# Práctico 1.

14 de abril de 2022

## 1. Análisis descriptivo y gráficos.

**ratings.** Cargar el paquete «languageR», que contiene la base de datos «ratings», con datos de tipos de ratings para 81 palabras sobre plantas y animales. Las variables son las siguientes: (1) Word: token (palabra); (2) Class: tipo semántico (planta o animal); (3) Length: largo de la palabra (en cantidad de letras); (4) Frequency: log(frecuencia) de la palabra; (5) meanWeightRating: rating subjetivo medio del peso del referente; (6) meanSizeRating: rating subjetivo medio del tamaño del referente; (7) meanFamiliarity: rating subjetivo medio de la familiaridad de la palabra (frecuencia subjetiva); (8) SynsetCount: logaritmo del número de Synsets (conjuntos de sinonimia) en la cual está listada en WordNet; (9) FamilySize: logaritmo de número de palabras complejas en las que ocurre; (10) DerivEntropy: entropía derivacional de la palabra (mide complejidad morfológica); (11) Complex: si la palabra es compleja o simple; (12) FreqSingular: frecuencia de ocurrecias de la palabra en singular; (13) FreqPlural: frecuencia de ocurrencias de la palabra en plural; (14) ringfl: log(FreqSingular / FreqPlural).

1. Inspeccionar el tipo de cada variable.
2. Para la variable «Length» calcular: (i) las distribuciones de frecuencias absolutas y relativas; (ii) las siguientes medidas: media; mediana; moda; media $\alpha-podada$ ($\alpha = 0{,}1$); mínimo; máximo; rango; rango intercuartil; percentiles del: 0 %, 10 %, 25 %, 50 %, 75 %, 90 %,100 %; varianza; desvío; MAD; MADN; curtosis, asimetría.
3. Usando la función «describe» del paquete «psych», calcular estadísticos descriptivos para todas las variables cuantitativas.
4. Calcular estadísticos segmentando por tipo de palabra («Class») usando: (i) la función «describeby» del paquete «psych»; (ii) la función «aggregate»; (iii) las función «tapply».
5. Graficar las variables: «Class», «complex», «Length», «Frequency», «SynsetCount», «FamilySize», «DerivEntropy». Elegir el gráfico apropiado para cada tipo de variable.
6. Realizar Boxplots para las variables: «Length», «Frequency», «SynsetCount», «FamilySize», «DerivEntropy».
7. Utilizando el comando «chart.Correlation» del paquete «PerformanceAnalytics» hacer un dispersograma para las variables: «Frequency», «FamilySize», «SynsetCount», «Length», «DerivEntropy».

**dative.** A partir de la base de datos «dative»: (i) Crear una tabla de contingencia cruzando las variables: «AnimacyOfRec» «AccessOfRec» y «RealizationOfRecipient»; (ii) Graficar dicha tabla mediante el comando «mosaic» del paquete «vcd».

**pym.** Cargar el paquete «Rling». En dicho paquete están las bases de datos «pym_high» y «pym_low». Se trata de un grupo de palabras a las que se midieron las siguientes variables: (1) syl: número de sílabas; (2) let: número de letras; (3) imag: índice subjetivo (1 a 7) de imaginabilidad (score promedio); conc: índice subjetivo (1 a 7) de concretud (score promedio); assoc: número promedio de asociaciones para cada palabra. La base «pym_high» coresponde a 50 palabras de alta frecuencia y la base «pym_low» corresponde a 51 palabras de baja frecuencia. Se quiere investigar las siguientes preguntas: (i) ¿la "asociatividad" de una palabra depende de su nivel de frecuencia?; (ii) ¿la "concretud" de una palabra depende de su nivel de frecuencia?. Para comparar ambas variables según su nivel de frecuencia inspeccionar estadísticos descriptivos y boxplots.

