Apuntes para el curso de estadística descriptiva con una breve introducción al diseño de experimentos

Jimmy Rodríguez Bolaños
jrodriguezb@utn.ac.cr
Área de Matemática y Estadística
Universidad Técnica Nacional, Sede Atenas - Costa Rica
Agosto, 2017.
Versión 0.2

Índice de contenidos

Índice de figuras

Capítulo 0

Prefacio

El presente documento busca ser un material de *ayuda didáctica o de consulta* para profesores y de *apoyo* para estudiantes del curso de *Estadística Descriptiva* que se ofrece en la Universidad Técnica Nacional. Su fin es exclusivamente didáctico, académico o formativo, sin ningún afán de lucro de parte del autor.

Sigue los contenidos propuestos en el programa ME005. En algunos casos se ofrece una *ampliación* en algunos conceptos que considero son de mayor atención, como por ejemplo el concepto sobre *error estándar de la media* y su relación con el *intervalo de confianza*.

Los capítulos del 1 al 5 se basaron en varias publicaciones, a saber, el libro de *Elementos de estadística descriptiva* de Miguel Gómez Barrantes (UNED, tercera edición, 2011), en los apuntes de *Introducción a la estadística descriptiva* de Javier Trejos Zelaya y Ericka Moya Vargas (Universidad Latina de Costa Rica. Editorial Sello Latino, San Jose. 2004) y en el libro de *Estadística y muestreo* de Ciro Martínez Bencardino (Ecoe, Bogotá, Colombia, décima tercera edición 2012).

El capítulo 6 sobre *probabilidad* se basó mayormente en el libro de James Victor Uspensky *Introduction to Mathematical Probability* (Mc Graw- Hill, 1937) y en el libro de *Probabilidad y estadística para ingeniería y ciencias*, novena edición de Ronald Walpole, Raymond Myres y Sharon L. Myers (Pearson, 2012).

El capítulo 7 sobre *regresión y correlación lineal* se basó en gran parte en el libro de Ciro Martínez Bencardino.

Para el capítulo 8, el cual trata de ser una breve *introducción* al diseño experimental, se advierte que no forma parte del actual programa de estadística descriptiva, pero se ofrece como apoyo para aquellos estudiantes que actualmente estén desarrollando sus anteproyectos de graduación. Este capítulo, el cual se encuentra todavía en *construcción*, se basa inicialmente en el libro *Análisis y diseño de experimentos* de Humberto Gutiérrez Pulido y Román de la Vara Salazar (Mc Graw- Hill, segunda edición, 2008) y será ampliado en una futura versión.

A los autores anteriores se debe en gran medida este material, a ellos su debido reconocimiento y gratitud.

Algunos de los datos usados en este documento para realizar los ejemplos fueron tomados de fuentes primarias de sitios web de instituciones públicas . A ellos su debido reconocimiento según sea el caso.

Este material se incluye en el *repositorio institucional* de la UTN y busca el apoyo de profesores y estudiantes (quienes así lo deseen), para ser ampliado y mejorado, ya que de momento, es solo una versión preliminar y en definitiva, requiere mejoras.

Se utilizó (y se recomienda) el programa de software libre *R Commander* para la creación de las diferentes gráficas y la realización de cálculos necesarios para hacer el análisis estadístico de los diferentes ejemplos a lo largo de este documento. Se espera en una versión futura incluir un anexo sobre el uso de este excelente software de análisis estadístico

J.A Rodríguez.
Atenas, Agosto 2017.

Capítulo 1

Principios de la Investigación Estadística: Sobre el concepto y elementos de la Estadística

1.1. Definición y aplicabilidad de la Estadística en la investigación

1.1.1. Estadística Descriptiva

Para efectos del curso de estadística descriptiva, se definirá *Estadística* como aquella ciencia que trata sobre la recolección, procesamiento, análisis e interpretación de datos obtenidos de una población para propósitos de investigación. La palabra *estadística* tiene probablemente su origen en el término alemán *stara*, que significa *estado*.

Si con los datos recolectados de una muestra de una población se pretende **describir** algún *factor o factores* que afecten a los individuos seleccionados, sin llegar a hacer *conclusiones o generalizaciones* sobre la población estudiada, se recurre al uso de *técnicas* de estadística descriptiva para su análisis. Para este tipo de casos se suelen usar estimaciones y técnicas tales como: obtención de medidas de posición central, por ejemplo la *media o promedio*, el análisis frecuencial de datos, cálculos de medidas de dispersión como la *desviación estándar*, el uso de *gráficos*, análisis correlacional, etc., con los cuales se pretende hacer una **descripción estimativa de los factores** que afectan a los individuos que componen esa muestra.

La estadística descriptiva usualmente se utiliza en proyectos de investigación donde se requiere **ahondar** sobre ciertos *factores* de interés para el investigador sobre la población que desea estudiar. A este tipo de investigaciones se le da el nombre de *investigación transversal* y se caracteriza por hacerse en periodos de tiempo relativamente cortos, con el objetivo de **profundizar** sobre ciertos *factores* que pueden estar afectando a la población de estudio, caracterizando a la muestra mediante el uso de *estadísticos* como el *promedio*, la *variabilidad*, etc.

Ejemplo 1.1

Mencione dos temas de investigación de tipo *transversal* que hagan uso de estadística descriptiva para su análisis. Recuerde que este tipo de investigación busca *evidenciar* los *factores* que sean *característicos* a la muestra o grupo de estudio, como lo son el promedio, la variabilidad de los datos, etc.

Solución

Un tema podría ser la *"Caracterización socioeconómica de la población estudiantil de primer ingreso de la U.T.N., en la Sede de Atenas para el año 2017 según el ingreso económico familiar"*.

Otro tema podría ser el *"Uso de la enzima X como catalizador alternativo en la fermentación de yogurt elaborado en la planta de la empresa XYZ en marzo-junio de 2017"*.

1.1.2. Estadística Inferencial

Para temas de investigación donde se desea hacer *generalizaciones* a partir de los datos obtenidos para **explicar las causas** de algún fenómeno, *validar las conclusiones* obtenidas a partir de los resultados y que en general buscan obtener *conclusiones* sobre los *fenómenos o comportamientos* que afectan a la población de estudio, se recurre principalmente a la aplicación de *Estadística Inferencial*. Para hacer las inferencias o generalizaciones en Estadística Inferencial se requiere del uso de cálculos probabilísticos, los cuales posteriormente permiten tomar decisiones sobre hipótesis que se hacen sobre los datos.

Se suele utilizar en investigaciones donde se desea *abarcar* lo más posible a la población. A este tipo de investigaciones se le conoce como *investigación longitudinal* y se caracteriza por hacerse en *periodos largos de tiempo*, en poblaciones de gran tamaño y buscan **explicar las causas** sobre *comportamientos o fenómenos* que puede tener la población que se desea investigar.

Ejemplo 1.2

Mencione dos temas de investigación de tipo *longitudinal* que hagan uso de estadística inferencial para su análisis. Recuerde que este tipo de investigación busca *explicar* los *fenómenos o comportamientos* observados en una muestra y busca generalizarlos a toda la población.

Solución

Como ejemplos de este tipo de investigación en el cual se requiere del uso de *Estadística Inferencial*, pueden mencionarse aquellas investigaciones en las que se desarrollan tratamientos de quimioterapias para la cura de ciertos tipos de cáncer como la leucemia, o bien, una investigación que pretenda indagar si la *heredabilidad* es causa significativa en la mortalidad de lechones con menos de 5 días de nacidos, tomando como base a las camadas de las principales razas de cría de cerdos utilizados en granjas de Centroamérica.

1.1.3. Algunas técnicas usadas en estadística inferencial

Algunas de las *técnicas* usadas en Estadística Inferencial para el análisis *causal* en una investigación pueden ser: el *análisis de componentes principales*, por sus siglas *ACP*, la cual es una técnicaque permite *reducir* el número de *varia-*

bles de una investigación, perdiendo un mínimo de información, y suele usarse en investigaciones *multidimensionales*.

Otra *técnica* muy común utilizada en Estadistica Inferencial es la *prueba de hipótesis*, que consiste en la formulación de dos hipótesis llamadas *nula* y *alternativa*, en la cual por medio de cálculos probabilísticos se *acepta o rechaza* algunas de esas hipótesis.

1.1.4. Resumen sobre los tipos de investigaciones en estadística descriptiva y diferencial

En resumen: las investigaciones de carácter *descriptivo* suelen utilizar *técnicas* de estadística descriptiva para tratar de *correlacionar* ciertos *factores* de interés para el investigador y suelen ser de tipo *transversal*, de bajo costo y donde usualmente se busca *profundidad* en los *factores* que intervienen en el tema a investigar. En cambio, en las investigaciones *causales*, es decir, donde se desean buscar las *causas* que afectan a la población de estudio, se caracterizan por ser investigaciones de tipo *longitudinal* , donde se hace uso de *técnicas* de Estadística Inferencial para **explicar las causas** del tema a investigar. Las investigaciones *causales* requieren de más *tiempo y recursos* que una investigación *descriptiva*.

1.2. Elementos básicos de la Estadística: unidad estadística, población, variable estadística y muestreo.

1.2.1. Definición de unidad estadística

Toda investigación posee un *objeto o sujeto* el cual se desea investigar. Al *objeto o sujeto* de estudio en Estadística se le conoce como *unidad estadística* y debe estar definida en *tiempo y lugar*.

Por ejemplo, si un tema de investigación es: *"Caracterización socioeconómica de la población estudiantil de primer ingreso de la U.T.N. en la Sede de Atenas para el año 2017 según el ingreso económico familiar"*, la *unidad estadística* del tema anterior sería *cada estudiante* de primer ingreso de la U.T.N., en la Sede de Atenas durante el año 2017.

Como otro ejemplo, si el tema de investigación es: *"Uso de la enzima X como catalizador alternativo en la fermentación de yogurt elaborado en la planta de la empresa XYZ en marzo-junio de 2017"*, la *unidad estadística* del tema anterior sería *cada unidad de yogurt* elaborada en la planta de la empresa XYZ en marzo-junio de 2017.

Ejercicio para la clase

Determine la unidad estadística para un tema de investigación titulado *"Uso del fungicida orgánico X como alternativa para el control de la roya (hemileia vastatrix) en fincas de café de la zona de los Santos durante el año 2017"*.

1.2.2. Definición de población

Al conjunto de **todas las unidades estadísticas** que intervienen en el tema a investigar se le conoce como *población*.

Una población puede ser *finita o infinita*. Una población es finita si tiene un número *limitado* de unidades estadísticas, mientras que una infinita tiene un número *ilimitado* de unidades estadísticas.

Por ejemplo, si se considera la población de primer ingreso a la U.T.N., durante el primer cuatrimestre de 2017, su población será *finita*, por estar definida en un *tiempo y lugar específicos*. En cambio, si se considera la población de equinos a los que se podría implementar y evaluar una técnica que permita obtener suero de caballos que han sido inmunizados con veneno de serpientes venenosas, está población sería infinita, ya que dicho experimento podría *repetirse indefinidamente* bajo las mismas condiciones.

Si un tema de investigación se titula *"Caracterización socioeconómica de la población estudiantil de primer ingreso de la U.T.N. en la Sede de Atenas para el año 2017 según el ingreso económico familiar"*, la *población* constaría de **todos** los estudiantes de primer ingreso de la U.T.N. en la Sede de Atenas durante el 2017 y sería una población finita.

Si un tema de investigación se titula " *Implementación y evaluación de la técnica de separación del suero equino de la sangre de caballos inmunizados con veneno de serpientes venenosas*", la población constaría de todos los equinos a los que se les podría implementar dicha técnica de separación, siendo esta población *infinita*.

1.2.3. Paramétros poblacionales y estimadores muestrales

Los *parámetros poblacionales* o simplemente *parámetros* se definen como aquellas *medidas* que describen numéricamente las *características* de una población, como lo son el promedio poblacional μ , la varianza poblacional σ^2, la desviación estándar poblacional σ, etc.

Los *estimadores puntuales* o solamente *estimadores* son las descripciones numéricas de las características estudiadas a las unidades estadísticas en una muestra, a saber, el promedio muestral $\overline{x}$, la varianza muestral s^2, la desviación estándar muestral s, etc.

El *error* que existe entre un parámetro y su estimador recibe el nombre de *error muestral* E_m.

El error muestral E_m debe ser fijado por el investigador. Este error muestral E_m permite hallar el *tamaño muestral* n necesario para poder relacionar las características de la investigación con los objetivos planteados en el estudio.

1.2.4. Definición de variable estadística

Una *variable estadística* se define como aquella *característica que se observa sobre las unidades estadísticas*. Cada variable estadística asigna a cada unidad estadística ya sea *un valor*, o bien, *un atributo*. Si a la variable se le asigna un *valor*, este debe de ir acompañado por las *unidades* para ese valor.

Si la variable es *temperatura*, *las unidades* para esa variable sería en *grados centígrados* $\left({}^0C\right)$. Las unidades a usar deben ser las del S.I.

1.2.5. Clasificación de las variables estadísticas

Las *variables estadísticas* o simplemente *variables* se clasifican en: *cuantitativas* y *cualitativas*.
Las *variables cuantitativas* se clasifican a su vez en *continuas y discretas*.

Una *variable cuantitativa continua* es aquella que admite valores en $\mathbb{R}$. En cambio, las *variables cuantitativas discretas* admiten valores en $\mathbb{N} \cup \{0\}$.

Están además las *variables cualitativas* que se definen según las *cualidades* o *atributos* a observar en las unidades estadísticas. Estas a su vez se clasifican en *nominales, ordinales y binarias*.

Las *variables cualitativas nominales* tienen como atributo la cualidad de *dar nombre* al dato observado en las unidades estadísticas, las *variables cualitativas ordinales* tienen como atributo la cualidad de *dar orden* al dato observado y las *variables cualitativas binarias* tienen como atributo la cualidad de *presencia o ausencia* del rasgo que interesa observar en las unidades estadísticas.

Ejemplos para la clase.

Clasifique las siguientes variables en variables cuantitativas, ya sea continua o discreta o bien, en variables cualitativas, ya sea nominal, ordinal o binaria.

Valor o atributo a observar.

1. Temperatura de la cobertura de un helado en un túnel de congelación.
2. Peso de un lechón al nacer.
3. Color de la carne molida a los 3 días de elaborada.
4. Posición del Club Sport Herediano en la tabla de posiciones del Torneo de Verano 2017 en la fecha 22.
5. Cantidad de hectáreas sembradas de teka en Abangares, Guanacaste.
6. Número de teléfono de un cliente bancario.
7. Sexo de un bovino.
8. Número de trabajadores independientes que cotizan a la CCSS en la zona de San Carlos.
9. Número de cédula de un estudiante becado de la U.T.N.
10. Nota obtenida al final del curso de estadística.
11. Lugar de nacimiento de una persona.
12. Posición obtenida por un competidor en la maratón de Boston.
13. Sabor de un helado hecho a base de nueces, macadamia y avena.
14. Estado civil de una persona.
15. Éxito o fracaso en la inseminación de una vaca por parte de un estudiante de asistencia veterinaria.

1.2.6. Distribuciones unidimensionales, bidimensionales y multivariantes en la investigación estadística

Las *características* a observar en una investigación se entenderán como los *rasgos, cualidades o propiedades* que se hagan sobre las *unidades estadísticas*.

Si se va a analizar o a describir *una* característica de manera independiente, se hará referencia a esa característica como una *distribución unidimensional* .

Si se busca relacionar *dos* características entre sí, se hará referencia a la relación entre esas dos características como una *distribución bidimensional*.

En caso de que se busque relacionar más de dos características, se dirá que la distribución será *multivariante* .

1.2.7. Definición de muestreo y tipos muestreos

Se define *muestreo* como la *técnica* utilizada para tomar una *muestra* de la *población*. El *tamaño muestral* se define como el *número de unidades estadísticas* extraídas de una población. El *muestreo* se utiliza principalmente cuando la población a estudiar es *infinita o muy grande*.

Entre los tipos de *muestreos* que más se utilizan está el *muestreo aleatorio*. El *muestreo aleatorio* es una de las técnicas más usadas para *muestrear*, ya que plantea la gran ventaja de que evita *sesgos* en la población de estudio, es decir, establecer inferencias que serían válidas solamente para una *población específica* en lugar de toda la *población en general*. El *muestreo aleatorio* requiere que la población presente *homogeneidad* en su composición, para evitar los *sesgos* como se mencionó anteriormente.

Se tienen los siguiente tipos de *muestreos aleatorios*: *muestreo aleatorio simple al azar*, *muestreo aleatorio estratificado*, *muestreo aleatorio sistemático* y el *muestreo aleatorio por conglomerado*.

En el *muestreo simple al azar* **todas** las unidades estadísticas tienen la misma probabilidad de ser seleccionadas. Se utiliza en poblaciones que sean homogéneas.

Si una población resulta muy *heterogénea* suele usarse el *muestreo aleatorio estratificado*, es decir , se divide a la población en *grupos o estratos* que resulten *homogéneos* entre sí, para luego seleccionar una *muestra aleatoria simple al azar* **dentro** de cada grupo o estrato. Esto se realiza con el fin de que la *variabilidad* en la muestra refleje diferencias **detectables** *entre los grupos*.

En el *muestreo aleatorio sistemático*, si se sabe que la población a estudiar es de tamaño N y se va a tomar una muestra de tamaño n, se debe definir primero lo que en esta técnica de muestreo se llama el *espaciamiento* k.
El espaciamiento se define como $k = \frac{N}{n}$. Después para lograr que toda unidad en la población tenga *igual probabilidad de seleccionarse*, se debe escoger un número al azar entre 1 y k, de esta forma *cada elemento de la población* tendrá una probabilidad igual a $\frac{1}{k}$ de ser seleccionada. Después se selecciona un número al azar entre 1 y k llamado a. Por último, la muestra estará constituida por los elementos que lleven los números de identificación $a, a+k, a+2k, a+3k, \ldots, a+(n-1)k$.

A manera de ejemplo para explicar el *muestreo aleatorio sistemático*, si la población es de $N = 4000$ y se va a tomar una muestra de $n = 400$, el espaciamiento será $k = \frac{4000}{400} = 10$. Luego se toma un número al azar entre 1 y 10, por ejemplo el 6. Por lo tanto la muestra estará constituida por los elementos identificados con los índices 6, 16, 26, 36, 46, 56, 66, 76, 86 ,. . . , 3996.

En el *Muestreo por Conglomerado* se selecciona al azar *grupos o conglomerados de elementos de la población*, para luego tomar *todos los elementos de cada conglomerado* para construir la muestra global.

En el *muestreo por conglomerado*, la *unidad de muestreo* contiene un conjunto de unidades de estudio, las cuales son observadas en su totalidad para aquellos conglomerados seleccionados al azar.

Tarea moral # 1

Traer ejemplos donde se utilice cada uno de los muestreos vistos en clase.

1.3. Fuentes y técnicas de recolección de información en Estadística

El primer paso después de conocer el *problema* que se desea investigar, es establecer las fuentes de información a las que se recurrirá para la recolección de los datos.

Existen dos tipos de *fuentes de información*: *fuentes primarias* y *fuentes secundarias*.

Las *fuentes primarias* son aquellas fuentes que *publican o suministran* datos recolectados *por ellos mismos*.
Un ejemplo de *fuente primaria* es el *Instituto Nacional de Estadística y Censos (INEC)*, el cual publica en línea datos obtenidos en los diferentes censos y encuestas que realizan a nivel nacional. Pueden visitar el siguiente *link* http://www.inec.go.cr/encuestas/encuesta-nacional-de-hogares-productores y ver los datos recolectados y resumidos en archivos de *Excel* con información sobre la última *Encuesta Nacional de Hogares Productores*.

Las *fuentes secundarias* son fuentes que *usan datos* que han sido recolectados y publicados por otras fuentes. Algunos ejemplos de fuentes secundarias son: *revistas indexadas*, *tesis* , *libros* y *este documento que estas usando*.

1.4. Principales técnicas de recolección de información

Para la recolección de información sobre el tema de investigación a tratar, se utilizan principalmente las siguientes técnicas de recolección de información: *la entrevista*, *el cuestionario*, *la observación* y el *registro o bitácora*.

En la entrevista, el entrevistador realiza un cuestionario con antelación para hacerle preguntas al o a los entrevistados. Puede ser administrado ya sea de manera personal o bien por *videochat*. Tiene la ventaja que se tiene mayor control sobre la muestra, pero puede tener la desventaja que el *entrevistado* mienta a la hora de responder a las preguntas y *sesgar* por lo tanto los resultados.

El *cuestionario* suele consistir en una serie de preguntas que se realizan por escrito hacia los *sujetos* de estudio. Estos *sujetos* responden al cuestionario de manera escrita y usualmente anónima. Nuevamente la información que se recolecte por medio de esta técnica dependerá de la *sinceridad* con la que los sujetos respondan a las preguntas. Posee la ventaja que puede administrarse a una gran cantidad de sujetos de estudio.

La *observación* consiste en *observar* a las unidades estadísticas de interés y realizar anotaciones sobre lo observado, que usualmente consiste en mediciones o conteos. Tiene la ventaja de la objetividad pero podría tener la desventaja de cometer *errores* a la hora de realizar las observaciones.

Por último, el *registro o bitácora* consiste en hacer observaciones sobre las unidades estadísticas, llevando un registro escrito de los hechos que son de interés y que por lo general son de *carácter obligatorio*. Presenta la ventaja de que la información registrada es información *real y objetiva*, además de ser una técnica de recolección de información de bajo costo. Tiene la desventaja que la información puede estar incompleta y desactualizada de no llevarse este *registro* al día.

1.5. Ejemplos de técnicas de recolección de información

Todos los ejemplos presentados en esta sección fueron tomados de *Introducción a la Estadística* de Javier Trejos Zelaya y Ericka Moya Vargas. Universidad Latina de Costa Rica. 2004.

Ejemplo 1.3
Si se quiere hacer una encuesta tal que la persona deba pensar durante un tiempo la respuesta, o deba consultar alguna documentación (facturas, recibos, etc.), será preferible utilizar un cuestionario.

Ejemplo 1.4

Si se quiere hacer una encuesta sobre personas asalariadas y se dispone de una lista de los lugares de trabajo de interés, es preferible hacer una entrevista en el lugar de trabajo que en el domicilio, ya que será más fácil localizar al informante.

Ejemplo 1.5

Para obtener información sobre hábitos de compra, publicidad por medio de afiches o presentación en vitrinas o estantes, por ejemplo, es preferible la entrevista personal (por ejemplo, en la calle), abordando al entrevistado cuando recién ha tenido contacto con el tema de la encuesta.

Ejemplo 1.6

Para calcular el porcentaje de hogares que escucharon un mensaje publicitario por radio o televisión, es recomendable utilizar la entrevista telefónica.

1.6. Fases de una investigación estadística típica

Cuando se desea dar *respuesta* a algún *problema de investigación* se suelen recurrir a las siguientes fases para definir y completar una investigación estadística:

1. Planteamiento del problema a investigar. Es aquí donde inicia la investigación. Es donde el *investigador* se plantea a sí mismo *un problema* al cual desea dar solución por medio *del método científico y técnicas estadísticas*.. Es en esta parte donde se define cuál será el *objeto o sujeto* a observar..

2. Diseño del *instrumento* de recolección de información para la investigación. Este estará basado según la o las *técnicas de información* elegidas, ya sea una *encuesta*, una *entrevista*, etc.

3. Aplicación del *instrumento* diseñado para la recolección de los *datos* .

4. Preparación de los datos e información recolectada. Es en este punto donde se *clasifica y codifica* los datos y la información recolectada según sea la pertinencia que tengan con el *tema y objetivos* a tratar en el *problema de investigación*.

5. Análisis e interpretación de los datos. Es en esta parte donde se aplican las distintas *técnicas estadísticas* ya sean técnicas de estadística descriptiva o estadística inferencial sobre los datos obtenidos, los cuales han sido previamente *clasificados y codificados* para su análisis e interpretación para la obtención de *resultados*.

6. Como último paso se *divulgan los resultados* obtenidos de la investigación.

Entre las principales finalidades de una *investigación estadística* se pueden mencionar:

1. Determinar cuáles son los valores *típicos* que intervienen en los procesos o fenómenos a investigar.
2. Determinar *cambios* en el proceso o fenómeno a investigar, principalmente los cambios que presentados a través del tiempo.
3. Comparar *dos o más* procesos o fenómenos, estimando si existe *correlación* entre *dos o más variables* en los datos observados de esos procesos o fenómenos.

1.7. Ejercicios

1. Considere el siguiente tema de investigación: *Cantidad de estudiantes de primer ingreso matriculados en la Universidad Técnica Nacional durante el primer ciclo de 2017, según la provincia de procedencia.*

Responda:

a. ¿Cuál es la unidad estadística del tema a investigar?

b. Liste dos variables (*una cualitativa y otra cuantitativa*) que podrían ser usadas en la investigación.

c. ¿Cuál es la población estadística del estudio?

d. ¿La población es finita o infinita?

2. Clasifique las siguientes variables en variables cuantitativas, ya sea continua o discreta o bien, en variables cualitativas, ya sea nominal, ordinal o binaria.

Valor o atributo a observar.
a. Peso de los becerros al nacer en la finca *El Ospino* en marzo de 2017.

b. Estado civil de los profesores del curso de estadística en enero de 2017.

c. Ingresos familiar de los estudiantes matriculados en el curso de estadística descriptiva durante el I ciclo de 2017.

d. Razas de ganado criados en fincas de la zona de Guanacaste y San Carlos durante el año 2017.

e. Número de cédula de clientes nuevos del banco XYZ durante mayo 2017.

f. Cantidad de fanegas entregadas a CoopeAtenas en la cosecha 2017-2018.

g. Sexo de los lechones nacidos en la finca *El Guayabal* en el mes de junio de 2017.

h. Número de integrantes del núcleo familiar de estudiantes de primer ingreso matriculados en la Sede de Atenas en el I ciclo de 2017.

i Presencia o ausencia de la bacteria *campylobacter* en muestras de salchichón de pollo elaborado en la planta XYZ después de una desinfección con cloro.

j. Número de muertes en carretera durante el 2017 en accidentes de tránsito en Costa Rica.

3. Mencione una característica típica de las fuentes secundarias de información.

4. Dé un ejemplo donde se utilice el registro como instrumento de recolección de datos.

5. Dé el nombre de dos fuentes primarias y dos fuentes secundarias de información que usted podría utilizar para realizar una investigación sobre la elaboración de un fungicida orgánico para el tratamiento y control de la roya, en fincas cafetaleras de la zona de Grecia durante el año 2017.

