

# PROVISIONAL. Taller de problemas inferencia 2023 MAT3 GIN

## Estadística

### Contenidos

<b>1 Taller INDIVIDUAL Problemas evaluable 22-23: Estadística Inferencial</b>	<b>1</b>
1.1 Problema 1: Contraste de parámetros de dos muestras. Test AB. (6 puntos)	1
1.2 Problema 2: Bondad de ajuste. La ley de Benford. (4 puntos)	2
1.3 Problema 3: Homegeneidad e independencia. (3 puntos)	2
1.4 Problema 4: Contraste de proporciones de dos muestras independientes. (3. puntos)	2
1.5 Problema 5 : Contraste de proporciones de dos muestras emparejadas. (2. puntos)	3

## 1 Taller INDIVIDUAL Problemas evaluable 22-23: Estadística Inferencial

**Cada apartado es 1 punto. Total 18 puntos**

Se trata de resolver los siguientes problemas y cuestiones en un fichero Rmd y su salida en un informe en html, word o pdf o escrito manualmente y escaneado.

### 1.1 Problema 1: Contraste de parámetros de dos muestras. Test AB. (6 puntos)

Se quiere evaluar dos interfaces gráficas para un vídeo juego la tipo A que es la actual y una nueva tipo B. Se selecciona dos muestras de jugadores independientes la primera prueba la interfaz A y la segunda la B. En cada muestra se observa el tiempo utilizado para completar una fase del juego en minutos. Las muestras son de tamaños  $n_A = 1000$  y  $n_B = 890$ .

Los datos están adjuntos a los enunciados, en la carpeta **datasets** en un ficheros **AB.csv** que contienen las variables tiempo y muestra que vale A o B.

1. Cargad de datos y calculad estadísticos descriptivos básicos y diagramas de caja e histogramas muestrales, utilizad la función **density**, comparativos de las dos muestras.
2. Estudiad si podemos aceptar que las muestras son normales con el test K-S-L, Ardenson-Darling test, Shapiro-Wilks y Dagostino-Pearson.
3. Calcular el estadístico de contraste del test K-S-L para la muestra A y comprobad el resultado.
4. Comprobad con el test de Fisher de razón de varianzas si las varianzas de las dos muestras son iguales contra que son distintas. Tenéis que resolver el test de Fisher con R y de forma manual y el de Flinger de R y decidir utilizando el  $p$ -valor.
5. Con la información anterior elegid el contraste adecuado para saber si hay evidencia de que la la nueva interfaz mejora el tiempo de la actual. Resolver manualmente definiendo adecuadamente las hipótesis y decidid según el  $p$ -valor.
6. Calculad e interpretad el intervalo de confianza de los estadísticos de los test de medias y el de Fisher de los apartados 4 y 5.

## 1.2 Problema 2: Bondad de ajuste. La ley de Benford. (4 puntos)

La ley de Benford es una distribución discreta que siguen las frecuencias de los primeros dígitos significativos (de 1 a 9) de algunas series de datos curiosas.

Sea una v.a.  $X$  con dominio  $D_X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$  diremos que sigue una ley de Benford si

$$P(X = x) = \log_{10} \left( 1 + \frac{1}{x} \right) \text{ para } x \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}.$$

Concretamente las probabilidades son

```
## [1] 0.30103000 0.17609126 0.12493874 0.09691001 0.07918125 0.06694679 0.05799195
## [8] 0.05115252 0.04575749
```

	Díg. 1	Díg. 2	Díg. 3	Díg. 4	Díg. 5	Díg. 6	Díg. 7	Díg. 8	Díg. 9
prob	0.30103	0.1760913	0.1249387	0.09691	0.0791812	0.0669468	0.0579919	0.0511525	0.0457575

En general esta distribución se suele encontrar en tablas de datos de resultados de observaciones de funciones científicas, contabilidades, cocientes de algunas distribuciones ...

1. Contrastar con un test  $\chi^2$  si el primer dígito significativo de los cubos de los números naturales del 1 al 1000 sigue esa distribución.
2. Contrastar con un test  $\chi^2$  si que el segundo dígito significativo de los cubos los números naturales del 1 al 1000 sigue una uniforme discreta de los diez dígitos del 0 al 9.
3. Calcular manualmente el estadístico y el  $p$ -valor de los dos contrastes anteriores.
4. Dibujad con R para los apartados 1 y 2 los diagramas de frecuencias esperados y observados. Comentad estos gráficos.

## 1.3 Problema 3: Homogeneidad e independencia. (3 puntos).

Queremos analizar los resultados de aprendizaje con tres tecnologías. Para ello se seleccionan grupos de 4 Grados (Grado1, Grado2, Grado3, y Grado4) de 50 estudiantes y se les somete a evaluación después de un curso que se encuentran en los datos adjuntos `datasets/tecnologias_4_grados.csv`.

Se pide

1. Discutid si hacemos un contraste de independencia o de homogeneidad de las distribuciones de las notas por tecnología. Escribid las hipótesis del contraste.
2. Interpretad la función `chisq.test` y resolved el contraste.
3. Calculad las frecuencias teóricas como producto de los vectores marginales y calculad el estadístico de contraste y el  $p$ -valor.

## 1.4 Problema 4: Contraste de proporciones de dos muestras independientes. (3. puntos)

Queremos comparar las proporciones de aciertos de dos redes neuronales que detectan si una foto con un móvil de una avispa es una [avispa velutina o asiática](#) o si es una avispa común. Esta avispa es una especie invasora y peligrosa por el veneno de su picadura. Para ello disponemos de una muestra de 1000 imágenes de insectos etiquetadas como avispa velutina y no velutina.

[Aquí tenéis el acceso a los datos](#). Cada uno está en fichero selecciona 500 fotos de de forma independiente para el algoritmo 1 y el 2. Los aciertos están codificados con 1 y los fallos con 0.

Se pide:

1. Cargad los datos desde el servidores y calcular el tamaño de las muestras y la proporción de aciertos de cada muestra.
2. Contrastad si hay evidencia de que las las proporciones de aciertos del algoritmo 1 son mayores que las del algoritmo 2. Definid bien las hipótesis y las condiciones del contraste. Tenéis que hacer el contraste con funciones de R y resolver el contraste con el  $p$ -valor.
3. Calculad el intervalo de confianza para la diferencia de proporciones **pág 187 tema 4: CH** que vimos de forma manual en teoría.

### 1.5 Problema 5 : Contraste de proporciones de dos muestras emparejadas. (2. puntos)

En el problema anterior hemos decidido quedarnos con el mejor de los algoritmos y mejorarlo. Pasamos las mismas 1000 imágenes a la version\_beta del algoritmo y a la version\_alpha. [Aquí tenéis el acceso a los datos en el mismo orden para las 1000 imágenes](#). Cada uno está en fichero los aciertos están codificados con 1 y los fallos con 0.

1. Cargad los datos desde el servidores y calcular el tamaño de las muestras y la proporción de aciertos de cada muestra.
2. Contrastad si hay evidencia de que las las proporciones de aciertos del algoritmo alfa son iguales que las del algoritmo beta. Definid bien las hipótesis y las condiciones del contraste. De forma manual como se explicó en **teoría pág 246 tema 4: CH** y resolver con el  $p$ -valor.