**GJT.** Cargar el archivo «DeKeyser_2000.RData». Mirar el box-plot de los scores de juicios de gramaticalidad (GJTScore) según la edad de los alumnos sea mayor o menor a 15 años (Status). ¿Las medianas y la dispersión de ambos grupos parecen iguales? ¿Las distribuciones parecen simétricas? ¿hay outliers?

## 2. Distribuciones de probabilidad.

1. Graficar las siguientes distribuciones: $X \sim Bi\,(8,p)$, según $p = \{0{,}1; 0{,}5; 0{,}8\}$, $Pois\left(\frac{5}{3}\right)$, $N\left(0; \sigma^2 = 0{,}81\right)$; $Gamma\,(3,2)$, $\chi^2_{(3)}$.

2. Suponiendo que una variable aleatoria se distribuye como normal estándar $N\,(0,1)$, calcular las siguientes probabilidades: $P\,(Z \geq 1{,}23)$;$P\,(Z < 1{,}23)$;$P\,(Z > -0{,}45)$;$P\,(0{,}5 \leq Z \leq 1{,}45)$;$P\,(Z = 1{,}53)$;$P\,(-1 \leq Z \leq 1)$;$P\,(-2 \leq Z \leq 2)$; $P\,(-3 \leq Z \leq 3)$.

3. Suponiendo que una variable aleatoria se distribuye como binomial $X \sim Bi\,(5; 0{,}7)$, calcular las siguientes probabilidades: $P\,(X = 3)$; $P\,(X > 3)$; $P\,(X < 3)$; $P\,(X \leq 3)$; $P\,(X \geq 3)$; $P\,(2 \leq X \leq 4)$.

4. Suponiendo que una variable aleatoria se distribuye como $X \sim Gamma\,(3,2)$, calcular las siguientes probabilidades: $P\,(X \geq 6)$; $P\,(6 \leq X \leq 8)$; $P\,(X \leq 12)$.

5. Suponiendo que una variable aleatoria se distribuye como $X \sim Pois\,(10)$, calcular las siguientes probabilidades: $P\,(X = 4)$; $P\,(X < 7)$; $P\,(X \leq 9)$; $P\,(X \geq 3)$; $P\,(X > 6)$.

6. Calcular los siguientes percentiles bajo la distribución normal estándar $N\,(0,1)$: $P\,(Z \geq z) = 0{,}005$; $P\,(Z \leq z) = 0{,}025$; $P\,(Z \geq z) = 0{,}05$; $P\,(Z \leq z) = 0{,}975$; $P\,(Z \leq z) = 0{,}5$.

7. Tomar una muestra aleatoria de $n = 300$ de las siguientes distribuciones: $N\,(0,4)$; $Exp\,(0{,}5)$; $F_{(5,4)}$; $t_{(4)}$. Graficar el histograma para cada muestra.

## 3. Distribuciones muestrales, intervalos de confianza, muestreo simple al azar.

1. En L2, la complejidad se concibe como la habilidad de usar un lenguaje más elaborado, tomando riesgos y, por ende, apelando a un conocimiento menos automatizado. Suponga que para una población de alumnos de inglés L2 un índice de complejidad sigue una distribución normal con esperanza de $\mu = 72$ y desvío $\sigma = 4$. (i) ¿Cuál es la probabilidad de que un alumno elegido al azar de esta población tenga un índice de más de *74*?; (ii) ¿Cuál es la probabilidad de que en una muestra aleatoria de *64* alumnos el promedio de los índices sea mayor que *74*?; (iii) ¿Si la distribución del índice no fuera normal cambiaría algo en la respuesta dada en (i) y (ii)?

2. En la población de alumnos de nivel intermedio en español L2 el promedio del número de palabras por unidad de habla es $\mu = 13{,}3$ con un desvío de $\sigma = 1{,}12$. (i) Si se tomaran repetidas muestras de tamaño *36* de esta población, ¿cuál es la probabilidad de que estas muestras tengan una media muestral de entre *13* y *13.6*?; (ii) Construir un intervalo para $\bar{X} = 13$ que incluya al *95 %* de las medias de muestras de tamaño *36*; (iii) ¿Cuál debería ser el tamaño de las muestras para que el *95 %* de las medias muestrales caigan en le intervalo *[13.1, 13.5]*?

