NOM:_____COGNOMS:_____

(***Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs***)

Problema 1 (B1-B2)

Totes les qüestions valen 1 punt

Malgrat els esforços de Twitter per controlar els missatges generats per robots (*bots*), cada cop més estan presents a la xarxa. Per intentar identificar de forma més precisa aquest missatges, Twitter ha contactat amb 2 empreses externes (Emp1 i Emp2) que es dediquen a detectar aquests missatges. La probabilitat de que un *tweet* (missatge) hagi estat generat per un Bot (B) és de **0.1**.

A més tenim les següents probabilitats:

- La probabilitat que l'empresa 1 classifiqui un missatge com a bot (B1) és de **0.17**
- La probabilitat que l'empresa 1 classifiqui un tweet generat per un bot (B) com a bot (B1) és de **0.8**
- La probabilitat que l'empresa 2 classifiqui un tweet generat per un bot (B) com a bot (B2) és de **0.78**
- La probabilitat que un tweet sigui un bot (B) i, a més, ambdues empreses el classifiquin com a bot és de **0.076**
- En un tweet que no és un bot ($\bar{B}$), la probabilitat que l'empresa 1 el classifiqui com a bot (B1) i l'empresa 2 (B2) el classifiqui com no bot ($\overline{B2}$) és **0.095**.
- En global, la probabilitat d'encert (correcta classificació) d'Emp2 és **0.0754** superior a la probabilitat d'encert d'Emp1.

Pista: Construeix l'arbre de probabilitat i afegeix valors a mesura que vas fent els següents apartats.

1. Quina és la probabilitat que l'empresa 1 classifiqui com a *bot* ($B1$) un tweet que no havia estat generat per un bot ($\bar{B}$)?

2. Quina és la probabilitat que un tweet generat per un bot (B) i que l'empresa 1 l'ha classificat com un bot (B1), l'empresa 2 també el classifiqui com un bot (B2)?

3. Quina és la probabilitat que un tweet sigui un bot i l'empresa 1 el classifiqui malament ($\overline{B1}$) però l'empresa 2 el classifiqui correctament (B2)?

4. Quina és la probabilitat que l'empresa 2 classifiqui correctament ($\overline{B2}$) un tweet que no és un bot ($\bar{B}$)?

5. Un tweet concret, l'empresa 1 el classifica com a bot (B1) i l'empresa 2 el classifica com a no bot ($\overline{B2}$). Quina és la probabilitat que realment no provingui d'un bot ($\bar{B}$)?

Un aspecte important a tenir en compte a l'hora de valorar si un *tweet* ha estat generat per un *bot* és el temps que fa que l'usuari que va publicar el *tweet* va crear el seu compte. Aquest temps (en mesos) es distribueix segons la funció de densitat descrita a continuació:

$$f(t) = k \cdot e^{-\frac{t}{100}} \qquad per\ 0 < t < 150$$

6. Troba el valor de la constant *k* per a que *f(t)* sigui una funció de densitat.

7. Un algoritme classifica com a probable usuari "fals" aquell que fa menys de mig any que va crear el compte. Calcula la probabilitat de que un usuari estigui classificat com a tal.

Les dues empreses de l'inici cobren per tweet analitzat. El preu per tweet analitzat depèn del nombre de tweets generats per bots reals detectats i va des de 0.01 € a 0.03 € i. La funció de probabilitat conjunta dels costos unitaris de cada empresa estan a la següent taula.

Costos Unitaris		**Empresa 2**		
		0.01 €	**0.02 €**	**0.03 €**
Empresa 1	**0.01 €**	0.1	0.05	0
	0.02 €	0.1	0.2	0.15
	0.03 €	0	0.1	0.3

8. Troba el cost unitari (per tweet) esperat per ambdues companyies. (4 decimals de precisió)

9. Twitter encarrega l'anàlisi d'1 milió de Tweets a l'empresa 1 i de 2 milions de tweets a l'empresa 2. Troba el cost unitari (per tweet) esperat dels 3 milions de tweets. (4 decimals de precisió)

10. Sense fer cap càlcul digues si creus que la covariància entre aquestes dues variables és positiva, negativa o nul·la i argumenta-ho.