6. Menciones dos técnicas de recolección de información que podrían ser usadas en el siguiente tema de investigación: *Elaboración de un fungicida orgánico para el tratamiento y control de la roya en fincas cafetaleras de la zona de Grecia durante el año 2017.*

7. Mencione dos técnicas estadísticas que se emplean en *estadística descriptiva* y en *estadística inferencial.*

8. Menciones el tipo de muestreo a utilizar en las siguientes temas de investigación:

a. *Presencia o ausencia de la bacteria campylobacter en muestras de salchichón de pollo elaborado en la planta XYZ después de un desinfección con cloro durante el mes de octubre de 2017.*

b. *Uso de la enzima X como catalizador alternativo en la fermentación de yogurt elaborado en la planta de la empresa XYZ en marzo-junio de 2017.*

c. *Caracterización socioeconómica de la población estudiantil de primer ingreso de la U.T.N. para el año 2017 según el ingreso económico familiar del estudiante.*

d. *Elaboración de un fungicida orgánico para el tratamiento y control de la roya en fincas cafetaleras de la zona de Grecia durante el año 2017.*

Capítulo 2

Formas de presentación de los resultados

.
Después de haber sido recolectada la información por medio del instrumento o instrumentos aplicados según la técnica seleccionada, se continúa con el procesamiento de la información.

Después de haber sido procesada la información, se debe escoger *cómo serán presentados* los resultados . La manera en *cómo* se presenten estos resultados dependerá de la naturaleza de las variables y de la información que se desprenda de los datos recogidos.

2.1. Presentación textual

La *presentación textual* suele utilizarse cuando la información con la que se cuenta es poca, permitiendo de esta manera resaltar los resultados más importantes de la investigación. Consiste en presentar los datos o cifras dentro de un texto, siendo esta forma de presentar los resultados muy utilizado en revistas y periódicos. Siempre debe indicarse la fuente de donde se tomaron los datos o la información con la que se elaboraron los resultados.

Ejemplo 1

En comparación con 2016, la economía costarricense experimentó en 2017 una reducción en la tasa de crecimiento, pues alcanzó apenas el 2.5 %, mientras que el promedio anual entre 2005 y 2016 habría sido del 4.9 %.
Fuente: *Estado de la Nación 2017*.

2.2. Presentación semitabular

La *presentación semitabular* consiste en introducir los resultados obtenidos mediante un breve texto y a continuación mostrar un resumen de estos resultados por medio de una *tabla*. Debe incluirse la fuente de donde se tomaron los datos o la información con la que se elaboraron los resultados. Se usa cuando se desea mostrar pocos resultados.

Ejemplo 2

El precio en ventanilla del dólar durante el mes de abril de 2017 presentó una variación importante con respecto a los meses anteriores en los diferentes bancos estatales y privados, como se puede apreciar a continuación en la siguiente tabla:

Nombre del banco	Precio promedio del dólar en ventanilla antes de abril de 2017	Precio promedio del dólar en ventanilla en abril de 2017
BCR	565	585
BNCR	564	590
Popular	567	595
Bancrédito	560	598
Banca Privada	567	597

Fuente: Página web del Banco de Costa Rica abril 2017.

2.3. Presentación tabular

La presentación tabular es aquella presentación que muestra ya sea una *tabla* o un *cuadro*. La tabla o cuadro debe ir acompañado siempre de los *elementos obligatorios* que debe llevar toda tabla o cuadro.

Una tabla se define como *un arreglo de datos cuantitativos interrelacionados y distribuidos en filas y columnas junto con los elementos obligatorios que debe llevar toda tabla.*

Una tabla o un cuadro presenta la ventaja de que resume la información sin necesidad de ninguna explicación textual.

2.3.1. Elementos obligatorios que debe tener toda tabla

Una *tabla* debe contener los siguientes elementos de manera *obligatoria*: *una columna matriz, encabezados y contenido.*

1. *Columna matriz*: Es la primera columna ubicada a la izquierda de la tabla y contiene *la variable principal* que se usó para clasificar los datos.

2. *Encabezados:* Es la parte de la tabla en la que están situados el resto de las columnas, que describen en forma general la o las *clasificaciones secundarias* que se hicieron para los datos .

3. *Contenido:* Es la parte de la tabla que *contiene las cifras* de los datos que se recolectaron. Se debe indicar la fuente de donde se tomaron los datos o la información usada.

4. Su forma debe respetar la norma establecida por *APA sexta edición.*

A continuación se muestra un ejemplo de una tabla.

Ejemplo 2.1

A continuación se muestra una tabla con los datos recolectados por el MAG en diferentes fincas de la zona de San Carlos, sobre el número de nacimientos de lechones según su sexo, durante el mes de agosto de 2017.

Nombre de la finca	Número de lechones nacidos	
	Machos	Hembras
El Ospino	100	104
El Jicaral	85	97
Nandayure	166	201
La Ceiba	54	66

Fuente: Oficina de información del MAG agosto 2017.

La columna matriz sería *Nombre de la finca*, el encabezado sería *Número de lechones nacidos* el cual a su vez se divide en los encabezados *Machos* y *Hembras*, la fuente sería *Oficina de información del MAG agosto 2017* y cumple con los dispuesto en su forma según APA sexta edición.

2.3.2. Definición y componentes de un cuadro

Un cuadro se define como una tabla junto con los siguientes elementos de carácter obligatorio: número de cuadro, título y fuente. Opcionalmente puede tener una nota introductoria y una nota al pie de página.

1. Número de cuadro: Se usa siempre que haya más de un cuadro dentro del documento donde el mismo se presenta; este número es importante para identificarlo o ubicarlo en una publicación.

2.Título: Es una breve explicación de la naturaleza, clasificación y referencia en el tiempo de los datos presentados, cuándo y dónde se recolectaron, cómo y bajo qué criterios se clasificaron. De ir siempre en letras mayúsculas.

3. Nota introductoria: Es una frase, generalmente entre paréntesis o guiones, colocada debajo del título. Explica o provee información relacionada con el cuadro; por ejemplo, se puede utilizar para: indicar las unidades, dar más claridad al titulo, prevenir sobre limitaciones y establecer base para hacer comparaciones.

4. Nota al pie de página: Es una frase que explica o aclara cierta cifra o clasificación, su función es más específica que la de la nota introductoria. Para indicar la nota al pie se utilizan llamadas de atención (números o símbolos como / ó *).

5. Fuente: específica de dónde se tomaron los datos o la información recolectada.

Ejemplo 2.2

Con base en el ejemplo 2.1 elabore un cuadro .

Solución

CUADRO # 1

CUADRO SOBRE EL NÚMERO DE NACIMIENTOS DE LECHONES SEGÚN SU SEXO DURANTE EL MES DE AGOSTO DE 2017 EN FINCAS DE LA ZONA DE SAN CARLOS.

Nombre de la finca	Número de lechones nacidos	
	Machos	Hembras
El Ospino	100	104
El Jicaral	85	97
Nandayure	166	201
La Ceiba	54	66

Fuente: Oficina de información del MAG agosto 2017.

Ejemplo 2.3

A continuación se muestra un cuadro con información acerca del nivel de pobreza en los hogares costarricenses durante 2014 y 2015, según la encuesta nacional de hogares del INEC.

CUADRO # 2

CUADRO SOBRE EL NIVEL DE POBREZA EN LOS HOGARES COSTARRINCENSES SEGÚN LA ENCUESTA NACIONAL DE HOGARES DEL INEC PARA LOS AÑOS 2014 Y 2015.

Nivel de pobreza	2014		2015	
	Absoluto	Relativo	Absoluto	Relativo
Total de hogares	627866	100	656445	100
Pobreza extrema	39095	6.2	45146	6.9
Necesidades básicas no satisfechas	88831	14.1	96331	14.7
Hogares no pobres	499940	79.6	514968	78.4

Fuente: Encuesta nacional de hogares del INEC para los años 2014 y 2015.

Ejercicio para la clase

Indique todas las componentes en el cuadro #2 y luego interprete los resultados.

2.4. Presentación gráfica

Otra forma de presentar los datos de una investigación estadística es por medio de la representación gráfica.

Se define una gráfica como *la representación visual de datos estadísticos mediante el uso de líneas, superficies o volúmenes.*

Se define *gráfico* como *aquella gráfica* que además contiene los siguientes elementos obligatorios: *número de gráfico, título y fuente.* Opcionalmente puede tener una nota introductoria, una nota al pie de página, leyenda, escala y título para los ejes.

El número de gráfico, título, fuente, nota introductoria y pie de página se definen de forma similar que para los elementos obligatorios de un cuadro.

La *leyenda* se utiliza cuando existen varias series de datos en la misma gráfica y suelen usarse símbolos o colores para cada serie representada en la gráfica.

La *escala* identifica la unidad de medida correspondiente a ambos ejes, por ejemplo, *1 cm = 1000 reses.*

Los *títulos de los ejes* se utilizan para *nombrar* a cada uno de los ejes.

En la figura (2.1) puede observarse los componentes y la estructura general de un gráfico.

Figura 2.1: Componentes de un gráfico

2.5. Tipos de gráficos más utilizados en estadística descriptiva: gráficos de barras, circulares y lineales

2.5.1. Gráficos de barras

Los gráficos de barras se utilizan para mostrar cómo está *distribuida la frecuencia absoluta o relativa de los datos*. Las variables para este tipo de presentación deben ser *cuantitativas discretas* o *cualitativas*.

Si la variable a representar es de tipo *cuantitativa discreta* o *temporales* se utilizan *barras verticales*. Si los datos son *cualitativos* se utilizan *barras horizontales*.

Las gráficas de barras tanto verticales u horizontales pueden ser de tipo: *simple, compuesto o comparativos*.

Se recomienda el uso de software para la creación de las gráficas.

2.5.2. Gráfico de barras simple horizontal

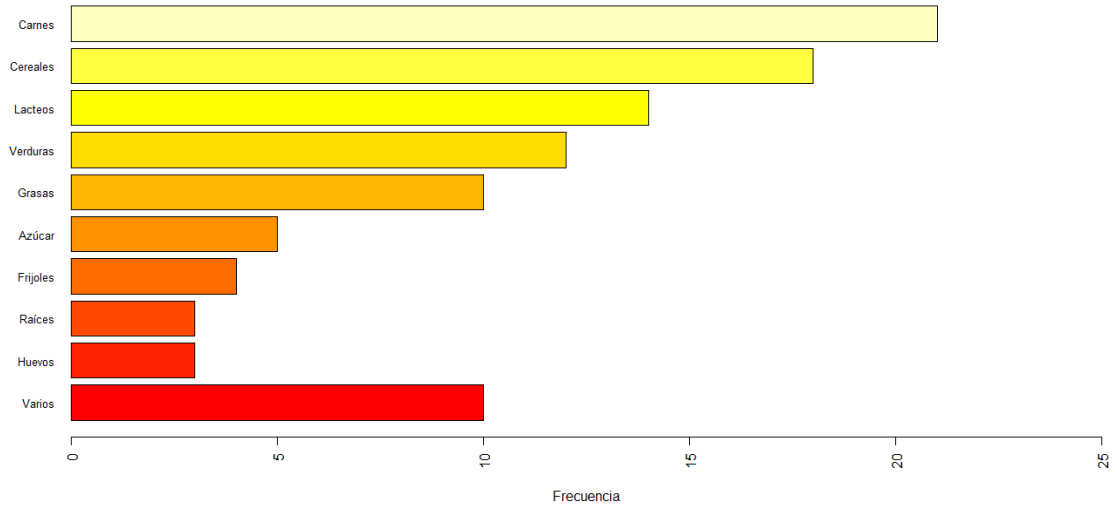
El gráfico de *barras simple horizontal* se usa cuando la variable es cualitativa y existe *solo un componente por categoría* el cual permite comparar las magnitudes para cada una de las categorías. Las barras deben estar siempre ordenadas de mayor a menor frecuencia, excepto para la categoría *varios* u *otros*, la cual siempre irá al final sin importar su magnitud.

Ejemplo 2.4

El gráfico #1 muestra la estructura de costos relativos para las diferentes categorías de la canasta básica de alimentos en abril de 2017, según el Ministerio de economía, Industria y Comercio (MEIC).

GRÁFICO # 1

ESTRUCTURA DE LOS COSTOS RELATIVOS PARA LAS DIFERENTES CATEGORÍAS DE LA CANASTA BÁSICA DE ALIMENTOS EN ABRIL DE 2017.



Fuente: Ministerio de Economía, Industria y Comercio (MEIC) abril 2017.

Figura 2.2: Costos para las categorías de la canasta básica de alimentos en abril de 2017, según el Ministerio de Economía, Industria y Comercio (MEIC)

Observe que la variable representada en el gráfico # 1 es de tipo *cualitativo* y por lo tanto deben usarse barras horizontales.

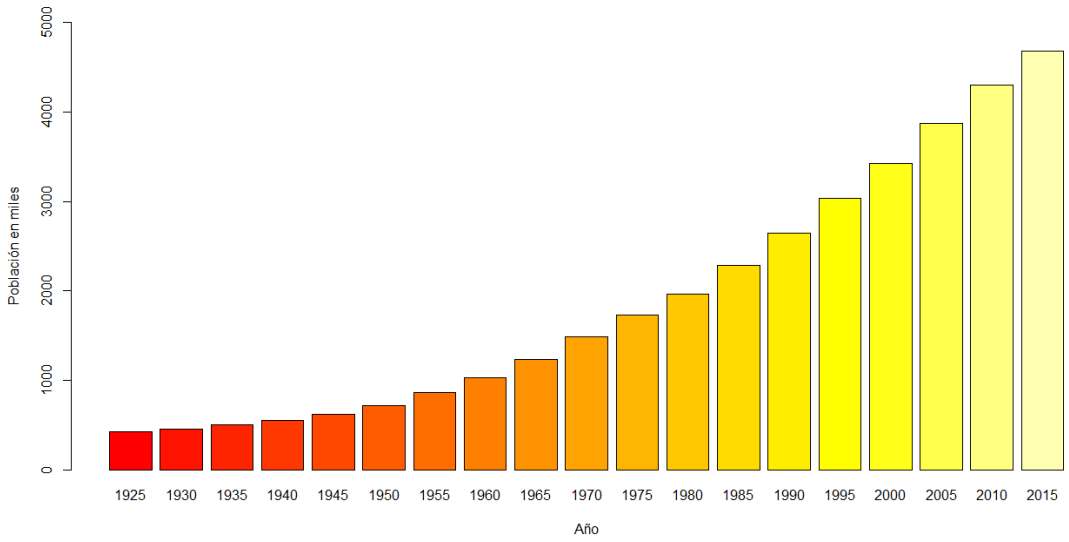
Una de las desventajas de utilizar gráficos para presentar la información es que los datos representados *se aproximan a ojo* , no puede calcularse los valores exactos a simple vista.

Ejemplo 2.5

El gráfico #2 muestra la evolución de la población en Costa Rica durante el periodo de 1925 a 2015, según datos del Instituto Nacional de Estadística y Censos (INEC).

GRÁFICO # 2

EVOLUCIÓN DE LA POBLACIÓN EN COSTA RICA DURANTE EL PERIODO DE 1925 A 2015 SEGÚN DATOS DEL INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS (INEC).
(Población en miles de habitantes a mediados de año.)



Fuente: Instituto Nacional de Estadística y Censos (INEC) 1925-2015.

Figura 2.3: Evolución de la población en Costa Rica durante el periodo de 1925 a 2015 según datos del Instituto Nacional de Estadística y Censos (INEC)

Observe que la variable representada en el gráfico #2 es de tipo *temporal* y la tendencia que muestra es de crecimiento en el tiempo.

Ejercicio para la clase

Con base en los resultados mostrados en los cuadros para los ejemplos 2.4 y 2.5, obtenga al menos dos posibles interpretaciones que se pueden desprender de dichos cuadros.

2.5.3. Gráfico de barras compuesto

El gráfico de *barras compuesto* y el *gráfico de barras comparativo* se utiliza cuando existe una *subcategorización* para la variable principal, donde se desea hacer una *comparación* o ver la *composición* que existe entre cada subcategoría .

Ejemplo 2.6

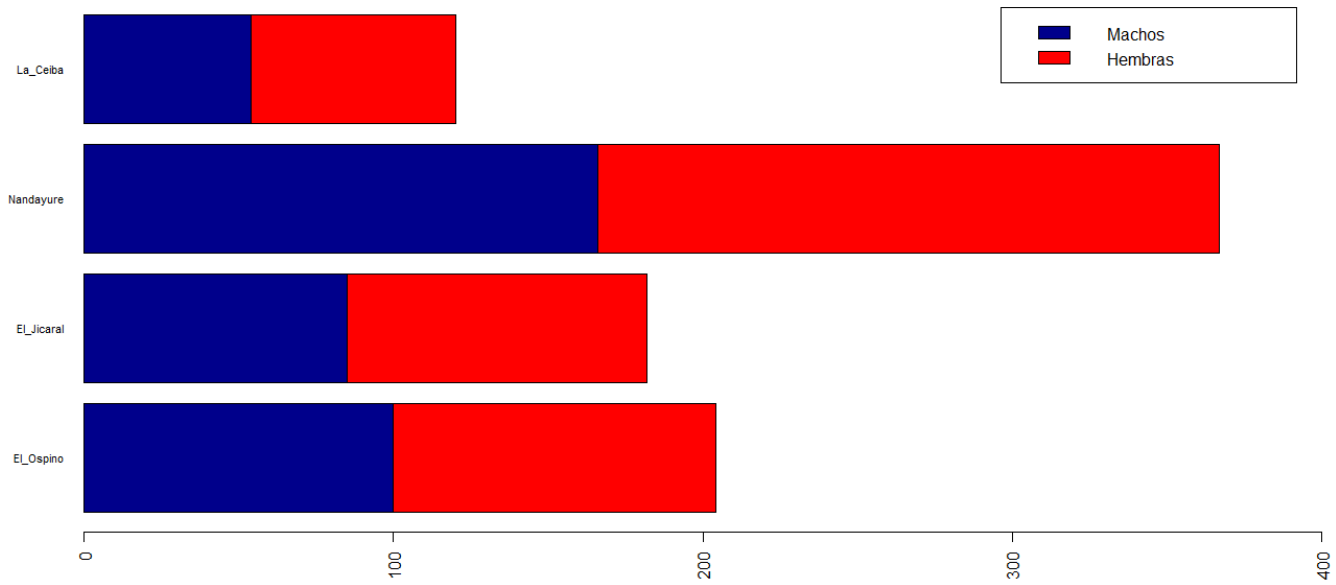
Haga una representación gráfica adecuada con base a los resultados mostrados en el cuadro #2.

Solución

Una buena elección para representar los resultados que se muestran en el cuadro #2 es la *gráfica de barras compuesta*, ya que nos permite ver la *composición* de la distribución de los datos según el sexo del lechón para cada finca, como se muestra en el gráfico #3.

GRÁFICO # 3

GRÁFICO PARA EL NÚMERO DE NACIMIENTOS DE LECHONES SEGÚN SU COMPOSICIÓN POR SEXO DURANTE EL MES DE AGOSTO DE 2017 EN FINCAS DE LA ZONA DE SAN CARLOS



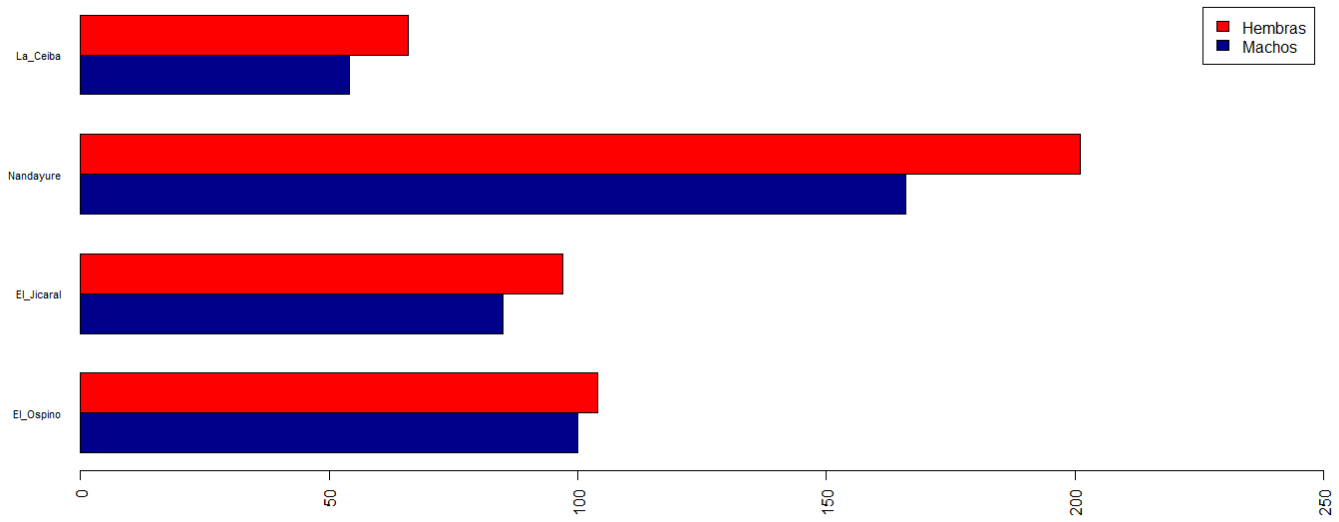
Fuente: Fuente: Oficina de información del MAG agosto 2017.

Figura 2.4: Número de nacimientos de lechones según su composición por sexo durante el mes de agosto de 2017 en fincas de la zona de San Carlos

Si en lugar de ver cómo está compuesta la población de lechones en cada finca según su sexo se desea hacer una *comparación* de la distribución de los lechones nacidos en cada finca según su sexo , se realiza entonces una *gráfica de barras comparativa*, como se muestra en el gráfico #4.

GRÁFICO # 4

GRÁFICO PARA EL NÚMERO DE NACIMIENTOS DE LECHONES MEDIANTE COMPARACIÓN POR SEXO DURANTE EL MES DE AGOSTO DE 2017 EN FINCAS DE LA ZONA DE SAN CARLOS



Fuente: Fuente: Oficina de información del MAG agosto 2017.

Figura 2.5: Número de nacimientos de lechones según su composición por sexo durante el mes de agosto de 2017 en fincas de la zona de San Carlos(INEC)

Ejercicio para la clase Interprete los resultados mostrados en los cuadros 3 y 4.

2.6. Gráfico circular

Se utiliza únicamente para variables de tipo *cualitativa nominal* o *cualitativa ordinal* y se usa para representar cómo está distribuida la frecuencia relativa para cada una de las categorías que en las que se clasificó la variable principal. Su orden debe aparecer de mayor a menor frecuencia en el sentido de las agujas del reloj si es una variable cualitativa nominal y si es ordinal hay que respetar el orden que se usó para la categorización de la variable.

Ejemplo 2.7

Con base en el nivel de pobreza para el año 2015 mostrado en el cuadro #2, realice un gráfico circular.

Solución

Note que la variable es de tipo cualitativa nominal.

GRÁFICO # 5

GRÁFICO PARA EL NIVEL DE POBREZA EN HOGARES COSTARRICENSES PARA EL AÑO 2015 SEGÚN EL INEC

Fuente: Oficina de información del MAG agosto 2017.

Figura 2.6: Número de nacimientos de lechones según su composición por sexo durante el mes de agosto de 2017 en fincas de la zona de San Carlos(INEC)

Ejemplo 2.8

Los datos que se dan en la siguiente tabla muestran los resultados del nivel académico para los empleados de la empresa XYZ en mayo de 2017.

Nivel de instrucción	*Número de empleados*
Primario	3
Secundario	8
Técnico	10
Universitario	4
Total	25

Fuente: Recursos Humanos empresa XYZ mayo de 2017.

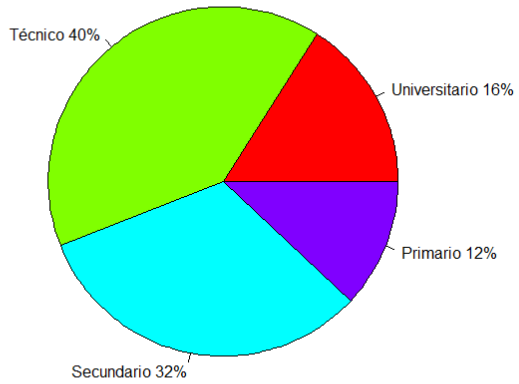
Haga un gráfico circular que muestre la distribución relativa en el nivel académico para los empleados de la empresa XYZ.

Solución

Note que la variable es de tipo cualitativa ordinal.

GRÁFICO # 6

GRÁFICO PARA EL NIVEL DE INSTRUCCIÓN DE LOS EMPLEADOS DE LA EMPRESA XYZ



Fuente: Recursos Humanos empresa XYZ mayo de 2017.

Figura 2.7: Nivel Académico alcanzado por empleados en la empresa XYZ a mayo de 2017

2.7. Gráfico lineal aritmético

El gráfico lineal aritmético se usa para representar variables de tipo *temporal* y permiten ver la tendencia de los datos a través del tiempo.

Ejemplo 2.8

La siguiente tabla muestra la evolución del *coeficiente de Gini* por persona en Costa Rica durante el período de 2010 a 2016, medidos en el mes de julio de cada año, según el INEC. El coeficiente de Gini mide la percepción que tienen las personas sobre distribución desigual de la riqueza en un país. Su valor varía entre 0 y 1. Cuanto más cerca esté de cero el coeficiente de Gini, mejor percepción en la distribución de la riqueza de un país tendrán las personas que

lo habitan. En cambio cuanto más cerca esté de 1 su valor, las personas perciben una peor distribución de la riqueza generada por el país donde viven.

