

**Estudi:** Grau en Llengües Aplicadas

**Nom de l'assignatura:** Mètodes Empírics per a l'Estudi del Llenguatge 2

**Codi:** 25667

**Docent Responsable:** Dr. Thomas Brochhagen

**Convocatòria:** Recuperació    **Data de l'examen:** 11/07/2022

**Curs:** 3er    **Trimestre:** 3er

**Nom i cognoms:**

---

### Fundamentos (1 punto por pregunta / 20 total)

1. De un ejemplo de una muestra en la cual el promedio y la mediana son iguales.
2. De un ejemplo de una muestra en la cual el promedio es menor que la mediana.
3. ¿Qué tipo de variable es la “edad de un sujeto” en un experimento? Justifique su respuesta.
4. ¿Qué tipo de variable es el “idioma materno de un sujeto” en un experimento? Justifique su respuesta.
5. ¿Qué tipo de variable es la “longitud de una palabra”? Justifique su respuesta.
6. Una muestra no-representativa, ¿puede ser completa? Justifique su respuesta.
7. De un ejemplo de un caso en el utilizaría el promedio para resumir una muestra, en comparación con la mediana; y vice-versa. Justifique su respuesta.
8. De un ejemplo de un caso en el cual utilizaría la desviación estándar para resumir una muestra, en vez de la varianza. Justifique su respuesta.
9. ¿Cuáles son los parámetros de la distribución Gaussiana? Explique, de manera intuitiva, lo que hace cada parámetro.
10. ¿Cuáles son los parámetros de la distribución Binomial? Explique, de manera intuitiva, lo que hace cada parámetro.
11. ¿Cuáles son los parámetros de la distribución de Poisson? Explique, de manera intuitiva, lo que hace cada parámetro.
12. ¿De qué manera se relacionan el tamaño de una muestra y el efecto que se estima a base de ella?
13. ¿En qué sentido es lineal una regresión lineal?
14. ¿Qué indica  $R^2$ ?
15. Mencione tres diferencias entre  $R^2$  y el Akaike Information Criterion (AIC).
16. Mencione tres maneras en las cuales se puede pre-procesar un corpus de texto antes de analizarlo.
17. La predicción de una regresión de Poisson, ¿puede ser 0? Justifique su respuesta.
18. La predicción de una regresión de Bernoulli, ¿puede ser 5? Justifique su respuesta.
19. La predicción de una regresión Normal, ¿puede ser 0? Justifique su respuesta.
20. Discuta una diferencia entre métodos descriptivos e inferenciales

## Caso de estudios I (3 puntos por pregunta / 30 total)

1. Quiere ver si una intervención mejora la pronunciación del dígrafo < th > en estudiantes principiantes del Inglés. Para ello, ha recogido muestras fonéticas de estudiantes del inglés de tres niveles –básico, intermedio y avanzado– antes y después de la intervención. ¿Es representativa esta muestra? Justifique su respuesta.
2. Quiere ver si una intervención mejora la pronunciación del dígrafo < th > en estudiantes principiantes del inglés. Para ello, ha recogido muestras fonéticas de estudiantes principiantes del inglés, antes y después de la intervención. ¿Es representativa esta muestra? Justifique su respuesta.
3. Después de recoger los datos, ha recodificado el resultado, codificando con 0 si no hubo mejora en la pronunciación y con 1 si la hubo. Para ver si la intervención surtió efectos, hace una regresión que tiene como predictores (i) si el estudiante participó en la intervención y (ii) a cuantas clases asistió. Inspeccione su modelo.
  - 3.1 ¿Qué tipo de regresión es?
  - 3.2 ¿Cuántos parámetros tiene el modelo?

```
##
## Call:
## glm(formula = pronunciacion ~ asistencia + intervencion, family = binomial(link = "logit"),
##      data = df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.834e-05  2.100e-08  2.100e-08  2.100e-08  3.377e-05
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -317.25  108057.12  -0.003   0.998
## asistencia      42.32   14410.03   0.003   0.998
## intervencion   211.88   71352.82   0.003   0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 7.0056e+01  on 69  degrees of freedom
## Residual deviance: 1.5039e-08  on 67  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 25
```

4. A base del modelo en (3): En palabras, ¿qué efecto tienen la intervención y el número de clases asistidas?
5. A base del modelo en (3): ¿Cual es la probabilidad de mejorar en la pronunciación del dígrafo < th > si se asiste a 8 clases y se participa en la intervención?
6. A base del modelo en (3): ¿Cual es la probabilidad de mejorar en la pronunciación del dígrafo < th > si se asiste a 10 clases pero no se participa en la intervención?
7. Para comprobar si el efecto de un predictor cambia en función a la presencia de otros, decide también hacer dos regresiones con sólo un predictor. En palabras, ¿cambió el efecto estimado de la asistencia? Si es así: ¿De qué manera?

```
##
## Call:
## glm(formula = pronunciacion ~ asistencia, family = binomial(link = "logit"),
##      data = df)
```

```
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -1.67791    0.07587    0.27765    0.54543    1.92651
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.7953      1.3796  -2.751  0.00594 **
## asistencia    0.7032      0.2023   3.475  0.00051 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 70.056  on 69  degrees of freedom
## Residual deviance: 48.548  on 68  degrees of freedom
## AIC: 52.548
##
## Number of Fisher Scoring iterations: 6
```

