Estadística para Contadores y Auditores con R

Oswaldo Navarrete Carreño María Alexandra Chavez P.

.

To my son,

without whom I should have finished this book two years earlier

Contents

List of Tables

List of Figures

¿A quién va dirigido este libro?

Este libro no es una introducción a la estadística. En la presente obra se intenta hacer un repaso de algunos temas de estadística que debe conocer quien desee hacer investigación en Contabilidad, en Auditoría o quizás en alguna ciencia social. Es probable que se omitan algunas cosas pero la retroalimentación de los lectores de esta obra será importante para su crecimiento.

En este texto se presentan, discuten y aplican los conceptos. La presentación de los conceptos es realizada pensando en un diálogo entre el autor y el lector, sin descuidar la formalidad de las expresiones matemáticas. Para la discusión y aplicación de los conceptos, se va mostrando al usuario como implementar el análisis estadístico en R.

Este libro fue desarrollado con el paquete `bookdown` que permite generar libros desde la consola de Rstudio una buena guía para empezar es **?**, la ventaja de usar este paquete es que se genera todo el contenido del libro (**?**), como índices, tablas, índices de tablas, índices de figuras mientras se desarrolla el análisis estadístico (**?**). Este paquete usa una variación de un lenguaje de marcado llamado Markdown (**?**)

Para aprovechar al máximo este libro se recomienda tener a mano una computadora con R y Rstudio instalados, a fin de poder ir ejecutando los códigos que se muestran. Los scripts y los conjuntos de datos que se presentan pueden ser descargados de `http://oswaldonavarrete.info/datos-libro/`

Aunque la obra tiene un enfoque práctico, el lector no debe olvidar que aprender a usar R no implica saber estadística y que los programas estadísticos no brindan soluciones si el usuario no conoce los conceptos que deben ser aplicados.

0.1 ¿Qué es R?

El análisis de datos requiere una inversión de tiempo considerable en obtener, fusionar, limpiar, transformar, ordenar, visualizar, analizar, modelar, evaluar y desarrollar modelos de los datos. Todas estas actividades generalmente implican interactuar con nuestros datos a un nivel sotisficado y elevado, esto implica usar un lenguaje mediante el cual expresamos nuestro trabajo. En este libro se presenta la escritura formal en un lenguaje de programación que sirve para el análisis de datos. Y es a través de este lenguaje que realizaremos el procesamiento de nuestros datos.

Los lenguajes de programación, como cualquier lenguaje, tienen sus reglas propias que muchas veces son más rígidas que las reglas de los lenguajes que usamos para comunicarnos diariamente. Estas reglas se dividen en sintaxis que se refiere a las reglas para producir oraciones y en semántica que se refiere al significado de cada palabra.

Existen muchos lenguajes de programación por ejemplo Python, Matlab, Wolfram, sin embargo en este libro se trabaja con un lenguaje de programación sencillo que ha crecido y evolucionado mucho los últimos años

y que proviene de la misma comunidad estadística.

R es un lenguaje y entorno para computación estadística y gráficos. En los últimos años el uso del programa estadístico R ha ido en aumento. Puede ser descargado de `https://cran.r-project.org/` (**?**).

En este texto suponemos que el lector no tiene conocimiento, familiaridad o experiencia con la sintaxis o la semántica de R. Estas se irán presentando a lo largo del libro, pensando exclusivamente en la programación sobre los datos. El entendimiento de la sintaxis y la semántica se dará por medio de los diferentes ejemplos que se presentan en el texto.

0.2 ¿Qué es Rstudio?

RStudio es un entorno de desarrollo integrado (IDE por sus siglas en inglés) que ayuda a explotar todas las capacidades de R. Rstudio se descarga de la página `https://www.rstudio.com/`. Puede ser instalado en casi todos los sistemas operativos de escritorio. También puede ser instalado en un servidor que funciones con el sistema operativo GNU/Linux. En la sección **??** se explica el uso de Rstudio.

0.3 Paquetes

La potencia de R nace de los usuarios y de la inmensa comunidad de usuarios que por si mismos extienden el programa, debido a su naturaleza de código abierto. Cualquier persona puede contribuir a R, mediante paquetes.

Un **paquete** es una colección de comandos para una tarea particular. Un **comando** es una instrucción en lenguaje de computadora, que sirve para indicarle a la computadora lo que debe hacer. Los paquetes generalmente necesitan de funciones o comandos de otros paquetes.

Los paquetes están disponibles principalmente en "la red completa de archivos R" (*the comprehensive R Archive Network* CRAN por sus siglas en inglés), en la actualidad existen cerca de 15 000 paquetes disponibles. Otros sitios donde se puede encontrar paquetes son el proyecto Bioconductor `http://www.bioconductor.org`, r-forge `https://r-forge.r-project.org/`, Github `https://github.com/languages/R` y la página de código de Google `https://code.google.com`

En este libro se usan los paquetes `dplyr`, `ggplot2`, `tidyr`, `BSDA`, `agricolae` y `Desctools`. Los tres primeros pertenecen al `tidyverse`, el `tidyverse` es una colección de paquetes diseñados para la ciencia de datos. Todos los paquetes comparten una filosofía de diseño, gramática y estructura subyacente de datos.

