

ESTADÍSTICA II (2021/22). Grado en Matemáticas
Examen final, 17 de enero de 2022

Nombre: _____

**En todos los contrastes de hipótesis se deberán plantear claramente las hipótesis nula H_0 y alternativa H_1 .
La nota máxima de este examen son 10 puntos.**

Problema 1: (2.5 puntos) Contrastar a nivel $\alpha = 0.1$ si la muestra

$$0.14 \quad -1.12 \quad -0.81 \quad 1.81$$

procede de una distribución $N(0,1)$.

Problema 2: (2 puntos) Sea $\mathbf{X} = (X_1, X_2, X_3)'$ un vector aleatorio normal tal que la correlación entre X_1 y X_2 es $\rho_{12} = 0.5$,

$$\begin{aligned} X_2|X_3 &\sim N(2X_3, 24), \\ X_3|X_1 &\sim N(2X_1 + 3, 14), \\ X_1 &\sim N(1, 4). \end{aligned}$$

Determinar la esperanza y la matriz de covarianzas de \mathbf{X} .

Problema 3: (1 punto) Para un vector de respuestas $\mathbf{Y} = (Y_1, \dots, Y_n)'$, se considera el siguiente modelo de regresión

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$, \mathbb{X} es la matriz del diseño y el vector de perturbaciones $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ satisface que $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ y $\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, siendo \mathbf{V} una matriz definida positiva conocida.

a) (0.5 puntos) Probar que

$$\hat{\boldsymbol{\beta}}_w = (\mathbb{X}'\mathbf{V}^{-1}\mathbb{X})^{-1}\mathbb{X}'\mathbf{V}^{-1}\mathbf{Y}$$

minimiza el criterio de mínimos cuadrados *ponderados* $\|\mathbf{V}^{-1/2}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta})\|^2$.

b) (0.5 puntos) Para $\hat{\sigma}^2 = (\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}_w)'\mathbf{V}^{-1}(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}_w)/(n - k - 1)$, probar que $(n - k - 1)\hat{\sigma}^2/\sigma^2$ sigue una distribución χ_{n-k-1}^2 .

Indicación: Se consideran probados todos los resultados teóricos que enunciamos en clase. No hace falta demostrarlos de nuevo y se pueden utilizar en la resolución del ejercicio.

Problema 4: (2 puntos) En la base de datos DASL (Data and Story Library) aparece un conjunto de datos que contiene la distancia de frenado (medida en pies) de un coche observada 3 veces para cada una de 5 velocidades diferentes (medidas en millas por hora). El objetivo era predecir la distancia de frenado en función de la velocidad. Al analizar los datos con R se obtuvieron los siguientes resultados:

```
reg <- lm(Stopping_Distance~Speed_mph,data=Datos)
summary(reg)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-65.933	12.475	-5.285	0.000148
Speed_mph	5.977	!!!!	!!!!!!	!!!!!!!

Residual standard error: 16.11 on 13 degrees of freedom

Multiple R-squared: 0.9695, Adjusted R-squared: 0.9671

Calcula un intervalo de confianza al 95% para la pendiente de la recta de regresión. A un nivel de significación del 5%, ¿podemos rechazar la hipótesis de que la velocidad no influye sobre la distancia de frenado?

Problema 5: (2.5 puntos) Cinar y Koklu (2019)¹ analizan dos variedades de arroz cultivadas en Turquía: Osmancik y Cammeo. En una imagen digitalizada de cada grano de arroz muestreado se midieron siete características morfológicas:

- **Area:** Número de píxeles dentro de la frontera del grano.
- **Perimeter:** Longitud aproximada de la frontera.
- **Major Axis Length:** Longitud del mayor eje de la elipse que mejor se ajusta a la imagen del grano.
- **Minor Axis Length:** Longitud del menor eje de la elipse.
- **Eccentricity:** Excentricidad de la elipse.
- **Convex Area:** Número de píxeles del cierre convexo de la imagen.
- **Extent:** Ratio del área de la imagen respecto a la del menor rectángulo que contiene al grano.

Al analizar estos datos con R se obtuvieron los siguientes resultados.

```
Datos <- read_excel("Rice.txt", header=TRUE)
Y <- Datos$CLASS == "Cammeo"
X <- Datos[,1:7]
ajuste <- glm(Y~AREA+PERIMETER+MAJORAXIS+MINORAXIS+ECCENTRICITY+CONVEX_AREA+EXTENT,
              data=Datos,family="binomial")
summary(ajuste)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-37.219666	53.085143	-0.701	0.48322
AREA	-0.006202	0.001541	-4.025	5.70e-05 ***
PERIMETER	-0.091859	0.030225	-3.039	0.00237 **
MAJORAXIS	0.151076	AAAA	BBBB	0.15306
MINORAXIS	-0.275577	0.263244	-1.047	0.29517
ECCENTRICITY	33.323432	52.242133	0.638	0.52356
CONVEX_AREA	0.009477	0.001447	6.551	5.73e-11 ***
EXTENT	-0.265793	0.964712	-0.276	0.78292

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5202.1 on 3809 degrees of freedom
 Residual deviance: 1356.3 on 3802 degrees of freedom
 AIC: 1372.3

- a) (1 punto) Calcular los términos AAAA y BBBB. Extraer conclusiones del contraste correspondiente.
- b) (1 punto) ¿Qué contraste resuelve el siguiente código? Determinar CC y DDDD. Decidir si se rechaza o no la hipótesis nula al nivel del 5%.

```
ajuste2 = glm(Y~AREA+PERIMETER+CONVEX_AREA,data=Datos,family="binomial")
anova(ajuste2,ajuste)
```

Analysis of Deviance Table

Model 1: Y ~ AREA + PERIMETER + CONVEX_AREA

Model 2: Y ~ AREA + PERIMETER + MAJORAXIS + MINORAXIS + ECCENTRICITY +
 CONVEX_AREA + EXTENT

	Resid. Df	Resid. Dev	Df	Deviance
1	3806	1516.2		
2	3802	1356.3	CC	DDDD

- c) (0.5 puntos) Explica para qué sirve el criterio AIC, calcula su valor para `ajuste2`, compáralo con el AIC de `ajuste` y extrae conclusiones. ¿Es coherente con la conclusión obtenida en el apartado (b)?

¹Cinar, I. and Koklu, M. (2019). Classification of Rice Varieties Using Artificial Intelligence Methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7, 3 188–194.