Maestría Análisis y Visualización de Datos Masivos – Big Data
Análisis e Interpretación de Datos

1. Se hizo una encuesta entre 25 familias sobre el número de hijos que tenías. Se obtuvieron los siguientes datos:

Num Hijos	Num Familias
0	5
1	6
2	8
3	4
4	2
Total	25

   a. Encuentre la media artimética
   b. Encuentre la varianza
   c. Encuentre la desviación estándar

   a. Encuentre la media

$$a = \frac{\sum_{i=1}^{5} x_i \cdot n_i}{n} = \frac{0\cdot 5 + 1\cdot 6 + 2\cdot 8 + 3\cdot 4 + 4\cdot 2}{25} = \frac{42}{25} = 1{,}68$$

   b. Encuentre la varianza

$$s^2 = \frac{1}{n}\sum {f_i x_i}^2 - \left(\frac{1}{n}\sum f_i x_i\right)^2 = 4.25 - (1.68)^2 = 1.4176$$

   c. Encuentre la desviación estándar

$$s = \sqrt{1.4176} = 1.19062$$

2. Para el siguiente conjunto de datos, encuentre los cuartiles:
-1, -3, 0, -1, -1, 5, 0, -3, 1, 2, 3, 3

Se ordenan los datos: -3, -3, -1, -1, -1, 0, 0, 1, 2, 3, 3, 5. Son 12 datos. $Q_2 = \frac{0+0}{2} = 0$.
Luego -3, -3, -1, -1, -1, 0, 0, 1, 2, 3, 3, 5. $Q_1 = \frac{-1-1}{2} = -1$, $Q_3 = \frac{2+3}{2} = 2.5$

3. Para el siguiente conjunto de datos, encuentre los cuartiles:
22, 20, 24, 30, 32, 28, 35

Se ordenan los datos: 20, 22, 24, 28, 30, 32, 35. Son 7 datos. $Q_2 = 28$.
Luego 20, 22, 24, 28, 30, 32, 35. $Q_1 = 22$, $Q_3 = 32$

4. La probabilidad de éxito de una determinada vacuna es 0.72 Calcula la probabilidad de a que una vez administrada a 5 pacientes

    a. Ninguno sufra la enfermedad; es decir todos estén sanos
    b. Todos sufran la enfermedad; es decir que no les haya funcionada la vacuna
    c. Dos de ellos contraigan la enfermedad
    d. Al menos una persona se enferme

a) Ninguno sufra la enfermedad; es decir todos estén sanos

Buscamos P(k=5)
p=0.72
q=1-0.72 =0.28
n=5

$$P(5) = \frac{5!}{5!\,(5-5)!}.(0.72)^5(0.28)^{5-5} = 0.1934$$

b) Todos sufran la enfermedad; es decir que no les haya funcionada la vacuna

Buscamos P(k=0)
p=0.72
q=1-0.72 =0.28
n=5

$$P(0) = \frac{5!}{0!\,(5-0)!}.(0.72)^0(0.28)^{5-0} = 0.0017$$

c) Dos de ellos contraigan la enfermedad

Buscamos P(k=2) con
p=0.28
q=1-0.28 =0.72
n=5

$$P(2) = \frac{5!}{2!\,(5-2)!}.(0.28)^2(0.72)^{5-2} = 0.2926$$

d) Al menos una persona se enferme, es decir, hay menos de 5 sanos
Buscamos P(k<5)=1-P(k=5)

p=0.72
q=1-0.72 =0.28
n=5

$$1 - P(5) = \frac{5!}{5!\,(5-5)!}.(0.72)^5(0.28)^{5-5} = 1 - 0.1934 = 0.8066$$

5. En una clase, el 55% son chicos y el 45% restante chicas. En el examen de una asignatura, han aprobado el 80% de los chicos y el 90% de las chicas.

   a. Sabiendo que un estudiante ha aprobado, calcula la probabilidad de que sea chica
   b. Calcula la probabilidad de que, al elegir un estudiante al azar, haya aprobado.

Vamos a suponer 100 estudiantes. 55 chicos, 45 chicas. Y llenamos la tabla:

	Chicos (H)	Chicas (M)	Total
Aprueba (+)	H+=	M+=	E=
No Aprueba (-)	H-=	M-=	S=
Total	H=55	M=45	100

80% de los chicos aprueba o sea 44, 90% de las chicas aprueba 40.5

	Chicos (H)	Chicas (M)	Total
Aprueba (+)	H+=55(0.8)=44	M+=45(0.9)=40.5	
No Aprueba (-)			
Total	H=55	M=45	100

Y completamos la tabla:

	Chicos (H)	Chicas (M)	Total
Aprueba (+)	H+=55(0.8)=44	M+=45(0.9)=40.5	E=44+40.5=84.5
No Aprueba (-)	H-=55-44=11	M-=45-40.5=4.5	S=11+4.5=15.5
Total	H=55	M=45	100

a) Sabiendo que un estudiante ha aprobado, calcula la probabilidad de que sea chica

$$P(M|+) = \frac{P(M\ y\ +)}{P(+)} = \frac{\frac{40.5}{100}}{\frac{84.5}{100}} = \frac{40.5}{84.5} = 0.4793$$

b) Calcula la probabilidad de que, al elegir un estudiante al azar, haya aprobado.

