=2,76$

4.33 $n=178$

4.35 $d=2,562$ $L_1=82,153$ $L_2=87,277$
Se estima que el verdadero promedio del porcentaje de calcio del compuesto se encuentra entre 82,153% y 87,277%, a un nivel de significancia del 10%.

4.37 $n=735$

4.39 $\bar{x} = 64,28$ $s=18,818$ $t=2,5706$ $d=19,748$ $L_1=44,5$ $L_2=84,0$
El punto promedio de vaporización de las sustancias se estima que se encuentra entre 44,5 y 84,0, con un nivel de confianza del 95%.

CAPÍTULO 5

- 5.1 $H_0: \mu_A = \mu_B$ $H_1: \mu_A \neq \mu_B$ $T_{obs} = 3,16$ $T_{teo} = \pm 2,1448$
Rechazar H_0 . Sí existe una diferencia significativa en el promedio de trazas de plomo determinado por ambos métodos, a un nivel de confianza del 95%.
- 5.3 $H_0: \mu = 73,2$ $H_1: \mu > 73,2$ $Z_{obs} = 2,73$ $Z_{teo} = 1,64$
Rechazar H_0 . Se tiene evidencia estadística, para asegurar que la media de la población en realidad es mayor a 73,2 puntos, con nivel de confianza del 95%.
- 5.5 $H_0: \mu_A = \mu_B$ $H_1: \mu_A \neq \mu_B$ $T_{obs} = 0,118$ $T_{teo} = \pm 3,1693$
No rechazar H_0 . No existe evidencia estadística, para suponer que existe una diferencia significativa entre ambos métodos de determinación del colesterol, a un nivel de confianza del 99%.
- 5.7 $H_0: \mu = 32,3$ $H_1: \mu \neq 32,3$ $Z_{obs} = -3,55$ $Z_{teo} = \pm 1,96$
Rechazar H_0 . El promedio resulta significativamente diferente a 32,3, con nivel de confianza del 95%.
- 5.9 $H_0: \mu = 800$ $H_1: \mu < 800$ $T_{obs} = -1,341$ $T_{teo} = -2,1318$
No rechazar H_0 . No hay suficiente evidencia estadística para asegurar que el rendimiento de la planta ha disminuido, a un nivel de confianza del 95%.
- 5.11 $H_0: P = 0,10$ $H_1: P \neq 0,10$ $Z_{obs} = 2,51$ $Z_{teo} = \pm 1,96$
Rechazar H_0 . La proporción de artículos defectuosos es significativamente diferente al 10%, a un nivel de confianza del 95%.
- 5.13 $H_0: \mu = 5$ $H_1: \mu < 5$ $Z_{obs} = -3,44$ $Z_{teo} = -1,96$
Rechazar H_0 . Si hay suficiente evidencia estadística que indique que el contenido de oxígeno disuelto es menor que 5 ppm, a un nivel de confianza del 97,5%.
- 5.15 $H_0: P_1 = P_2$ $H_1: P_1 > P_2$ $Z_{obs} = 0,807$ $Z_{teo} = 1,64$
No rechazar H_0 . No existe suficiente evidencia estadística para suponer que el primer método de impregnar nubes es mejor que el segundo, con un nivel de confianza del 95%.
- 5.17 $H_0: P_1 = P_2$ $H_1: P_1 \neq P_2$ $Z_{obs} = -0,618$ $Z_{teo} = \pm 2,57$

No rechazar H_0 . Los resultados no proporcionan suficiente evidencia para concluir que hay relación entre el uso de la aspirina y la frecuencia de la formación de coágulos postoperatorios, a un nivel de confianza del 99%.

5.19 $H_0:\mu=16$ $H_1:\mu<16$ $Z_{obs}=-12,07$ $Z_{teo}=-2,06$
Rechazar H_0 . Existe evidencia estadística para concluir que el peso neto es menor al indicado en el envase del producto, con un nivel de confianza del 98%.

5.21 $H_0:P=0,90$ $H_1:P<0,90$ $Z_{obs}=-5,185$ $Z_{teo}=-1,96$
Rechazar H_0 . Se puede concluir que la afirmación del fabricante resulta falsa, a un nivel de confianza del 97,5%. La proporción de manchas removidas, 0,68, resulta significativamente inferior al 90%, dada por el fabricante.

