Práctica de Probabilidades y Estadística II (PYE2)

Inferencia Estadística

Juan A Fdez del Pozo (CIG-UPM)

10/02/2021

Outline

Outline

Introducción

Definición

- La Práctica de Probabilidades y Estadística II es un proyecto de análisis de datos mediante técnicas de inferencia estadística para obtener conclusiones.

Procedimientos de Inferencia Estadística

- Objetivos: Muestreo / Diseño_{no}
- Métodos: Paramétrico / No.Paramétrico
- Información: Clásico / Bayesiano

La Práctica consta de 5 partes

- Identificación de Modelo y Muestreo,
- Estimación Clásica (puntual, intervalos),
- Estimación Bayesiana (puntual, intervalos),
- Contrastes (paramétricos y no paramétricos) y
- Regresión lineal simple (estimación y contraste).

Introducción

Normas de Entrega del informe de las prácticas ∼ Guía de PYE2

Técnica: Trabajo en Grupo. **Evaluación**: continua y sólo prueba final No presencial. Informe y Test **Calificación**: APTO / NO-APTO. Necesario práctica APTA para aprobar la asignatura

- El trabajo se realiza en grupos de 4 alumnos. Todos los componentes de los grupos de práctica deben pertenecer al mismo grupo de clase
- Grupos: hasta el 28 de febrero, enviar un email con los datos de los miembros del grupo a <jafernandez@fi.upm.es>, se dará de alta el grupo con un número *xxx*
- Se asignará un conjunto de datos específico a cada grupo y un enunciado común para todos los grupos, PYE2DataSet(*xxx*).csv
- Entregas: Moodle, antes de la semana 15 (20/05, 23:55), doc-pdf, formato: portada (Título, fecha, id-dataset, datos de miembros del grupo (nombre, apellidos, matrícula y correo@alumnos.ump.es), índice de contenidos, figuras y tablas, Anexos: descriptiva y código). El informe de la práctica debe editarse con un editor de texto y no pueden ser realizadas a mano
- Tutorías: no presenciales, en MS Teams, con asistencia de todo el grupo

Entorno R

Datos → Entorno software

Entorno

- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`
- RStudio URL `https://rstudio.com/about/`

Sesión

- R, RStudio
- install.packages("package.name"), library(package.name), library(help=package.name)
- ?topic
- sink("resultados.txt") ... run script ... sink()
- fit <- lm(some ~ model); png(filename="your/file/location/name.png"); plot(fit); dev.off(); print(fit); summary(fit)
- q()

Paquetes: título

- base: The R Base Package (. . .)
- tidyverse: Easily Install and Load the 'Tidyverse'
- broom: Convert Statistical Analysis Objects into Tidy Tibbles
- plyr: Tools for Splitting, Applying and Combining Data
- ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics
- lattice: Trellis Graphics for R
- Rmisc: Ryan Miscellaneous (utilities)
- DescTools: Tools for Descriptive Statistics
- MASS: Support Functions and Datasets for Venables and Ripley's MASS
- car: Companion to Applied Regression
- EstimationTools: Maximum Likelihood Estimation for Probability Functions from Data Sets

Paquetes: título

- ISwR: Introductory Statistics with R
- IndependenceTests: Non-Parametric Tests of Independence Between Random Vectors
- lmtest: Testing Linear Regression Models
- boot: Bootstrap Functions
- vcd: Visualizing Categorical Data
- rcompanion: Functions to Support Extension Education Program Evaluation
- FSA: Simple Fisheries Stock Assessment Methods
- psych: Procedures for Psychological, Psychometric, and Personality Research
- e1071: Misc Functions of the Department of Statistics[a]

[a] Probability Theory Group (Formerly: E1071), TU Wien

Data-analysis: model-estimation, model-test

Referencias `http://www.r-tutor.com/r-introduction`
maximum-likelihood (ML), confidence intervals (CIs), least squares (LS) and ML, expectation-maximization (EM), Metropolis–Hasting (M–H)

- https://en.wikipedia.org/wiki/Maximum_likelihood_estimation
  https://en.wikipedia.org/wiki/Method_of_moments_(statistics)
- https://en.wikipedia.org/wiki/Student%27s_t-test,
  https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test,
  https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test
- https://en.wikipedia.org/wiki/Analysis_of_variance,
  https://en.wikipedia.org/wiki/One-way_analysis_of_variance,
  https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance,
- https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test,
  https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test,
  https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test,

Outline

Datos

Datos: PYE2DataSet(*xxx*).csv, Data ¡- read.csv(file=paste("PYE2DataSet",*xxx*,".csv",sep="")), header=TRUE)

Variables

data.frame ∼ dim(Data): 10000 x 10, names(Data):
"name", "Sex", "Nation", "sleeptime", "steps", "height", "weight" "Age",
"name": identificador-clave
"Sex" "Nation": nominales
"sleeptime", "steps", "height", "weight", "Age": intervalo

Datos

Proceso del Estudio:

