Esta entrega de problemas se puntúa sobre 10. Aporta 1 punto a la nota final global y no es recuperable. La entrega es indivdual y se entrega en papel que proporciona el profesor ¡¡¡Suerte!!!

1) Consideremos los siguientes datos de una muestra aleatoria simple:

$$-4, -3, -2, -1, 0, 1, 2, 3, 4,$$

Estimad el error estándar de la media aritmética (**0.5 puntos**)

Solución:

La doy solo con R vosotros tenéis que hacerla de forma manual (con calculadora convencional)

```
> x=(-4):4
> mean(x)

[1] 0

> sd(x)

[1] 2.738613

> n=length(x)
> n

[1] 9

> sqrt(60/8)

[1] 2.738613

> sqrt(60/9)

[1] 2.581989

> sd(x)

[1] 2.738613

> estandard_error=sd(x)/sqrt(length(x))
> estandard_error

[1] 0.9128709

> sqrt(60/(8*9))

[1] 0.9128709
```

2) Se realiza el siguiente contraste de hipótesis

$$\begin{cases} H_0: & \mu = 3 \\ H_1: & \mu > 3 \end{cases}$$

resultado que el p-valor del contraste es 0.10 ¿Para qué niveles de significación aceptaríamos la hipótesis nula? (**0.5 puntos**)

Solución:

Aceptamos las hipótesis nula del contraste para todos los niveles de significación α menores que 0.10

3) Nuestro jefe nos has dicho que pagará una encuesta para saber cuál es el porcentaje de sus clientes que están interesados en un nuevo producto. Desconocemos totalmente el posible porcentaje de clientes interesados. El jefe se pregunta cuál debe ser el tamaño de la muestra para tener un error del $\pm 1\,\%$ con un nivel de confianza del $95\,\%$. Se pide que contestemos suponiendo el peor de los casos el que $p = 0.5$. (**1 punto**)

Solución:

La amplitud del intervalos es $A_0 = 0.02$, $1-\alpha = 0.95$, $1-\alpha = 0.025$, $1-\frac{\alpha}{2} = 0.975$ $z_{0.975} = 1.959964$ con las tablas 1.96.

Supuesto el peor caso que la proporción poblacional sea $p = 0.5$ sabemos que el tamaño de la muestra buscado es

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}}{A_0^2} \right\rceil \approx \left\lceil \frac{1.96^2}{0.02^2} \right\rceil = 9604.$$

Con código R

```
> n=ceiling(qnorm(1-0.05/2)^2/0.02^2)
> n

[1] 9604
```

4) Continuando con la encuesta el jefe ha decidido, por el momento, pagar una muestra de tamaño 200. En esa muestra 108 clientes están interesado en el producto. Ahora el jefe nos pregunta

a) Contrastad que la proporción de clientes interesados es mayor del $60\,\%$ contra que es menor. Resolved el contraste utilizando el p-valor. (**2 puntos**)

b) Dad el intervalo de confianza del $95\,\%$ asociado al contraste, e interpretarlo. (**1 punto**)

Solución:

Apartado a)

Sea p la proporción de clientes interesados queremos compararla con $p_0 = 0.6$. Nos piden que contrastesmos

$$\begin{cases} H_0: & p \geqslant 0.6 \\ H_1: & p < 0.6 \end{cases}$$

El tamaño de la muestra es $n = 200$ y la proporción muestral es $\hat{p} = \frac{108}{200} = 0.54$.

Utilizaremos el estadistico $Z = \frac{\hat{p}-p_0}{\sqrt{p_0 \cdot (1-p_0) n}}$ que en nuestra caso toma un valor

$$z_0 == \frac{0.54 - 0.6}{\sqrt{0.6 \cdot (1-0.6)200}} = -1.732051 \approx= -1.73.$$

Para esta hipótesis alternativa el p-valor es $P(Z < -1.73) = 1 - P(Z < 1.73) = 0.0418$. Es un valor inferior a 0.05 (aunque cercano) hay cierta evidencia para rechazar la hipótesis nula de que el producto interesa al $60\,\%$ o más de los clientes.

Apartado b)

El intevalor de confianza unilateral en este caso es

$$\left(-\infty, \hat{p} + z_{1-\alpha} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}\right) \approx \left(-\infty, 0.54 + 1.65 \cdot \sqrt{\frac{0.54 \cdot (1-0.54)}{200}}\right) = (-\infty, 0.5981493)\,.$$

El intervalo de confianza no contiene la proporción $p = 0.6$ que confirma que hay cierta evidencia de queesa porporción no supera 0.6 aunque sea por pocas centésimas.