$$(A|B) = \frac{P(A) * P(B|A)}{P(A) * P(B|A) + P(B)P(B|B) + P(C)\,P(C|B)} = \frac{P(A) * P(B|A)}{P(B)}$$

$$P(B) = P(A) * P(B|A) + P(B)P(B|B) + P(C)\,P(C|B)$$
$$P(+) = P(H) * P(+|H) + P(M)P(+|M)$$
$$P(+) = 0.55(0.80) + 0.45\ (\ 0.90) = 0.44 + 0.405 = 0.845$$

6. Se tiene el siguiente modelo lineal $\hat{y} = a_0 + a_1\, x$ que se pretende ajustar al siguiente conjunto de datos:

x	0	1	2	5	7
y	1	4	5	6	9

a. Encuentre el coeficiente de regresión lineal a_1
b. Encuentre el R^2

a) Encuentre el coeficiente de regresión lineal a_1

	x	y	(x-E[x])^2	(x-E[x])*(y-E[y])
	0	1	9	12
	1	4	4	2
	2	5	1	0
	5	6	4	2
	7	9	16	16
Sum	15	25	34	32
E[x]=	3	5		

$$a_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{32}{34} = 0.9412$$

b) Encuentre el R^2

	x	y	(y-E[y])^2	y^	(y-y^)^2
	0	1	16	2.176470588	1.384083045
	1	4	1	3.117647059	0.778546713
	2	5	0	4.058823529	0.885813149
	5	6	1	6.882352941	0.778546713
	7	9	16	8.764705882	0.055363322
Sum	15	25	34	25	3.882352941
E[x]=	3	5			

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{3.88}{34} = 0.848$$

7. La compañía Nobb Door fabrica puertas para vehículos recreativos. La compañía tiene dos propósitos en conflicto: desea construir puertas lo más pequeñas posible para ahorrar material pero, para conservar su buena reputación con el público, se siente obligada a fabricar puertas con la altura suficiente para que el 95% de la población adulta de Estados Unidos pueda pasar sus marcos. Con el fin de determinar la altu¬ra con la cual fabricar las puertas, la Nobb está dispuesta a suponer que la altura de la gente adulta de Es¬tados Unidos está distribuida normalmente con una media de 73 pulgadas (1.85 m), con una desviación estándar de 6 pulgadas (15.24 cm). ¿Qué tan altas deberán ser las puertas que fabrica la compañía Nobb?

X=desconocida
μ = 185 cm
σ = 15.24 cm
95% equivale a z= 1.645

$$1.64 = \frac{x - 185}{15.24}$$

Despejando x
(1.645)(15.24) +185 = 210 cm
Las puertas deberán ser de 210.00 cm

8. Se desea estimar el peso promedio de los sacos que son llenados por un nuevo instrumento en una industria. Se conoce que el peso de un saco que se llena con este instrumento es una variable aleatoria con distribución normal. Si se supone que la desviación típica del peso es de 0,5 kg. Determine el tamaño de muestra aleatoria necesaria para determinar una probabilidad igual a 0,95 de que el estimado y el parámetro se diferencien modularmente en menos de 0,1 kg.

$d = 0{,}1$
$\sigma = 0{,}5$
$1 - \alpha = 0{,}95$
$1 - \alpha/2 = 0{,}975$
$Z_{1-\alpha/2} = 1{,}96$

$$n = \left(\frac{\sigma Z_{1-\frac{\alpha}{2}}}{d}\right)^2 = \left(\frac{(0{,}5)(1{,}96)}{0{,}1}\right)^2 = 96{,}4$$

9. En el proyecto de Al Haouz en Marruecos, se ha calculado que cerca del 30% (0,3) de los niños de la zona del proyecto padecen de malnutrición crónica. Este dato se basa en estadísticas nacionales sobre malnutrición en las zonas rurales. Encuentre el tamaño de muestra para saber el porcentaje de niños con malnutrición crónica con un error del 5%.

$$n = \frac{(1.96)^2(0.3)(0.7)}{(0.05)^2} = 323$$

10. Se estudiaba la altura de los individuos de una ciudad, obteniéndose en una muestra de tamaño 20 los siguientes valores: media de 168.0 cm y una desviación, estándar de 7 cm. Calcular un intervalo de confianza del 95 % para la altura de los individuos de la ciudad.

σ=7

n=20

$\bar{x}$ =168

$z_{1-\alpha/2}$=1.96

$$\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$168 - 1.96\frac{7}{\sqrt{20}} < \mu < 168 + 1.96\frac{7}{\sqrt{20}}$$

El intervalo va de 164.94 a 171.06

11. Un informe indica que el precio medio del billete de avión entre Monterrey y Guadalajara es, como máximo, de $120 con una desviación típica de $40. Se toma una muestra de 100 viajeros y se obtiene que la media de los precios de sus billetes es de $128. ¿Podemos decir que el precio de los vuelos ha aumentado con un nivel de confianza del 95%? Utilice el método de la z_{crit}.