Año	*Coeficiente de Gini por persona*
2010	0.507
2011	0.515
2012	0.515
2013	0.522
2014	0.516
2015	0.516
2016	0.521

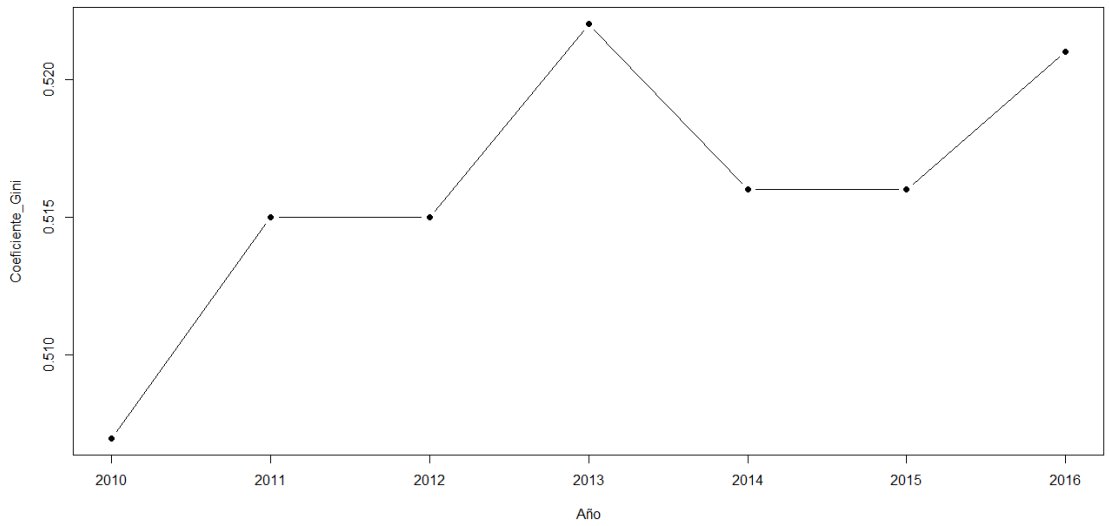
Fuente: INEC.

Represente mediante un gráfico lineal aritmético la tendencia del coeficiente de Gini por persona en Costa Rica durante el período 2010 a 2016 e interprételo.

Solución

GRÁFICO # 7

GRÁFICO PARA LA EVOLUCIÓN DEL COEFICIENTE DE GINI POR PERSONA EN COSTA RICA DURANTE EL PERÍODO 2010 A 2016



Fuente: INEC.

Figura 2.8: Evolución del coeficiente de Gini por persona en Costa Rica durante el período 2010-2016.

Como puede apreciarse claramente en el gráfico, la tendencia en general que muestra el coeficiente de Gini por persona es a aumentar a través del tiempo. Por lo tanto, se interpreta que las personas perciben una mayor desigualdad en la distribución de la riqueza en el país al ir pasando el tiempo.

Capítulo 3

Tasas e índices

Las tasas e índices se suelen expresar en términos relativos, es decir en cantidades por 100 o por 1000.

Una tasa se define como la *frecuencia relativa* expresada en términos de cantidades por 100 o por 1000 de un fenómeno en el tiempo, generalmente de un año.

Un índice se define como aquella *comparación* , usualmente en cantidades por 100, que se realiza sobre variaciones que afectan a ciertos fenómenos a lo largo del tiempo o del espacio.

3.1. Tasas: vitales, de crecimiento y desempleo

Un ejemplo de *tasa vital* es la *tasa bruta de natalidad*, que corresponde al número de nacimientos vivos ocurridos durante el año Z en una población dividido por el total de la población a mitad del año Z y multiplicado por 1000. De esta manera, *la tasa bruta de natalidad* corresponde al número de nacimientos por cada 1000 habitantes.

Por ejemplo, según datos del INEC el número de nacimientos ocurridos en Costa Rica durante el 2016 fue de 70004 nacimientos vivos. La población del país para junio de ese año fue de 4836438. Por lo tanto la tasa bruta de natalidad T_n para 2016 fue de:

$$\mathrm{T}_n = \frac{total\,de\,nacimientos\,en\,2016}{total\,de\,poblacin\,delpas\,en\,2016} * 1000 = \frac{70004}{4836438} \approx 14.47$$

La tasa de natalidad anterior T_n se interpreta de que en el país hubo 14,47 nacimientos por cada 1000 habitantes durante el 2016.

La tasa bruta de natalidad es una de las llamadas *tasas vitales* que se usan en demografía. Las *tasas vitales* que se usan en demografía son: *tasa bruta de natalidad*, *tasa bruta de mortalidad* y la *tasa de crecimiento natural*.

La tasa bruta de mortalidad T_m se define como el número de defunciones en un año Z, dividido entre la población total a mediados del año Z.

El total de defunciones para 2016 fue de 22603, por lo tanto la tasa bruta de mortalidad T_m para 2016 fue de:

$T_m = \frac{22603}{4836438} * 1000 \approx 4.67.$

La tasa anterior se interpreta de que hubo 4.67 muertes por cada 1000 habitantes.

Por último, la *tasa de crecimiento natural* T_c se define como la tasa bruta de natalidad T_n para el año Z menos la tasa bruta de mortalidad T_m. De esta forma, la tasa de crecimiento natural T_c para el año 2016 es:

$T_c = 14.47 - 4.67 = 9.8.$

La *tasa de crecimiento natural* significa que la población para el año 2016 creció a ritmo de 9.8 personas por cada 1000 habitantes.

Otra tasa muy usada en medicina es la *tasa de mortalidad infantil* T_{mi}. Esta tasa se define como el número de niños nacidos muertos entre el total de nacimientos en un año y multiplicado por 1000.

El número de nacimientos en Costa Rica para el año 2016 fue de 71819 mientras que el total de niños nacidos muertos fue de 557. Por lo tanto la tasa de mortalidad infantil en 2016 fue :

$T_{mi} = \frac{557}{71819} * 1000 \approx 7.8.$

La tasa anterior se interpreta de que hubo 7.8 niños nacidos muertos por cada 1000 nacimientos. La tasa anterior es la tasa de mortalidad infantil más baja de la historia en Costa Rica y tiene su explicación en los avances médicos y una baja en la natalidad de la población.

Otra tasa, usada en economía, es la tasa de desempleo T_d, la cual se define como el número de personas sin trabajo y que anda en busca de trabajo, dividido entre el total de la población económicamente activa, esto es, el total de personas con trabajo más las desempleadas y multiplicado por 100.

El total de desempleados en Costa Rica durante el primer trimestre de 2017 según el INEC era de 217.623, mientras que la población que contaba con trabajo era de 2165980 personas. Por lo tanto, la tasa de desempleo en Costa Rica para el primer trimestre de 2017 fue de :

$T_d = \frac{217623}{2165980+217623} * 100 = \frac{217623}{2383603} * 100 \approx 9.13\,\%.$

La tasa anterior se interpreta de que 9 de cada 100 personas económicamente activa se encuentra sin trabajo.

3.2. Índices

Por otra parte, los índices son muy utilizados, por ejemplo en economía. Uno de ellos es *el índice de precios al consumidor.*

El BCCR define el índice de precios al consumidor como *"el gasto que los hogares realizan en determinada canasta de consumo. De esta forma, al comparar estos montos para dos momentos diferentes en el tiempo se obtiene, en forma ponderada, el crecimiento promedio de los precios al consumidor en ese lapso."*.

(ver: *http://indicadoreseconomicos.bccr.fi.cr/indicadoreseconomicos*).

Vamos a explicar con el siguiente ejemplo como calcular el *índice de precios* para ciertos productos que se consumen en los hogares costarricenses y que son adquiridos en las ferias del agricultor.

En la siguiente tabla se muestra la evolución de los precios de seis productos que se venden en las ferias del agricultor durante los años 2014, 2015 y 2016.

Productos	*2014*	*2015*	*2016*
Chile dulce	250	250	225
Papas	625	800	650
Huevos	2000	1900	1500
Frijoles rojos	1000	1200	1200
Lechuga	275	300	275
Tomate	785	1075	1050
Total	4935	5525	4900

Nota al pie: Información obtenida de *https://www.simacr.go.cr/index.php/ferias-a/precios-de-referencias*
Fuente: Sistema de información de mercados agroalimentarios SIMA .

Calcule el índice de precios para el año 2016, tomando como año base 2014.

Solución

Para calcular el índice de precios para al año 2016 se divide el precio total de los productos de 2016 entre el precio total de esos productos en 2014 y su resultado se multiplica por 100 . De esta manera: $I_{p2016} = \frac{4900}{4935} * 100 \approx 99.29$

El resultado anterior se interpreta de que en 2016 los precios fueron un 0.71 % menor comparado con los precios del 2014.

Otro índice económico de interés es el *índice de precios ponderado*, el cual mide las variaciones en los precios de una manera *más real*.

El índice de precios ponderado toma en cuenta no solo el precio del producto, sino también la cantidad comprada.

Considere a continuación la siguiente tabla. Calcule el índice de precios ponderados para 2015 y 2016, tomando como base el año 2014.

Solución

Producto	2014		2015		2016	
	p_1	q_1	p_2	q_2	p_3	q_3
Chile dulce	250	3	250	3	225	4
Papas	625	3	800	2	650	4
Huevos	2000	1	1900	1	1500	1.5
Frijoles rojos	1000	1	1200	1	1200	1
Lechuga	275	1	300	1	275	1
Tomate	785	2	1075	1.5	1050	1

Fuente:Elaboración propia.

El costo de adquirir los productos para cada año fue:

$$C(2014) = 250 * 3 + 625 * 3 + 2000 * 1 + 1000 * 1 + 275 * 1 + 785 * 2 = 6685.$$

$$C(2015) = 250 * 3 + 800 * 2 + 1900 * 1 + 1200 * 1 + 300 * 1 + 1075 * 1.5 = 7362.5.$$

$$C(2016) = 225 * 4 + 650 * 4 + 1500 * 1.5 + 1200 * 1 + 275 * 1 + 1050 * 1 = 8000.$$

Por lo tanto el índice de precios ponderado I_{pp} para los años 2015 y 2016 tomando como año base 2014 son:

$$I_{pp2015} = \frac{7362.5}{6685} * 100 = 110.$$

$$I_{pp2015} = \frac{8000}{6685} * 100 = 120.$$

Ejercicio para la clase

Interprete los dos índices anteriores.

3.3. Densidad poblacional

La densidad poblacional D_p se define como el número de habitantes de un territorio a mitad del año Z dividido entre el área de dicho territorio.

Para mitad del año 2016 la población de Costa Rica fue de 4857000. El área total de país es de $51100\,km^2$. Por lo tanto la densidad poblacional D_p para Costa Rica en el año 2016 fue:

$D_d = \frac{4857000}{51100} = 95.04$ habitantes por km^2.

3.4. Producto interno bruto (PIB) y producto interno per cápita

El *producto interno bruto* (PIB) para Costa Rica lo define el BCCR como *"el valor, a precios de productor, de la producción de bienes y servicios llevada a cabo en el territorio nacional en un período determinado , menos el valor, a precios de comprador, del consumo intermedio utilizado en esa producción. En la generación del producto pueden haber participado factores de producción pertenecientes a extranjeros."*

(ver: *http://indicadoreseconomicos.bccr.fi.cr/indicadoreseconomicos*).

El PIB usualmente se mide en el período de un año.

El *producto interno per cápita* P_{ipc} se define como el PIB dividido entre la población total del país durante un determinado año.

.
Como ejemplo, el PIB de Costa Rica para el año 2016 según el BCCR fue de 27.069.118,9 millones de colones.

Por lo tanto, el producto interno per cápita en colones durante el 2016 fue de $P_{ipc} = \frac{27069119000000}{4857000} = 5573000$
.

Así, el P_{ipc} fue de aproximadamente 5573000 colones por persona o unos \$10000 por habitante.

El producto interno per cápita mide de cierta forma el poder de compra o adquisitivo de las personas en un país. Cuanto más alto sea el P_{ipc}, mayor poder de adquirir bienes y servicios tendrán sus habitantes. Los habitantes en Costa Rica, en general, tienen un *bajo nivel* de poder adquisitivo.

.

Capítulo 4

Distribución de frecuencias para variables cuantitativas y cualitativas

4.1. Definición de los componentes de una distribución de frecuencias

En general, una distribución de frecuencias está compuesta por:

1. Una clase o categoría .

2. Una frecuencia absoluta y una frecuencia relativa.

3. Una frecuencia acumulada absoluta y una frecuencia acumulada relativa.

La *categoría o clase* c_i es aquel *grupo* en el que se realizan las observaciones de las variables a estudiar.

La *frecuencia absoluta* es el *número de unidades estadísticas contadas* para cada una de las categorías o clases. Se suele representar con f_i.

La *frecuencia relativa* es *la frecuencia absoluta* f_i *de la clase* c_i dividida entre el *total de observaciones de todas las clases* y multiplicada por 100. Se suele representar con f_r

La *frecuencia acumulada* es la *frecuencia absoluta* de una *categoría* c_i *sumándole* las *frecuencias absolutas* de cada una de las categorías $c_1, c_2, ..., c_{i-1}$. Se representa con $F \downarrow$ si es la *frecuencia acumulada menos* de o con $F \uparrow$ si es *la frecuencia acumulada más de.*

La *frecuencia acumulada relativa* es la *frecuencia relativa* de una *categoría* c_i sumándole las *frecuencias relativas* de cada una de las categorías $c_1, c_2, ..., c_{i-1}$. Se representa con $F_r \downarrow$ si es *la frecuencia acumulada relativa menos de* o con $F_r \uparrow$ si es la *frecuencia acumulada relativa más de.*

4.1.1. Distribución de frecuencias para una variable cualitativa

Una distribución de frecuencias de una variable cualitativa contiene el total de observaciones hechas para cada categoría c_i de la variable. Si la variable es nominal las categorías se colocan ordenadas por magnitud de mayor a menor. Si la variable es ordinal, entonces las categorías se colocan según el orden que tienen. El cálculo de frecuencias acumuladas

para variables cualitativas se hace únicamente si la variable es ordinal.

Ejemplo 4.1
Se le pregunta a 20 personas cuál es su marca de galleta preferida en octubre de 2016, como parte de un estudio de mercado por parte de la empresa XYZ. Los datos recolectados se muestran a continuación :

Chiky	Yipi	Chiky	Cremitas	Maria
Maria	Yipi	Maria	Cremitas	Chiky
Maria	Festival	Chiky	Maria	Festival
Recreo	Tentación	Yipi	Cremitas	Yipy

Haga un cuadro donde se muestre la distribución de frecuencias para la tabla anterior. Observe que la variable *marca de galleta preferida* es una variable *cualitativa nominal*.

Solución

Se cuentan las *frecuencias* para cada una de las marcas de galleta, para así obtener la *frecuencia absoluta* f_i. La categoría o clase será *marca de la galleta* . El cuadro de distribución de frecuencias queda de la siguiente forma:

CUADRO # 1
CUADRO DE DISTRIBUCIÓN DE FRECUENCIAS PARA DIFERENTES MARCAS DE GALLETAS EN OCTUBRE DE 2016.

Marca de la galleta	f_i	f_r
Maria	5	25
Chiky	4	20
Yipi	4	20
Cremitas	3	15
Festival	2	10
Recreo	1	5
Tentación	1	5

Fuente: Encuesta de opinión por parte de la empresa XYZ en octubre de 2016.

Obsérvese que no se obtuvieron las *frecuencias acumuladas* por ser la variable de tipo *cualitativa nominal*.

Ejemplo 4.2

Se pregunta a 25 personas de la empresa XYZ cuál es su nivel de instrucción, ya sea primario, secundario, técnico o universitario, en mayo de 2017. Las resultados se muestran en la siguiente tabla:

Primario	Técnico	Técnico	Secundario	Técnico
Secundario	Universitario	Primario	Técnico	Secundario
Universitario	Secundario	Técnico	Universitario	Secundario
Técnico	Técnico	Secundario	Secundario	Secundario
Primario	Técnico	Técnico	Universitario	Técnico

Haga un cuadro donde se muestre la distribución de frecuencias para la tabla anterior.

Solución

De forma similar al *ejemplo 1*, se hace el conteo frecuencial para cada nivel de instrucción mencionados por los empleados. En este ejemplo, la variable *nivel de instrucción* es una variable de tipo *ordinal*.Además la elaboración del cuadro de distribución de frecuencias requiere que se incluyan las *frecuencias acumuladas más y menos de*. El cuadro de distribución de frecuencias queda de la siguiente forma:

CUADRO # 2
CUADRO DE DISTRIBUCIÓN DE FRECUENCIAS PARA DIFERENTES NIVELES DE INSTRUCCIÓN DE LOS EMPLEADOS DE LA EMPRESA XYZ EN MAYO DE 2017.

Nivel de instrucción	f_i	f_r	$F\downarrow$	$F_r\downarrow$	$F\uparrow$	$F_r\uparrow$
Primario	3	12	3	12	25	100
Secundario	8	32	11	44	22	88
Técnico	10	40	21	84	14	56
Universitario	4	16	25	100	4	16
Total:	25	100	-	-	-	-

Fuente: Recursos Humanos empresa XYZ mayo de 2017.

4.1.2. Distribución de frecuencias para variables cuantitativas discretas

Una distribución de frecuencias de variables cuantitativas discretas contiene el número de observaciones hechas para cada valor que asume la variable.

Ejemplo 4.3

En la empresa *Dos Cipreses* se contaron el número de llamadas diarias recibidas en su central de ventas de lunes a viernes, entre las 7:00 a.m. y las 5:00 p.m., para los meses de marzo a mayo de 2017. Los resultados del número de llamadas diarias realizadas a la central se muestran en la siguiente tabla:

Número de llamadas diarias	frecuencia número de llamadas
102	1
103	6
104	8
105	10
106	12
107	8
108	7
109	6
110	2

Fuente: Empresa Dos Cipreses.

Haga un cuadro donde se muestre la distribución de frecuencias para la tabla anterior.

Solución

Con base en la información de la tabla anterior se obtiene el siguiente cuadro de distribución de frecuencias.

CUADRO # 3
CUADRO DE DISTRIBUCIÓN DE FRECUENCIAS PARA EL NÚMERO DE LLAMADAS RECIBIDAS EN LA EMPRESA DOS CIPRESES EN MARZO-MAYO DE 2017.

Número de llamadas	f_i	f_r	$F\downarrow$	$F_r\downarrow$	$F\uparrow$	$F_r\uparrow$
102	1	1.67	1	1.67	60	100
103	6	10	7	11.67	59	98.33
104	8	13.33	15	25	53	88.33
105	10	16.67	25	41.67	45	75
106	12	20	37	61.67	35	58.33
107	8	13.33	45	75	23	38.33
108	7	11.67	52	86.67	15	25
109	6	10	58	96.67	8	13.33
110	2	3.33	60	100	2	3.33

Fuente: Recursos humanos empresa Dos Cipreses marzo-mayo de 2017.

Ejercicio para la clase.

Argumente al menos dos hipótesis que se podrían obtener de los resultados mostrados en el cuadro # 3.

4.1.3. Distribución de frecuencias para una variable cuantitativa continua

Una variable cuantitativa continua puede tomar valores arbitrarios en un intervalo real, es decir, valores que admiten cifras decimales.
Cuando se determinan los límites de las clases hay que tomar en cuenta que un límite de clase puede coincidir con un dato observado.
Las representaciones decimales pueden no ser exactas y estar afectadas por el sistema de redondeo utilizado.

4.1.4. Precisión y redondeo en los datos de una variable continua

Para efectos del curso se utilizará el *redondeo usual*, es decir, se trata de redondear una cifra al número superior o bien, que su valor no varíe, dentro de la precisión establecida.

Ejemplo 4.4

Redondee las siguientes cifras sabiendo que su *precisión* se encuentra en las centésimas: 1.632, 1.817, 1.765, 1.774, 1.85334.

Solución

Recuerde que en el redondeo usual, la *precisión* considera como *cifras significativas exactas* donde esté indicada la precisión. En nuestro ejemplo, la precisión se encuentra en las centésimas. Así para el valor 1.632, la parte 1.63 se considera como un valor *exacto* en su medición.

Recordemos las reglas del redondeo usual:

1. Si la cifra que antecede a la precisión es cinco o mayor a cinco, aumenta en 1 la cifra donde está la precisión.

2. Si la cifra que antecede a la precisión es menor a cinco, la cifra donde está la precisión no cambia.

De esta manera para el valor 1.632 al aplicar el redondeo usual queda como 1.63.

El valor 1.817 queda redondeado a 1.82.

El valor 1.765 queda redondeado a 1.77.

El valor 1.774 queda redondeado a 1.77.

El valor 1.85334 queda redondeado a 1.85.

4.1.5. Límites para la clase c_i

Para cada clase c_i, esta estará definida en forma de intervalo por su *límites inferior* l_i y su *límite superior* l_s.

La amplitud para cada clase se define como $c = \frac{M-m}{k}$, donde M es el valor máximo para de los datos, m es el valor mínimo de los datos y k es el número de clases.

La amplitud ajustada c_a se define como el valor inmediato superior al valor de c, respetando la precisión que se tenga de los datos.

El punto medio para cada clase c_i se define como $PM = \frac{l_s+l_i}{2}$.

El valor elegido para k dependerá de la finalidad del estudio, el grado de variabilidad que presente los datos, la necesidad de comparar los resultados de la investigación con otros estudios realizados previamente y que tengan el mismo número de clases, etc.

Se recomienda que el número de clases k cumpla $5 \leq k \leq 16$.

Ejemplo 4.5

Se registran las alturas de 35 personas adultas en *cm* en el centro de salud de Grecia durante enero de 2017. La precisión de los datos está en las *unidades*. Construya un cuadro de distribución de frecuencias para las alturas que se muestran a continuación:

168	173	161	178	183
155	168	172	180	195
163	153	174	170	175
178	169	160	173	170
152	145	165	177	190
170	165	163	181	199
166	165	182	155	165

Suponga que el número de clases a usar es $k = 7$.

Solución

Basados en la tabla anterior se obtiene que:

1. $M = 199$ y $m = 145$.

2. La amplitud es: $c = \frac{M-m}{k} = \frac{199-145}{7} = \frac{54}{7} = \approx 7.71$.

3. El valor inmediato superior será $c_a = 8$. De esta manera cada clase c_i tendrá una amplitud de 8.

4. La clase c_1 se calcula de la siguiente manera : $l_{i1} = m$ y $l_{s1} = m + c_a$. En nuestro caso $c_1 = [145, 153[$.

5. Luego de manera similar: $c_2 = [153, 161[$, $c_3 = [161, 169[$, $c_4 = [169, 177[$, $c_5 = [177, 185[$, $c_6 = [185, 193[$ y $c_7 = [193, 201[$

El cuadro de distribución de frecuencias queda de la siguiente forma:

CUADRO # 4
CUADRO DE DISTRIBUCIÓN DE FRECUENCIAS PARA LAS ALTURAS DE 35 PERSONAS REGISTRADAS EN EL CENTRO DE SALUD DE GRECIA EN ENERO DE 2017.

Clases para las alturas	f_i	f_r	$F\downarrow$	$F_r\downarrow$	$F\uparrow$	$F_r\uparrow$	PM
$[145, 153[$	2	5.71	2	5.71	35	100	149
$[153, 161[$	4	11.43	6	17.14	33	94.29	157
$[161, 169[$	10	28.57	16	45.71	29	82.86	165
$[169, 177[$	9	25.71	25	71.42	19	54.29	173
$[177, 185[$	7	20	32	91.42	10	28.58	181
$[185, 193[$	1	2.87	33	94.29	3	8.58	189
$[193, 201[$	2	5.71	35	100	2	5.71	197

Fuente: Oficina de información Centro de Salud de Grecia enero de 2017.

NOTAS **IMPORTANTES** A LA HORA DE HACER EL CUADRO DE DISTRIBUCIÓN DE FRECUENCIAS

1. Todos los datos registrados se suponen que ya han sido **REDONDEADOS**.

2. Siempre hay que redondear la **amplitud ajustada** c_a hacia arriba.

3. Para todas las frecuencias relativas se usaran **dos cifras decimales** aplicando el redondeo usual de ser necesario .

4. El punto medio PM se calcula con **TODAS** sus cifras decimales.

Tarea moral # 2

Repita el ejemplo anterior usando un valor de $k = 10$.

Ejemplo 4.6

Se mide el consumo en *gigabites* de 50 clientes del servicio de *internet celular* durante el mes de abril de 2017 de la operadora *Mólestar*. La precisión de las mediciones está en las décimas. Haga el cuadro de distribución de frecuencias basándose en los siguientes datos:

1.6	1.7	1.6	1.7	1.8
2.5	1.6	3.7	1.8	1.9
1.0	0.8	1.7	1.9	0.7
4.2	3.3	5.8	0.4	1.3
2.5	4.1	3.0	3.0	2.6
1.7	2.8	8.4	1.4	2.0
10.2	4.6	6.6	1.6	0.9
2.7	5.8	3.4	1.4	2.5
2.2	4.6	4.6	5.0	1.0
1.2	3.2	1.8	2.0	2.0

Suponga un valor de $k = 9$.

Solución

Basados en la tabla anterior se obtiene que:

1. $M = 10.2$ y $m = 0.4$.

2. La amplitud es: $c = \frac{M-m}{k} = \frac{10.2-0.4}{9} = \frac{9.8}{9} = \approx 1.088$.

3. Por redondeo hacia arriba $c_a = 1.1$. De esta manera cada clase c_i tendrá una amplitud de 1.1.

4. La clase c_1 se calcula de la siguiente manera : $l_{i1} = m$ y $l_{s1} = m + c_a$. En nuestro caso $c_1 = [0.4, 1.5\,[$.

5. Luego de manera similar: $c_2 = [1.5, 2.6[$, $c_3 = [2.6, 3.7[$, $c_4 = [3.7, 4.8[$, $c_5 = [4.8, 5.9[$, $c_6 = [5.9, 7.0[$, $c_7 = [7.0, 8.1[$, $c_8 = [8.1, 9.2[$, $c_9 = [9.2, 10.3[$

.
El cuadro de distribución de frecuencias queda de la siguiente forma:

CUADRO # 5
CUADRO DE DISTRIBUCIÓN DE FRECUENCIAS PARA EL CONSUMO EN GIGABITES DE 50 USUARIOS DEL SERVICIO CELULAR DE LA EMPRESA MÓLESTAR EN ABRIL DE 2017.

Clases para los gigas consumidos	f_i	f_r	$F\downarrow$	$F_r\downarrow$	$F\uparrow$	$F_r\uparrow$	PM
$[0.4, 1.5[$	10	20	10	20	50	100	0.95
$[1.5, 2.6[$	20	40	30	60	40	80	2.05
$[2.6, 3.7[$	8	16	38	76	20	40	3.15
$[3.7, 4.8[$	6	12	44	88	12	24	4.25
$[4.8, 5.9[$	3	6	47	94	6	12	5.35
$[5.9, 7.0[$	1	2	48	96	3	6	6.45
$[7.0, 8.1[$	0	0	48	96	2	4	7.55
$[8.1, 9.2[$	1	2	49	98	2	4	8.65
$[9.2, 10.3[$	1	2	50	100	1	2	9.75

Fuente: Oficina de información empresa Mólestar abril de 2017.

