Prácticas de Probabilidades y Estadística II

Febrero - 2022

Índice

1. Resumen general del trabajo **1**
- 1.1. Normas . . . 1
- 1.2. Entorno . . . 2
- 1.3. Datos . . . 2

2. Guión **4**
- 2.1. Partes del trabajo práctico de Probabilidades y Estadística II . . . 4
- 2.2. REFERENCIAS: . . . 5

3. Estudio **6**
- 3.1. Parte 1: Identificación de Modelo y Muestreo . . . 7
- 3.2. Parte 2: Estimación Clásica (puntual, intervalos) . . . 8
- 3.3. Parte 3: Estimación Bayesiana (puntual, intervalos) . . . 9
- 3.4. Parte 4: Contrastes (paramétricos y no paramétricos) . . . 10

4. Bibliografía **11**

5. Teoría **11**

6. Estadística con R **11**

7. Entorno R **11**

1. Resumen general del trabajo

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

1.1. Normas

Técnica: Trabajo en Grupo.
Evaluación: continua y sólo prueba final No presencial. Inforem y Test
Calificación: APTO / NO-APTO. Necesario entregar la 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, las entregas se realizarán como muy tarde 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.upm.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, petición de tutoría vía email al profesor.
- Revisión: en MS Teams, con la asistencia de todo el grupo de prácticas petición de revisión vía email al profesor.

1.2. Entorno

Software recomendado:

- R Core Team (2021). 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 de R:

- R, RStudio
- `install.packages(''package.name''), library(package.name), library(help=package.name)`
- Documentación de los paquetes y funciones `?topic`
- Redireccionado de la salida de R a un fichero
  `sink(''resultados.txt'') ...run your script.R ...sink()`
- `fit` $<-$ `lm(some` $\sim$ `model);`
- Guardar graficos en ficheros: `png(filename=''your/file/location/name.png''); plot(fit); dev.off();`
- Presentar la salida en la consola de R: `print(fit); summary(fit)`
- q()

1.3. Datos

Datos: PYE2DataSet(xxx).csv,

`Data` $<-$ `read.csv( file=paste("PYE2DataSet",`xxx`,".csv",sep=), header=TRUE)`

data.frame $\sim$ `dim(Data)`: 10000 x 10, `names(Data)`:

"name", "Sex", "Nation", "sleeptime", "steps", "heigh", "weigh", "Age", donde:

"name": identificador-clave; "Sex", "Nation": nominales; "sleeptime", "steps", "heigh", "weigh", "Age": intervalo

Proceso del Estudio:

- Población: 10000 filas, 8 variables
- Descriptiva de la población, preliminar a la realización de inferencias
- Muestras de tamaño 20, (importante! $\rightarrow$) set.seed(2021) (análisis reproducible)
  `S <- sample(1:dim(Data)[1],20); Data20 <- Data[S,]`
- Descriptiva de las muestras, ajuste de distribución (estimación y gráficas) y distribución en el muestreo (un estadístico es una variable aleatoria que toma valores en diferentes muestras)

- Inferencias: puntual, intervalo, Bayesiana, regresión simple

Descriptiva e Informe:

- summary(Data); summary(Data200)
- hist, stem.leaf, barplot, plot, boxplot
- Interpretación, comentario, conclusión, descripción
- Funciones: sink, par(mfrow=c(1,2))

2. Guión

2.1. Partes del trabajo práctico de Probabilidades y Estadística II

1. Parte 1: Identificación de Modelo y Muestreo

a) Ajuste de Modelo

1) Breve descripción de las variables `sleeptime` y `steps`
2) Ajustar todos los datos de `sleeptime` a una d. Normal, una d. Gamma y una d. Exponencial. Mostrar los 3 resultados: estimadores de los parámetros respectivos, histograma de los datos con la curva de densidad del modelo correspondiente. Usar el *Test de Kolmogorov-Smirnov* para analizar el ajuste realizado (un p-value menor de 0.1 indica que el ajuste no es bueno).

b) Muestreo

1) Se toman muestras del conjunto de datos con tamaño 20 para la variable `Age`
2) Con 30, 50 y 100 muestras, es decir, calcular las 30, 50 y 100 Medias muestrales y representar `hist` y `boxplot`. Ajustar a la distribución normal cada uno de los vectores de medias (variable aleatoria muestral)
3) Con 30, 50 y 100 muestras, es decir, calcular las 30, 50 y 100 Varianzas muestrales y representar `hist` y `boxplot`. Ajustar a la distribución normal cada uno de los vectores de varianzas (variable aleatoria muestral)
4) Con 30, 50 y 100 muestras, es decir, calcular las 30, 50 y 100 proporcioń muestral de Mujeres/Varones y representar `hist` y `boxplot`. Ajustar a la distribución normal cada uno de los vectores de proporciones (variable aleatoria muestral)

2. Parte 2: Estimación Clásica (puntual, intervalos)

a) Puntual

1) Estimar media y varianza de las variables `sleeptime` y `steps`. Primero con todos los datos y segundo con una muestra de tamaño 20.
2) Estimar media y varianza de las variables `sleeptime` y `steps` entre las Mujeres. Primero con todos los datos y segundo con una muestra de tamaño 20.
3) Estimar media y varianza de las variables `sleeptime` y `steps` entre los Varones. Primero con todos los datos y segundo con una muestra de tamaño 20.