- Población: 10000 filas, 8 variables
- Descriptiva de la población
- Muestras de tamaño 20:50, set.seed(2022),
- Descriptiva de las muestras y distribución en el muestreo.
- Inferencias: puntual, intervalo, Bayesiana, regresión simple

Acceso a datos

- Fuente: los datos se descargan de Moodle con el número asignado del grupo
- Formato: fichero csv con nombre de columnas
- Funciones para resumenes y gráficas. El informe incluye descriptiva del conjunto y visualización del análisis *summary, hist, steam.leaf, boxplot, plot,. . .*

Report: numeric, plot, text

Descriptiva e Informe

- summary
- hist, stem.leaf, barplot, plot, boxplot
- Interpretación, comentario, conclusión, descripción
- Funciones: sink, par(mfrow=c(1,2))

Outline

Identificación del Modelo

- Descriptiva: histograma, tallo y hoja, barra, caja
- Ajuste a la Distribución Normal, Gamma, Exponencial

— Tareas —

Descripción de variables y ajuste a un modelo de distribución. Variables: Data$sleeptime y Data$steps.

- Descriptiva: summary, hist, boxplot, skewness, kurtosis,. . . (paquete e1071)
- Ajuste: fitdistr(, c("normal","gamma","exponential")) (paquete MASS)
- $\sim$ Distribución Exponencial Gamma Normal
- $\sim$ Distribución Exponencial Gamma Normal
- Test de Kolmogorov-Smirnov: ks.test
- Gráfica: hist, dens, dens.teórica

Muestreo y Distribución Muestral

- Muestra y Población
- Muestra Aleatoria Simple: MAS
- Estratificado, Conglomerados y Sistemático
- Media, Varianza, Proporción

— Tareas —

Muestreo del conjunto de datos Data$Age. Distribución en el muestreo de media y varianza. Muestras de tamaño 20 de Data, calcular media y varianza muestrales.

- 30, 50 y 100 muestras de tamaño 20 de Data, calcular las medias y representar hist y boxplot, ajustar a la distribución normal
- idem con la Varianza muestral
- idem con la Proporción de Mujeres/Varones muestral

Métodos

- Método de Máxima Verosimilitud
- Bootstrap
- Propiedades de los estimadores

— Tareas —

Estimación clásica de media y varianza. Paquetes *stats*, *EstimationTools*. Variables Data$sleeptime y Data$step. Datos agrupados con el factor Data$Sex, con niveles {"M", "V"}.

- Estimar media y varianza de Data$sleeptime y Data$steps
- Estimar media y varianza de Data$sleeptime y Data$steps, entre las mujeres
- Estimar media y varianza de Data$sleeptime y Data$steps, entre los varones

Parámetros

- Medias, con varianza conocida y desconocida, $\sim$ normales o muestras grandes
- Medias y Proporciones en general ($\rightarrow$ paquete boot)
- Varianzas $\sim$ normales

— Tareas —

Estimación clásica del intervalo de confianza para la media, varianza, proporción, dif de medias y razón de varianzas, al nivel de confianza 90%, 95% y 99%. Variables Data$sleeptime y Data$step. [1], pág 59

- IC para medias, con varianza conocida y desconocida
- IC para varianzas

Idem en problaciones generales ($\rightarrow$ paquete boot)

Parámetros

- Diferencia de medias, con varianzas iguales y desiguales, $\sim$ normales
- Diferencia de medias y de proprociones en general ($\rightarrow$ paquete boot)
- Razón de Varianza en poblaciones normales

— Tareas —

Estimación clásica del intervalo de confianza para la media, varianza, proporción, dif de medias y razón de varianzas, al nivel de confianza 90%, 95% y 99%. Variables Data$sleeptime y Data$step.

- IC para dif de medias, con varianza conocida y desconocida
- IC para razón de varianzas

Idem en problaciones generales ($\rightarrow$ paquete boot)

Concepto y análisis secuencial

- Estimación puntual y por intervalos de una proporción
- Estimación puntual de la media, con varianza conocida y desconocida, $\sim$ normales

— Tareas —

La proporción p_e de individuos de nacionalidad española en la población está entre 25% y el 35%. En una muestra de 20 personas hay nE españoles. Suponer que la $p_e \sim \beta(\alpha = 5, \beta = 10)$, con función de densidad $f(x; \alpha; \beta) = \frac{1}{B(\alpha,\beta)x^{\alpha-1}(1-x)^{\beta-1}}$ y moda $= \frac{\alpha-1}{\alpha+\beta-1}$.

- Obtener la p_e tras la información aportada por la muestra
- Obtener el IC con 95% de confianza para la p_e
- La estatura (DAta$height) del grupo de españoles, franceses e italianos sigue una $N(170, 7)$. Estimar con la muestra anterior, la estatura con varianza conocida.