Ejercicio para la clase.

Obtenga al menos dos hipótesis que se podrían inferir de los resultados mostrados en el cuadro # 5.

4.1.6. Representación Gráfica: Histogramas, polígonos de frecuencias y ojivas de frecuencias

Las representaciones gráficas tienen la ventaja de que permiten *resumir la información de los datos obtenidos* en resultados *visuales* que no requieren de ningún cálculo.

4.1.7. El histograma

1. Cada clase está representada por un rectángulo de altura proporcional a la frecuencia de la clase. Los rectángulos van continuos y, en algunas ocasiones, tienen ancho proporcional a la amplitud de la clase.

2. En el eje de las abscisas (*eje horizontal*), se colocan los límites reales de las clases, donde se coloca el límite de cada rectángulo.
3. En el eje de las ordenadas (*eje vertical*) se coloca una escala para las frecuencias de las clases.

4. El histograma tiene la utilidad de que permite *observar si existe normalidad* en los datos recolectados.

Para efectos del curso, se recomienda el uso del programa *R Commander* para la creación de gráficos y en general, para cualquier análisis estadístico donde el uso de software sea requerido.

5. Además si los datos se van a agrupar me da una estimación del número k de clases a utilizar.

Ejemplo 4.7

Realice el histograma para los cuadros de frecuencias 3 y 5. Además realice una interpretación para cada uno de los histogramas.

Solución

Observe que la variable para el ejemplo del cuadro de frecuencias # 3 es de tipo *discreta*. Mediante el uso de *R Commander* su histograma queda como se muestra en la figura (4.1):

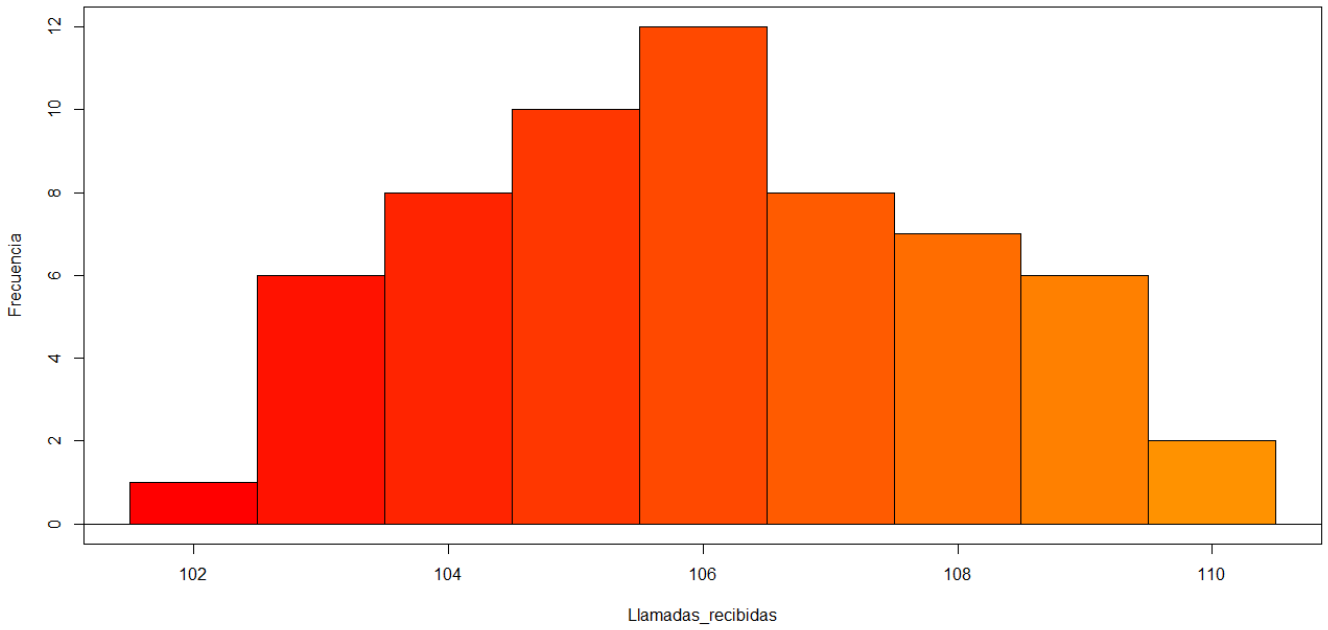
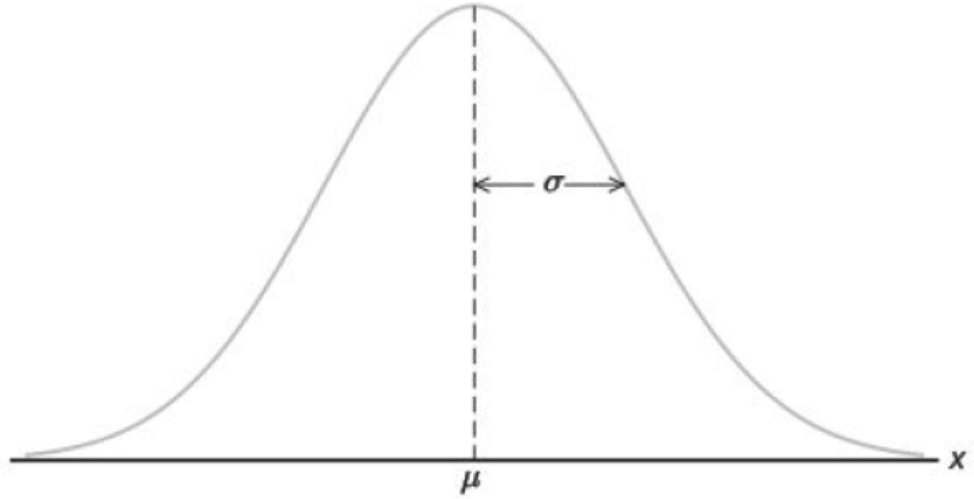


Figura 4.1: Histograma que muestra la distribución de llamadas recibidas en la central de ventas de la empresa Dos Cipreses

Una interpretación que puede hacerse sobre el histograma de la figura (4.1), es que se observa que los datos presentan *normalidad*. El término *normalidad* hace referencia a que los datos se *distribuyen* de manera muy cercana a la *curva de distribución normal*. La curva de distribución normal tiene la forma de una *campana* y su gráfica se presenta en la figura (4.2). La *normalidad* es uno de los requisitos que se pide que cumplan los datos para aplicar ciertas técnicas de análisis estadístico, principalmente en *estadística inferencial*, *siendo esta la distribución de probabilidad continua más importante en estadística*.

Por lo tanto, los datos mostrados en el histograma de la figura (4.1) presentan normalidad en su distribución.



Fuente: Imagen obtenida de *Probabilidad y estadística para ingeniería y ciencias* de Walpole, Myers y Myers, página 173, novena edición. Editorial Pearson. México 2012.

Figura 4.2: Curva de distribución normal

Algunas características de la curva de distribución normal son:

1. Presenta simetría con respecto a la media μ.
2. Tiene un máximo en μ.
3. Presenta puntos de inflexión en $\mu - \sigma$ y en $\mu + \sigma$.

Para el ejemplo del cuadro de frecuencias # 5 , cuya variable es de tipo *continua*, su histograma queda como se muestra en la figura (4.3):

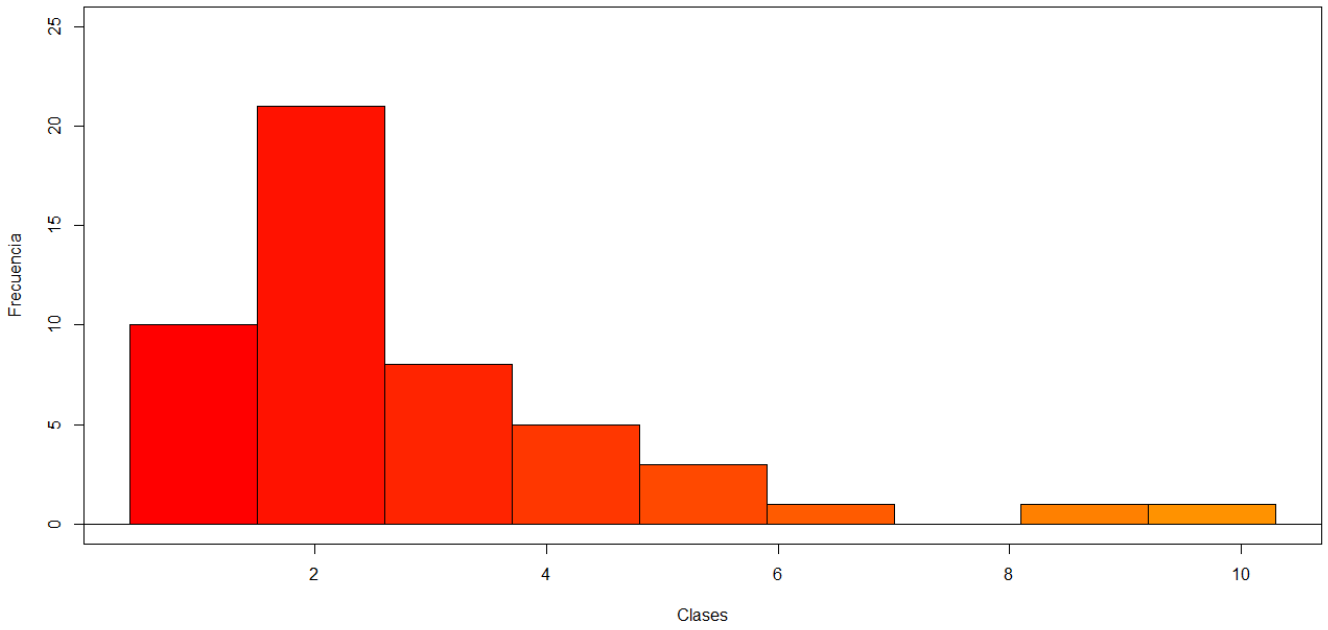


Figura 4.3: Histograma que muestra la distribución en el consumo de *gigabites* por clientes de internet celular de la empresa Mólestar.

Se interpreta a simple vista que los datos mostrados en el histograma de la figura (4.3) **NO** presentan normalidad, es decir, no se asemeja a la distribución que tiene la curva de distribución normal mostrada en la figura (4.2).

4.2. Polígonos de frecuencias

Los *polígonos de frecuencias* se usan para representar gráficamente *variables continuas* en datos agrupados.

Su gráfica consta de líneas rectas que unen los puntos consecutivos de pares ordenados (x, y) , donde x es la marca de cada clase , es decir, el punto medio de la clase, y el valor de y será la frecuencia para esa clase.

Ejemplo 4.8

Considérese los datos agrupados en el cuadro # 5 para las alturas de pacientes de la CCSS de la zona de Grecia. Realice el polígono de frecuencias para el cuadro mencionado anteriormente.

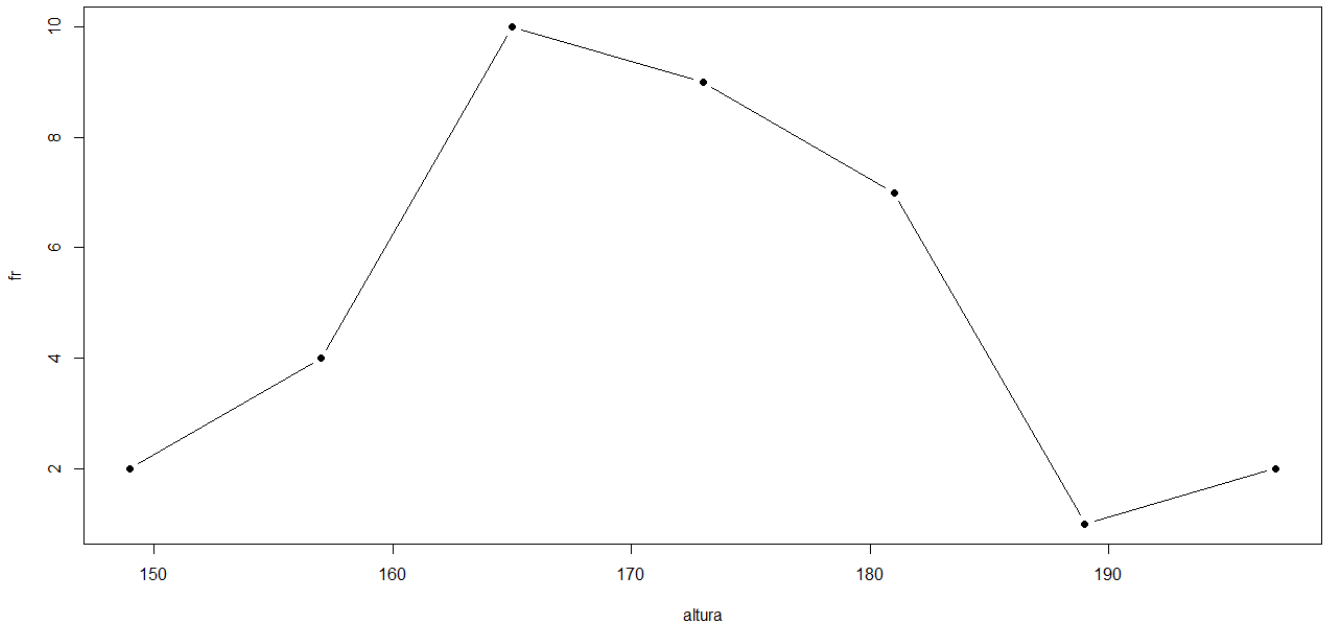


Figura 4.4: Polígono de frecuencias para datos agrupados de las alturas de pacientes de la CCSS de la zona de Grecia

El *polígono de frecuencias* permite hacer estimaciones a simple vista, sobre los estadísticos para la media poblacional μ y la desviación estándar poblacional σ para las alturas, en general, de los habitantes de la región de Grecia.

4.3. Ojivas

Las ojivas muestran de manera gráfica las *frecuencias acumuladas más o menos de* para variables continuas en datos agrupados.

Su utilidad radica en que permite determinar el número de valores que se encuentran por debajo de una cifra en particular.

Su gráfica consta de líneas rectas que unen los puntos consecutivos de pares ordenados (x, y) , donde x es la marca de cada clase , es decir, el punto medio de la clase, y el valor de y será la frecuencia acumulada para esa clase.

Ejemplo 4.9

Considérese nuevamente los datos agrupados en el cuadro # 5 para las alturas de pacientes de la CCSS de la zona de Grecia.

Realice la ojiva para el cuadro mencionado anteriormente utilizando para ello la *frecuencia acumulada menos de*

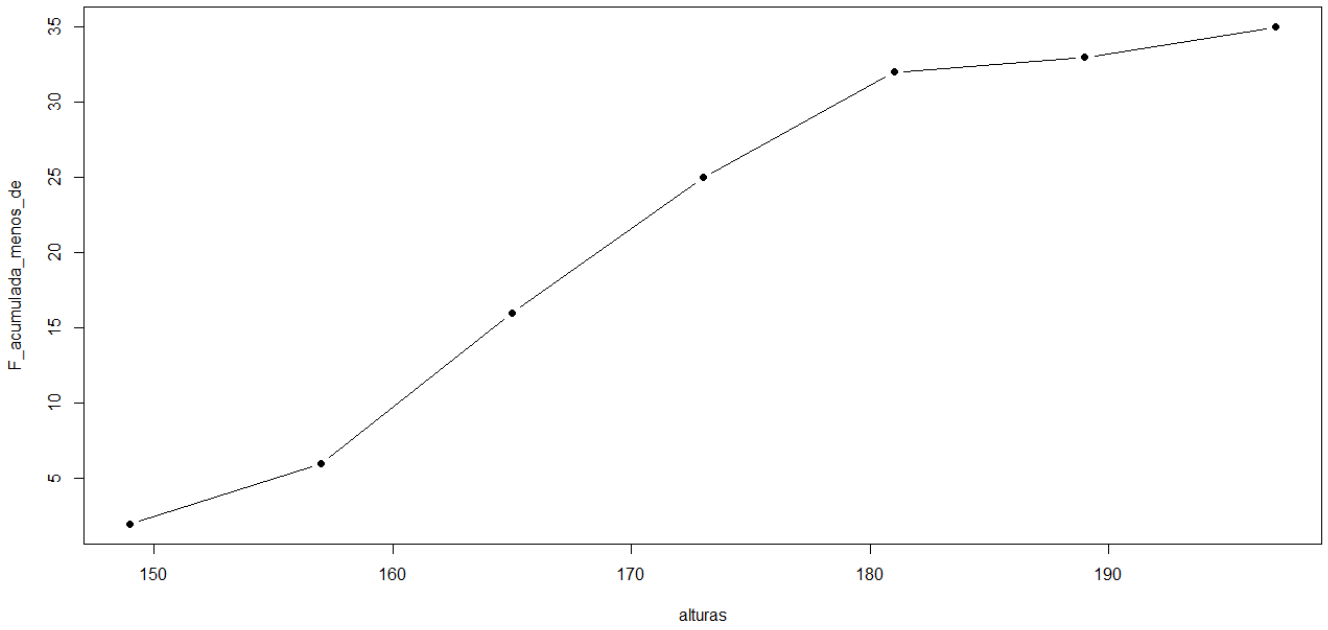


Figura 4.5: Ojiva para datos agrupados de las alturas de pacientes de la CCSS de la zona de Grecia según su frecuencia acumulada menos de

Tarea moral #3
Realice la ojiva para los datos del problema anterior, donde se muestre en una misma gráfica la *frecuencia acumulada más de* y la *frecuencia acumulada menos de* y mencione en qué punto se intersecan ambas ojivas. El punto donde se interquen debe corresponder al valor de la *mediana*.

4.4. Diagramas de tallo-hoja

Los diagramas de *tallo-hoja* (en inglés *stem and leaf diagram*) permite obtener las frecuencias de los valores para la distribución y a su vez una representación gráfica de esa distribución. Las variables deben ser de tipo *cuantitativo*. Para construirlo, basta separar en cada dato el último dígito de la derecha, esto si fuera a una *hoja*, del bloque de cifras restantes que formaran el tallo. Esta representación es similar a la información que puede proporcionar un *histograma*, pero siendo su construcción más sencilla.

Ejemplo 4.10

Realice el diagrama de tallo-hoja del consumo de *gigabites* por clientes de internet celular de la empresa Mólestar a una hoja. Debe considerar que las unidades para el cálculo de cada tallo-hoja se hace según la precisión indicada, en este caso en las cifras decimales.

Tallo	Hoja
0*	4789
1*	0023446666777788899
2*	0002555678
3*	002347
4*	12666
5*	088
6*	6
8*	4
10*	2

Figura 4.6: Diagrama de tallo-hoja para el consumo de internet celular de 50 clientes de la empresa Mólestar

Observe que cada tallo puede representar una *clase* si se deseara agrupar los datos para su estudio. Además se observa del diagrama que la mayoría de los datos se encuentran entre los 0.4 y 4.6 *gigabites*. Lo anterior podría suponer que como la mayoría de los usuarios de internet tienen un consumo menor a 5 *gigas* al mes, la empresa podría ofrecer paquetes y ofertas para ese grupo de personas en un rango de hasta 5 o 6 gigas y así potencialmente aumentar sus ganancias.

Ejemplo 4.11

Considérese las alturas de pacientes de la CCSS de la zona de Grecia. Realice el diagrama tallo-hoja para dichos alturas. Recuerde que la precisión está en las unidades.

Tallo	Hoja
15*	2355
16*	013355556889
17*	00023345788
18*	0123
19*	059

Figura 4.7: Diagrama de tallo-hoja para las alturas de 35 personas obtenidas en el centro de salud de Grecia

Del diagrama de tallo-hoja anterior puede notarse que la mayoría de las alturas de estas personas está en un rango de entre 152 y 178 cm, lo cual podría interpretarse como el rango *normal* de altura en el que podría esperarse se ubiquen la mayoría de las personas en la comunidad de Grecia.

El *diagrama de tallo-hoja* se usaba principalmente en los años 1980 y 1990, cuando las computadoras no tenían la capacidad de procesamiento para realizar gráficos pero sí el de imprimir los números en pantalla. Este tipo de diagrama puede decirse que está *en desuso*, ya que las computadoras actuales pueden realizar fácil y rápidamente las distintas gráficas que sean requeridas.

Ejercicio para la clase

Obtenga dos posibles hipótesis de los resultados mostrados en las figuras (4.6) y (4.7).

Capítulo 5

Medidas de posición central y de variabilidad o de dispersión

5.1. Medidas de posición central para datos no agrupados

Las medidas de posición central buscan *resumir* en un solo valor la *distribución* de los datos. Este valor usualmente se encuentra en la *posición central* de la distribución, de ahí su nombre.

Las medidas de posición central más usadas son: *la media o promedio*, *la moda* y *la mediana*.

La *moda*, representada por Mo, es aquel *elemento* de una distribución X que muestra la *mayor frecuencia*. Tiene la ventaja de que no se ve afectada por los valores altos o bajos presentes en la distribución. En distribuciones basadas en un número bajo de observaciones puede darse el caso de que la moda Mo no exista, o bien, no sea única, con lo cual presenta la limitación de requerir un número elevado de observaciones (*datos*) para que este valor se pueda manifestar de una manera más evidente.

La *mediana*, denotada Me, será aquel valor de una distribución X de datos, ordenada de menor a mayor, que estará ubicada a la *mitad* de la distribución. Si el número de elementos n de una distribución es impar, entonces su mediana será :

$$Me = x_{\frac{n+1}{2}}. \tag{5.1}$$

Si el número de elementos n de la distribución es par, entonces la mediana será :

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}. \tag{5.2}$$

Por último, la *media o promedio* es la medida de posición central más utilizada. El promedio, denotado como $\overline{x}$ de una distribución de n elementos $x_1, x_2, ..., x_n$ se define como:

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{5.3}$$

Ejemplo 5.1

Obtenga la moda, la mediana y el promedio para la distribución de las alturas de 35 pacientes de la CCSS de la zona de Grecia.

Solución

Para el ejemplo de las alturas de los 35 pacientes, se denotará su distribución por $A = \{145, 152, 153, 155, 155, 160, 161, 163, 163, 165, 165, 165, 165, 166, 168, 168, 169, 170, 170, 170, 172, 173, 173,174, 175, 177, 178, 178, 180, 181, 182, 183, 190, 195, 199\}$

Vemos que en la distribución A el valor de mayor frecuencia es 165 (se repite 4 veces en la distribución), por lo tanto $Mo = 165$ cm.

Como la distribución A posee 35 elementos, su valor $n = 35$. Su mediana estará ubicada en la posición número $\frac{35+1}{2} = 18$ de la distribución A, cuyo elemento está dado por $a_{18} = 170$. Por lo tanto $Me = 170$ cm.

Por último, el promedio de la distribución A será $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{145+152+...+199}{35} = \frac{5958}{35} \approx 170.23$ cm.

Ejemplo 5.2

Obtenga la moda, la mediana y la media para la distribución en el consumo de *gigabites* de 50 clientes del servicio de internet celular de la empresa Mólestar.

Solución

Para el consumo de *gigabites* de 50 clientes en el servicio de internet celular de la empresa Mólestar, se denotará su distribución por G={ 0.4, 0.7, 0.8, 0.9, 1.0, 1.0, 1.2, 1.3, 1.4, 1.4, 1.6, 1.6, 1.6, 1.6, 1.7, 1.7, 1.7, 1.7, 1.8, 1.8, 1.8, 1.9, 1.9, 2.0, 2.0, 2.0, 2.2, 2.5, 2.5, 2.5, 2.6, 2.7, 2.8, 3.0, 3.0, 3.2, 3.3, 3.4, 3.7, 4.1, 4.2, 4.6, 4.6, 4.6, 5.0, 5.8, 5.8, 6.6, 8.4, 10.2}.

Se observa que en la distribución G, esta presenta dos valores con la mayor frecuencia en la distribución, 1.6 y 1.7, con una frecuencia de 4 para cada valor. Por lo tanto la distribución G es *bimodal*, con $Mo_1 = 1.6$ Gb y $Mo_2 = 1.7$ Gb.

Como la distribución G tiene 50 elementos, su valor $n = 50$. Así, su mediana será la suma de los elementos en las posiciones $\frac{50}{2} = 25$ y $\frac{50}{2} + 1 = 26$, los cuales serán $g_{25} = 2.0$ y $g_{26} = 2.0$. Por lo tanto su mediana es $Me = \frac{2.0+2.0}{2} = 2.0$ Gb.

La media de la distribución G será $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{0.4+0.7+...+10.2}{50} = \frac{139.8}{50} \approx 2.80$ Gb.

Ejercicio para la clase

Use los resultados de los valores obtenidos en los ejemplos (5.1) y (5.2) y redacte al menos cinco hipótesis que podrían obtenerse de esas medidas de posición central. Además determine cuál de las tres medidas de posición central representa de mejor manera a cada distribución

5.2. Medidas de posición central para valores agrupados

Para los datos que estén agrupados en *clases*, su *media o promedio* se calcula mediante la fórmula:

$$\overline{x} = \sum_{i=1}^{k} \frac{PM_i \cdot f_i}{n} \tag{5.4}$$

Donde:

PM_i es el punto medio de la clase i.
k es el número de clases.
f_i es la frecuencia absoluta para la clase i.
n es el número total de observaciones.

La mediana Me para datos agrupados se aproxima mediante la fórmula:

$$Me = L_i + \left(\frac{\frac{n}{2} - F_a}{f_i} \right) \cdot c \tag{5.5}$$

donde:

f_i es igual a la frecuencia absoluta de la clase que contiene a la mediana.
n es el número total de observaciones.
L_i es el límite inferior de la clase que contiene a la mediana.
F_a es la frecuencia acumulada anterior a la clase que contiene la mediana.
$c =$ es la amplitud de la clase que contiene a la mediana.

Por último, la moda Mo para datos agrupados se define como el punto medio PM de la clase con la mayor frecuencia absoluta.

Ejemplo 5.3

Calcule la moda , la mediana y el promedio para los datos agrupados en el ejemplo (4.5).

