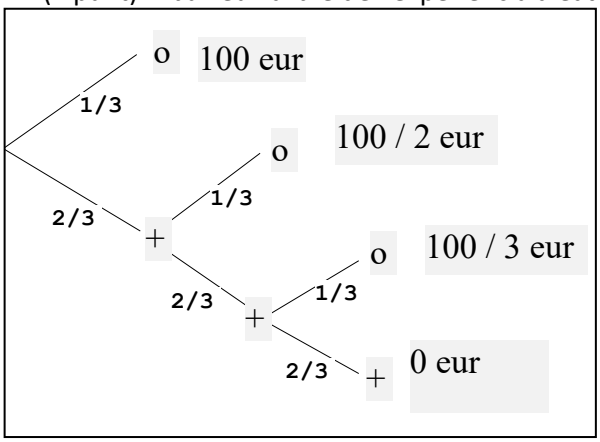
NOM:_____COGNOM:_____

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

Problema 1 (B1 B2)

Volem estudiar les probabilitats i el que es pot guanyar en un joc que consisteix en llençar una moneda desequilibrada (creu surt el doble de cops de cara), un màxim de tres intents, seguint els següents passos: si surt cara en el primer intent es guanyen 100 eur i acaba el joc, i si surt creu hi ha un segon intent; si surt cara en el segon intent es guanya la meitat i si surt creu hi ha un tercer intent; si surt cara en el tercer es guanya una tercera part, i, per tant, si surten tres creus no es guanya res.

1.- (1 punt) Dibuixeu l'arbre de l'experiència aleatòria i expliciteu el conjunt de resultats i les seves probabilitats



Ω = { o, +o, ++o, +++} o bé Ω = { o_100eur, +o_50eur, ++o_33.33eur, +++_0eur}

Amb probabilitats P(o)=1/3=0.333 P(+o)=2/3*1/3=2/9=0.222 P(++0)=2/3*2/3*1/3=4/27=0.148 P(+++)=$(2/3)^3$=8/27=0.296

2.- (1 punt) Calculeu (expressant-ho formalment) la probabilitat de guanyar el màxim i la de no guanyar res

P("màxim") = P(o) = **0.33**

P("res") = P(+++)=$(2/3)^3$=8/27 = **0.296**

3.- (1 punt) Calculeu (expressant-ho formalment) la probabilitat de guanyar alguna cosa?

P("alguna") = 1-P("res") = 1 – 0.296 = **0.704**

4.- (1 punt) Quina és la probabilitat de guanyar alguna cosa si a la primera tirada ha sortit creu?

P("alguna"|"1a+") = P("alguna","1a+") / P("1a+") = (2/3*1/3+2/3*2/3*1/3) / (2/3) = (10/27) / (2/3) = 5/9 = **0.556**

5.- (1 punt) Quines haurien de ser les probabilitats de cara i creu d'una moneda amb la qual la probabilitat de no guanyar res fos del 50%

P(+++) = 0.50 = $(P(+))^3$ -> P(+)=$(0.5)^{1/3}$ -> **P(+) = 0.794 o 0.8** i **P(o) = 0.206 o 0.2** (només 1 de cada 5 surt cara)

6.- (2 punts) Definiu la variable aleatòria que indica les probabilitats de les diferents quantitats a guanyar (incloent no guanyar res), i calculeu-ne l'esperança i la desviació

K	P_X (n)
0	8/27 = 0.296
100 / 3	4/27 = 0.148
100 / 2	2/9 = 6/27 = 0.222
100	1/3 = 9/27 = 0.333

E(X) = 0*8/27 + 100/3*4/27 + 100/2*2/9 + 100*1/3 = 4000/81 = **49.3 eur**

V(X) = $(0-49.3)^2*8/27+(100/3-49.3)^2*4/27+(100/2-49.3)^2*2/9+(100-49.3)^2*1/3$ = 1613,25 eur^2
-> sqrt(1613,25) = **40.16 eur**