Estadística. Conceptos Básicos.

En nuestra vida diaria es común escuchar el término **estadística** en situaciones como las tasas de desempleo, el índice de pobreza, el saldo promedio de nuestra cuenta de ahorros, el número de goles realizados en la LigaPro durante el fin de semana, etc. Aunque los ejemplos descritos no corresponden a una forma incorrecta de ver las estadísticas, en este texto se pensará a la estadística como un conjunto de métodos que se utilizan para **recoger, clasificar, resumir, organizar, presentar, analizar e interpretar información numérica** (**?**).

? manifiestan que "la estadística es la ciencia de diseñar estudios o experimentos, recoger datos y modelar o analizar los datos con el propósito de tomar decisiones o realizar descubrimientos científicos cuando la información disponible es tanto limitada como variable. En otras palabras la estadística es la ciencia de *aprender de los datos*".

En las empresas la estadística es usada para tomar decisiones como los productos y las cantidades que deben ser producidas, la frecuencia con la que una maquinaria debe recibir mantenimiento, el tamaño del inventario, la forma de distribuir los productos, y casi todos los aspectos relativos a sus operaciones (**?**). En el estudio de las finanzas, la contabilidad, la economía y otras ciencias sociales la motivación para usar estadística radica en entender como funcionan los sistemas económicos, financieros o contables (**?**).

0.4 Estadística descriptiva e inferencial

La estadística puede ser usada de dos formas. La primera, cuando se describen y se presentan los datos. Y la segunda es cuando los datos son utilizados para hacer inferencias sobre características del ambiente o del entorno de donde se seleccionaron los datos o sobre el mecanismo subyacente que generó los datos. La primera forma recibe el nombre de **estadística descriptiva** y la segunda se conoce como **estadística inferencial** (**?**).

En la estadística descriptiva se utilizan métodos numéricos y gráficos para encontrar patrones y características de los datos a fin de resumir la información y presentarla de una forma significativa. Mientras que en la estadística inferencial se utilizan los datos para tomar decisiones, hacer estimaciones, pronósticos o predicciones y generalizaciones sobre el entorno del que fueron obtenidos los datos o el proceso que los generó (**?**).

Sea en estadística descriptiva o en estadística inferencial, el primer paso siempre va a ser obtener información de alguna característica, medida o valor que nos interese de un grupo de elementos. Esa característica, medida o valor de interés para el investigador recibe el nombre de **variable** (**?**).

0.5 Tipos de Variables

Muchos autores presentan algunas clasificaciones para las variables, en el texto trabajamos con una clasificación que se ajusta a las necesidades de la investigación en las áreas de nuestro interés. Según esta clasificación hay dos grandes grupos de variables: cuantitativas y cualitativas. Las primeras son las que toman valores **numéricos**. Mientras que las cualitativas toman valores que describen una **cualidad** o **categoría** (**?**).

Las variables cuantitativas se clasifican a la vez en **continuas** que se presentan cuando las observaciones pueden tomar cualquier valor dentro de un subconjunto de los números reales, ejemplos de variables cuantitativas continuas son: edad, altura, temperatura y peso. Las **discretas** son aquellas cuya característica principal es que las observaciones pueden tomar un valor basado en un recuento de un conjunto de valores enteros distintos. Ejemplos de variables cuantitativas discretas son: número de hijos, número de comprobantes de venta emitidos en un mes, número de clientes haciendo fila durante una hora en un banco, etc (**?**).

0.5.1 Niveles de medición

Las variables tienen cualquiera de los siguientes niveles de medición:

1. Ordinal
2. Nominal
3. Intervalo
4. Razón

En el nivel ordinal las observaciones toman valores que se ordenan o clasifican de forma lógica, por ejemplo las tallas de ropa (pequeña, media, grande, extra grande), la frecuencia con la que se hace una actividad (nunca, casi nunca, a veces, casi siempre, siempre). Por otro lado, en el nivel nominal las observaciones toman valores que no se pueden organizar de forna lógica, por ejemplo el sexo, el color de ojos, la marca de ropa favorita. Si se usan números en variables con nivel de medición nominal, estos números son usados solo para clasificar (**??**).

En el nivel de intervalo existe diferencia significativa entre valores pero el cero no representa la ausencia de la característica un ejemplo es la temperatura medida en grados Farenheit, el nivel de intervalo no solo clasifica y ordena las mediciones además indica que las distancias entre cada intervalo en la escala son equivalentes. Finalmente en el nivel de razón el 0 es significativo y la razón entre dos números es significativa, un ejemplo es la temperatura medida en grados Kelvin.

0.6 Otros conceptos importantes

Existen algunos conceptos que son importantes y que se deben conocer al momento de realizar análisis estadístico de datos.

- **Población**: una población es el conjunto de todos los sujetos u objetos de interés en una investigación o análisis. Por ejemplo si se desea analizar la intención de voto en una ciudad para las próximas elecciones seccionales, la población serían todas las personas en edad de votar empadronadas en la ciudad.
- **Muestra**: es la parte de la población que es analizada, dicho de otra forma una muestra es un subconjunto de la población. Sigamos con el ejemplo de la intención de voto, aunque el investigador quisiera no puede acceder a toda la población ya sea por cuestiones de tiempo o dinero y por esta razón