No se pide pero R coprobamos los resultados

```
> n=200
> phat=108/200
> p0=0.6
> z0=(phat-p0)/sqrt(p0*(1-p0)/n)
> z0

[1] -1.732051

> pvalue=pnorm(z0)
> pvalue

[1] 0.04163226

> IC=c(-Inf,phat+qnorm(1-0.05)*sqrt(phat*(1-phat)/n))
> IC

[1]     -Inf 0.597968
```

5) El `data frame` `datos_vuelos` contiene información del retraso en minutos de vuelos de varias compañías aéreas diferentes.

```
> head(datos_vuelos)

    retraso compania
1 12.954091       C1
2  7.940958       C1
3  3.439275       C1
4 13.664147       C1
5  4.930072       C1
6 12.952991       C1

> str(datos_vuelos)

'data.frame':        240 obs. of  2 variables:
 $ retraso : num  12.95 7.94 3.44 13.66 4.93 ...
 $ compania: Factor w/ 2 levels "C1","C2": 1 1 1 1 1 1 1 1 1 1 ...

> table(datos_vuelos$compania)

 C1  C2
120 120

> aggregate(retraso~compania,data=datos_vuelos,FUN=mean)

  compania   retraso
1       C1  9.508184
2       C2 11.006678

> aggregate(retraso~compania,data=datos_vuelos,FUN=sd)
```

```
  compania  retraso
1       C1 3.723884
2       C2 4.058315

> var.test(retraso~compania)

        F test to compare two variances

data:  retraso by compania
F = 0.84198, num df = 119, denom df = 119, p-value = 0.3495
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.586802 1.208119
sample estimates:
ratio of variances
          0.841978

> retrasoC1=datos_vuelos$retraso[datos_vuelos$compania=="C1"]
> retrasoC2=datos_vuelos$retraso[datos_vuelos$compania=="C2"]
> ks.test(retrasoC1,"pnorm",mean(retrasoC1),sd(retrasoC1))

        One-sample Kolmogorov-Smirnov test

data:  retrasoC1
D = 0.061626, p-value = 0.7522
alternative hypothesis: two-sided

> ks.test(retrasoC2,"pnorm",mean(retrasoC2),sd(retrasoC2))

        One-sample Kolmogorov-Smirnov test

data:  retrasoC2
D = 0.041381, p-value = 0.9863
alternative hypothesis: two-sided
```

Contestad a las siguientes cuestiones justificando que parte del código utilizáis

1) Enunciar las hipótesis necesarias para el contraste de medias y discutid, utilizando sólo los resultados del código, si se cumplen las condiciones necesarias para realizar este contraste. (**1.5 punto**)

2) Contrastad si las medias de retraso son iguales contra que son distintas; definid el contraste y resolved calculando el p-valor[1]. (**1.5 punto**)

Solución:

Apartado a)

Tenemos un contraste de medias de poblaciones independientes, necesitamos normalidad de la distribución del retraso en cada población tamaños muestrales grandes. También tenemos que saber si aplicamos el test suponiendo igualdad entra las varianzas de ambas poblaciones.

Denotemos por μ_1 y μ_2 las medias poblacionales de los retrasos de las compañías C1 y C2, igualmente denotemos por σ_1^2 y σ_2^2 las varianzas poblacionales de los retrasos en cada compañía.

El test de igualdad de varianzas (razón de varianzas) es

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Con R se resuelve con la función

[1]Podéis aproximar la t de Student por una normal estándar.

```
> var.test(retraso~compania)

        F test to compare two variances

data:  retraso by compania
F = 0.84198, num df = 119, denom df = 119, p-value = 0.3495
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.586802 1.208119
sample estimates:
ratio of variances
          0.841978
```

El p valor obtenido es muy alto así que aceptamos la igualdad de varianzas.

Para la normalidad hacemos los `ks.test` (hubiera sido más correcto utilizar el `lillie.test`[2] que es el `ks.test` específico para la distribución normal). En cada población aceptamos con p valores altos la normalidad.

Apartado b)

El test de igualdad de medias es

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Bajo estas condiciones el estadístico de contraste es

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1-1)\cdot\tilde{S}_1^2+(n_2-2)\cdot\tilde{S}_1^2}{n_1+n_2-2}\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

que sigue una ley de distribución $t_{n1+n2-2} = t_{138}$.

Donde, de la función `table` y de las dos funciones `aggregate` se deduce que los tamaños muestrales son $n_1 = n_2 = 120$, las medias muestrales son $\overline{x}_1 = 9.508184$ y $\overline{x}_2 = 11.006678$, por último las desviaciones típicas muestrales son $\tilde{S}_1 = 3.723884$ y $\tilde{S}_2^2 = 4.058315$. Así es valor del estadístico de contraste es

$$t = \frac{9.508184 - 11.006678}{\sqrt{\frac{(120-1)\cdot 3.723884^2+(120-1)\cdot 4.058315^2}{120+120-2}\cdot\left(\frac{1}{120}+\frac{1}{120}\right)}} = -2.9803.$$

El p-valor es $2\cdot P(t_{238} > |-2.9803|) = 2\cdot(1 - P(t_1 38 \leqslant 2.9803)$. Para hacerlo manualmente con las tablas aproximamos t_{138} por una normal estándar Z así el p-valor es , aproximadamente $2\cdot(1-P(Z\leqslant 2.98)) = 2\cdot(1-0.9986) = 0.0028$.