NOM:_____COGNOMS:_____

(***Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs***)

Problema 2 (B3-B4)

Per decidir sobre normatives de control de contaminació un dels elements més controvertits són els diversos indicadors del nivell d'emissions de gasos contaminants dels vehicles. Una determinada marca indica per a un del seus models que la mesura d'opacitat (absorció en un filtre en acceleració forçada seguint una escala anomenada de Bacharach, en la qual quant més alt més contaminant) segueix una distribució Normal centrada a 1.5 i amb desviació 0.6

1.- (0.5 punts) Calculeu la probabilitat que un d'aquests vehicles doni una mesura d'opacitat per sobre de 2

2.- (1 punt) Indiqueu el valor d'opacitat pel qual s'assegura que un 97.5% d'aquests vehicles no el superarà. I entre quins valors, centrats en la mitjana, es pot assegurar la opacitat amb una probabilitat del 95%

3.- Habitualment, enlloc de fer una sola mesura, es realitzen repetides mesures independents. Si en realitzem 4, indiqueu:

a) (1 punt) la probabilitat que la mitjana aritmètica d'aquestes 4 repeticions doni un valor d'opacitat mitjana per sobre de 2 (expliciteu la variable, model i paràmetres que useu pel càlcul d'aquesta probabilitat i compareu-la amb la de l'apartat 1)

b) (1 punt) la probabilitat que en la mitat d'aquestes repeticions la mesura estigui per sota de 2 (expliciteu la variable, model i paràmetres que useu pel càlcul d'aquesta probabilitat)

4.- (0.5 punts) Ara ens interessa el nombre de repeticions que cal fer fins obtenir mesures d'opacitat superior a 2. Indiqueu la variable, model, paràmetres, esperança i variància de la variable nombre de repeticions fins a una primera mesura d'opacitat superior a 2. I el mateix per la variable nombre de repeticions fins a dues mesures d'opacitat superiors a 2

5.- (1 punt) En una determinada estació d'ITV tenen estudiat que vehicles amb mesures molt altes d'opacitat (superiors a 5) cada any en passen una mitjana de 8. Calculeu la probabilitat que un any en passin només 6, i l'esperança del nombre de mesos en que no hi passi cap d'aquests vehicles

6.- Es decideix prendre 17 mesures a un vehicle per fer inferència del seu nivell d'opacitat, i els resultats han estat:
Opa = c(0.2, 0.6, 0.8, 0.9, 1.2, 1.3, 1.5, 1.5, 1.5, 1.5, 1.6, 1.6, 2.0, 2.0, 2.2, 2.3, 2.8) sum(Opa) = 25.5 sum(Opa*Opa) = 45.07

a) (1 punt) Calculeu una estimació puntual de la mitjana, la desviació i l'error estàndard. Interpreteu-los.

b) (1 punt) Calculeu una estimació per interval amb una confiança del **95%** per a la mitjana de la opacitat. Interpreteu-lo.

c) (1 punt) Quantes mesures hauríem de repetir si haguéssim volgut una amplada de 0.5 (0.25 cada costat de l'interval) per a l'interval de confiança de la mitjana d'opacitat al 95% i suposant una desviació poblacional coneguda i igual a 0.7.

d) (1 punt) A partir d'aquestes 17 dades i sabent que un valor inferior a 2.0 equival a una marca positiva, i si no a una de negativa, calculeu un interval de confiança al 95% de la proporció de positius.

e) (1 punt) Realitzeu la prova amb risc 5% de si la proporció de positius es d'un 50% o superior i indiqueu: hipòtesis, estadístic, punt/s crític/s, resolució i conclusió.

NOM:_____COGNOMS:_____

(***Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs***)

Problema 3 (B5-B6)

Uns alumnes estudien les diferències de la valoració donada a (A) jocs de pagament i (B) jocs gratuïts, prenent una mostra (seleccionats de la recopilació feta per la revista "Supergamer") de 30 jocs de cada classe que tinguin almenys 50 valoracions dels usuaris. Els usuaris valoren els jocs en una escala entera de 1 a 5, i aquests alumnes recullen el nombre de valoracions i la valoració mitjana (Y) per cada joc.