7.- (3 punts) Considereu ara una altra variable aleatòria indicant el nombre de creus obtingudes en l'experiència aleatòria anterior. Indiqueu la taula de probabilitat d'aquesta nova variable, i la seva esperança i desviació. I també la taula de probabilitats conjuntes (d'aquesta variable nombre de creus i l'anterior de guanys), i calculeu la covariància i la correlació, i comenteu la relació entre les dues variables relacionant-ho amb el valor de la correlació entre elles

K	P_Y (n)
0	1/3 = 9/27 = 0.333
1	2/9 = 6/27 = 0.222
2	4/27 = 0.148
3	8/27 = 0.296

E(Y) = 0*1/3 + 1*2/9 + 2*4/27 + 3*8/27 = **1.4 creus**
V(Y) = $(0-1.4)^2*9/27+(1-1.4)^2*2/9+(2-1.4)^2*4/27+(3-1.4)^2*8/27$ = 1.5 eur^2
-> sqrt(1.5) = **1.22 creus**

	0	100/3	100/2	100	
0	**0**	**0**	**0**	**1/3=9/27=0.333**	1/3
1	**0**	**0**	**2/9=6/27=0.222**	**0**	2/9
2	**0**	**4/27=0.148**	**0**	**0**	4/27
3	**8/27= 0.296**	**0**	**0**	**0**	8/27
	8/27	4/27	2/9	1/3	

Cov(X,Y) = (0-1.4)(100-49.3)*1/3 + (1-1.4)(50-49.3)*2/9 + (2-1.4)(100/3 -49.3)*4/27 + (3-1.4)(0-49.3)*8/27 = **-48.465**
Cor(X,Y) = Cov(X,Y) / (sqrt(V(X))*sqrt(V(Y))) = -48.465 / (40.16*1.22) = **- 0.99**

La relació és inversa (com més creus surten, menys guanys s'esperen)
La correlació és de -0.99, molt propera a -1, però no exactament -1 tot i que totes les probabilitats estan a la diagonal inversa

Pr2 NOM:_____COGNOMS:_____

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

Un determinat algorisme consisteix en N etapes que cal superar satisfactòriament. Com es veu al pseudocodi, l'etapa *i*-èssima es computa a la funció `proces(i)`, que retorna un indicador d'estat `Z`, que es avaluat a `condicio(Z)` per determinar si l'etapa s'ha superat. La funció `proces` fa uns càlculs molt complexos, i l'etapa no es supera el 40% de les vegades que s'executa, obligant a repetir-ho provant una altra estratègia.

```
for (i=0; i<N; i++) {
   repeat {
       Z=proces(i);
   } until (condicio(Z));
}
```

Quan la funció `proces` s'ha d'executar novament a la mateixa etapa es genera un indicador independent de l'anterior. Les etapes diferents també generen indicadors independents.

1. Fixem-nos en una etapa *i* fixa. Anomeneu B_i a la variable aleatòria que compta el nombre de cicles del bucle `repeat`. Quin és el model probabilístic per a aquesta variable, i perquè? Expresseu la funció de probabilitat de B_i i digueu quin és el nombre esperat de vegades que la funció `proces` serà emprada.

   B_i segueix un model Geomètric, amb paràmetre p=0.60, perquè el nombre de cicles correspon al nombre d'iteracions (almenys una) fins a que s'assoleix la condició de sortida, i aquesta condició és independent de cicles anteriors, segons l'enunciat. Si no es supera el 40% de les vegades, vol dir que la condició és certa amb probabilitat 0.60.

   $$p_{Bi}(k) = (1-p)^{k-1}p = q^{k-1}p, \quad k = 1, 2, 3, \ldots.$$

   El nombre de vegades que la funció és utilitzada coincideix amb el nombre d' iteracions, per tant el seu valor esperat és:
   $\frac{1}{p}$ =1.67.