8. Para ver si el efecto de un predictor cambia en función a la presencia de otros, decide también hacer dos regresiones con sólo un predictor. En palabras, ¿cambió el efecto estimado de la intervención? Si es así: ¿De qué manera?

```
##
## Call:
## glm(formula = pronunciacion ~ intervencion, family = binomial(link = "logit"),
##      data = df)
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -1.23462    0.00005    0.00005    0.00005    1.12126
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.1335     0.3660   0.365   0.715
## intervencion  20.4325  2803.4177   0.007   0.994
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 70.056  on 69  degrees of freedom
## Residual deviance: 41.455  on 68  degrees of freedom
## AIC: 45.455
##
## Number of Fisher Scoring iterations: 19
```

9. ¿Cual de los tres modelos es mejor? justifique su respuesta

10. A base de tus respuestas anteriores: ¿Que sugieren los datos en cuanto a la efectividad de la intervención? ¿Debería implementarse en futuros cursos de inglés? Justifique su respuesta.

## Caso de estudios II (3 puntos por pregunta / 30 total)

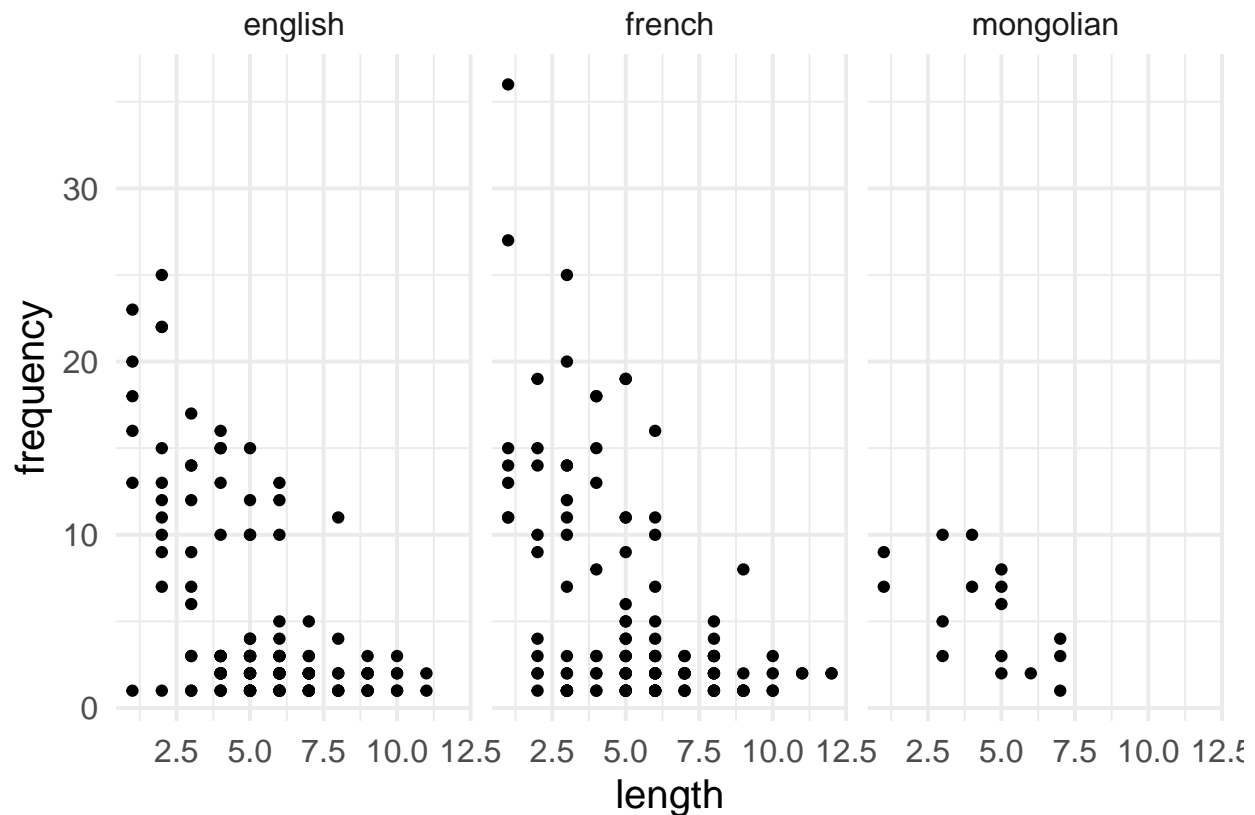
Para este caso de estudios, los datos fueron recogidos de corpora de texto en tres idiomas: inglés, francés, y mongol. Después de normalizar los datos, para cada idioma, se midió la frecuencia de los lexemas en el corpus

y su longitud para ver si la Ley de Abreviación de Zipf se cumple a través de diferentes lenguajes naturales.