No se pedía pero la solución con es

```
> t.test(retrasoC1,retrasoC2,var.equal = TRUE)

        Two Sample t-test

data:  retrasoC1 and retrasoC2
t = -2.9803, df = 238, p-value = 0.003178
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4890048 -0.5079821
sample estimates:
mean of x mean of y
 9.508184 11.006678
```

[2]El El Kolgomorov-Smirnov-lilliefors test de normalidad

Haciendo los cálculos con R

```
> m1=mean(retrasoC1)
> m1

[1] 9.508184

> m2=mean(retrasoC2)
> m2

[1] 11.00668

> sd1=sd(retrasoC1)
> sd1

[1] 3.723884

> sd2=sd(retrasoC2)
> sd2

[1] 4.058315

> n1=length(retrasoC1)
> n1

[1] 120

> n2=length(retrasoC2)
> n2

[1] 120

> T.test=(m1-m2)/sqrt((((n1-1)*sd1^2+(n2-1)*sd2^2)/(n1+n2-2))*(1/n1+1/n2))
> T.test

[1] -2.980283

> 2*(1-pt(abs(-2.9803),n1+n2-2))

[1] 0.003178241

>
```

6) Se simula que el tiempo en segundos transcurrido entre dos reservas de vuelos de avión en un mismo día podría seguir una distribución exponencial con $\lambda = 1/5$. Se toma una muestra de 10 tiempos en segundos.

```
 [1]  1.6  1.8  2.8  3.9  4.3  4.7  4.8  7.3  8.7 11.1

 [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

 [1] 0.2738510 0.3023237 0.4287909 0.5415940 0.5768379 0.6093722 0.6171071
 [8] 0.7677637 0.8244796 0.8913909

[1] 0.273851
```

Vuelo	1	2	3	4	5	6	7	8	9	10
Retraso	0.50	1.40	1.60	2.20	2.40	3.70	3.90	4.50	5.20	7.10

```
        One-sample Kolmogorov-Smirnov test

data:  retraso
D = 0.27385, p-value = 0.3722
alternative hypothesis: two-sided
```

1) ¿Cuál es y qué parámetros tiene la función de distribución teórica propuesta? Escribid correctamente la función de distribución. (**0.5 puntos**)

2) Contrastar la hipótesis del enunciado con el test KS, al nivel de significación $\alpha = 0.1$. (**1.5 puntos**)

Solución:

Apartado a) La distribución es una exponencial del parámetro $\frac{1}{5}$, por lo tanto su función de distribución es

$$F_X(x) = \begin{cases} 1 - e^{-\frac{1}{5}\cdot x} & \text{si } x > 0. \\ 0 & \text{en otro caso} \end{cases}$$

Apartado b)

El contraste es

$$\begin{cases} H_0: & \text{Los datos provienen de una distribución} Exp(\frac{1}{5}). \\ H_1: & \text{Los datos NO provienen de una distribución} Exp(\frac{1}{5}) \end{cases}$$

Calculemos el estadístico del test K-S.

i	retraso x_i	$Fn(x_i) = \frac{i}{n=10}$	$Fx(x) = 1 - e^{-\frac{1}{5}\cdot x_i}$	$F_1 = \lvert Fx(x_i) - \frac{i-1}{10} \rvert$	$F_2 = \lvert Fx(x_i) - \frac{i}{10} \rvert$	máx$\{F1, F2\}$
1	1.60	0.10	0.27	0.27	0.17	0.27
2	1.80	0.20	0.30	0.20	0.10	0.20
3	2.80	0.30	0.43	0.23	0.13	0.23
4	3.90	0.40	0.54	0.24	0.14	0.24
5	4.30	0.50	0.58	0.18	0.08	0.18
6	4.70	0.60	0.61	0.11	0.01	0.11
7	4.80	0.70	0.62	0.02	0.08	0.08
8	7.30	0.80	0.77	0.07	0.03	0.07
9	8.70	0.90	0.82	0.02	0.08	0.08
10	11.10	1.00	0.89	0.01	0.11	0.11
						$D_{10} = 0.2739$

Ahora mirando en las tablas de los valores críticos del test K-S tenemos que $D_{n=10,\alpha=0.1} = 0.368$.

Como $D_{10} = 0.2739 \not> D_{n=10,\alpha=0.1} = 0.368$, no podemos rechazar la hipótesis nula de que los datos provengan de una población $Exp(-\frac{1}{5})$.

Con R

```
> retraso=c(1.6,1.8,2.8,3.9,4.3,4.7,4.8,7.3,8.7,11.1)
> retraso

 [1]  1.6  1.8  2.8  3.9  4.3  4.7  4.8  7.3  8.7 11.1

> n=length(retraso)
> Fn=(1:n)/n
> Fn

 [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

> Fx=pexp(retraso,1/5)
> Fx
```

```
 [1] 0.2738510 0.3023237 0.4287909 0.5415940 0.5768379 0.6093722 0.6171071
 [8] 0.7677637 0.8244796 0.8913909

> F1=abs(Fx-c(0,Fn[-10]))
> F2=abs(Fx-Fn)
> max(pmax(F1,F2))

[1] 0.273851

> ks.test(retraso,"pexp",1/5)

        One-sample Kolmogorov-Smirnov test

data:  retraso
D = 0.27385, p-value = 0.3722
alternative hypothesis: two-sided
```