2. Calculeu la probabilitat que l'etapa *i* la funció `condicio` sigui avaluada almenys 4 vegades.

   $$P(B_i \geq 4) = 1 - [P(B_i = 1) + P(B_i = 2) + P(B_i = 3)] = 1 - [p + q\,p + q^2 p]$$
   $$= 1 - [0.6 + 0.4\;0.6 + 0.16\;0.6] = 0.064$$

   També es pot trobar més fàcilment amb la funció de distribució:
   $$P(B_i \geq 4) = 1 - P(B_i \leq 3) = 1 - F_{Bi}(3) = 1 - (1-q^3) = 0.4^3 = 0.064$$

3. Volem estudiar el comportament de l'algorisme, concretament a les etapes 5, 10, 15, 20, 25, 30, 35 i 40. Quin és el nombre esperat d'etapes on el nombre de cicles serà superior a 3? Trobeu la probabilitat que observem com a molt a una etapa aquest fet (més de 3 cicles). *Ajut: definiu una variable aleatòria adient, especificant el seu model.*

   Definirem la variable S per comptar quantes vegades ocorrerà que el nombre de cicles serà superior a 3, en vuit etapes (no té importància si són aquestes o qualssevol altres, perquè sempre són independents entre sí). Com és el mateix esdeveniment que el de la pregunta anterior, ja sabem que la probabilitat d'èxit és 0064.

   Per tant, es tracta d'una variable aleatòria amb distribució Binomial, i paràmetres n=8 i p'=0.064. S ~ B(8, 0.064)

   El nombre esperat demanat és E(S) = $n\,p'$ = 0.512

   La probabilitat demanada és $P(S \leq 1) = (1-p')^8 + 8(1-p')^7 p' = 0.5891 + 0.3223 = 0.9114$.

4. Suposem que l'algorisme precisa exactament N=100 etapes. A) Digueu quin és el model adequat per descriure el nombre total d'execucions de la funció `proces` (amb els seus paràmetres); B) Proposeu un model que pugui aproximar raonablement el model exacte, i expliqueu perquè es pot suposar raonable.

   A) Binomial negativa(r=N=100, p=0.6). Perquè és el total acumulat de N variables B_i de model geomètric, amb idèntic paràmetre p, independents entre sí; o bé, perquè expressa el nombre total d'intents fins a arribar al r-èssim èxit, que és la definició clàssica de la binomial negativa.

   B) Una variable aleatòria per el nombre total d'execucions de la funció (T) pot ser aproximada per un model Normal, donat que és una suma de 100 variables, que és un nombre prou alt, encara que la variable B_i és discreta i molt asimètrica, característiques que suposen que la convergència a la Normal requereix almenys un nombre similar. Es pren per a μ el valor esperat de T (segons el model de la binomial negativa), i el mateix per a σ^2:

   $$T \approx N\left(\mu = \frac{r}{p}, \sigma^2 = \frac{qr}{p^2}\right) = N(\mu = 166.7, \sigma^2 = 111.11)$$

5. Amb el model aproximat anterior, trobeu una fita superior pel nombre de vegades que s'executa `proces` i que sigui vàlida amb un error de l'1%. Per a N=200, la fita seria el doble d'aquesta? Justifiqueu-ho.

   Trobem un valor f tal que P(T > f) = 0.01. O de forma equivalent, el percentil 99: $F_T(f) = 0.99$

   Utilitzem la forma estàndard de T: $P\left(Z < \frac{f-\mu}{\sigma}\right) = 0.99$, i busquem a les taules el percentil 99 de Z, que és 2.33. Llavors, $f = \mu + 2.33\sigma = 166.7 + 2.33\sqrt{111.11} = 191.23$

   Per tant, menys de 1 de cada 100 vegades el nombre d'execucions de `proces` serà superior a 192 (amb un bucle de 100 etapes). Si fossin 200 no seria el doble, perquè la desviació tipus de T no seria tampoc el doble encara que si ho seria la mitjana. Concretament, la fita seria: $333.33 + 2.33\sqrt{222.22} = 368$.

6. Hem afegit una variable per comptar el nombre de vegades que s'executa `proces` i hem fet córrer l'algorisme, obtenint que han calgut 145 avaluacions, amb N=100.
   a. Amb aquestes dades, doneu una estimació puntual per a la probabilitat que la funció `proces` superi l'etapa corresponent.
   b. Quin és l'error tipus de l'estimació anterior? Expliqueu el significat de l'error tipus trobat.
   c. Estimeu amb un interval de confiança del 95% la probabilitat que una avaluació de la funció `condicio` retorni CERT.
   d. Expliqueu com s'interpreta l'interval que heu trobat a l'apartat anterior, i si és compatible amb les assumpcions inicials del problema.