b) Intervalo

1) Estimación del intervalo de confianza para la media, varianza, proporción, al nivel de confianza 90 %, 95 % y 99 %, para las variables `sleeptime` y `step` entre según niveles {"M", "V" del factor `Sex`, con una muestra de tamaño 20. Primero suponer normalidad y segundo usar *Bootstrap* para el caso de poblaciones de distribución general o arbitraria. Para la media suponer primero varianza conocida y segundo desconocida.
2) Estimación del intervalo de confianza para la diferencia de medias, razón de varianzas, proporción, al nivel de confianza 90 %, 95 % y 99 %, para las variables `sleeptime` y `steps` según niveles {"M", "V" del factor `Sex`, con una muestra de tamaño 20. Primero suponer normalidad y segundo usar *Bootstrap* para el caso de poblaciones de distribución general o arbitraria. Para la diferencia de medias suponer primero varianzas conocidas y segundo desconocidas e iguales.

3. Parte 3: Estimación Bayesiana (puntual, intervalos)

a) 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 $f(x;\alpha;\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ y moda $= \frac{\alpha-1}{\alpha+\beta-1}$.

- Obtener la p_e tras la información aportada por la muestra (la distribución a posteriori)
- Obtener el IC con 95 % de confianza para la p_e, (los cuantiles que dejan a derecha e izquierda el 0.025 de probabilidad según la distribución a posteriori)
- La estatura (variable `height`) del grupo de españoles, franceses e italianos sigue una $N(170, 7)$. Estimar con la muestra anterior, la estatura media con varianza conocida.

En esta parte, se trata de implementar en un script de R el procedimiento de inferencia manual, con las formulas de clase y según los ejemplos de clase, no se usarán paquetes R (*https://cran.r-project.org/web/views/Bayesian.html*)

4. Parte 4: Contrastes (paramétricos y no paramétricos)

a) Tomar dos muestras de tamaño 20 de la variable `IMC`: $Sample_1$ y $Sample_2$

- Contrastar si la media μ_1 de $Sample_1$ es $Q_1 \leq \mu_1$, con varianza desconocida (Q_1 : cuartil 1 de la muestra)
- 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 $Sample_1$ y $Sample_2$ respectivamente, con varianzas desconocidas
- Contrastar si $\sigma_1^2/\sigma_2^2 = 1$, con $Sample_1$ y $Sample_2$ respectivamente

(Q_1 y Q_3 : cuartiles 1 y 3 de la muestra)

b) Tomar una muestra de tamaño 20 de la variable `IMC`, *Sample*, con nivel de significación $\alpha = 0{,}05$

- Contrastar la normalidad de *Sample*, mediante el test de Pearson y el test de Kolmogorov-Smirnov
- Contrastar la independencia de *Sample*, mediante el test de Durbin-Watson. Se trata de ver si hay dependencia de `IMC` respecto a algunas variables de conjunto de datos. Sugerencia: paquete `lmtest` y función `dwtest()`, es decir, se toma una muestra de tamaño 20 del conjunto de datos y se proponen algunas variables independientes de las que puedea depender `IMC`, y tras hacer el test se saca una conclusión.
- Contrastar la homogeneidad de *Sample*, mediante el test de Wilcoxon. Se trata de ver si varias muestras provienen de la misma población, es decir, tomamos dos muestras de tamaño 20 de la variable `IMC` (de la misma población) y tras hacer el test se saca una conclusión.
  Sugerencia: paquete `stats` y función `wilcox.test()`

Nota: en R un modelo de dependencia se define: $W \sim X + Y$, es decir, $IMC \sim height + weight$

5. Parte 5, Regresión lineal simple (estimación y contraste)

a) Estimación del modelo de regresión simple
Con las variables Data$height y Data$weight tomar una muestra de tamaño 20 y estimar un modelo de regesión simple.

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

b) Contraste de regresión
Para el modelo de regresión Data$weight $\sim$ Data$height

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

2.2. REFERENCIAS:

`https://fhernanb.github.io/Manual-de-R/`
`http://rcompanion.org/documents/RHandbookProgramEvaluation.pdf`
`http://rcompanion.org/handbook/`

3. Estudio

El estudio consiste en 4 partes. En esta sección se enuncia el análisis y la lista de tareas. El conjunto de datos se referencia como *Data* y las variables *Data$NombreVariable*. Se sugieren algunos paquetes y funciones que implementan los análisis y producen las salidas (texto, tablas, gráficas). Las salidas o resultados del análisis se deben recoger en la memoria o informe con sus correspondientes comentarios, explicaciones y conclusiones. Cada parte tiene un código (script) que debe incluirse en la memoria o informe para que se pueda reproducir el resultado, y debe ir comentado de modo eficaz para que se pueda leer. Al generar una muestra del conjunto de datos se debe fijar la semilla con `set.seed(2022)` al comienzo de cada parte.