Solución

La *clase modal* sería la clase $[161, 169[$, ya que es la clase que cuenta con la mayor frecuencia (10). El punto medio de la clase es $PM = 165$. Por lo tanto su moda es $Mo = 165$ cm.

La clase que contiene a la mediana será aquella clase donde ocurra *la primer frecuencia acumulada menos de* que supere el 50 % de los datos observados. Del cuadro # 4 se aprecia que la clase $[169, 177[$ será la clase que contiene a la mediana, ya que es en ella donde ocurre la primer *frecuencia acumulada menos de* superior al 50 % de los datos de la distribución.

De esta manera $Me = 169 + \left(\frac{\frac{35}{2}-16}{9}\right) \cdot 8 \approx 170.33$.

Por último, el promedio será $\overline{x} = \sum_{i=1}^{7} \frac{PM_i \cdot f_i}{35} = \frac{149 \cdot 2 + 157 \cdot 4 + 165 \cdot 10 + 173 \cdot 9 + 181 \cdot 7 + 186 \cdot 1 + 197 \cdot 2}{35} = \frac{5980}{35} = 170.86$.

Ejercicio para la clase

Compare los resultados obtenidos del ejemplo (5.1) con los del ejemplo (5.3), y mencione si se pueden apreciar diferencias entre los datos agrupados y los no agrupados. Además mencione cuál medida de posición central representa de mejor manera a la distribución de datos, tanto en los datos agrupados como no agrupados.

Ejemplo 5.4

Calcule la moda , la mediana y el promedio para los datos agrupados en el ejemplo (4.6).

Solución

La *clase modal* será la clase $[1.5, 2.6[$, ya que es la clase que cuenta con la mayor frecuencia (20). El punto medio de la clase es $PM = 2.05$. Por lo tanto su moda es $Mo = 2.05$ Gb.

La clase que contiene a la mediana será aquella clase donde ocurre *la primer frecuencia acumulada menos de* que supere el 50 % de los datos observados. Del cuadro # 5 se aprecia que la clase $[1.5, 2.6[$ será la clase que contiene a la mediana, ya que es en ella donde ocurre la primer *frecuencia acumulada menos de* superior al 50 % de los datos de la distribución.

De esta manera $Me = 1.5 + \left(\frac{\frac{50}{2}-10}{20}\right) \cdot 1.1 \approx 2.33$ Gb.

Por último, el promedio será $\overline{x} = \sum_{i=1}^{9} \frac{PM_i \cdot f_i}{50} = \frac{0.95 \cdot 10 + 2.05 \cdot 20 + 3.15 \cdot 8 + 4.25 \cdot 6 + 5.35 \cdot 3 + 6.45 \cdot 1 + 7.55 \cdot 0 + 8.65 \cdot 1 + 6.45 \cdot 1 + 9.75 \cdot 1}{50} = \frac{142.1}{50} = 2.84$ Gb.

Ejercicio para la clase

Compare los resultados obtenidos del ejemplo (5.2) con los del ejemplo (5.4), y mencione si se pueden apreciar diferencias entre los datos agrupados y los no agrupados. Además mencione cuál medida de posición central representa de mejor manera a la distribución de datos, tanto en los datos agrupados como no agrupados.

5.3. Medidas de variabilidad para datos no agrupados

La experimentación científica recurre usualmente a los *diseños experimentales* para aceptar o rechazar hipótesis que se plantean durante una investigación científica. Los *diseños experimentales* permiten obtener datos que serán posteriormente procesados, analizados e interpretados. Estos *diseños experimentales* tienen la característica de que pueden *repetirse* y permite *comparar* los resultados obtenidos *entre* cada una de las veces que el experimento sea repetido. Cada vez que se repite un experimento, los resultados van a *variar* en mayor o menor medida de los experimentos que se hayan realizado anteriormente.

Una vez que se han procesado los datos para su análisis y se han calculado sus valores de posición central, como se vio en la sección 5.2, surge la necesidad de medir *cuán dispersos* están los datos, con respecto a los valores de posición central que buscan representar la distribución de esos datos.

Por lo anterior, no solo tiene importancia saber un promedio sino también conocer *cuánto varían* los datos alrededor de él, si se concentran o dispersan alrededor del promedio calculado.

Datos tomados para distintos grupos de una misma población podrían por ejemplo, tener *el mismo promedio*, pero diferente *variabilidad*. Lo anterior pone en evidencia que los valores de posición central no son suficientes para carac-

terizar a una distribución de datos.

Las *medidas de variabilidad* más utilizadas en estadística son: *el recorrido o amplitud*, *la desviación estándar*, *la varianza y el coeficiente de variación*.

5.3.1. El recorrido o amplitud

Una forma de estimar la variabilidad de los datos en una distribución, es considerando su *recorrido o amplitud*, al tomar la diferencia entre los valores extremos de la distribución.

Su fórmula será:

$$A = M - m \tag{5.6}$$

Donde M será el *valor máximo* en la distribución y m el valor mínimo.

5.3.2. La desviación estandar

La *desviación estándar* es la medida de *variabilidad* más usada en estadística. Esta medida de dispersión nos indica cuánto se alejan, en promedio, los valores de la distribución de los datos con respecto a la media.

Si se va a calcular la *desviación estándar muestral* s, de una muestra de tamaño n, su fórmula viene dada por:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}} \tag{5.7}$$

Donde $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ representa como se dijo en la sección 5.2, el *promedio muestral*.

En cambio, para estimar la *desviación estándar poblacional* σ, de una población de tamaño N se utiliza la fórmula:

$$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}} \tag{5.8}$$

Donde $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ representa la *media poblacional*.

5.3.3. Varianza

La *varianza muestral* $Var\,(s)$ se define como:

$$Var\,(s) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} = s^2 \tag{5.9}$$

Note que la varianza muestral $Var\,(s)$ es la *desviación estándar muestral* al cuadrado.

La varianza poblacional $Var\,(\sigma)$ se define como:

$$Var\,(\sigma) = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} = \sigma^2 \tag{5.10}$$

Observe que la *varianza poblacional* $Var\,(\sigma)$ es la *desviación estándar poblacional* al cuadrado.

5.3.4. Coeficiente de variación

El *coeficiente de variación* indica la *dispersión relativa* existente entre la desviación estándar de una distribución y su promedio.

El coeficiente de variación CV se define como:

$$CV = \frac{s}{\overline{x}} \cdot 100 \tag{5.11}$$

Su importancia radica en que es una medida *independiente de las unidades* que se estén usando y la magnitud en general que presentan los datos en la distribución, es decir, si se tiene que la desviación estándar es grande porque los datos de la distribución son grandes también, al dividirse entre el promedio ese factor es eliminado. El CV no tiene unidades.

Ejemplo 5.5

Con base en los datos del ejemplo (4.5) calcule la amplitud, la desviación estándar, la varianza y el coeficiente de variación, usando para ello algún programa que el estudiante domine. Además interprete si el promedio es la mejor elección de medida de posición central para representar la distribución de las alturas. Interprete los resultados para la dispersión que se muestra en la distribución de los datos de las alturas según su desviación estándar.

Solución

Se mostrarán unicamente las respuestas, para corroborar los resultados obtenidos.

1. $A = 54$ cm.
2. $s = 11.80$ cm.
3. $Var\,(s) = 139.24\ cm^2$.
4. $CV = 6.93\,\%$.

Para saber si el promedio es la mejor elección para resumir la distribución de las alturas, se observa que el coeficiente de variación CV tiene una *dispersión relativa baja*, de menos del 7 %, con lo cual el *promedio* resulta la mejor elección para representar la distribución de los datos de las alturas.

Se interpreta que los datos de la distribución de las alturas se alejan, en promedio, 11.80 cm de la media de la distribución, la cual es de 170.23 cm.

Ejemplo 5.6

Con base en los datos del ejemplo (4.6) calcule la amplitud, la desviación estándar, la varianza y el coeficiente de variación, usando para ello algún programa que el estudiante domine. Además interprete si el promedio es la mejor elección de medida de posición central para representar la distribución de las alturas. Interprete los resultados para la dispersión que muestran la distribución de los datos de las alturas según su desviación estándar.

Solución

Se mostrarán unicamente las respuestas, para corroborar los resultados obtenidos.

1. $A = 9.8$ Gb.

2. $s = 1.97$ Gb.
3. $Var\,(s)$ =3.88 Gb^2.
4. $CV = 70.29\,\%$.

Observe que ahora el coeficiente de variación tiene una *dispersión relativa alta*, de más del 70 %, con lo cual el promedio **NO** resulta la mejor elección para representar la distribución en el consumo de internet celular con base en la muestra seleccionada. La moda tampoco sería la mejor opción, ya que como se vio en el ejemplo (5.2), esta distribución es *bimodal*. Por lo tanto el mejor valor de posición central para representar la distribución G es la *mediana* Me.

5.4. Medidas de variabilidad para datos agrupados

La medida utilizada para estimar la variabilidad en datos agrupados por clases es la *desviación estándar para datos agrupados*.

Su fórmula viene dada por:

$$s = \sqrt{\sum_{i=1}^{k} \frac{(PM_i - \overline{x})^2 \cdot f_i}{n-1}} \tag{5.12}$$

Donde:

PM_i es el punto medio de la clase i.
$\overline{x}$ es el promedio para datos agrupados.
f_i es la frecuencia para la clase i.
k es el número de clases.
n es el número total de observaciones.

Ejemplo 5.6

Considere los datos agrupados en el ejemplo (4.6).y estime su desviación estándar.

Solución

La desviación estándar según la ecuación (5.12) será:

$$s = \sqrt{\sum_{i=1}^{9} \frac{(PM_i - \overline{x})^2 \cdot f_i}{50-1}} = \sqrt{\Big((0.95 - 2.84)^2 \cdot 10 + (2.05 - 2.84)^2 \cdot 20 + (3.15 - 2.84)^2 \cdot 8 + ... + (9.75 - 2.84)^2 \cdot 1\Big)} =$$
$$\sqrt{\frac{175.44}{49}} = 1.89 \text{ Gb.}$$

5.5. El error típico de la media

Los datos que se recolectan en una investigación suelen basarse en una *muestra* extraída de una o varias poblaciones. Lo anterior se justifica desde el punto de vista estadístico, debido a que la caracterización que se haga de la distribución de los datos muestrales, como las medidas de posición central o de variabilidad, tenderán a parecerse al de la *población* que se esté investigando. Por eso, cuanto más grande sea el tamaño de la muestra, mayor su parecido con la caracterización de la población estudiada.

Es por lo anterior, que cabe preguntarse cuál sería el *error* que se comete al calcular un *promedio* obtenido de una muestra, con respecto al valor *real* del promedio para la población. A este error se le conoce como *error típico de la media* o *error estándar de la media* y su fórmula está dada por:

$$ESM = \frac{s}{\sqrt{n}} \tag{5.13}$$

Donde:

s es la desviación estandar muestral.
n es el tamaño muestral.

De esta manera, el error estándar de la media *ESM cuantifica la variación promedio* que sufre la media muestral $\overline{x}$ con respecto a la media poblacional μ, en palabras sencillas, es una *medida del error* que se comete al tomar la media calculada en una muestra como estimación de la media de la población .

Ejemplo 5.7

Considere los datos del ejemplo (4.5).y estime su error estándar de la media ESM.

Solución

El error estándar de la media ESM es:

$ESM = \frac{s}{\sqrt{n}} = \frac{11.80}{\sqrt{35}} = 1.99$ cm.

5.6. El intervalo de confianza

A partir del error estándar se construye el *intervalo de confianza*. Este *intervalo de confianza* permite estimar de manera probabilística *el valor de la media poblacional* μ dentro de un intervalo. El intervalo de confianza más utilizado en

estadística es el *intervalo de confianza al 95 % I_{95}*. Lo anterior quiere decir que existiría un 95 % de probabilidad de que el valor de la media poblacional μ se encuentre en dicho intervalo de confianza I_{95}, donde los límites inferior l_i y superior l_s están datos por :
$l_i = \overline{x} - 1.96 \cdot ESM$
$l_s = \overline{x} + 1.96 \cdot ESM$

Ejemplo 5.8

Considere los datos del ejemplo (4.5) y calcule el intervalo de confianza al 95 % I_{95}. Interprete el resultado

Solución

El error estándar de la media como se vio en el ejemplo (5.7) es de 1.99 cm. Por lo tanto su intervalo de confianza al 95 % I_{95} es:

$l_i = \overline{x} - 1.96 \cdot ESM = 170.23 - 1.96 \cdot 1.99 = 166.33$
$l_s = \overline{x} + 1.96 \cdot ESM = 170.23 + 1.96 \cdot 1.99 = 174.13$

Por lo tanto su intervalo de confianza al 95 % I_{95} corresponde a:

$I_{95} = [166.33, 174.13]$.

Del resultado anterior se interpreta que el promedio de altura μ para la población de Grecia tendrá una probabilidad del 95 % de que oscile entre 166.33 cm y 174.13 cm.

5.7. Coeficiente de asimetría A_s

El *coeficiente de asimetría* A_s permite determinar si los datos están *distribuidos de manera uniforme*, sin necesidad de realizar ningún tipo de gráfica.

La asimetría presenta tres estados diferentes:

1. *Asimetría positiva*: esta ocurre cuando $\overline{x} > Me > Mo$. Presenta un alargamiento de sus datos hacia la derecha.

2. *Simetría*: esta ocurre cuando $\overline{x} = Me = Mo$. Su forma presenta simetría.

3. *Asimetría negativa*: esta ocurre cuando $\overline{x} < Me < Mo$. Presenta un alargamiento de sus datos hacia la izquierda.

La figura (5.1) muestra gráficamente las tres diferentes formas de asimetría que existen.

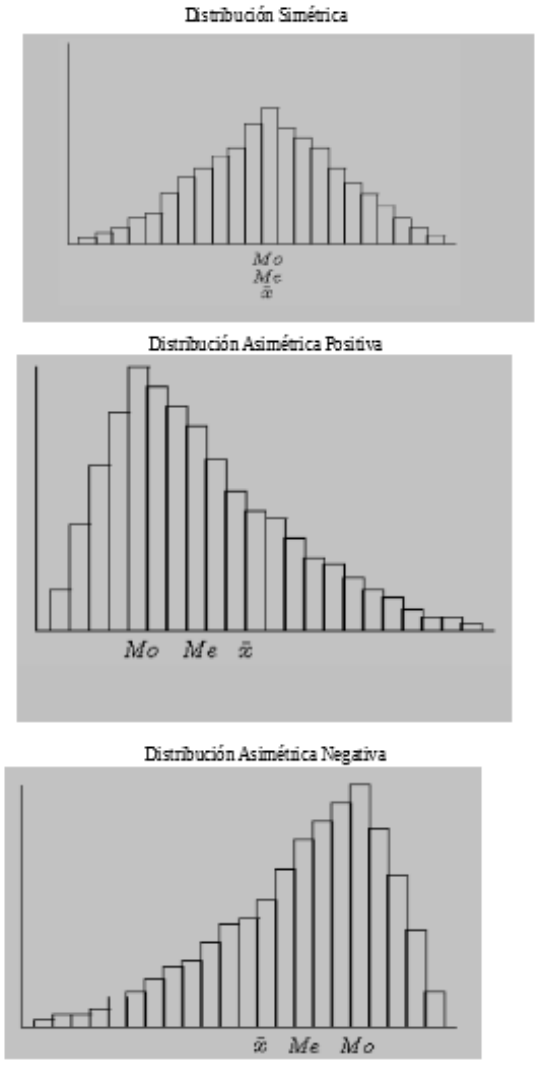


Figura 5.1: Formas para los tres tipos de asimetrías en una distribución de datos

Fuente: *Introducción a la estadística descriptiva*. Javier Trejos Zelaya - Ericka Moya Vargas. Universidad Latina de Costa Rica. Editorial Sello Latino, San Jose. 2004

La fórmula para el coeficiente de asimetría A_s viene dado por:

$$A_s = \frac{3(\overline{x} - Me)}{s} \tag{5.14}$$

1. Si $A_s = 0$ presenta una distribución simétrica.
2. Si $A_s > 0$ presenta una distribución asimétrica positiva.
3. Si $A_s < 0$ presenta una distribución asimétrica negativa.

Ejemplo 5.9

Calcule el coeficiente de asimetría A_s para los datos del ejemplo (4.5). Mencione qué tipo de asimetría presenta. Realice una gráfica lineal aritmética para apreciar si presenta algún tipo de *alargamiento* o por el contrario, presenta simetría en su forma.

Solución

El coeficiente de asimetría A_s viene dado por:

$A_s = \frac{3(\overline{x} - Me)}{s} = \frac{3(170.23 - 170)}{11.80} = 0.058.$

El coeficiente de asimetría cumple que $A_s > 0$, por lo tanto su distribución es asimétrica positiva, la cual puede apreciarse en la siguiente gráfica lineal aritmética.

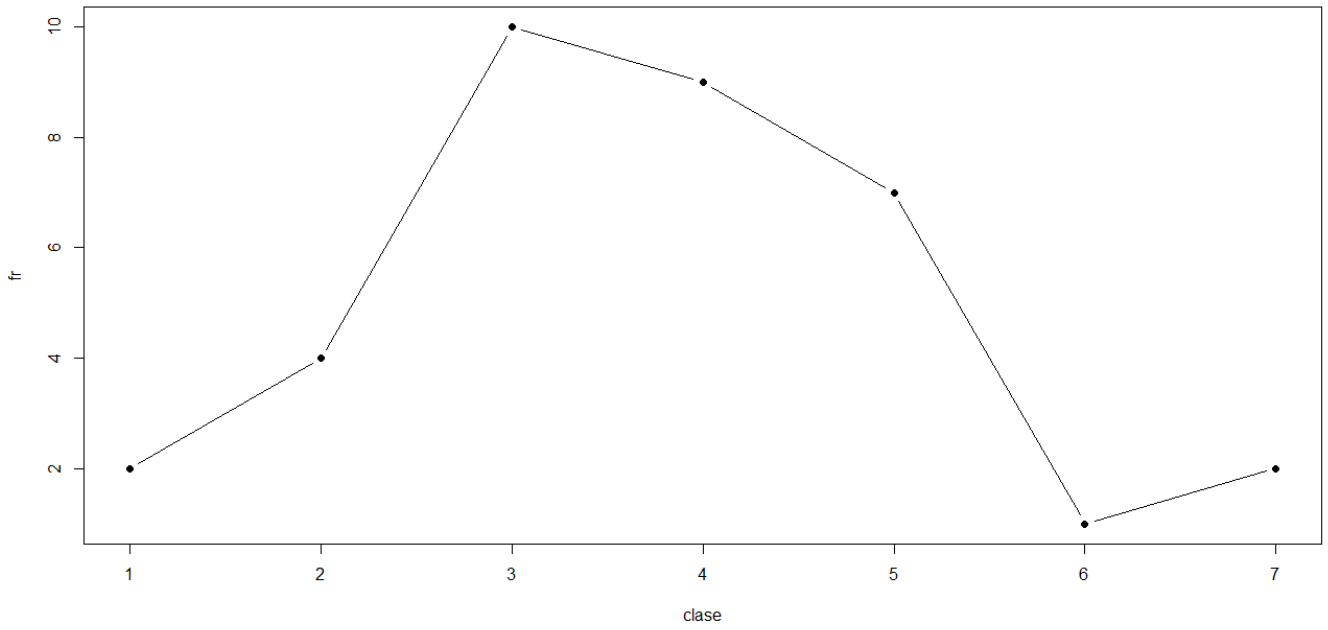


Figura 5.2: Gráfica lineal aritmética para la distribución de clases del ejemplo (4.5)

Puede notarse en la figura (5.2) que la distribución de las clases muestra un ligero alargamiento hacia la derecha, coincidiendo con el valor calculado para el coeficiente de asimetría A_s.

Se advierte que existen otras fórmulas para el cálculo del coeficiente de asimetría A_s, pero para efectos del curso solo se usará la que se da en la ecuación (5.14).

Para más información acerca de algunas otras fórmulas para calcular la asimetría en una distribución de datos, ver [4]

5.8. Cuantiles para datos no agrupados

El *cuantil* se define como aquel valor de posición P_m tal que:

$$P_m = a_{\frac{m(n+1)}{100}} \tag{5.15}$$

Donde:

m es el número de percentil.
n es el número total de observaciones.

Para ciertos *cuantiles* en particular se utiliza la siguiente notación:

1. Q_i denota al *cuartil* i.
2. K_i denota al *quintil* i.
3. D_i denota al *decil* i.
4. En general, P_i denota al *percentil* i.

Si el resultado es fraccionario, deben buscarse las posiciones para la parte entera y su sucesor, hallar su diferencia multiplicada por la parte fraccionaria. El calculo anterior se suma al valor de posición para la parte entera para hallar el percentil buscado.

5.8.1. El rango intercuartílico

El rango intercuartílico se define como Q_R y está dado por la ecuación:

$$Q_R = Q_3 - Q_1 \tag{5.16}$$

5.9. Cuantiles para datos agrupados

5.9.1. Cuantiles para datos agrupados de variable discreta

Para los cuantiles en datos agrupados de variable discreta se define como el valor P_m tal que:

$$P_m = PM_m \tag{5.17}$$

Donde:

PM_m= es el valor del punto medio para la posición igual o superior a $a_{\frac{m(n+1)}{100}}$ de la distribución.

5.9.2. Cuantiles para datos agrupados de variable continua

En caso de que la variable sea continua el percentil P_m se define como:

$$P_m = L_i + \left(\frac{\frac{m}{100}\cdot n - F_a}{f_i}\right)\cdot c \tag{5.18}$$

Donde:

m es el número de percentil.
L_i es el valor inferior de la clase donde se ubica el percentil.
n es el número de observaciones.
F_a es la *frecuencia acumulada menos* de la clase anterior a la clase que contiene al percentil.
c es la amplitud de la clase.
f_i es la frecuencia absoluta de la clase que contiene al percentil.

Ejemplo 5.10

Calcule Q_1, Q_2, Q_3, D_3, P_{43}, P_{82} y el rango intercuartílico Q_R de los datos del ejemplo (4.5) considerándolos como datos no agrupados.

Solución

1. $Q_1 = P_{25}$=$a_{\frac{25(35+1)}{100}} = a_9 = 163$ cm.
2. Q_2= $P_{50} = a_{\frac{50(35+1)}{100}} = a_{18} = 170$ cm.
3. Q_3 =$P_{75} = a_{\frac{75(35+1)}{100}} = a_{27} = 177$ cm.

4. D_3= P_{30}= $a_{10.8}$.

El valor de la parte entera es 10. Por lo tanto hay que hallar los valores en las posiciones 10 y 11, los cuales son $a_{10} = 165$ y $a_{11} = 165$. Luego se hace la diferencia $a_{11} - a_{10} = 0$. Por último se suma este resultado al valor de posición $a_{10} + 0 \cdot 0.8 = 165$ cm.

5. P_{43}= $\frac{43(35+1)}{100} = a_{15.48}$.

El valor de la parte entera es 15. Por lo tanto hay que hallar los valores en las posiciones 15 y 16, los cuales son $a_{15} = 168$ y $a_{16} = 168$. Luego se hace la diferencia $a_{16} - a_{15} = 0$. Por último se suma este resultado al valor de posición $a_{10} + 0 \cdot 0.48 = 168$ cm.

6. $P_{82} = a_{\frac{82(35+1)}{100}} = a_{29.52}$ = El valor de la parte entera es 29. Por lo tanto hay que hallar los valores en las posiciones 29 y 30, los cuales son $a_{29} = 180$ y $a_{30} = 181$. Luego se hace la diferencia $a_{30} - a_{29} = 1$. Por último se suma este resultado al valor de posición $a_{30} + 1 \cdot 0.52 = 180 + 0.52 = 180.52$ cm

7. $Q_R = Q_3 - Q_1 = 177 - 163 = 14$ cm .

Ejemplo 5.11

Calcule Q_1, Q_2, Q_3, D_3, P_{43}, P_{82} y el rango intercuartílico Q_R de los datos del ejemplo (4.5) como datos agrupados.

1. $Q_1 = P_{25} = a_{\frac{25}{100}\cdot 35} = a_{8.75}$. Como este valor no aparece en la columna de las *frecuencias acumuladas menos* del cuadro # 4, se selecciona el valor que este inmediatamente anterior, en este caso $F_a = 6$.

La clase donde estará ubicado el Q_1 será por lo tanto $[161, 169[$. De esta manera :

$$P_{25} = 161 + \left(\frac{\frac{25}{100}\cdot 35 - 6}{10}\right)\cdot 8 = 163.2 \text{ cm.}$$

2. $Q_2 = P_{50} = a_{\frac{50}{100}\cdot 35} = a_{17.50}$. Como este valor no aparece en la columna de las *frecuencias acumuladas menos* del cuadro # 4, se selecciona el valor que este inmediatamente anterior, en este caso $F_a = 16$.

La clase donde estará ubicado el Q_2 será por lo tanto $[169, 177[$. De esta manera :

$$P_{50} = 169 + \left(\frac{\frac{50}{100}\cdot 35 - 16}{9}\right)\cdot 8 = 170.33 \text{ cm.}$$

3. $Q_3 = P_{75} = a_{\frac{75}{100}\cdot 35} = a_{26.25}$. Como este valor no aparece en la columna de las *frecuencias acumuladas menos* del cuadro # 4, se selecciona el valor que este inmediatamente anterior, en este caso $F_a = 25$.

La clase donde estará ubicado el Q_3 será por lo tanto $[177, 185[$. De esta manera :

$$P_{75} = 177 + \left(\frac{\frac{75}{100}\cdot 35 - 25}{7}\right)\cdot 8 = 178.43 \text{ cm.}$$

4. $D_3 = P_{30} = a_{\frac{30}{100}\cdot 35} = a_{10.50}$. Como este valor no aparece en la columna de las *frecuencias acumuladas menos* del cuadro # 4, se selecciona el valor que este inmediatamente anterior, en este caso $F_a = 6$.

La clase donde estará ubicado el D_3 será por lo tanto $[161, 169[$. De esta manera :

$$P_{30} = 161 + \left(\frac{\frac{30}{100}\cdot 35 - 6}{10}\right)\cdot 8 = 164.6 \text{ cm.}$$

5.$P_{43} = a_{\frac{43}{100}\cdot 35} = a_{15.05}$. Como este valor no aparece en la columna de las *frecuencias acumuladas menos* del cuadro # 4, se selecciona el valor que este inmediatamente anterior, en este caso $F_a = 6$.