5.23 $H_0:\mu=55$ $H_1:\mu<55$ $T_{obs}=-0,702$ $T_{teo}=-1,3664$
No rechazar H_0 . Existe evidencia estadística para concluir que la dureza media del agua en la salida de la planta, no es inferior a 55 ppm, a un nivel de confianza del 90%.

5.25 $H_0:\mu=2$ $H_1:\mu\neq 2$ $T_{obs}=-1,72$ $T_{teo}=\pm 1,3178$
Rechazar H_0 . Se tiene evidencia estadística para concluir con un nivel de confianza del 90%, que el peso promedio de las cajas es diferente a las especificaciones de producción.

5.27 $H_0:\mu=5$ $H_1:\mu>5$ $Z_{obs}=15,57$ $Z_{teo}=1,28$
Rechazar H_0 . Se puede concluir que el límite tolerable de estroncio 90, no se cumple para esta región del país, con un nivel de confianza del 90%.

5.29 $H_0:\mu_1=\mu_2$ $H_1:\mu_1\neq\mu_2$ $T_{obs}=0,2275$ $T_{teo}=\pm 2,9768$
No rechazar H_0 . No existe evidencia estadística para concluir que la determinación del contenido del calcio de la sangre por ambos métodos resulta significativamente diferente, a un nivel de confianza del 99%.

5.31 $H_0:\mu=0,067$ $H_1:\mu<0,067$ $T_{obs}=0,1273$ $T_{teo}=-1,7959$
No rechazar H_0 . Se tiene evidencia estadística para concluir que la absorción promedio es 0,067, con un nivel de confianza del 95%.

5.33 $H_0:P=0,05$ $H_1:P>0,05$ $Z_{obs}=3,08$ $Z_{teo}=2,25$
Rechazar H_0 . Existe suficiente evidencia estadística para concluir, que lo afirmado por la compañía es falso, por lo cual la proporción de tortas de pollo

que se presentan algún defecto es superior al 5%, con un nivel de confianza del 98%.

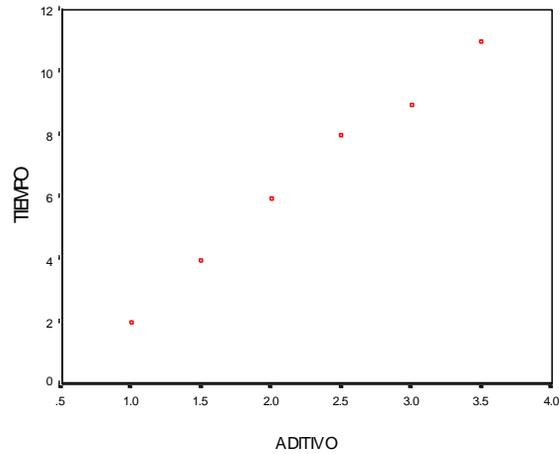
- 5.35 $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ $Z_{obs} = 3,48$ $Z_{teo} = \pm 1,645$
Rechazar H_0 . Sí existe una diferencia significativa en el promedio de resistencia a la tensión de los dos tipos de acero, a un nivel de confianza del 90%.
- 5.37 $H_0: P_1 = P_2$ $H_1: P_1 \neq P_2$ $Z_{obs} = -1,02$ $Z_{teo} = \pm 2,57$
No rechazar H_0 . No existe una diferencia significativa en la proporción de defectuosos producidos por los procesos de fabricación, a un nivel de confianza del 99%.
- 5.39 Del análisis de variancia de una vía, razón de Fisher resulta 11,567 y la probabilidad observada 5,366E-11. Por tanto, se puede concluir que existe una diferencia significativa en el contenido de nitrógeno de un lote a otro, con un nivel de significancia del 5%.
- 5.41 $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ $t_{obs} = -9,138$ $t_{teo} = \pm 2,3060$
Rechazar H_0 . Si hay una diferencia significativa en las determinaciones promedios de los analistas, a un nivel de confianza del 95%.
- 5.43 $H_0: \mu_A = \mu_B$ $H_1: \mu_A \neq \mu_B$ $t_{obs} = 2,69$ $t_{teo} = \pm 1,7459$
Rechazar H_0 . Existe una diferencia significativa de las determinaciones de cobre de los analistas, a un nivel de confianza del 95%.
- 5.45 $H_0: P_1 = P_2$ $H_1: P_1 \neq P_2$ $Z_{obs} = -0,39$ $Z_{teo} = \pm 1,96$
No rechazar H_0 . Se concluye que no existe una diferencia significativa entre las proporciones de las poblaciones, a un nivel de confianza del 95%.
- 5.47 $H_0: P_1 = P_2$ $H_1: P_1 \neq P_2$ $Z_{obs} = 1,315$ $Z_{teo} = \pm 2,57$
No rechazar H_0 . Se concluye que no existe una diferencia significativa entre las proporciones de las poblaciones, a un nivel de confianza del 95%.