3. En una población de niños de entre tres y cinco años la cantidad de errores morfológicos en sesiones observacionales se supone Normal con una media de $\mu = 16$ y un desvío de $\sigma = 2$. (i) ¿Cuál es la probabilidad de que haya alumnos con con entre *3* y *8* errores?; (ii) ¿Cuál es la probabilidad de que muestras de tamaño *5* tengan una media muestral de entre *14* y *18*?; (iii) Construir un intervalo que incluya al *95 %* de las medias de muestras de tamaño *10* para $\bar{X} = 16$.

4. Se tomaron *5* muestras de tamaño *10* de una población de niños de tres años. La siguiente tabla indica la cantidad de palabras nuevas aprendidas por los niños. (i) Asumiendo que dicha variable aleatoria tiene distribución Normal, construir cinco intervalos de confianza de nivel *0.95* para la variable a partir de cada una de las *5* muestras; (ii) supongamos de estudios anteriores se conoce que la esperanza poblacional es de $\mu = 112$. ¿cuántos de estos intervalos contienen al parámetro $\mu$?

| | m1 | m2 | m3 | m4 | m5 |
|---|---|---|---|---|---|
| 1 | 97 | 177 | 97 | 101 | 137 |
| 2 | 117 | 198 | 125 | 114 | 118 |
| 3 | 140 | 107 | 62 | 79 | 78 |
| 4 | 78 | 99 | 120 | 120 | 129 |
| 5 | 99 | 104 | 132 | 115 | 87 |
| 6 | 148 | 121 | 135 | 117 | 110 |
| 7 | 108 | 148 | 118 | 106 | 106 |
| 8 | 135 | 133 | 137 | 86 | 116 |
| 9 | 126 | 126 | 126 | 110 | 140 |
| 10 | 121 | 115 | 118 | 119 | 98 |
| $\bar{x}$ | 116.9 | 132.8 | 117 | 106.7 | 111.9 |
| $s$ | 21.7 | 32.62 | 21.28 | 14.13 | 20.46 |

Cuadro 1: Cantidad de palabras nuevas en niños de tres años.

5. Se tomó una muestra de *100* estudiantes de español L2 y se calculó un índice medio muestral de fluidez en la oralidad de *36.22* con un desvío de *0.105*. (i) Obtener un intervalo de nivel 99 % de confianza a partir de dicha muestra; (ii) ¿Qué tamaño de muestra se debería usar si se quiere que la longitud del intervalo sea a lo sumo 0.04?

6. Se sabe que en un determinado test tipo «cloze» de 50 «blancos a llenar» de español L2 los alumnos comenten un diez por ciento de errores. ¿Cuál es la probabilidad de que un alumno cometa al menos diez errores? Calcular dicha probabilidad; (i) usando la Binomial; (ii) usando la aproximación a la Normal; (iii) usando la aproximación a la Normal con corrección de continuidad.

7. Se observa que en un corpus del español (de millones de palabras), los sustantivos terminados en -*a* son femeninos en un *96.3 %*. Se extrae una muestra de 200 sustantivos terminados en -*a* y se observa que 180 son femeninos. Calcular un intervalo de confianza del 95 % para la proporción del corpus. ¿Dicha proporción está incluida en el intervalo?

8. Sea el siguiente vector que representa una población finita. (i) Seleccionar una muestra de 20 valores usando muestreo simple al azar; (ii) calcular el estimador insesgado de la media poblacional bajo este tipo de muestreo y su varianza; (iii) ¿qué tamaño muestral es necesario para tener un margen de error del *5 %* como máximo?

```
U = c(12,10,13,11,14,12,14,8,15,13,13,11,6,8,9,10,12,7,11,12,7,14,14,13,13,15,7,15,6,10,8,6,8,15,10,10,6,12,10,14,11,13,5,11,14,7,9,13,16,15)
```