La clase donde estará ubicado el P_{43} será por lo tanto $[161, 169[$. De esta manera :

$$P_{43} = 161 + \left(\frac{\frac{43}{100}\cdot 35 - 6}{10}\right)\cdot 8 = 168.24 \text{ cm.}$$

6. $P_{82} = a_{\frac{82}{100}\cdot 35} = a_{28.7}$. Como este valor no aparece en la columna de las *frecuencias acumuladas menos* del cuadro # 4, se selecciona el valor que este inmediatamente anterior, en este caso $F_a = 25$.

La clase donde estará ubicado el P_{82} será por lo tanto $[177, 185[$. De esta manera :

$$P_{82} = 177 + \left(\frac{\frac{82}{100}\cdot 35 - 25}{7}\right)\cdot 8 = 181.23 \text{ cm.}$$

7. $Q_R = Q_3 - Q_1 = 178.43 - 163.2 = 15.23$ cm .

5.10. Desviación cuartil y gráficos de cajas de dispersión

5.10.1. Desviación cuartil Q_D

La desviación cuartil es la medida de variabilidad asociada a la mediana. Se denota por Q y de define como la diferencia entre el tercer y el primer cuartil entre 2, es decir:

$$Q_D = \frac{Q_3 - Q_1}{2} \tag{5.19}$$

La desviación cuartil Q_D en una *distribución normal* equivale a 0.6745 veces la desviación estándar.

Ejemplo 5.12

Halle la desviación cuartil para los datos no agrupados del ejemplo (4.6). Interprete el resultado.

Solución

Los cuartiles Q_1 y Q_3 para los datos del ejemplo (4.6) sin agrupar serían:

$Q1 = P_{25} = a_{\frac{25(50+1)}{100}} = a_{12.75}$. El valor entero es 12. Por lo tanto hay que hallar las posiciones a_{12} y a_{13}, las cuales corresponden a $a_{12} = 1.6$ y $a_{13} = 1.6$. La diferencia entre ambos es de cero. Por lo tanto $Q_1 = 1.6$ Gb.

$Q3 = P_{75} = a_{\frac{75(50+1)}{100}} = a_{38.25}$. El valor entero es 38. Por lo tanto hay que hallar las posiciones a_{38} y a_{39}, las cuales corresponden a $a_{38} = 3.4$ y $a_{39} = 3.7$. La diferencia entre ambos es $a_{39} - a_{38} = 3.7 - 3.4 = 0.3$. Por lo tanto $Q_3 = a_{38} + 0.3 \cdot 0.25 = 3.48$ Gb.

Por último, la desviación cuartil Q_D será:

$Q_D = \frac{Q_3 - Q_1}{2} = \frac{3.48 - 1.6}{2} = 0.94$ Gb.

La desviación cuartil Q_D anterior se interpreta de manera que los datos de la distribución G se distancian de la mediana Me en promedio 0.94 Gb. Recordemos que la mediana de los datos de la distribución G del ejemplo (4.6) es $Me = 2.0$ Gb.

Ejercicio para la clase

Halle la desviación cuartil Q_D para los datos agrupados del ejemplo (4.6). Interprete el resultado.

5.10.2. El diagrama de cajas de dispersión

El *diagrama de cajas* es una buena elección gráfica para mostrar la dispersión de los datos en una distribución.

Para su construcción se recomienda el uso de software, que en nuestro caso particular sería *R Commander*. Los valores a considerar en una *caja de dispersión* se muestran en la figura (5.3).

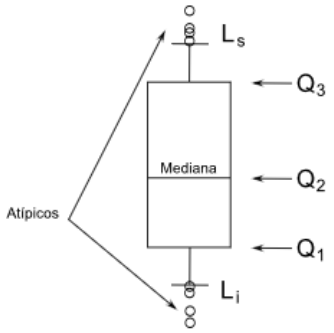


Figura 5.3: Esquema de un diagrama de dispersión y sus componentes.

Fuente: *https://upload.wikimedia.org/wikipedia/commons/6/67/Diagrama_de_caja.svg*

Donde:

1. L_s es el valor para el límite superior de la caja de dispersión y se calcula con la ecuación:

$$L_s = Q_3 + 2Q_D \tag{5.20}$$

2. L_i es el valor para el límite inferior de la caja de dispersión y se calcula con la ecuación :

$$L_i = Q_1 - 2Q_D \tag{5.21}$$

3. Q_1, Q_2 y Q_3 son los cuartiles de la distribución y Q_D es la desviación cuartil.

4. Los valores que estén por encima de L_s o por debajo de L_i se llaman *valores atípicos*.

Ejemplo 5.13

Para el siguiente ejemplo se considera la base de datos de *R Commander* llamado *dataset*, usando el archivo *chickwts* como ejemplo. En este archivo se recogen los pesos finales de 71 polluelos en gramos, según el tipo de dieta seguida durante un periodo de 1 semana. Los valores que toma la variable *Feed* (alimentación) son: *horsebean* (habas), *linseed*, (linaza), *soybean* (soja), *sunflower* (girasoles), *meatmeal* (harina de carne) y *casein* (caseína).

Haga un análisis por grupo que permita evaluar las diferencias del peso en función de la alimentación usando diagramas de caja. Interprete cuál sería la alimentación que produce los polluelos con un mejor y un peor peso final.

Solución

El diagrama de dispersión en función de la alimentación recibida por los polluelos se muestra en la figura (5.4).

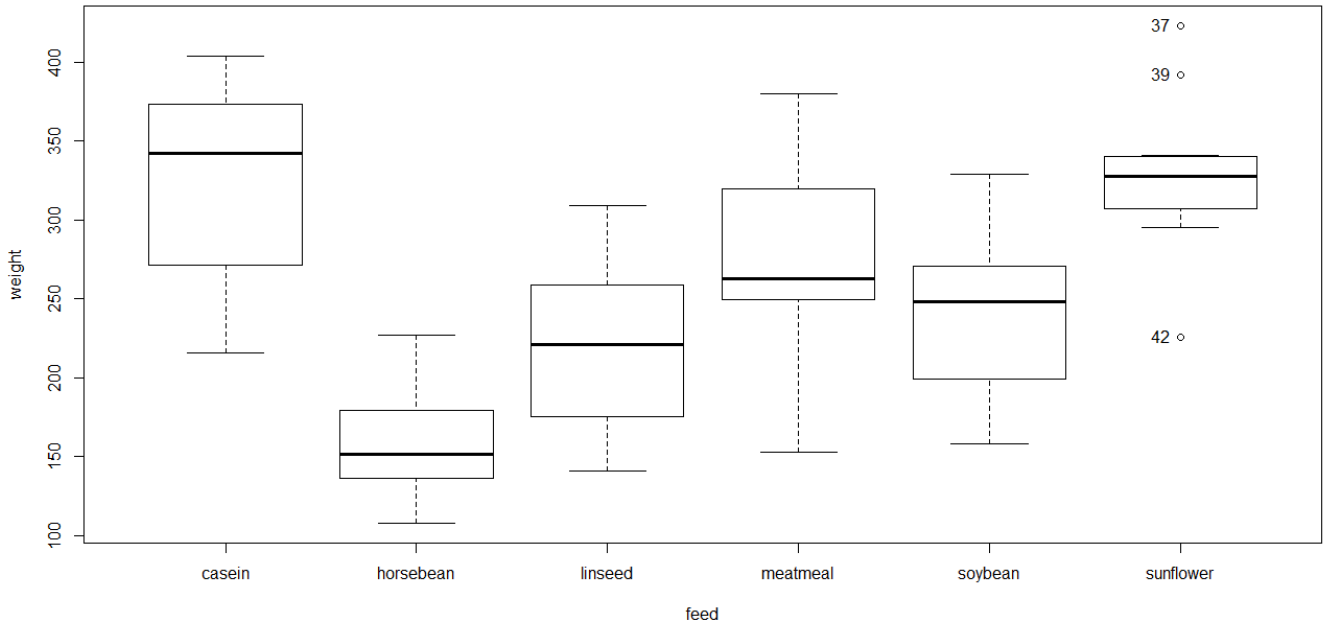


Figura 5.4: Diagrama de dispersión en función de la alimentación recibida durante una semana a 71 polluelos

Fuente: Base de datos *dataset* de *R commander*.

Se observa que los valores de la variable *peso* están más concentrados para la dieta *sunflower* (girasoles) alrededor de su mediana. También hay que notar que éste es el único grupo en el que se dan *valores atípicos superiores* para los polluelos #37 y #39. y un *valor atípico inferior* para el polluelo #42. Por otro lado, se observa que la mayor dispersión de los datos se produce en la dieta *casein* (caseína). Una evaluación inicial parece indicar que la dieta que produce polluelos con un mayor peso promedio es la basada en *girasoles*, ya que los pesos que consigue están más concentrados entorno a uno de los valores más altos de las medianas con respecto a las demás cajas. Se observa también que la dieta con *horsebean* (habas) está entre las concentraciones más bajas en torno al peso, sugiriendo que fue la dieta que dio los polluelos de menor peso en promedio alrededor de la mediana.

Ejemplo para la clase

Realice el diagrama de cajas de dispersión para el ejemplo (4.7) e interprete los resultados.

Capítulo 6

Probabilidades

6.1. Principios para el cálculo de eventos en experimentos estadísticos

1. Al conjunto de todos los resultados posibles de un experimento estadístico se le llama el *espacio muestral* y se denota con la letra S. A cada elemento a del espacio muestral se le llama *punto muestral*. Un espacio muestral puede estar escrito por *extensión*, es decir enumerando cada uno de los elementos que lo compone, o bien por *comprensión*, mencionando la cualidad que deben cumplir los elementos del espacio muestral.

Ejemplo 6.1

Se lanzan dos dados. Si el evento es que los dados sumen 5, escriba el espacio muestral para el evento anterior por extensión y por compresión.

Solución

a. El espacio muestral S escrito por *extensión* es $S = \{(1,4),(2,3),(3,2),(4,1)\}$.

b. El espacio muestral S escrito por *comprensión* es $S = \{(x,y)/x+y=5, x,y \in \mathbb{N}\}$

Ejercicio para la clase

Se lanzan dos dados. Si el evento es que los dados sumen 6 o menos, escriba el espacio muestral para el evento anterior por extensión y por comprensión.

2. Un *evento* se define como cualquier subconjunto del espacio muestral S, incluyendo al conjunto $\emptyset$ y al mismo espacio muestral S.

Ejemplo 6.2

Se define el espacio muestral $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, a, b, c, d, e, f, g, h, i\}$. Determine el evento A por extensión y por comprensión sobre S si :

a. El evento A es un número par.

b. El evento A es un número mayor o igual a 4.

c. El evento A son las tres primeras letras del abecedario.

d. El evento A son las últimas tres letras del abecedario.

e. El evento A son las soluciones para la ecuación $x^2 - x - 2$.

Solución

Se darán solamente las respuestas a las preguntas dando los conjuntos por *extensión*, dejando al estudiante los conjuntos escritos por *compresión* para realizarlos en clase.

a. $A = \{2, 4, 6, 8, 10\}$.

b. $A = \{4, 5, 6, 7, 8, 9, 10\}$.

c. $A = \{a, b, c\}$.

d. $A = \emptyset$.

e. $A = \{2\}$.

3. La no ocurrencia del evento A y denotado como $\overline{A}$ con respecto al espacio muestral S recibe el nombre *complemento de* A y se define como todos los elementos de S que no están en A.

Ejemplo 6.3

Halle el complemento $\overline{A}$ de los eventos del ejemplo 6.2.

Solución

Se darán solamente las respuestas a las preguntas, dejando al estudiante los razonamientos necesarios que conduzcan a la solución.

a. $\overline{A} = \{1, 3, 5, 7, 9, a, b, c, d, e, f, g, h, i\}$.

b. $\overline{A} = \{1, 2, 3, a, b, c, d, e, f, g, h, i\}$.

c. $\overline{A} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, d, e, f, g, h, i\}$.

d. $\overline{A} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, a, b, c, d, e, f, g, h, i\}$.

e. $\overline{A} = \{1, 3, 4, 5, 6, 7, 8, 9, 10, a, b, c, d, e, f, g, h, i\}$.

4. La intersección de dos eventos A y B , denotado con el símbolo $A \cap B$, se define como el evento que contiene *todos los eventos* comunes de A y B.

Ejemplo 6.4

Sea $A = \{a, b, c, d, e, f\}$, $B = \{a, 2, c, 4, e, 6\}$, $C = \{1, 2, 3, 4, 5, 6\}$. Hallar:

a. $A \cap B$.

b. $A \cap C$.

c. $B \cap C$.

d. $A \cap B \cap C$.

Solución

Se darán solamente las respuestas.

a. $A \cap B = \{a, c, e\}$.

b. $A \cap C = \emptyset$.

c. $B \cap C = \{2, 4, 6\}$.

d. $A \cap B \cap C = \emptyset$.

5. Dos eventos son *mutuamente excluyentes* si $A \cap B = \emptyset$.

Ejemplo 6.5 .

Del ejemplo (6.4) los eventos A y C son eventos *mutuamente excluyentes*, al igual que los eventos A, B y C.

6. La unión de dos eventos A y B, denotado $A \cup B$, se define como el evento que contiene todos los elementos que pertenecen a A, o a B o a ambos.

Ejemplo 6.6

Sea $A = \{x \in \mathbb{R} / -5 \leq x \leq 10\}$, $B = \{x \in \mathbb{Z} / -5 \leq x < 10\}$, $C = \left\{x \in \mathbb{Z} / x^3 - 2x^2 + x = 0\right\}$. Halle:

a. Escriba los conjuntos A, B y C por *extensión* o en notación de intervalo según sea el caso.

b. $A \cup B$.

c. $A \cup C$.

d. $B \cup C$.

e. $A \cup B \cup C$.

f. $(A \cap B) \cup C$.

g. $A \cap (B \cup C)$.

Solución

a. $A = [-5, 10]$, $B = \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $C = \{0, 1\}$.

b. $A \cup B = A = [-5, 10]$.

c. $A \cup C = A = [-5, 10]$.

d. $B \cup C = B = \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

e. $A \cup B \cup C = A = [-5, 10]$.

f. $(A \cap B) \cup C = (B) \cup C = B = \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

g. $A \cap (B \cup C) = A \cap (B) = B = \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Tarea moral # 4.

Realizar los ejercicios 2.1 al 2.20 de la novena edición de *Probabilidad y estadística para ingeniería y ciencias* de Ronald Walpole. Pearson 2012, paginas 42 a la 44.

7. *Regla de la multiplicación*: si un evento A_1 se puede ejecutar de A_2 formas diferentes y A_2 se puede ejecutar de A_3 maneras diferentes y así sucesivamente hasta llegar al evento A_{n-1} que puede ejecutarse de A_n maneras diferentes entonces la serie de eventos A_1, A_2, ..., A_n se puede ejecutar de $E = A_1 \cdot A_2 \cdot ... \cdot A_n$ formas diferentes.

Ejemplo 6.7 .

Determine el tamaño del espacio muestral S al lanzar cuatro monedas al aire.

Solución

La primer moneda puede caer de $A_1 = 2$ formas diferentes, la segunda moneda puede caer en $A_2 = 2$ formas diferentes, la tercera moneda puede caer en $A_3 = 2$ formas diferentes y la cuarta moneda puede caer de $A_4 = 2$ formas diferentes. Usando la regla de multiplicación se tiene que $E = 2 \cdot 2 \cdot 2 \cdot 2 = 16$. Por lo tanto el tamaño del espacio muestral S será 16.

Ejercicio para la clase

Obtenga el espacio muestral S al lanzar cuatro monedas al aire. Suponga que el evento C es que la moneda caiga " Corona " y el evento E que caiga " Escudo ".

Ejemplo 6.8

Existen 4 maneras diferentes de ir del punto A al punto B y 6 maneras de ir del punto B al punto C. ¿De cuántas maneras se puede ir del punto A al punto C si siempre se debe pasar por B?
Solución

Por la *regla de la multiplicación*, existirán $E = 4 \cdot 6 = 24$ maneras diferentes de ir de A a C pasando siempre por B.

Ejemplo 6.9

De una baraja de 52 cartas se sacan dos cartas sin repetición. ¿De cuántas maneras distintas se pueden sacar dos cartas cualesquiera sin devolver la carta nuevamente a la baraja?.

Hay 52 formas diferentes de sacar la primer carta y 51 formas diferentes de sacar la segunda carta. Por lo tanto hay $E = 52 \cdot 51 = 2652$ formas diferentes de sacar dos cartas cualesquiera de la baraja.

Ejemplo 6.10

En una urna hay 5 bolas blancas, 12 negras, 6 azules y 8 verdes. Se saca una bola de la urna fijándose en su color para luego ser devuelta a la misma urna. Si esta acción se repite 3 veces, ¿de cuántas formas distintas se pueden obtener tres bolas de cualquier color en esos 3 intentos ?

Solución

La urna tiene un total de 31 bolas. La primera vez que se saca una bola hay 31 formas diferentes, al igual que la segunda y tercera vez que se saca, ya que la bola es devuelta a la urna. Al usar la regla de *la multiplicación* hay $E = 31 \cdot 31 \cdot 31 = 27791$ formas diferentes de obtener tres bolas de cualquier color en esos 3 intentos.

Ejercicio para la clase

Misma pregunta que en el ejemplo 6.10, pero sin devolver la bola a la urna. R/ 26970

6.2. Permutaciones y combinaciones

6.2.1. Definición de permutación

Una permutación puede definirse como un *arreglo* de n objetos distinguibles, los cuales al ser tomados *todos a la vez* se pueden ordenar de $n!$ formas diferentes.

El número total de arreglos que se pueden hacer a partir de n objetos distinguibles es :.

$$P(n) = n! \tag{6.1}$$

Ejemplo 6.11

Halle el número total de permutaciones que se pueden hacer con las letras a, b y c.

Solución

Se pueden obtener las siguientes permutaciones con las letras a, b y c:

abc, bac, cab, acb, bca y cba.

Note que $3! = 6$, donde además puede apreciarse que el arreglo en que se disponen las letras depende del orden de las letras.

Ejemplo 6.12

Cuántas permutaciones se pueden hacer con los dígitos 1, 2, 3 y 4. Enumérelas todas y compruebe que el número total de permutaciones está dado por la ecuación (6.1).

Solución

Se pueden obtener las siguientes permutaciones con los dígitos 1, 2 , 3, 4:

1234, 1243, 1342, 1324, 1423, 1432

2134, 2314, 2143, 2314, 2431, 2413

3124, 3142, 3214, 3241, 3421, 3412 .

4321, 4312, 4213, 4231, 4123, 4132.

En total se pueden hacer 24 permutaciones con los dígitos 1, 2 , 3, 4.

Además por la ecuación (6.1) $P(4) = 4! = 24$, confirmando el resultado anterior.

Ejemplo 6.13

En una mesa con seis sillas se van a sentar seis personas. ¿De cuántas maneras diferentes podrán sentarse a esas seis personas alrededor de la mesa? .
Solución

Se podrán sentar de $P(6) = 6! = 720$ formas diferentes.

6.2.2. Permutación con repetición

Si en un arreglo de n objetos hay r de ellos que se repiten, su número de permutaciones con repetición está dada por la fórmula:

$$P_r(n) = \frac{n!}{r!} \tag{6.2}$$

Ejemplo 6.14

Cuantas palabras diferentes se pueden formar con la palabra " SELE ". Enumérelas primero y luego use la ecuación (6.2) para corroborar el resultado.

Como en la palabra " SELE " se repite la " E " dos veces, el número de permutaciones será menor a 24. Enumerando los resultados se tiene:

SELE, *SEEL*, *SLEE*, *ESEL*, *EESL*, *ELSE*,

LESE, *LEES*, *LSEE*, *LSEE*, *EELS*, *ELES*.

Usando la ecuación (6.2) se tiene que $P_2(4) = \frac{4!}{2!} = 12$, resultado que concuerda con las permutaciones enumeradas anteriormente.

En general, si en una permutación de n elementos hay r_1, r_2, ..., r_m elementos que se repiten, entonces su fórmula viene dada por :

$$\binom{n}{r_1, r_2, \ldots, r_m} = \frac{n!}{r_1! \cdot r_2! \cdot \ldots \cdot r_m!} \tag{6.3}$$

Donde $r_1 + r_2 + ... + r_m = n$.

Ejercicio para la clase

Cuántas permutaciones diferentes se pueden formar con la palabra " MATEMATICA " $R/151200$.

6.2.3. Permutaciones de n objetos tomando r de ellos a la vez

Si se tienen n objetos distinguibles y se toman r de ellos a la vez, el total de permutaciones que se pueden hacer viene dada por la fórmula:

$$_nP_r = \frac{n!}{(n-r)!} \tag{6.4}$$

Ejemplo 6.15

Se va a hacer una rifa sacando dos bolas sin devolverlas a una urna que contiene 50 bolas numeradas del 1 al 50.

a. Calcule el número total de permutaciones que se pueden hacer para la rifa.

b. ¿Cuántas permutaciones se pueden hacer si se sabe que las dos bolas que se sacan son números pares?

Solución

a. ${}_{50}P_2 = \frac{50!}{48!} = 2450$ permutaciones en total.

b. Hay 25 bolas marcadas con números pares. Por lo tanto el total de permutaciones que existen para sacar dos bolas pares de la urna sigue la regla del producto, por lo tanto $E_{dos \cdot bolas \cdot pares} = 25 \cdot 24 = 600$.

6.2.4. Definición de combinación

Sea A un conjunto no vacío. Una combinación de m elementos de un conjunto A de n elementos se define como cualquier *subconjunto* que pueda formarse a partir de los m elementos seleccionados, sin considerar el orden en que se tomen y sin permitir la repetición de un mismo elemento y sin hacer distinciones entre sus elementos.

Si el conjunto A posee n elementos de los cuales se van a tomar m elementos , el número total de combinaciones posibles sin repetición viene dado por la fórmula :

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{6.5}$$

Ejemplo 6.16

Sea $A = \{a, b, c, d, e\}$. Halle el número total de combinaciones tomando dos elementos a la vez sin repetición.

Solución

Primero se hallarán el número total de subconjuntos que pueden formarse tomando dos elementos a la vez del conjunto A sin repetición, para después corroborarlo por medio de la ecuación (6.5).

a. $\{a, b\}$, $\{a, c\}$, $\{a, d\}$, $\{a, e\}$, $\{b, c\}$, $\{b, d\}$, $\{b, e\}$, $\{c, d\}$, $\{c, e\}$, $\{d, e\}$.

b. $\binom{5}{2} = \frac{5!}{2!(5-2)!} = 10$.

Como se observa para ambos casos el resultado es el mismo, 10.

Ejemplo 6.17

En una finca experimental se va a realizar una prueba seleccionando 3 tipos de variedades de pasto al azar de entre un total de 10 variedades para mejorar su resistencia a las sequías.

¿De cuántas maneras pueden seleccionarse los 3 tipos de variedades de pasto?

Solución

El total de combinaciones posibles está dado por la ecuación (6.5) dando como resultado:

$\binom{10}{3} = \frac{10!}{3!\cdot 7!} = 120$ combinaciones diferentes.

Ejemplo 6.18

De una baraja de 52 cartas se sacan 5 cartas sin devolverlas al mazo. Halle el número de combinaciones que se pueden hacer :

a Al sacar 5 cartas cualesquiera.

b. Si hay un par entre las 5 cartas.

c. Si hay dos pares entre las 5 cartas.

d. Sacar un *full house*, es decir un trío y un par.

Solución.

a. El número de combinaciones que se pueden hacer con 5 cartas de una baraja de 52 es $\binom{52}{5} = 2598960$ combinaciones posibles.

b. Hay $\binom{4}{2} = 6$ formas de sacar dos cartas de un mismo valor. Luego existen 13 valores diferentes para ese par de cartas. Las otras 3 cartas deben ser diferentes a las dos cartas de un mismo valor que ya se tienen , por lo tanto se pueden elegir entre 12 valores diferentes para cada una de las 3 cartas, existiendo por lo tanto $\binom{12}{3} = 220$ combinaciones diferentes para esas 3 cartas. Por último, como hay 4 palos diferentes para cada una de esas 3 cartas, por la regla del producto se tiene que hay $4^3 = 64$ combinaciones diferentes. Así el número de combinaciones para obtener un par serían $C_{par} = \binom{4}{2} \cdot 13 \cdot \binom{12}{3} \cdot 4^3 = 1098240$.

c. Hay $\binom{4}{2} = 6$ formas de sacar dos cartas de un mismo valor y $\binom{4}{2} = 6$ formas de sacar otras dos cartas de un mismo valor diferentes al otro par obtenido. Como hay 13 valores diferentes para cada una de esos 2 pares, existen $\binom{13}{2} = 78$ combinaciones posibles. Además la última carta debe ser diferente a los valores de los dos pares obtenidos, por lo tanto hay 44 formas diferentes de sacar la quinta carta. Así el número de combinaciones para obtener dos pares sería $C_{dos \cdot pares} = \binom{13}{2} \cdot \binom{4}{2} \cdot \binom{4}{2} \cdot 44 = 123552$.

d. Hay $\binom{4}{3} = 4$ formas diferentes de sacar tres cartas iguales de una misma figura y se pueden obtener de 13 formas diferentes (ya que hay 13 valores diferentes). Hay $\binom{4}{2} = 6$ formas de obtener un par. Como no se puede tener un par usando el valor de la carta que hace el trío, existen $\binom{12}{1} = 12$ formas diferentes de sacar el par. Por lo tanto el total de combinaciones para hacer un *full house* es $C_{full \cdot house} = 13 \cdot \binom{4}{3} \cdot \binom{4}{2} \cdot \binom{12}{1} = 3744$ formas diferentes de tener un *full house*.

Ejercicio para la clase

De una baraja de 52 cartas se sacan 3 cartas sin devolverlas al mazo. Halle el número de combinaciones que se pueden hacer si se obtiene un par. $R/3744$.

Tarea moral # 5

Realizar los ejercicios 2.21 al 2.48 de la novena edición de *Probabilidad y estadística para ingeniería y ciencias* de Ronald Walpole. Pearson 2012, paginas 51 a la 52.

6.3. Conceptos básicos sobre teoría de la probabilidad

Nota: los siguientes apartados se basaron en *Introduction to mathematical probability* de J. V. Uspensky. McGraw-Hill. New York. 1937.