   Estimació puntual de π, probabilitat que la funció `proces` superi l'etapa corresponent:

   P = N/M = 100/145 = 0.69

   Error tipus de la proporció: $\sqrt{\frac{P(1-P)}{M}} = \sqrt{\frac{0.69\ 0.31}{145}} = 0.0384$. IC(π, 95%) = $0.69 \pm z_{0.975}\ 0.0384$ = [61.4, 76.5]%

   Hem obtingut a la mostra observada que la condició s'avalua satisfactòriament el 69% de les vegades. Però aquesta mesura té un error de mostratge de gairebé un 4% (és a dir: una altra mostra de la mateixa grandària podria exhibir una diferència d'aquesta magnitud amb la nostra estimació, simplement per atzar).

   (S'ha de precisar que *una altra mostra* hauria de ser de 145 iteracions, i el que podria variar seria el nombre d'etapes superades).

   Amb confiança del 95%, creiem que la probabilitat π es podria situar entre 0.614 i 0.765. Això voldria dir que a l'inici podríem haver sobreestimat la probabilitat de no superar la condició establerta (en principi, 40%).

També hem incorporat al programa anterior codi que obté el temps de CPU emprat per l'ordinador a la funció `proces`:

```
for (i=j=s=sq=0; i<N; i++) {
   repeat {
      t = CPU.time( Z=proces(i) );
      s += t;
      sq += t*t;
      j++;                    // el comptador d'iteracions
   } until (condicio(Z));
}
```

A més del resultat `j=145`, també hem obtingut `s=1125` (segons) i `sq=11600` (s^2).

7. Doneu una estimació puntual i per interval de confiança del 95% per al temps que en mitjana empra l'execució de la funció `proces`.

   Sigui Q la variable "temps d'execució de la funció `proces`", i μ=E(Q) el paràmetre d'interés.
   Estimació puntual: $\bar{x} = \frac{1}{M}\sum q_i = \frac{1125}{145} = 7.76$ seg

   Càlcul de la desviació tipus mostral: $s_Q = \sqrt{\frac{\sum q_i^2 - M\,\bar{x}^2}{M-1}} = \sqrt{\frac{11600 - 145 \cdot 60.2}{144}} = 4.466$ seg

   IC(μ, 95%) = $\bar{x} \pm \frac{z_{0.975} s_Q}{\sqrt{M}} = 7.76 \pm 1.96 \frac{4.466}{\sqrt{145}} = (7.03, 8.49)$ seg

8. Sabent que el temps de CPU no pot ser negatiu, raoneu si es pot admetre que la variable recollida (`CPU.time`) es distribueix com a Normal, i també perquè (o perquè no) es necessita Normalitat de la variable per a calcular intervals de confiança per al valor esperat.

   Encara que les estimacions anteriors tenen error de mostratge, possiblement són representatives dels valors reals dels paràmetres. Si les prenem com a reals, una variable Normal amb mitjana 7.76 i desviació 4.47 tindria una probabilitat bastant gran de prendre valors negatius. Per tant, és quasi segur que la variable temps és asimètrica per la dreta, ja que esporàdicament es podria veure algun valor molt alt, però mai un molt allunyat per la banda esquerra. En principi, per a fer aquests IC es precisa Normalitat, però com la mostra és gran l'estimació de σ és fiable i per tant no seria cap problema.