A: $\sum y_{A,i} = 136.0517$ $\sum y_{A,i}^2 = 617.0739$ B: $\sum y_{B,i} = 135.2725$ $\sum y_{B,i}^2 = 610.0058$

L'objectiu és comprovar si els jocs de pagament o gratuïts tenen diferent valoracions. Expliqueu pas a pas i detalladament com arribeu a la resposta, que ha d'incloure a més a més un interval de confiança al 95% i la seva interpretació. (3pt)

Calia posar un nombre mínim de valoracions als jocs que anaven a ser inclosos a l'estudi? Quina relació té aquest criteri amb les premisses de la prova que heu fet anteriorment? (1pt)

Un dels autors de l'estudi escriu al final: "El valor P obtingut, 4%[1], ens diu que la probabilitat d'equivocar-nos rebutjant la igualtat de mitjanes és només un 4%". Podeu comentar si aquesta afirmació és correcta o tractar de millorar-la? (1pt)

[1] Suposem que aquest hagués estat el valor P a la solució.

La segona part de l'estudi tracta de veure si un joc amb més valoracions té millor (o pitjor) valoració mitjana. S'agafa la mostra de jocs gratuïts, dels quals s'obtenen uns estadístics del logaritme natural del nombre de valoracions (X), ja que hi ha grans diferències entre uns jocs i altres:

$$\bar{x} = 6.489353 \quad \max = 8.451694 \quad s_x = 1.337217 \quad \text{cov} = 0.01695832$$

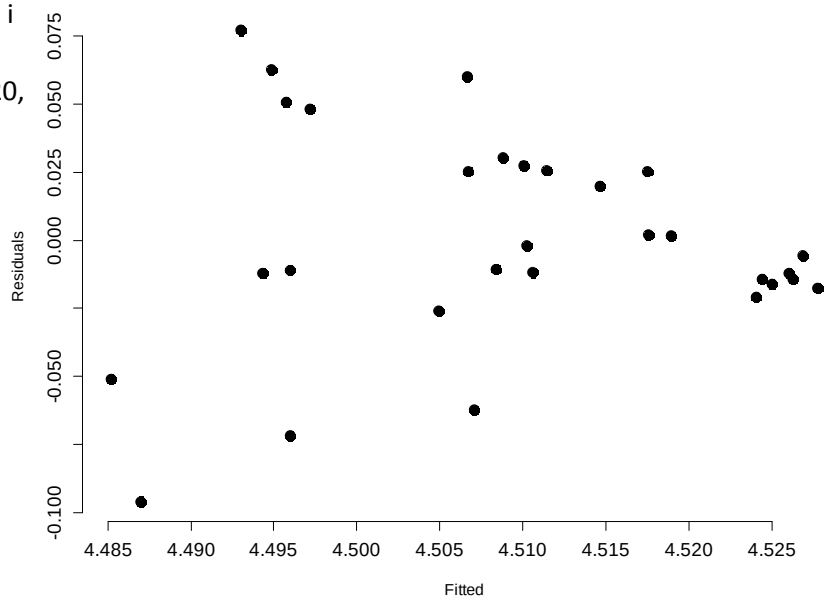
(1pt) Amb aquesta informació, justifiqueu que l'equació de la recta que relaciona X amb Y és: Y = 4.44754 + 0.009484 X

(0.5pt) Si es duplica el nombre de valoracions d'un joc, quin increment (puntual) podem esperar a la variable resposta?

(1pt) Trobeu l'error tipus associat al pendent, i resoleu la prova d'hipòtesis per determinar si amb més valoracions esperem una valoració del joc més alta. Estimeu per IC al 95% de confiança el pendent.

(1pt) Quant val el coeficient de determinació d'aquest estudi, i com s'interpreta?

(1pt) El joc nº 20 de la mostra va tenir 1590 valoracions i una mitjana de 4.542767. Amb aquesta informació, localitzeu al gràfic del costat quin punt és el del joc nº 20, justificant la resposta.



(0.5pt) A partir del gràfic, comenteu les premisses del model que veieu.