6.3.1. Definición de Teoría de la Probabilidad

La *teoría de la probabilidad* se define como aquella rama de la matemática aplicada que tiene como objeto de estudio la *ocurrencia de eventos*, cuya causa es debida al *azar*. La frase *evento al azar* se entiende como aquel evento que resulta *incierto* y en el que no existe *voluntad alguna* para que este llegue a ocurrir.

6.3.2. Conceptos básicos: eventos, eventos mutuamente excluyentes y espacio muestral en el cálculo de probabilidades

En la *experimentación científica* , con la ayuda de hipótesis y ciertos conceptos artificiales, es posible derivar leyes que pueden ser aplicadas al mundo real, en muchos fenómenos naturales, con un sorprendente grado de precisión.

De esta forma un *evento* A puede ocurrir bajo ciertas condiciones, pero pudiendo aparecer también otras condiciones B, C, D, etc. Así, para que ocurra el evento A existirán condiciones que están completamente fuera de nuestro control o de nuestro conocimiento. Por lo tanto si el evento A llega a ocurrir o no, diremos que el evento A es un *evento probable*.

Como ejemplo supongamos que en una urna hay dos bolas: una blanca y otra negra. Supondremos además que las dos bolas son similares en todo, excepto en el color. El evento A se definirá como el evento de sacar la bola blanca de la urna y el evento B sacar la bola negra. De manera instintiva sabemos que ambos eventos son *igualmente probables*. Pero si en la urna hubiesen 10 bolas blancas y una negra, se sabe que la *probabilidad* de que ocurra el evento A sería *mayor* a que ocurriera el evento B.

6.3.3. Definición de eventos mutuamente excluyentes

Dos eventos A y B se definen como *mutuamente excluyentes* si *ambos* eventos no pueden ocurrir *al mismo tiempo*. De esta manera si los eventos A_1, A_2, ..., A_n son eventos *mutuamente excluyentes*, se puede suponer que la probabilidad de que ocurra el evento A_i con $1 \leq i \leq n$ es $P(A_i) = \frac{1}{n}$, es decir los eventos A_1, A_2, ..., A_n son eventos *equiprobables*.

Un evento A puede ocurrir en varias formas mutuamente excluyentes denotados como a_1, a_2, ...,a_m. Lo anterior significa que si el evento A ocurre entonces *uno y solo uno* de los eventos a_1, a_2, ...,a_m ocurre también y viceversa, es decir, si alguno de los eventos a_1, a_2, ...,a_m ocurre fue porque el evento A ocurrió también.

Como ejemplo, suponga que el evento A consiste en sacar un *as* de una baraja. De esta manera el evento A puede ocurrir de cuatro maneras *mutuamente excluyentes*, a saber: un as de corazón, un as de trébol, un as de espadas o un as de diamante.

Suponga ahora que el evento A puede ocurrir bajo las formas a_1, a_2, a_3,..., a_m de manera *mutuamente excluyente*, junto con las formas a_{m+1}, a_{m+2} ,..., a_n en las que **no** ocurre el evento A.

El conjunto que contiene las condiciones $a_1, a_2, ..., a_m, a_{m+1}, ..., a_n$ que hacen posible que ocurra o no el evento A se define como *el espacio muestral* denotado S, donde $S = \{a_1, a_2, ..., a_m, a_{m+1}, ..., a_n\}$. A cada elemento del espacio muestral S se le llama *punto muestral*.

El espacio muestral S se puede mostrar en forma de diagrama, llamado *diagrama de árbol*.

Ejemplo 6.18

Se realiza un experimento que consiste en lanzar un dado y si sale un 5, se lanza otra vez el dado anotando el segundo resultado. Si no sale un 5 en el primer lanzamiento, se sacará una bola de una urna que contiene 3 bolas de diferente color: Blanco (B), Negro (N) y Azul (A).

a. Realice el *diagrama de árbol* para el experimento anterior y mencione de forma explícita su espacio muestral S.

b. Cuál es la probabilidad de obtener un número par en el segundo intento.

c. Cuál es la probabilidad de obtener un número par en el primer intento y una bola azul o negra en el segundo intento.

Solución

a. En la figura (6.1) se muestra el diagrama de árbol para el ejemplo (6.18).

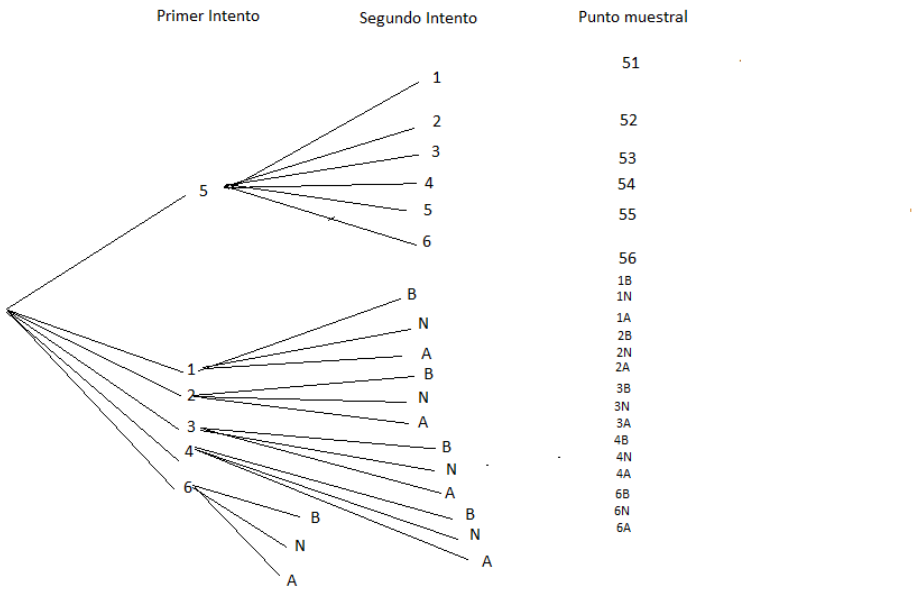


Figura 6.1: Diagrama de árbol para el ejemplo 6.18

Fuente: Datos ejemplo 6.18.

El conjunto $S = \{51, 52, 53, 54, 55, 56, 1B, 1N, 1A, 2B, 2N, 2A, 3B, 3N, 3A, 4B, 4N, 4A, 6B, 6N, 6A\}$.

b. Existen 3 maneras de obtener un número par en el segundo intento. Como el espacio muestral S tiene 21 elementos, la probabilidad buscada es $P = \frac{3}{21} = \frac{1}{7}$.

c. Hay que hallar la probabilidad de obtener un número par junto a la probabilidad de sacar una bola azul o negra. Existen 6 maneras de obtener un número par cuya bola sea azul o negra ($2A$, $4A$, $6A$, $2N$, $4N$ y $6N$). Por lo tanto la probabilidad buscada es $P = \frac{6}{21} = \frac{2}{7}$

.

6.3.4. Definición clásica de probabilidad

Por lo tanto, si de acuerdo con las condiciones dadas en el espacio muestral S hay n casos exhaustivos, mutuamente excluyentes e igualmente probables, y una cantidad m de ellos son favorables para que ocurra el evento A, entonces la probabilidad matemática de que ocurra A se define como $P(A) = \frac{m}{n}$.

Si se tienen los eventos a_1, a_2, ..., a_n, los cuales son mutuamente excluyentes, se define como evento *exhaustivo* al hecho de que necesariamente *uno* de los eventos a_1, a_2,..., a_n deberá forzosamente de ocurrir.

Por ejemplo, si el evento A consiste en sacar un *rey* de una baraja de 52 cartas, entonces la probabilidad de que ocurra A es:

$P(A) = \frac{4}{52} = \frac{1}{13}$.

Observe que existen 4 formas mutuamente excluyentes de sacar un *rey*.

Se denota la probabilidad de que ocurra el evento A como $P(A)$, la probabilidad de que ocurra el evento A o B o *ambos* se denota como $P(A+B)$.

En general, $P(A_1 + A_2 + ... + A_n)$ se interpretará como la probabilidad de que ocurra *al menos* un evento A_i, donde $1 \leq i \leq n$.

6.4. Propiedades básicas de las probabilidades

6.4.1. Probabilidad total

Teorema 6.1 *Teorema de la probabilidad total*

La probabilidad de que ocurra un evento A es la suma de las probabilidades de sus formas mutuamente excluyentes a_1, a_2, ..., a_m. Matemáticamente $P(A) = P(a_1) + P(a_2) + ... + P(a_m)$.

Para entender los conceptos que se han expuesto hasta ahora, considérese el siguiente ejemplo.

Ejemplo 6.19

Se lanzan dos dados al aire ¿Cuál es la probabilidad de que sumen 5? .

Solución

Como cada dado tiene 6 caras, la totalidad de eventos que se pueden esperar al lanzar los dos dados son $6 \cdot 6 = 36$ (recuerde la *regla del producto*).

De todos los eventos posibles al lanzar los dados, interesan solo aquellos donde se obtenga *una suma* de 5. Si llamamos al evento A a la suma al lanzar los dados, hay que calcular $P(A = 5)$.

Para hacer el cálculo anterior considere los eventos en los que que $A = 5$ mostrados a continuación:

Dado 1	Dado 2
1	4
2	3
3	2
4	1

Como cada uno de los eventos anteriores son *mutuamente excluyentes y exhaustivos* , por el teorema (6.1) se tiene que $P(A=5)=\frac{4}{36}=\frac{1}{9}$.

Si los eventos a_1, a_2,..., a_n no son sólo mutuamente excluyentes, sino también *exhaustivos*, es decir, que uno de ellos debe necesariamente de ocurrir, la probabilidad de que *uno* de ellos ocurra será igual a 1, de modo que se cumple :

$$P(a_1)+P(a_2)+...+P(a_n)=1 \tag{6.6}$$

Si el evento A puede o no ocurrir , se denotará la *no* ocurrencia del evento A como $\overline{A}$. Si A y $\overline{A}$ son eventos mutuamente excluyentes y exhaustivos entonces:

$$P(A)+P\left(\overline{A}\right)=1 \tag{6.7}$$

Para estos casos se suele usar la notación $P(A)=p$ y $P\left(\overline{A}\right)=q$ y por lo tanto:

$$p+q=1 \tag{6.8}$$

Ejemplo 6.20

En una urna hay 5 bolas blancas (B), 4 negras (N), 3 verdes (V) y 2 azules (A). Si se sacan dos bolas de la urna sin devolverlas, halle la probabilidad de $P(B)$, $P(N)$, $P(V)$ y $P(A)$.

Solución

Primero se calculará el total de permutaciones al sacar dos bolas de la urna, sin sacar una misma bola dos veces seguidas.

En la urna hay un total de 14 bolas y se van a sacar 2. Existen ${}_{14}P_2 = 182$ permutaciones diferentes . Por lo tanto hay un total de $E = 182$ permutaciones sin sacar una misma bola dos veces seguidas.

Ahora hallemos de cuántas formas se puede obtener una bola blanca en el primer o segundo intento

Si se supone que sale una bola blanca en el primer intento, entonces hay $13_{P1} = 13$ formas distintas de sacar una bola en el segundo intento. Como hay 5 bolas blancas, el número total de permutaciones al sacar una bola blanca en el primer o segundo intento sería $E_B = 5 \cdot 13 = 65$.

Mediante un razonamiento similar se obtiene que el número de permutaciones para obtener una negra azul en el primer o segundo intento serían $E_N = 52$, el número de permutaciones para sacar una bola verde sería $E_V = 39$ y el número de permutaciones para sacar una bola azul sería $E_A = 26$.

Por lo tanto :

a. $P(B) = \frac{65}{182} = \frac{5}{14}$.

b. $P(N) = \frac{52}{182} = \frac{2}{7}$.

c. $P(V) = \frac{39}{182} = \frac{3}{14}$.

d. $P(V) = \frac{26}{182} = \frac{1}{7}$.

¿Cómo puedes verificar que la respuesta anterior es correcta?

6.4.2. Probabilidad en eventos simultáneos: uso de la probabilidad condicional

Si los eventos A y B pueden ocurrir de manera *simultánea*, la probabilidad de la ocurrencia de los eventos A y B viene dada por el producto de la *probabilidad no condicional* del evento A por la *probabilidad condicional* de B, suponiendo que el evento A ocurrió.

Por *probabilidad condicional* se entiende como la probabilidad de que un evento suceda dado que otro ya ocurrió.

De esta manera la probabilidad de que los eventos A y B ocurran simultáneamente está dada por la fórmula:

$$P(AB) = P(A) \cdot P(B/A) \tag{6.9}$$

Donde:

$P(AB)$ es la probabilidad de que ocurran simultáneamente el evento A y el evento B.

$P(A)$ es la probabilidad de que ocurra el evento A.

$P(B/A)$ es la probabilidad de que ocurra B si ha ocurrido el evento A y recibe el nombre de probabilidad condicional

6.4.3. Probabilidad condicional $P(B/A)$

La probabilidad de que ocurra B sabiendo que ha ocurrido A viene dada por la fórmula:

$$P(B/A) = \frac{P(AB)}{P(A)} \tag{6.10}$$

6.4.4. La ley de la suma

Dados dos eventos A y B, la probabilidad de que suceda al menos uno de ellos está dada por la *ley de la suma*, la cual establece que:

$$P(A \vee B) = P(A) + P(B) - P(AB) \tag{6.11}$$

La notación $P(A \vee B)$ se lee " La probabilidad de obtener el evento A o B pero no ambos ".

El símbolo $\vee$ es una " o " exclusiva.

Nótese que la ecuación (6.10) no es más que un despeje de la ecuación (6.9).

Ejemplo 6.21

Con base en el ejemplo 6.20, halle :

a. La probabilidad de sacar una bola azul si se sabe que se sacó una bola verde.

b. La probabilidad de sacar una bola negra en el segundo intento si se sabe que se ha sacado una bola blanca en el primer intento, esto es, $P(N/B)$

c. $P(V \vee A)$.

Solución

a. Hay que hallar la probabilidad condicional $P(A/V)$. Como ya ocurrió el evento A se tiene que: $P(A/V) = \frac{2}{13}$.

b. Como el evento B ya ocurrió, $P(N/B) = \frac{4}{13}$.

c. Por la ecuación (6.11) se tiene que $P(V \vee A) = P(V) + P(A) - P(VA)$.

Por los resultados del ejemplo 6.20 se tiene que $P(V) = \frac{3}{14}$ y $P(A) = \frac{1}{7}$.

Haciendo uso de la ecuación (6.10) se tiene que :

$P(VA) = P(V) \cdot P(A/V) = \frac{3}{14} \cdot \frac{2}{13} = \frac{6}{182} = \frac{3}{91}$.

Por lo tanto $P(V \vee A) = \frac{3}{14} + \frac{1}{7} - \frac{3}{91} = \frac{59}{182} \approx 0.3242$.

Ejercicio para la clase

Hallar :

a. $P(B \vee N)$. R/ $\frac{97}{182}$.

b. $P(B + N)$. R/ $\frac{137}{182}$.

Primera pregunta reto

Nueve pasajeros se montan en un tren que consta de tres vagones. Cada pasajero elige al azar el vagón en el cuál desea tomar asiento. Hallar :

a. El tamaño del espacio muestral. $R/n = 6561$

b. Que tres pasajeros suban un vagón cualquiera. $R/\frac{1792}{2187}$.

c. Que tres pasajeros se monten por vagón. $R/\frac{1120}{2187}$.

d. Que dos personas suban a un vagón, tres en otro y cuatro en el último. $R/\frac{280}{729}$.

6.4.5. Definición de eventos independientes

Dos eventos A y B son independientes si se cumplen una de las siguientes dos siguientes condiciones:

1. $P(B/A) = P(B)$.

2. $P(A/B) = P(A)$.

Se interpreta de lo anterior, que si se cumple $P(B/A) = P(B)$, entonces la probabilidad condicional $P(B/A)$ puede calcularse *independientemente* del resultado que haya ocurrido del evento A.

Por lo anterior, si dos eventos A y B son simultáneos e independientes su probabilidad $P(AB)$ está dada por:

$$P(AB) = P(A) \cdot P(B) \tag{6.12}$$

Ejemplo 6.22

En una urna hay cuatro tiquetes con los números 112, 121, 211 y 222. Si se saca un tiquete de la urna, ¿cuál es la probabilidad de que el primero, segundo o tercer dígito del número en el tiquete sea un 1? (Ejemplo tomado de *Introduction to mathematical probability* de James Victor Uspensky, paginas 34 y 35. Mc Graw-Hill. New York, 1937)

.
Suponga que el evento A será el evento de sacar un 1 en el primer dígito, el evento B el de sacar un 1 en el segundo dígito y el evento Cde sacar un 1 en el tercer dígito .

1. Calcule $P(A)$, $P(B)$ y $P(C)$.

2. Calcule $P(AB)$, $P(AC)$ y $P(BC)$.

3. ¿Son los eventos A y B eventos independientes? Justifique su respuesta.

4. ¿Son los eventos A y C eventos independientes? Justifique su respuesta.

5. ¿Son los eventos B y C eventos independientes? Justifique su respuesta.

6. ¿Son los eventos A, B, y C eventos independientes? Justifique su respuesta.

Solución

1. Las probabilidades para $P(A)$, $P(B)$ y $P(C)$ serían:

a. $P(A) = \frac{2}{4} = \frac{1}{2}$

b. $P(B) = \frac{2}{4} = \frac{1}{2}$

c. $P(C) = \frac{2}{4} = \frac{1}{2}$

2. Calculemos ahora las probabilidades para los eventos simultáneos $P(AB)$, $P(AC)$ y $P(BC)$.

a. $P(AB) = \frac{1}{4}$. (solo hay un tiquete que tiene un 1 en el primer y segundo dígito de un total de 4)

b. $P(AC) = \frac{1}{4}$.

c. $P(BC) = \frac{1}{4}$.

3. Para saber si los eventos A y B son eventos *independientes*, hay que verificar que dichos eventos cumplan con la definición de *eventos independientes* , es decir, que $P(B/A) = P(B)$ o que $P(A/B) = P(A)$.

Haciendo uso de la ecuación (6.10) se tiene:

a. $P(B/A) = \frac{P(AB)}{P(A)} = \frac{1/4}{1/2} = \frac{1}{2} = P(B)$.

b. $P(A/B) = \frac{P(AB)}{P(B)} = \frac{1/4}{1/2} = \frac{1}{2} = P(A)$.

Por lo tanto, al cumplirse la definición de eventos independientes, se concluye que A y B son *eventos independientes*

4. Para saber si los eventos A y C son eventos *independientes*, hay que verificar que dichos eventos cumplan con la definición de *eventos independientes* , es decir, que $P(C/A) = P(C)$ y que $P(A/C) = P(A)$.

Haciendo uso de la ecuación (6.10) se tiene:

a. $P(C/A) = \frac{P(AC)}{P(A)} = \frac{1/4}{1/2} = \frac{1}{2} = P(C)$.

b. $P(A/C) = \frac{P(AC)}{P(C)} = \frac{1/4}{1/2} = \frac{1}{2} = P(A)$.

Por lo tanto, al cumplirse la definición de eventos independientes, se concluye que A y C son *eventos independientes* .

5. Se deja de ejercicio al lector.

6. Para verificar si los tres eventos A, B y C son *eventos independientes* , estos deben cumplir que $P(ABC) = P(A) \cdot P(B) \cdot P(C)$.

Observe primero que $P(ABC) = 0$, ya que *no existe ningún* tiquete que tenga un 1 simultáneamente en el primero, segundo y tercer dígito.

Luego:

$P(ABC) = P(A) \cdot P(B) \cdot P(C) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \neq 0$.

De esta manera se comprueba que los eventos A, B y C **NO** son eventos independientes.

6.5. Teorema de Bayes

Teorema 6.2 *Si los eventos A_1, A_2,..., A_k cumplen que $A_1 \cup A_2 \cup ... \cup A_k = S$, entonces para cualquier evento A que pertenezca a S se tiene que:*

$$P(B_r/A) = \frac{P(B_r)\cdot P(A/B_r)}{\sum_{i=1}^{k} P(B_i)\cdot P(A/B_i)}$$

El teorema 6.2 recibe el nombre de *Teorema de Bayes*.

Ejemplo 6.22

Un finquero tiene 3 lecherias B_1, B_2 y B_3. La lechería B_1 produce un 30 % de su producción total de leche, la lechería B_2 un 45 % y la lechería B_3 un 25 %. Se sabe por los registros que se llevan en la finca que hay un 2 % de probabilidad de que la leche de la lechería B_1 de positivo en la prueba de antibióticos, la lechería B_2 un 3 % y la lechería B_3 un 2 % de dar positivo en las pruebas de control de antibióticos.

.
Si se hace un muestreo al azar de control de antibióticos y da positivo, ¿cuál es la probabilidad de haya sido la lechería B_3?.

Lo primero que hay que hacer es hallar la probabilidad $P(A)$ de que una de las lecherías dé positivo en el control de antibióticos mediante la fórmula $P(A) = \sum_{i=1}^{3} P(B_i) \cdot P(A/B_i)$.

De esta manera:

$P(A) = P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) + P(B_3) \cdot P(A/B_3)$, donde :

$P(B_1) \cdot P(A/B_1) = (0.30) \cdot 0.02 = 0.006.$

$P(B_2) \cdot P(A/B_2) = (0.45) \cdot 0.03 = 0.0135.$

$P(B_3) \cdot P(A/B_3) = (0.25) \cdot 0.02 = 0.005.$

Por lo tanto $P(A) = 0.006 + 0.0135 + 0.005 = 0.0245.$

El resultado anterior se interpreta de que existe un 2.45 % de probabilidad de que alguna lechería dé positivo en el control de antibiótico.

Por último hay que hallar la probabilidad de que si el control de antibiótico dio positivo, sea de la lechería B_3, es decir, $P(B_3/A)$. Este último resultado se obtiene al aplicar el teorema 6.2. Por lo tanto se tiene que :

$P(B_3/A) = \frac{P(B_3/A)}{P(B_1)\cdot P(A/B_1)+P(B_2)\cdot P(A/B_2)+P(B_3)\cdot P(A/B_3)} = \frac{0.005}{0.025} = \frac{10}{49} \approx 0.2041.$

.
Lo anterior quiere decir que si una lechería dio positivo en el control de antiobióticos, hay un 20.41 % de probabilidad de que haya sido la lechería B_3.

Tarea moral # 6

Realizar los ejercicios 2.95 al 2.102 de la novena edición de *Probabilidad y estadística para ingeniería y ciencias* de Ronald Walpole. Pearson 2012, paginas 76 a la 77.

Capítulo 7

Regresión y correlación lineal

7.1. Concepto sobre regresión y correlación lineales

El concepto de *correlación* busca estimar si existe asociación *entre variables* de una misma unidad de estudio y de existir esta asociación *medir* su grado o intensidad.

El término *regresión* busca establecer una relación entre las variables por medio de *funciones* que permitan predecir *una* de ellas (la variable dependiente), conociendo el valor de las otras (las variables independientes).

7.2. Diagrama de dispersión

Un diagrama de dispersión representa en un plano de dos dimensiones a los pares ordenados de dos variables cuyos datos fueron obtenidos por medio de alguna técnica de muestreo estadístico.

El diagrama de dispersión lo que busca es mostrar alguna *tendencia* en la nube de datos, por ejemplo, si presenta una tendencia *lineal*, *cuadrática*, *exponencial*, etc, en su forma.

Ejemplo 7.1

Se realiza un estudio entre 30 alumnos para determinar si existe algún grado de correlación entre su coeficiente intelectual (variable independiente) y la nota final obtenida en el curso de cálculo (variable dependiente). Los resultados se muestran a continuación:

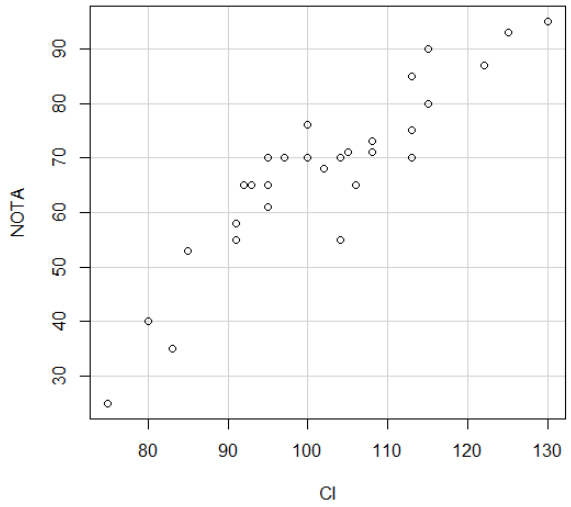
CI	*Nota final cálculo*
95	65
104	70
91	58
95	61
108	71
80	40
85	53
93	65
100	70
115	80
113	85
106	65
91	55
95	70
92	65
105	71
83	35
130	95
113	70
100	76
102	68
97	70
104	55
102	68
115	90
125	93
122	87
108	73
75	25
113	75

Fuente: Datos para el ejemplo 7.1 .

Realice mediante *R Commander* el diagrama de dispersión para los datos anteriores y estime la *tendencia* que muestran los datos. Suponga que la variable dependiente es *NOTA* y la independiente es *CI*.

Solución

El diagrama de dispersión se muestra en la figura (7.1).



Fuente: Datos ejemplo 7.1.

Figura 7.1: Diagrama de dispersión para las notas de un curso de cálculo (NOTA) en relación con con el coeficiente intelectual (CI) de 30 alumnos.

Se observa de la figura (7.1) que la nota final de curso será mayor cuanto mayor sea el coeficiente intelectual del estudiante.

7.3. Covarianza

Según Martínez (2012, p. 797) la covarianza " determina la variación conjunta de dos variables. Definido como la media aritmética del producto de las desviaciones entre los valores que toman las variables y sus medias aritméticas. Puede tomar signo positivo o negativo indicándonos una relación positiva o negativa".

Su fórmula viene dada por :

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n} - \overline{x} \cdot \overline{y} \tag{7.1}$$

Ejemplo 7.2

Hallar la covarianza del ejemplo 7.1.

Solución

Mediante el *R Commander* se obtiene la siguiente salida, según se muestra en la figura (7.2):

```
            CI     NOTA
CI    171.8862 188.1172
NOTA  188.1172 250.6713
```

Fuente: Datos ejemplo 7.1.

Figura 7.2: Salida *R Commander* para la covarianza del ejemplo 7.1

Por lo tanto el valor de la covarianza sería $S_{CI \cdot NOTA} \approx 188.11$.