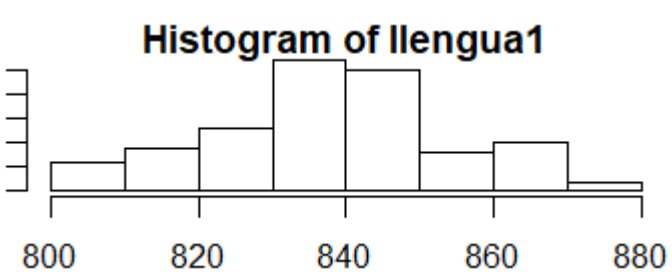
NOM:_____COGNOM:_____

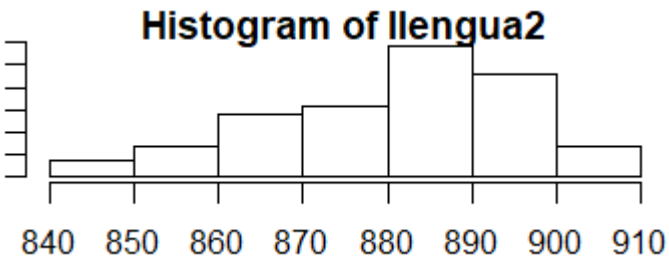
(*Contesteu cada pregunta en el seu lloc. Expliciteu i justifiqueu els càlculs*)

Problema 3 (B5-B6)

Un grup d'estudiants de la UPC volen fer un estudi per comparar si hi ha diferència en l'ús de les vocals entre el català i el castellà. Per fer-ho cerquen de manera aleatòria en el repositori de la UPC 100 documents en cadascun dels dos idiomes. Per cada document, compten el nombre de vocals que hi ha en els primers 2000 caràcters sense tenir en compte els espais. Amb aquesta resposta es defineixen les corresponents variables aleatòries: T pels documents en català i S pels documents en castellà, obtenint-se les dades següents:

$\bar{T}$= 837.4 i s_T= 16.5523 $\bar{S}$= 881.28 i s_S= 15.3149





1. Indiqueu quin és l'histograma corresponent a T i a S. Raoneu la resposta. (1 punt)

A partit dels valors de les mitjanes trobats en l'apartat anterior, es pot deduir que l'histograma corresponent a llengua1 és el de la variable nombre de vocals en català (T) perquè el seu valor mitjà és 837,4 i el corresponent a llengua2 és el de la variable nombre de vocals en castellà (S) perquè el seu valor mitjà és 881,28.

2. Es tracta de dues mostres independents o aparellades? Raoneu la resposta (1 punt)

Es tracta de mostres independents, ja que cada idioma s'escullen documents diferents.

Es vol estudiar si les dues llengües usen el mateix nombre de vocals o si és el castellà el qual n'usa més. Per contestar els següents apartats considereu un risc α=0.05 i una confiança del 95%.

3. a) Indiqueu les hipòtesis i les premisses de la prova a realitzar. (0'5 punts)

$\begin{cases} H_0: \mu_T = \mu_S \\ H_1: \mu_T < \mu_S \end{cases}$, test unilateral. Dues mostres aleatòries simples independents. N més gran o igual a 100.

b) Indiqueu la fórmula de l'estadístic i quina és la distribució d'aquest quan se suposa que les dues respostes tenen la mateixa mitjana (0'5 punts)

$$Z = \frac{\bar{T}-\bar{S}}{\sqrt{\frac{s_T^2}{n_T}+\frac{s_S^2}{n_S}}}$$, Z~ N(0,1)

c) Calculeu l'estadístic i raoneu si podem rebutjar la hipòtesi nul·la. (1 punt)

$$Z = \frac{\bar{T}-\bar{S}}{\sqrt{\frac{s_T^2}{n_T}+\frac{s_S^2}{n_S}}} = -19.45851$$

Com que $z_{0'05}$ = -1.645, es té que l'estadístic -19,46 < -1,645, és a dir, és més petit que el punt crític i per tant es tenen evidències per rebutjar que el nombre de vocals emprades per les dues llengües siguin iguals.