Tipos

- Contrastes para los parámetros: μ, σ, $\mu_1 - \mu_2$, σ_1^1/σ_2^2
- Contrastes en una población
- Contrastes en dos poblaciones

— Tareas —

Tomar dos muestras de tamaño 20 de Data$IMC, $Sample_1$ y $Sample_2$

- Contrastar si la media μ_1 de $Sample_1$ es $Q_1 \leq \mu_1$, con varianza desconocida
- Contrastar si la media μ_1 de $Sample_1$ es $\mu_1 \leq Q_3$, con varianza desconocida
- Contrastar si la varianza σ^2 de $Sample_1$ es mayor que 1.0, con media desconocida
- Contrastar si $\mu_1 - \mu_2 = 0$, con varianzas desconocidas
- Contrastar si $\sigma_1^2/\sigma_2^2 = 1$

Tipos

- Contrastes de Distribución: χ^2 de Pearson, Kolmogoriv-Smirnov y normalidad
- Contrastes de Independencia: identificación, rachas y autocorrelación
- Contrastes de Homogeneidad: Wilcoxon, tablas de contingencia, datos atípicos

— Tareas —

Tomar una muestra de tamaño 20 de Data$IMC, *Sample*

- Contrastar la normalidad de *Sample*, nivel de significación $\alpha = 0.05$, mediante el test de Pearson y Kolmogorov-Smirnov
- Contrastar la independencia de *Sample*, mediante el test de Durbin-Watson
- Contrastar la homogeneidad de *Sample*, mediante el test de Wilcoxon

Parte 5

Estimación del modelo de regresión simple

Regresión simple

m ¡- lm(Y $\sim$ Y), plot(m), print(m), summary(m), anova(m)

- Hipótesis (linealidad, residuos $\sim N$, homocedasticidad residuos, independencia residuos)
- Metodología, trasformaciones. Estimación y propiedades
- Predicciones e Intervalo de Confianza para predicciones

— Tareas —

Variables Data$height y Data$weight. Tomar una muestra de tamaño 20. Estimar el modelo, hacer predicciones y calcular el IC $\alpha = 0.95$ para las prediciones de min{Data$height}, media{Data$height} y max{Data$height}.

- Estimar el modelo de regresión simple de Data$weight según Data$height para los individuos en la muestra.
- Idem para los grupos de mujeres Data$Sex = M" y varones Data$Sex = V"
- Idem para Data$Age $\leq$ 30

Contrastes del modelo de regresión simple

- Linealidad, $\beta_1 \neq 0$
- Hipótesis (linealidad, residuos $\sim N$, homocedastidad residuos, independencia residuos) contrastadas con los residuos
- Interpretación

— Tareas —

Para el modelo de regresión Data$weight $\sim$ Data$height

- Contrastar la linealidad, $\beta_1 \neq 0$ en los modelos estimados (p-valor)
- Contrastar mediante el análisis de los residuos (plots)
- Interpretar el modelo: conclusiones

Outline

Referencias: Apuntes, Libros, Web, Software

Estadística. Modelos y Métodos. 1 Fundamentos.
Daniel Peña. Alianza Universidad Textos. 2 ed, 1998

SUMMARY AND ANALYSIS OF EXTENSION EDUCATION PROGRAM EVALUATION IN R.
SALVATORE S. MANGIAFICO. Rutgers Cooperative Extension. New Brunswick, NJ. VERSION 1.6.19.
http://rcompanion.org/documents/RHandbookProgramEvaluation.pdf

http://rcompanion.org/handbook/

R Reference Card by Tom Short, EPRI PEAC, tshort@epri-peac.com 2004-11-07.
Granted to the public domain. See www.Rpad.org for the source and latest version.
Includes material from R for Beginners by Emmanuel Paradis (with permission).

http://users.stat.ufl.edu/ rrandles/sta4210/Rclassnotes/introduction-to-R/R-reference-card.pdf

RStudio
https://rstudio.com/
https://rstudio.com/resources/cheatsheets/

Referencias: Apuntes, Libros, Web, Software


Summary and Analsysis of Extension Education Program Evalution in R. Salvatore S. Mangiafico. Rutgers Cooperative Extension. New Brunswick, NJ. Version 1.6.19.
http://rcompanion.org/documents/RHandbookProgramEvaluation.pdf
http://rcompanion.org/handbook/


R Tutorial
http://www.r-tutor.com/r-introduction


An R Companion for the Handbook of Biological Statistics. Salvatore S. Mangiafico
http://rcompanion.org/rcompanion/index.html


Handbook of Biological Statistics. John H. McDonald.
http://www.biostathandbook.com/index.html

Referencias: Apuntes, Libros, Web, Software

Lista de Paquetes R:

`https://cran.r-project.org/web/packages/`
base, boot, tidyverse, broom, car, EstimationTools, DescTools, e1071, FSA, ggplot2, lattice, MASS, ISwR, plyr, psych, rcompanion, IndependenceTests, lmtest, Rmisc, vcd

¿Comentarios y Preguntas?

GII - PYE2 - 2022

jafernandez@fi.upm.es