CAPÍTULO 6

6.1 a) Diagrama de dispersión

b) 0,9782

c) 11,9334



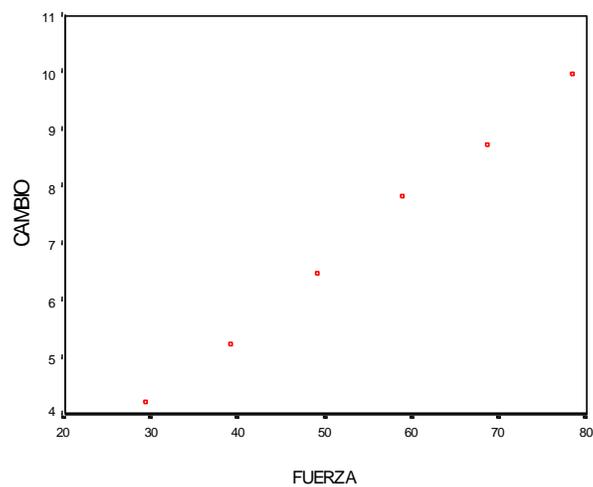
6.3 a) $a=-0,950$ $b=0,269$

b) Se espera que al incrementarse en 1% la humedad de almacenamiento, el contenido de humedad de la fibra sintética aumenta en 0,269%

c) $L_1=9,232$ $L_2=10,388$

d) $L_1=0,185$ $L_2=0,353$

6.5 a) Diagrama de dispersión



b) $a=0,720$ $b=0,118$

c) $L_1=0,111$ $L_2=0,126$

d) $H_0:\alpha=0$ $H_1:\alpha\neq 0$ $t_{\text{obs}}=4,548$ $t_{\text{teo}}=4,604$

No rechazar H_0 . Se comprueba la validez de la ley de Hooke, al no rechazarse la hipótesis que la constante es igual a cero, es decir, la recta debe pasar por el origen, con un nivel de significación del 99%.

e) $L_1=7,08$ $L_2=7,34$

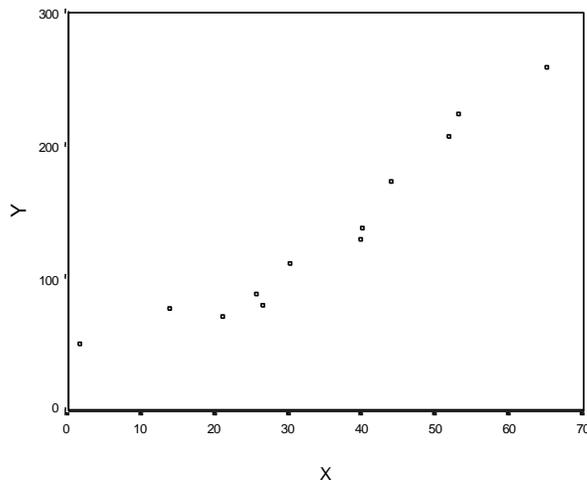
6.7 a) $r^2=0,9803$

b) $L_1=1,347$ $L_2=1,586$

c) $H_0:\beta=0$ $H_1:\beta\neq 0$ $t_{\text{obs}}=-17,290$ $t_{\text{teo}}=\pm 2,4469$

Rechazar H_0 . Existe evidencia estadística para concluir que existe algún grado de relación lineal entre el tiempo y los residuos de cloro, con un nivel de confianza del 99