També es vol estudiar si hi ha relació entre el nombre de vocals emprades en ambdues llengües. Per això se seleccionen aleatòriament textos de 2000 caràcters sense comptar espais de 20 textos en català i es tradueixen al castellà mitjançant un traductor automàtic. Anomenant les dues variables aleatòries *Nombre de vocals en català* (T) i *Nombre de vocals en castellà* (S), es té que:

```
> mean(T)
[1] 837.55
> mean(S)
[1] 879.4
```

```
> sd(T)
[1] 16.58304
> sd(S)
[1] 19.47846
```

```
> sum(T)
[1] 16751
> sum(S)
[1] 17588
> sum(T*S)
[1] 14731604
```

4. a) Calculeu la covariància i el coeficient de correlació (1 pt)

Cov(T,S) = 40.76842 i $r_{TS} = \frac{Cov(T,S)}{s(T)\cdot s(S)} = 0.1262132$

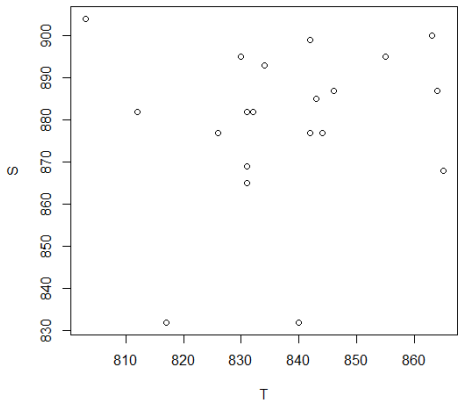
b) Interpreteu el coeficient de correlació lineal (0'5 punts)

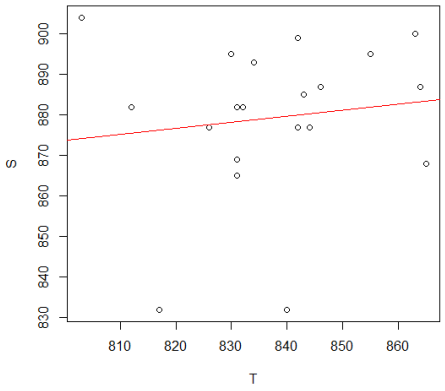
El valor del coeficient de correlació és molt proper a 0, per tant la relació entre les dues variables és molt feble / no hi ha un bon ajustament lineal / el soroll aleatori al voltant de la (suposada) recta és molt gran / el nombre de vocals d'un text en una llengua no serveix per predir be el nombre de vocals en l'altra.

5.a) Calculeu la recta de regressió del Nombre de vocals en castellà (S) respecte el Nombre de vocals en català (T) (1 pt)

b1= $\frac{cov(T,S)}{var(T)}$ = 0.1483 i b0= $\bar{S} - b1 \cdot \bar{T}$ = 755.2330. La recta de regressió és S= 0.1483T + 755.23

b) Representeu la recta de regressió en el gràfic següent. Expliciteu la representació de manera raonada. (1 pt)





En la recta es representa el punt (837.55, 879.4) corresponent a les mitjanes de les dues variables.
Per l'altre punt es calcula, per exemple, la imatge per T=850, 881.2.

c) És versemblant que el coeficient del pendent sigui zero? Realitzeu la prova d'inferència estadística apropiada per donar resposta a la pregunta. (1'5 punts)

Plantegem si el pendent de la recta de regressió és diferent de zero.

$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$, tenim que $T' = \frac{b1}{S_{b1}}$ = 0.540, T'~ $t_{18,}$ per tant el punt crític és 2,01

On $S_{b1} = \sqrt{\frac{S^2}{(n-1)\cdot s(T)^2}}$ =0.2746 (S^2 = 400.203)

A partir de la prova realitzada podem concloure que és versemblant que el coeficient de la recta de regressió sigui zero i per tant no sembla que el nombre de vocals en castellà estigui relacionat amb el nombre de vocals en català.

d) Calculeu l'interval de confiança pel pendent i interpreta'l. (1 punt)

$IC(\beta_1, 95\%) = (b_1 - t_{n-2,1-\alpha/2} \cdot S_{b1}\ ,\ b_1 - t_{n-2,1-\alpha/2} \cdot S_{b1}\ )$ = (0.1483 - 2,01 · 0.2746, 0.1483 + 2,01 · 0.2746) =
= (-0.4119, 0.6919)

Amb les dades disponibles, el pendent de la recta que relaciona les dues variables podria anar des de -0.41 (amb signe negatiu) a 0.69 (positiu), per tant qualitativament no resulta gens informatiu.