Enunciamos las hipótesis nula y alternativa:

H0: μ = 120

H1: μ ≠ 120

Para α = 0.05, le corresponde un valor crítico: $z_{\alpha/2}$ = 1.96.

$$z_{calc} = \frac{128 - 120}{\frac{40}{\sqrt{100}}} = \frac{8}{4} = 2$$

Como 2>1.96, rechazamos la hipótesis nula.

12. Un fabricante de lámparas eléctricas está ensayando un nuevo método de producción que se considerará aceptable si las lámparas obtenidas por este método dan lugar a una población normal de duración media 2400 horas, con una desviación típica igual a 300. Se toma una muestra de 100 lámparas producidas por este método y esta muestra tendrá una duración media de 2320 horas. ¿Se puede aceptar la hipótesis de que la duración promedio de las lámparas fabricadas por el nuevo proceso ha disminuido con un riesgo igual o menor al 5%? Utilice el método del valor p.

Enunciamos las hipótesis nula y alternativa:

H0: μ = 2400
H1: μ ≠ 2400
α = 0.05
zα = 1.96.

$$z_{calc} = \frac{2400 - 2320}{\frac{300}{\sqrt{100}}} = \frac{80}{30} = 2.6667$$

El valor p es 0.004. Como 2x0.004 << 0.05, rechazamos la hipótesis nula.

13. Dieta contra ejercicio. Se intenta averiguar con un 95% de nivel de confianza, si hay diferencia en la pérdida de peso entre las personas que hacen ejercicio y las que hacen dieta:

Sólo dieta
Pérdida de peso μ_1= 5.9 kg
Desviación estándar σ_1= 4.1 kg
Tamaño de muestra = n_1 = 42

Sólo ejercicio
Pérdida de peso μ_2= 4.1 kg
Desviación estándar σ_2= 3.7 kg
Tamaño de muestra = n_2 = 47

Hipótesis:
$H0: \mu_1 = \mu_2$
$H0: \mu_1 \neq \mu_2$
$\alpha = 0.05$

a) Suponiendo que las desviaciones estándar son iguales:

Desviación conjunta

$$S^2_{pooled} = \frac{(42-1)(4.1)^2 + (47-1)(3.7)^2}{42+47-2} = 15.1603$$

$$SE = \sqrt{\frac{15.1603}{42} + \frac{15.1603}{47}} = 3.8936\sqrt{\frac{1}{42} + \frac{1}{47}} = 0.8268$$

Con distribución Normal:

$z_{crit,1-\alpha/2} = 1.96$

$z_{calc} = \frac{1.8-0}{0.8268} = 2.1772$

Valor p=2x0.015=0.03

Como $z_{calc} > z_{crit}$ y $valor\ p = 0.03 < \alpha = 0.05$, se rechaza la hipótesis nula. Parece ser que la dieta funciona mejor.

Con t-student:

Grados de libertad=$42 - 47 - 2 = 87$

$t_{crit,87,1-\frac{\alpha}{2}} = 1.988$

$t_{calc} = \frac{1.8-0}{0.8268} = 2.1772$

Valor p=2x0.0161=0.0322

Como $t_{calc} > t_{crit}$ y $valor\ p = 0.0322 < \alpha = 0.05$, se rechaza la hipótesis nula. Parece ser que la dieta funciona mejor.

b) Suponiendo que las desviaciones estándar son diferentes (**opcional, no lo cubre el material**):

$SE1 = \frac{4.1}{\sqrt{42}} = 0.633$

$SE2 = \frac{3.7}{\sqrt{47}} = 0.540$

Desviación conjunta:

SE=$\sqrt{(0.633)^2 + (0.540)^2} = 0.83$

Con distribución Normal:

$z_{crit,1-\frac{\alpha}{2}} = 1.96$

$z_{calc} = \frac{1.8-0}{0.83} = 2.17$

Valor p=2x0.015=0.03

Como $z_{calc} > z_{crit}$ y $valor\ p = 0.03 < \alpha = 0.05$, se rechaza la hipótesis nula. Parece ser que la dieta funciona mejor.

Con t-Student:
Grados de libertad:

$$\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} = \frac{\left(\frac{4.1^2}{42}+\frac{3.7^2}{47}\right)^2}{\frac{\left(\frac{4.1^2}{42}\right)^2}{42-1}+\frac{\left(\frac{3.7^2}{47}\right)^2}{47-1}} = \frac{0.4783}{\frac{0.16}{41}+\frac{0.0848}{46}} = \frac{0.4783}{0.0057} = 83.14$$

$t_{crit,83,1-\frac{\alpha}{2}} = 1.989$

$t_{calc} = \frac{1.8-0}{0.83} = 2.17$

Valor p=2x0.0165=0.033

Como $t_{calc} > t_{crit}$ y $valor\ p = 0.033 < \alpha = 0.05$, se rechaza la hipótesis nula. Parece ser que la dieta funciona mejor.