3.1. Parte 1: Identificación de Modelo y Muestreo

Identificación del Modelo

- Descriptiva: histograma, tallo y hoja, barras, caja
- Ajustes a la Distribución: Normal, Gamma y Exponencial

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

1. Descriptiva: summary, hist, boxplot, skewness, kurtosis,... (paquete e1071)
2. Ajuste: fitdistr(x, c("normal",'gamma","exponential")) (paquete MASS)
3. Data$sleeptime $\sim$ Distribución Exponencial, Gamma, Normal
4. Data$steps $\sim$ Distribución Exponencial, Gamma, Normal
5. Test de Kolmogorov-Smirnov: ks.test
6. Gráfica: histograma de estimación de la densidad y densidad teórica de cada modelo

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

3.2. Parte 2: Estimación Clásica (puntual, intervalos)

1. Estimación Puntual
   Métodos

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

   Tareas:
   Estimación clásica de media y varianza. Paquetes stats funciones t.test() y var.test(), EstimationTools funciones maxlogL() y rcompanion. 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. Hacer las estimaciones con el conjunto de datos completo y con muestras de tamaño 20
   - Estimar media y varianza de Data$sleeptime y Data$steps, entre las mujeres. Hacer las estimaciones con el conjunto de datos completo y con muestras de tamaño 20
   - Estimar media y varianza de Data$sleeptime y Data$steps, entre los varones. Hacer las estimaciones con el conjunto de datos completo y con muestras de tamaño 20

2. Estimación por Intervalos, una población con muestras de tamaño 20
   Parámetros

   - Medias, con varianza conocida y desconocida, $\sim$ normales o muestras grandes $\rightarrow$ paquetes stats, t.test(), var.test() y rcompanion, groupwiseMean()
   - Medias y Proporciones en general ($\rightarrow$ paquete boot, boot(), boot.ci())
   - Varianzas $\sim$ normales

   Tareas: [10], pág 59
   Estimación clásica del intervalo de confianza para la media, varianza, proporción, al nivel de confianza 90 %, 95 % y 99 %. Variables Data$sleeptime y Data$step.

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

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

3. Estimación por Intervalos, dos poblaciones
   Parámetros

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

   Tareas: Estimación clásica del intervalo de confianza para la diferencia de medias, razón de varianzas, proporción, al nivel de confianza 90 %, 95 % y 99 %. Variables Data$sleeptime y Data$step entre niveles {"M", "V" del factor Data$Sex.

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

   Idem en problaciones generales ($\rightarrow$ paquete boot, función boot() y boot.ci())

3.3. Parte 3: Estimación Bayesiana (puntual, intervalos)

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 200 personas hay nE españoles. Suponer que la $p_e \sim \beta(\alpha = 5, \beta = 10)$, con función de densidad $f(x; \alpha; \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ 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 media con varianza conocida.

3.4. Parte 4: Contrastes (paramétricos y no paramétricos)

1. Contrastes Paramétricos [10], pág 477

   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 $\sigma_1^2/\sigma_2^2 = 1$

2. Contrastes No Paramétricos
   Tipos

   - Contrastes de Distribución: χ^2 de Pearson, Kolmogorov-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, $Sample$, nivel de significación $\alpha = 0{,}05$

   - Contrastar la normalidad de $Sample$, mediante el test de Pearson y el test de Kolmogorov-Smirnov
   - Contrastar la independencia de $Sample$, mediante el test de Durbin-Watson
   - Contrastar la homogeneidad de $Sample$, mediante el test de Wilcoxon

4. Bibliografía

Referencias

5. Teoría

[1] Peña, D. (2001). Fundamentos de Estadística. Alianza Editorial.

[2] Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P., Meester, L.E. (2005), A Modern Introduction to Probability and Statistics. Understanding Why and How. Springer.

[3] James, G., Witthen, D., Hastie, Tr., Tibshirani, R. (2018), An Introduction to Statistical Learning with Applications in R. Springer. https://www.ime.unicamp.br/ dias/Intoduction%20to%20Statistical%20Learning.pdf

[4] Heiberger, R.M., Hollanda, B. (2015), Statistical Analysis and Data Display. An Intermediate Course. Springer.

[5] Heumaann, Chr., Schomaker, M. (2016), Introduction to Statistics and Data Analysis with Exercises, Solutions and Applications in R. Springer.

[6] Wasserman, L. (2004), All of Statistics. A Concise Course in Statistical Inference. Springer.

[7] Fernández Cuesta, C. y Fuentes García, F. (1995). Curso de Estadística Descriptiva. Teoría y Práctica. Ed. Ariel.

6. Estadística con R

[8] Manual de R, Freddy HernÃ¡ndez, Olga Usuga, 2021-03-12, https://fhernanb.github.io/Manual-de-R/

[9] Introductory Statistics with R. Peter Dalgaard, Springer, Statistics and Computing, 2002. Paquete ISwR: https://cran.r-project.org/web/packages/ISwR/index.html

[10] 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/

[11] 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

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

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

7. Entorno R

[14] RStudio
https://rstudio.com/

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