7.4. Coeficiente de correlación lineal de Pearson

El *coeficiente de correlación lineal de Pearson*, denotado R , se define como el cociente de la covarianza S_{xy} entre el producto de las desviaciones estándar para ambas variables x e y. Su ecuación está dada por :

$$R = \frac{S_{xy}}{s_x \cdot s_y} \tag{7.2}$$

La ventaja del coeficiente de correlación de Pearson, o simplemente coeficiente de correlación, es que no posee unidades con lo cual no interesa las unidades de medición con las cuales se esté trabajando.

Para determinar el grado de intensidad en la correlación se utilizará el criterio dado por *Herrera y Carse* (2000), los cuales dicen que un valor de R entre 0 y 0.19 da una *correlación muy débil* (o entre -0.19 y 0 *negativamente muy débil*), si R está entre 0.20 y 0.39 entonces la *correlación es débil* (entre -0.39 y -0.20 entonces es *negativamente débil*), para el caso en que R se encuentre ubicado entre 0.40 y 0.69 se considera una *correlación moderada* (es *moderadamente negativa* si está entre -0.69 y -0.40), en caso que se R se encuentre entre 0.70 y 0.89 se dice que existe una *fuerte correlación* (si se encuentra entre -0.89 y -0.70 es *negativamente fuerte*) y finalmente si el valor de R está entre 0.90 y 1 es *muy fuerte* (entre -1 y -0.90 *negativamente muy fuerte*), recordando que, para cada uno de los coeficientes de correlación R, los valores extremos de $R = -1$ y $R = 1$ significan la correlación perfecta (o correlación *negativamente*

perfecta, según sea el caso).

Ejemplo 7.3

Calcule mediante el programa *R Commander* el coeficiente de correlación lineal de Pearson R para el ejemplo 7.1. y mencione el grado de intensidad que existe entre las variables.

Solución

Mediante el *R Commander* se obtiene la siguiente salida, según se muestra en la figura (7.3):

```
        Pearson's product-moment correlation

data:  CI and NOTA
t = 11.141, df = 27, p-value = 1.331e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8082280 0.9554306
sample estimates:
      cor
0.9062764
```

Fuente: Datos ejemplo 7.1.

Figura 7.3: Salida *R Commander* para el coeficiente de correlación de Pearson R del ejemplo 7.1

Por lo tanto, el valor para el coeficiente de correlación sería $R \approx 0.9063$, correlacionando la variable CI con la variable $NOTA$ con un grado de intensidad *muy fuerte* según el criterio de Herrera y Carse (2000).

7.5. Regresión lineal simple: el método de mínimos cuadrados

En los *análisis de regresión* se suele usar la *linea recta* por su simplicidad, además de modelar de buena manera muchos procesos y fenómenos presentes en la naturaleza. A este tipo de regresión se le llama *regresión lineal simple*

El modelo matemático que describe una relación lineal cuando se desea *estimar* la variable explicada y a través de de las variables conocidas x se da por medio de la siguiente relación:

$$\hat{y} = bx + c \tag{7.3}$$

Donde:

1. $\hat{y}$ es el *estimador* para la variable explicada y.

2. b es la pendiente de la recta de regresión, también llamado coeficiente angular.

3. c es el *coeficiente de posición* y da el valor donde la recta de regresión corta al *eje y*.

4. x es la variable *independiente o explicativa*.

En la práctica, la dificultad se centra en el criterio que permita obtener "el mejor ajuste", es decir, la recta que mejor se ajuste a los datos.

De esta manera los valores para b y c deben calcularse de modo que el valor estimado para $\hat{y_i}$ se aproxime lo más posible al valor observado y_i.

El *método* para hallar esa *recta que mejor se ajuste a los datos* recibe el nombre de *método de mínimos cuadrados*. Este método lo que pretende es minimizar las sumas de los cuadrados de las diferencias entre el valor estimado $\hat{y_i}$ y el valor observado y_i

Las siguiente es la fórmula para calcular el coeficiente b en la recta de regresión:

$$b = \frac{S_{xy}}{s_x^2} \tag{7.4}$$

Para hallar el coefiente c se usa la siguiente fórmula:

$$c = \frac{\sum_{i=1}^{n} y_i - b \cdot \sum_{i=1}^{n} x_i}{n} = \overline{y} - b \cdot \overline{x}$$

(7.5)

Ejemplo 7.4

Con base en los datos del ejemplo 7.1, calcule la recta de regresión mediante el uso de *R Commander*.

Solución

Por el ejemplo 7.3 sabemos que la correlación entre las variables CI y $NOTA$ es *muy fuerte*.

Mediante el *R Commander* se obtiene la siguiente salida, según se muestra en la figura (7.4):

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -44.06936   10.09215  -4.367 0.000167 ***
CI            1.09442    0.09823  11.141 1.33e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.936 on 27 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8213, Adjusted R-squared:  0.8147
F-statistic: 124.1 on 1 and 27 DF,  p-value: 1.331e-11
```

Fuente: Datos ejemplo 7.1.

Figura 7.4: Salida *R Commander* para la recta de regresión lineal 7.1

De la figura (7.4) se observa que $b = 1.0944$ y $c = -44.0694$. Para un mejor entendimiento supondremos que $y = NOTA$ y que $x = CI$

Por lo tanto la recta de regresión o de mejor ajuste es $\hat{y} = 1.0944x - 44.0694$.

Como $b = \frac{S_{xy}}{s_x^2}$, por los ejemplos anteriores tenemos que $S_{xy} = 188.11$ y $s_x^2 = 171.87$, con lo cual obtenemos que $b = \frac{188.11}{171.87} = 1.0944$, que es el mismo valor que se halló usando *R Comamnder*.

7.6. Coeficiente de determinación R^2

El coeficiente de correlación de Pearson al cuadrado, R^2, recibe el nombre de *coeficiente de determinación*. Diremos que existe una relación lineal entre las variables estudiadas si $R^2 \geq 0.9$. De no ser así, la relación será *no lineal* y debe usarse algún método de *regresión no lineal* en el caso que la intensidad de la correlación sea *fuerte o muy fuerte*.

Para el ejemplo 7.1 se tiene que la intensidad de la correlación entre las variables es fuerte y su $R^2 = (0.9063)^2 = 0.82$. Lo anterior sugiere que existe una fuerte correlación entre las variables mas la regresión lineal no sería la mejor opción para modelar el problema y debería estimarse utilizar alguna regresión no lineal.

7.7. Varianza residual, error estándar de estimación e intervalo de confianza

7.7.1. Varianza residual

La varianza residual se define mediante la fórmula:

$$Var_r(y) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{7.6}$$

La varianza residual $Var_r(y)$, también llamada varianza no explicada, determina cuán dispersos están los puntos con respecto a la recta de regresión lineal. Si la $Var_r(y)$ es cero, significa que todos los puntos están sobre la recta de regresión y el coeficiente de correlación R es igual a 1.

7.7.2. Error estándar de estimación

El error estándar de estimación se define como:

$$E_{est} = \sqrt{(Var_r(y))} \tag{7.7}$$

7.7.3. Intervalo de confianza

El intervalo de confianza con un nivel de confianza del 95 % para un valor específico x_k, tiene un límite inferior que está dado por la expresión :

$$l_i = \hat{y_k} - t_{\frac{\alpha}{2}(n-2)} \cdot E_{est} \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\right)} \tag{7.8}$$

Su límite superior será igual a :

$$l_s = \hat{y_k} + t_{\frac{\alpha}{2}(n-2)} \cdot E_{est} \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\right)} \tag{7.9}$$

Donde:

1. $\hat{y_k}$ es el valor estimativo de y_k.

2. $t_{\frac{\alpha}{2}(n-2)}$ es el valor t en la tabla de Student con $n-2$ grados de libertad y nivel de confianza α. El valor de $\alpha = 0.05$.

3. Los demás términos se definen como se han hecho anteriormente.

Ejemplo 7.5

Calcule la varianza residual $Var_r(y)$ y el error estándar de estimación E_{est} para los datos del ejemplo 7.1.usando *R Commander*

Usando el programa de R se tiene que $Var_r(y) \approx 43.30$. Por lo tanto $E_{est} = \sqrt{(Var_r(y))} = \sqrt{43.30} = 6.58$.

Ejercicio para la casa

Verificar los resultados del ejemplo 7.5 usando Excel.

Ejemplo 7.6

Cuál sería la nota final de curso para un estudiante con un $CI = 110$.

Solución

La recta de regresión dio $\hat{y} = 1.0944x - 44.0693$. Tomando $x = 110$ y sustituyéndolo en la recta de regresión se tiene que $\hat{y} = 76.31$.

Por lo tanto se esperaría que un estudiante con un coeficiente de 110 tenga una nota final de curso estimada de 76.31.

Ejemplo 7.7

Suponga que se desea establecer un intervalo de confianza al 95 % para estimar cuál debe ser la nota final de curso para un estudiante cuyo $CI = 110$. Determine los límites inferior y superior de dicho intérvalo.

Solución

Por el ejemplo 7.6 se tiene que $\hat{y} = 76.31$ y $x_k = 110$.

Como $n = 30$ y $\alpha = 0.05$ se tiene que el valor de $t_{\frac{\alpha}{2}\,(n-2)} = t_{0.025}$ con 28 grados de libertad. Por lo tanto $t_{0.025_{(28)}} = 2.048$.

Luego $\overline{x} = 101.9$, $\sum_{i=1}^{n} x_i^2 = 316493$, $\sum_{i=1}^{n} x_i^2 = 3057$.

Sustituyendo se tiene :

$$l_i = \hat{76.31} - 2.048 \cdot 6.58 \cdot \sqrt{\left(1 + \frac{1}{30} + \frac{(110-101.9)^2}{316493 - \frac{3057}{30}}\right)} = 62.6114.$$

$$l_i = \hat{76.31} + 2.048 \cdot 6.58 \cdot \sqrt{\left(1 + \frac{1}{30} + \frac{(110-101.9)^2}{316493 - \frac{3057}{30}}\right)} = 90.0086.$$

Por lo tanto el verdadero valor para la nota final de curso de un estudiante con un $CI = 110$ debe estar en el intervalo $[62.6114, 90.0086]$ con una confianza del 95 %, es decir, existe un 95 % de probabilidad de que un estudiante con un $CI = 110$ obtenga una nota final en el curso de cálculo entre 62 y 90 aproximadamente.

Ejemplo 7.8

Según el BCCR (*http://indicadoreseconomicos.bccr.fi.cr/indicadoreseconomicos/Cuadros/frmVerCatCuadro.aspx?idioma=1CodCuadro=%202980*) el PIB para los años 2012 a 2016 del país fueron (en billones de colones):

Año	*PIB (En billones de colones)*
2012	21.38
2013	22.75
2014	24.96
2015	26.85
2016	28.56

Fuente: *http://indicadoreseconomicos.bccr.fi.cr/indicadoreseconomicos/Cuadros/frmVerCatCuadro.aspx?idioma=1CodCuadro=%202980* .

Con base en la información anterior halle:

a. La recta de regresión lineal.

b. El valor a futuro del PIB para el año 2018.

Solución

a. Usando el programa *R Commander* se obtiene la función de regresión lineal $\hat{y} = 1.85x - 3693$.

b. El valor a futuro del PIB para el año 2018 según el resultado anterior será $\hat{y} = 1.85 \cdot 2018 - 3693 = 32.23$ billones de colones.

Obsérvese que el valor calculado mediante regresión lineal varía en menos de un 3 % según el valor esperado por el BCCR para el año 2018.

7.8. Nociones para regresiones no lineales

7.8.1. Regresión parabólica

Como se mencionó en las secciones anteriores, se debe observar la gráfica de dispersión y determinar qué tendencia muestran los datos.

Por lo general si la tendencia muestra una curva ascendente y luego una descendente o viceversa, estamos posiblemente frente a una regresión no lineal cuadrática.

Las curvas de grados mayores son poco usadas debido a la complejidad matemática que presentan, aunque se pueden obtener regresiones con polinomios de tercer grado en estudios sobre producción. (Martinez, 2012, p. 543).

Ejemplo 7.9

A continuación se muestran los datos que se obtuvieron según el precio X en dólares de un cierto modelo de teléfono celular en China y sus ventas en millones de unidades hacia los Estados Unidos.

X (En dólares)	*Y (En millones de unidades vendidas)*
97.64	5.98
92.94	14.53
90 .	22.22
85.29	29.91
80	37.62
74.12	46.15
68.82	52.99
64.12	58.97
58.24	64.95
52.35	70.94
47.06	76.07
41.76	81.20
36.47	84.62
31.18	88.89
25.88	92.31
19.41	95.73
77.65	41.03
83.53	33.33
96.47	9.40
88.82	24.78
100	0
96.17	7.27
91.95	15.76
88.98	21.82
85.17	28.48
80.93	36.36
78.81	39.39
74.58	46.06
70.33	52.12
64.41	61.21
56.78	70.30
49.58	78.18
42.80	84.85
36.86	89.70
31.78	93.33
26.27	97.58
21.18	100
61.44	64.84
68.22	55.76
53.81	73.33

Fuente: Datos tomados de *http://vivaelsoftwarelibre.com/regresion-polinomial-grafica-con-su-95-de-intervalo-de-confianza-y-prediccion-de-valores-usando-r-commander/* .

Nota al pie: Los datos fueron redondeados por mí.

a. Realice el diagrama de dispersión y determine su tendencia.

b. Calcule la ecuación para la regresión no lineal que mejor se ajuste al diagrama de dispersión anterior. Use en ambos casos *R Commander* .

c. Calcule el coeficiente de determinación múltiple R^2. Si $R^2 \geq 0.90$ la relación entre las variables se puede considerar cuadrática .

Solución

a. Como primer paso, se realizará el diagrama de dispersión de los datos para analizar la tendencia que muestran.

Puede observarse de la figura (7.5) que los datos muestran una tendencia parabólica decreciente.

b. Al realizar la aproximación mediante un polinomio cuadrático usando *R Commander* se tiene que $\hat{Y} = -0.009886X^2 - 0.007965X + 101.318990$.

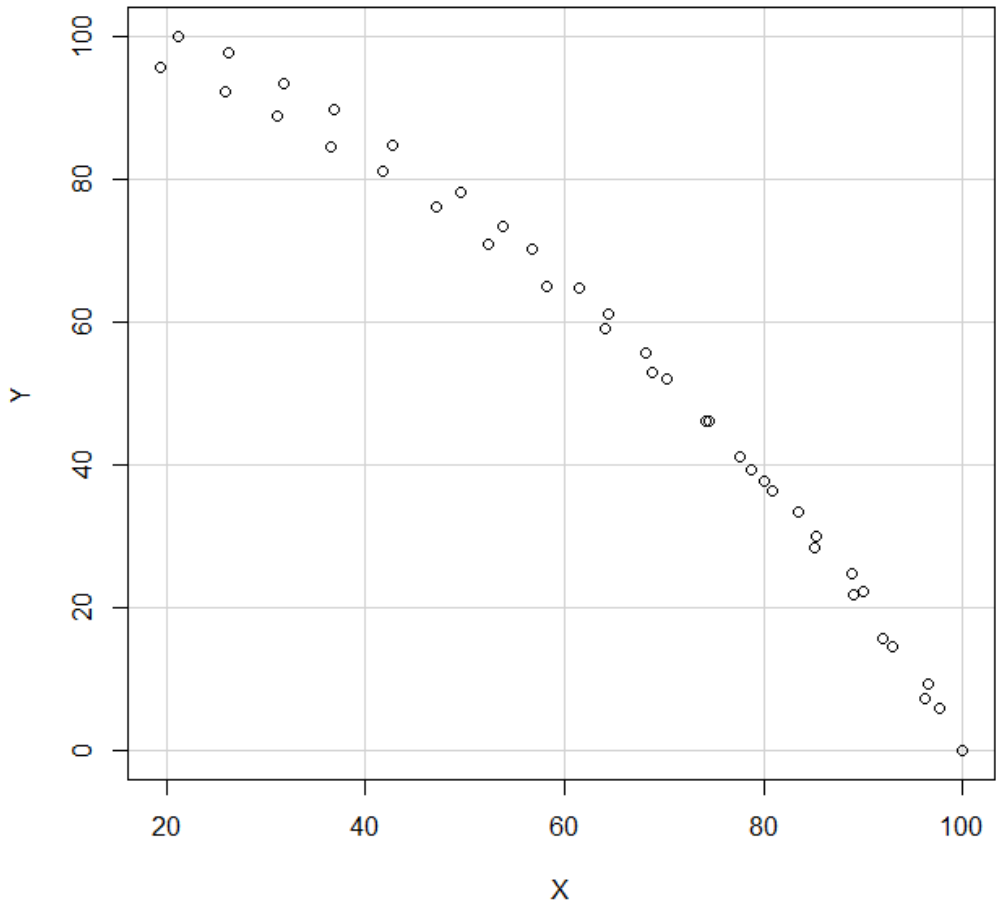
c. El coeficiente de correlación múltiple $R^2 = 0.9964$ (Calculado en *R Comamdner*). Por lo tanto la regresión anterior puede considerarse como cuadrática.

Como dato interesante un polinomio de tercer grado para los datos del ejemplo 7.9 no sería el indicado para estimar las unidades vendidas. El polinomio óptimo para el problema 7.9 sería el cuadrático obtenido en el punto *b*.

Ejercicio para la clase

a. Con base en la ecuación obtenida anteriormente (7.5), estime cuál sería el precio con el cuál el ingreso por ventas de ese modelo específico de celular serían máximas. Recuerde que $I = X\hat{Y}$, es decir, el ingreso es el precio multiplicado por la cantidad estimada de unidades vendidas $R/\$58.42$

b. Realice un análisis de regresión lineal y determine su ecuación y estime cuál sería el precio con el cuál el ingreso por ventas de ese modelo específico de celular serían máximas. Recuerde que $I = X\hat{Y}$, es decir, el ingreso es el precio multiplicado por la cantidad estimada de unidades vendidas $R/\$54.60$



Fuente: Datos ejemplo 7.9.

Figura 7.5: Diagrama de dispersión para los datos del ejemplo 7.9

Capítulo 8

Consideraciones a la hora de diseñar y analizar un diseño estadístico experimental

8.1. Créditos

Los conceptos aquí mostrados son una síntesis de *Análisis y diseño de experimentos* de Humberto Gutiérrez Pulido y Román de la Vara Salazar (2008). A ellos el crédito. (Los errores son de mi autoría).

8.2. El diseño experimental: aspectos fundamentales

Un *diseño estadístico experimental* permite determinar *qué* pruebas deben realizarse en un experimento y *cómo* han de llevarse a cabo dichas pruebas. Los diseños experimentales son usados principalmente en la industria para *resolver problemas* o *introducir mejoras* en los *procesos* o productos.

Cuando se desea mejorar un proceso se puede recurrir a la *observación*, haciendo un registro para su posterior análisis estadístico, el cual al final permitirá establecer mejoras en los procesos o productos . A este tipo de estrategia se le llama *estrategia pasiva*, ya que solo se *observa* el proceso, sin intervenir en el.

En cambio, si se provoca deliberadamente *cambios* al proceso para *identificar factores* que puedan mejorarlo, se dice que se está recurriendo a una *estrategia activa*. El diseño de experimentos consistirá en aquel conjunto de *técnicas activas* que a la hora de ser implementadas al proceso nos proporcione información que permita mejorarlo.

El uso de *métodos estadísticos* en el *diseño de experimentos* permite que la información recolectada, procesada y analizada se realice de manera *eficiente*, generando con ello nuevos conocimientos validados científicamente.

Los objetivos planteados en la investigación derivaran en una serie de *hipótesis* que especulan acerca de los posibles resultados del experimento. Si la hipótesis inicial puede ser *comparada* contra los datos obtenidos, se dice que existe un *proceso de deducción* de los datos , el cual que valida la hipótesis inicial.

Si por el contrario, la hipótesis inicial no puede ser comparada contra los datos, se dice que existe un *proceso inductivo* y la hipótesis inicial debe ser reformulada, repitiendo de nuevo el experimento, para nuevamente comparar y contrastar los datos viejos con los nuevos en un ciclo llamado *ciclo de realimentación*.

Este *ciclo de realimentación* permite hacer futuras modificaciones al proceso, generando con ello la obtención de nuevos conocimientos.

8.3. Elementos básicos en el diseño de experimentos

Se define *experimento* como aquel *cambio provocado de manera intencionada* a un proceso, el cual a su vez deberá producir un cambio sobre las *variables de respuesta* de dicho proceso. Los cambios provocados de manera intencional se hacen a través de las llamadas *variables o factores de control*. Debe tomarse en consideración que todo experimento que se realice a un proceso tiene una *variabilidad natural* que es debida a *factores propios del azar*, también llamados *factores no controlables*.

8.4. Elementos básicos de un experimento: variables de control, variables de respuesta, factores no controlables y factores estudiados.

Las *variables de control* , también llamado *factores de control*, se definen como aquellas variables que pueden *manipularse de forma intencionada y fijarse en cierto nivel durante el experimento* y son las responsables de introducir *variaciones observables durante el proceso*. Las *variables o factores de control* son los que permiten la obtención de información que lleva a la mejora de los procesos o los productos, observando el resultado que se obtiene en *las variables de respuesta*. Las variables de control se suelen denotar con la letra x

Las *variables de respuesta* se definen como el efecto medido *en los resultados* de cada prueba hecha al proceso durante el experimento. Las *variables de respuesta* son las que permiten hacer mejoras en los procesos o productos, y tienen estrecha relación con las características de calidad del proceso o producto. Se suelen denotar con la letra y.

Hay que destacar nuevamente que existen *factores o variables* que no pueden ser controlados durante el experimento. Estos *factores no controlables* introducirán *variabilidad* en el proceso. Esta variabilidad debida a estos *factores no controlables* (también llamados *factores debidos al azar*) debe ser tomada en cuenta. En los procesos industriales la variabilidad debe ser *pequeña*.

La variabilidad se define como las diferencias obtenidas al final del proceso al repetirse el experimento bajo las mismas condiciones.

Los *factores estudiados* se definen como aquellos factores o variables que han sido probados en al menos *dos niveles* durante la parte experimental y tratan de medir cómo afectan estos factores estudiados a las variables de respuesta.

8.5. Niveles y tratamientos en los experimentos

Los diferentes valores que se asignen a cada *factor estudiado* durante un experimento se llaman *niveles*.

Una *combinación de niveles* recibe el nombre de *tratamiento*.

Por ejemplo, si en un diseño experimental se tienen las variables de control *velocidad, temperatura y presión* en la fabricación remaches de acero, mientras que la variable de respuesta será la *resistencia térmica* que presenten esos remaches, los niveles y tratamientos corresponderán a los mostrados en la siguiente tabla:

Nivel velocidad	Nivel temperatura	Nivel de presión	Número de tratamiento	Resistencia térmica y
1	1	1	1	y_1
1	1	2	2	y_2
1	2	1	3	y_3
1	2	2	4	y_4
2	1	1	5	y_5
2	1	2	6	y_6
2	2	1	7	y_7
2	2	2	8	y_8

Fuente: Elaboración propia.

La tabla anterior nos dice que si se desea que cada una de las variables de control se convierta en un *factor estudiado* debe existir por lo menos *dos niveles* para cada variable de control.

Es recomendable antes de empezar propiamente con el diseño experimental, hacer un *prepilotaje*, realizando al menos dos tratamientos para un mismo nivel, y de esta forma hacer estimaciones sobre la variabilidad que podrían afectar a las variables de respuesta del proceso. Se espera que la variabilidad al repetir los tratamientos en un mismo nivel sea *pequeña*.

8.6. Errores a la hora de realizar un experimento: el error aleatorio y el error experimental

Parte de la variabilidad que se observa en las variables de respuesta está afectada, como se dijo anteriormente, por *factores no controlables*. Esta variabilidad que afecta a los resultados medidos en las variables de respuesta, no puede ser explicada por los factores estudiados y por lo tanto introduce *errores* en los resultados experimentales obtenidos. Este error debido a los factores no controlables recibe el nombre de *error aleatorio*.

El *error aleatorio* en procesos industriales debe ser *pequeño*.

El *error aleatorio* también absorberá los errores que el *experimentador* realice durante el experimento. Los errores que se comentan en las pruebas experimentales que sean propios de la persona encargada de realizarlos, recibe el nombre de *error experimental*.

Si predomina el *error experimental*, los resultados obtenidos carecerán de validez científica por su alta variabilidad y el experimento deberá desecharse o replantearse.

Existirán factores de diseño que son relativamente fáciles de controlar en un proceso, por ejemplo: temperatura de horneado, tiempo de fermentación, velocidad de un aspersor, etc. Existirán otros factores en el diseño que son difíciles de controlar, como lo son variables de tipo ambiental como la humedad, la temperatura ambiental, el humor de los operarios al hacer el proceso, etc. Estos factores que son difíciles de controlar se llaman *factores de ruido* y afectan las características de calidad del proceso o producto (variables de respuesta).

Todos los factores que afecten de manera significada a las variables de respuesta del proceso deben ser tomados en cuenta sin excepción. A este principio se le conoce como *principio de bloqueo*.

Bibliografía

[1] Gómez Barrantes, Miguel, *Elementos de estadística descriptiva*, tercera edición, UNED , San José, Costa Rica, 2011.

[2] Gutiérrez Pulido, Humberto y De la Vara Salazar, Román, *Análisis y diseño de experimentos*, segunda edición, Mc Graw-Hill , México, 2008.

[3] Herrera, J y Carse, L, «Guía de Aplicación de Pruebas Estadísticas en el Programa Systat 7.0 para Ciencias Biológicas y Forestales.», *Recuperado de http://pdf.usaid.gov/pdf$_d$ocs/PNACL878.pdf*

[4] Martínez Bencardino, Ciro, *Estadística y muestreo*, décima tercera edición, Ecoe, Colombia, Bogotá, 2012.

[5] Trejos Zelaya, Javier y Moya Vargas, Ericka, *Introducción a la estadística descriptiva*,Sello Latino, San José, Costa Rica, 2004

[6] Uspensky, James Victor, *Introduction to Mathematical Probability*, first edition, Mc Graw-Hill Book Company, USA, New York, 1937.

[7] Walpole, Ronald, Myers, Raymond y Myers, Sharon L, *Probabilidad y estadística para ingeniería y ciencias*, novena edición, Pearson, México, 2012.