6.9 a) Diagrama de dispersión



b) $a=12,033$ $b=3,567$

c) $r=0,956$

- 6.11 a) $a=76,7917$ $b=0,4028$
 b) $a=168,436$ $b=2,384$ c) $\hat{y} = 311$
 c) $r^2=0,502$
- 6.13 a) $a=27,7929$ $b=6,5973$
 b) $r=0,9538$. Existe una correlación positiva de 0,9538 entre el porcentaje de materia orgánica oxidada y el tiempo en días de exposición.
 c) 80,6%
 d) $S_e= 0,7857$
- 6.15 a) El porcentaje de impurezas del compuesto se aumenta 0,5% al modificarse la temperatura en un grado a la que se realiza la mezcla de sus componentes.
 b) Que existe un porcentaje de impurezas constante 2,2%.
 c) 37,2
- 6.17 a)) $r^2=0,122$
 b) $a=-5,22$ $b=2,496$
 c) Al aumentar el número atómico en una unidad la masa se incrementa en 2,496.
 d) $H_0:\beta=0$ $H_1:\beta\neq 0$ $t_{obs}=25,325$ $t_{teo}=\pm 3,1824$
 Rechazar H_0 . Existe evidencia estadística para concluir que existe algún grado de relación lineal entre las variables de estudio, a un nivel de confianza del 95%.
- 6.19 a) $a=4,209$ $b=0,928$
 b) $r=0,1782$
 c) El nivel de tensión constante a un nivel de viscosidad de cero.
 d) 5,137

6.21 a) $a=0,00714$ $b=0,0369$

b) Al aumentar la novacaína en un mg el nitrito de sódico se incrementa en 0,0369 ml

c) 0,4261

6.23 $y=a+bx$ $a=0,117$ $b=0,02975$ $r=0,9715$ $r^2=0,9438$
 $\log Y=a+bx$ $a=-1,841$ $b=0,099$ $r=0,988$ $r^2=0,977$

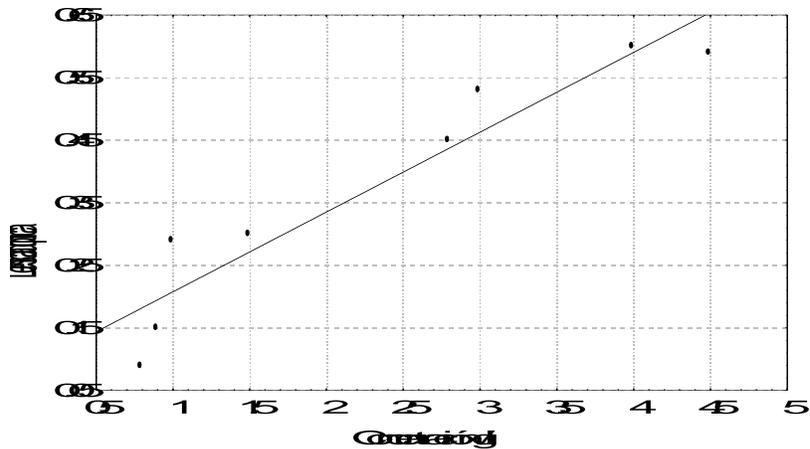
6.25 $a=0,054$ $b=0,0002$ $y=0,0596$

6.27 $r=-0,933$

6.29 8,358

6.31 b) $a=0,229$ $b=1,135$
c) 0,54

6.33 Diagrama de dispersión



BIBLIOGRAFÍA

- ALEGRE MARTÍN, J. *Aplicaciones Económicas de Estadística Descriptiva*. Universitat de les Illes Balears, Palma de Mallorca, 1999.
- ARCE CONSTANTINO REAL EULOGIO. *Introducción al Análisis Estadístico con SPSS para Windows*. Editorial Manantial, 2001.
- BAIRD D. C. *Experimentación*. Editorial Prentice-Hall, 1991.
- BERENSON Mark L. y David M. Levine. *Estadística para la Administración y Economía*, México: Mc Graw Hill, 1991.
- BERENSON, M. L. y LEVIN, D.M. *Estadística para Administración y Economía*. Editorial Prentice Hall, segunda edición, 2001.
- BRUNK, H.D., *Introducción a la Matemática Estadística*. Edit. Trillas 5a. Edición, 1991.
- COCHRAN. *Técnicas de Muestreo*. Editorial C.E.C.S.A., 1995.
- DEL PINO, M. GUIDO. *Estadística Teoría y Métodos*. Editorial Manantial, 1995.
- FERNÁNDEZ. M. FRANCISCO. *Introducción a la Estadística y sus Aplicaciones*. Editorial Piramide. 2001.
- FONTANET G., REIRA J., SALILLAS J. *Hacer Estadística Breda*. Editorial Addison-Wesley Iberoa, 1998.
- GÓMEZ BARRANTES, MIGUEL. *Elementos de Estadística Descriptiva*. Editorial UNED, tercera edición, 1998.
- HANKE, E. JOHN y REITSCH, G. ARTHUR. *Pronósticos en los Negocios*. Editorial Hispanoamericana, S. A., 1996.
- HINES, W-MONTGOMERY, D.C., *Probabilidad y Estadística para la Ingeniería y Administración*. 2ª ed. CECSA, México, 1995.
- KAZMIER L. Y A. Díaz Mata. *Estadística aplicada a la administración y a la economía*, México: Mc Graw Hill, 1993.
- KISH, L. *Diseño estadístico para la investigación*, Centro de Investigaciones Sociológicas, Madrid, 1995.
- KUEHL ROBERT O. *Diseño de Experimentos*. Editorial Thomson International, 2001.
- KURINCIC GABRIELA. *Estadística Herramientas de Inferencia*. Editorial Ediciones Cooperativas, 2001.