Aquí hay una muestra de los datos:

```
## frequency length language
## 1      1      5 english
## 2      3      6 english
## 3      2      9 english
## 4      3      6 english
## 5      4      8 english
## 6      1      5 english
```

Y aquí hay una visualización de los datos:



1. A base de la visualización: ¿Se parece cumplir la Ley de Abreviación de Zipf en las tres lenguas? Justifique su respuesta.
2. Para comprobar si se cumple la Ley de Abreviación de Zipf, hace una regresión con lengua y longitud como predictores. Según el modelo, ¿se cumple la Ley de Abreviación de Zipf en las tres lenguas? Justifique su respuesta.

```
##
## Call:
## lm(formula = frequency ~ length + language, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.5504 -3.4987 -0.9764  2.2826 25.1897
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      11.8127      0.7947  14.865 <2e-16 ***
## length           -1.2623      0.1238 -10.194 <2e-16 ***
## languagefrench    0.2600      0.6114   0.425  0.671
## languagemongolian -0.7736      1.3331  -0.580  0.562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.023 on 282 degrees of freedom
## Multiple R-squared:  0.2696, Adjusted R-squared:  0.2618
## F-statistic: 34.69 on 3 and 282 DF, p-value: < 2.2e-16
```

3. Para ver si el efecto de un predictor cambia en función a la presencia de otros, decide también hacer dos regresiones con sólo un predictor. En palabras, ¿Cambió el efecto estimado de la lengua? Si es así: ¿De qué manera?

```
##
## Call:
## lm(formula = frequency ~ language, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.438 -4.015 -3.015  1.800 30.800
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.0148     0.5048   9.934 <2e-16 ***
## languagefrench  0.1852     0.7139   0.259  0.796
## languagemongolian 0.4227     1.5508   0.273  0.785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.865 on 283 degrees of freedom
## Multiple R-squared:  0.0004067, Adjusted R-squared: -0.006658
## F-statistic: 0.05757 on 2 and 283 DF, p-value: 0.9441
```

4. Para ver si el efecto de un predictor cambia en función a la presencia de otros, decide también hacer dos regresiones con sólo un predictor. En palabras, ¿Cambió el efecto estimado de la longitud? Si es así: ¿De qué manera?

```
##
## Call:
## lm(formula = frequency ~ length, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.592 -3.324 -1.070  2.184 25.408
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.8457     0.7228  16.39 <2e-16 ***
## length       -1.2537     0.1230 -10.19 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.011 on 284 degrees of freedom
## Multiple R-squared:  0.2679, Adjusted R-squared:  0.2653
```

## F-statistic: 103.9 on 1 and 284 DF, p-value: < 2.2e-16

5. Compara los tres modelos en (2), (3) y (4). ¿Cual es el mejor modelo? Justifique su respuesta
6. A base de sus respuestas a las preguntas (1) y (5): ¿Qué aprendemos sobre si hay una diferencia entre lenguas en cuanto a su cumplimiento de la Ley de Abreviación de Zipf?
7. ¿Cuál es el error de predicción promedio del modelo en (2)?
8. ¿Cuánta varianza explica el modelo en (2)? ¿Es una cantidad aceptable? Justifique su respuesta
9. ¿Qué frecuencia predice el modelo de (2) para una palabra de 5 caracteres de longitud en mongol?
10. ¿Qué frecuencia predice el modelo de (4) para una palabra de 5 caracteres de longitud? ¿Es esta una predicción más acertada para datos del inglés o para datos del mongol? Justifique su respuesta.

### Teoría (5 puntos por pregunta / 20 total)

1. Explique, de manera intuitiva, cada una de las leyes de Zipf. Ofrezca una hipótesis de lo que causa cada una.
2. Mencione tres posibles causas por las cuales el estimado de un predictor puede indicar que no tiene un efecto sobre el resultado. Para cada posible causa, ofrezca una manera de corroborar si esta podría ser la causa.
3. Explique dos maneras de estimar el tamaño de muestra necesario para comprobar si un predictor tiene un efecto sobre un resultado.
4. Explique por qué hay que utilizar diferentes tipos de regresión, por ejemplo, Poisson, Gaussiana o Bernoulli, para diferentes tipos de datos.