- LIPSCHUTZ SEYMOUR. *Probabilidad Teoría y Problemas*. Editorial Schaum, 2001.
- LOHR SHARON L. *Muestreo Diseño y Análisis*. Editorial Thomson International, 2000.
- LUQUE MARTÍNEZ, T. *Técnicas de análisis de datos en investigación de mercados*. Ed. Pirámide, 2000.
- MALHOTRA, N.K. *Investigación de Mercados. Un Enfoque Práctico*. Prentice Hall, 1997.
- MENDENHALL WILLIAM. *Estadística Matemática con Aplicaciones*. Editorial Iberoamericana, 1994.
- MENDENHALL, W.; WACKERLY, D.; SCHEAFFER, R. *Estadística Matemática con Aplicaciones*, 2ª. ed., Iberoamericana, 1994.
- MILLER, I.-FREUND, J.-JOHNSON, R., *Probabilidad y Estadística para Ingenieros*. 3ª ed. Prentice Hall, México, 1992.
- MOREU JALON PEDRO. *Estadística Informatizada*. Editorial Paraninfo, 1999.
- NEWBOLD, P. *Estadística para los Negocios y la Economía*. Prentice Hall, Madrid, 1996.
- NEWBOLD, P. *Statistics for Business and Economics*, third edition, Prentice Hall, 1994.
- PEÑA, D. *Estadística, Modelos y métodos, Fundamentos*, 2ª. ed. revisada, Alianza, 1995.
- PEÑA, D. *Estadística. Modelos y Métodos. 2. Modelos lineales y series temporales*. Alianza Universidad, Madrid, 1992.
- PEÑA, D. y ROMO, J. *Introducción a la Estadística para las Ciencias sociales*. MacGraw Hill, Madrid., 1997.
- PEREZ LÓPEZ CESAR. *Técnicas de Muestreo Estadístico*. Editorial RA-MA, 1999.
- PÉREZ SUAREZ, R. *Análisis de datos económicos I. Métodos descriptivos*. Pirámide, Madrid, 1993.
- PÉREZ, C. *Estadística práctica con Statgraphics*. Prentice Hall, Madrid, 2001.
- PILAR PESTAÑA DE MARTÍNEZ. *ESTADÍSTICA: Conceptos básicos, terminología y metodología de la Estadística Descriptiva*. Colección Minerva, 2001.
- SMIRNOV N. DUNIN-BARKOVSKI. *Cálculo de Probabilidades y Estadística Matemática*, Editorial Paraninfo, 1998.
- SPIEGEL MURRAY R. *Estadística*. Editorial Schaum, 1991.

VISAUTA, B. *Análisis estadísticos con SPSS para Windows*. McGraw Hill, Madrid, 1997.

VISAUTA, B. *Análisis Estadístico con SPSS para Windows. Estadística Multivariante*, McGraw Hill, 1998.

WALPOLE, R.-MYERS, R., *Probabilidad y Estadística*. 3ª ed. McGraw-Hill, México, 1992.

WEBSTER, ALLEN. *Estadística Aplicada a los Negocios y la Economía*. Editorial McGraw-Hill, tercera edición, 2000

ZAR, J. H. *Biostatistical Analysis*. 3ª ed. Prentice Hall Inc, 1996.

ZIKMUND, WG. *Investigación de Mercados. Sexta Edición*. Prentice Hall, 1998.