ESTADÍSTICA
Tercer Curso de CC. AA. (2008-2009)

I. Modelos de diseño de experimentos de un factor

1.– Se quiere comparar la capacidad pulmonar en niños, adultos y ancianos, obteniéndose los siguientes resultados:

Niños	8,4	7,6	7,9	8,0	8,1
Adultos	8,7	8,1	8,5	8,2	8,0
Ancianos	7,4	7,8	7,3	7,6	8,0

Hacer un estudio completo.

2.– Las precipitaciones caídas en un país han disminuido de manera preocupante durante el último año. Antes de tomar ninguna medida se decide hacer un estudio previo para saber si el descenso de las lluvias se produjo de forma homogénea en todo el país. Para ello se seleccionan aleatoriamente cinco estaciones meteorológicas en cada una de las cuatro regiones del país, obteniéndose los siguientes porcentajes de disminución de las precipitaciones en cada una de ellas:

Región Este	Región Norte	Región Oeste	Región Sur
10,4	12,8	11,2	13,9
12,8	14,2	9,8	14,2
15,6	16,3	10,7	12,8
9,2	10,1	6,3	15,0
8,7	12,0	12,4	13,7

a) Proponer un modelo para comparar los porcentajes de disminución de las precipitaciones en las cuatro regiones.

b) ¿En qué zona parecen haber disminuido más las precipitaciones?

c) Obtener la tabla ANOVA y contrastar la hipótesis de que las medias de disminución del porcentaje de lluvias en el país fueron las mismas en las cuatro regiones (tomar $\alpha = 0{,}10$).

d) Efectuar comparaciones entre las medias de las diferentes regiones con un nivel de confianza global del 90 %.

e) Hacer un diagnóstico de las hipótesis del modelo mediante el análisis de los residuos.

3.– En un bosque próximo a una incineradora los árboles no crecen con normalidad. Se piensa que unos nuevos abonos americanos y australianos pueden ser la solución. Para ver si es efectiva esta medida se utiliza el abono americano en un tercio de los árboles, el abono australiano en otro tercio y para el tercio restante no se utiliza ningún abono. Después de 3 meses se han obtenido los siguientes resultados sobre el crecimiento en centímetros de 60 árboles en total:

$$\begin{array}{lcllcl} \bar{y}_{1\cdot} & = & 6{,}57 & s_1^2 & = & 0{,}70 \\ \bar{y}_{2\cdot} & = & 5{,}30 & s_2^2 & = & 0{,}65 \\ \bar{y}_{3\cdot} & = & 3{,}20 & s_3^2 & = & 0{,}50 \end{array}$$

¿Se puede afirmar que se obtienen diferencias en los resultados, con un nivel de significación 0,01? En caso necesario, efectuar una comparación de los crecimientos medios, con un nivel de significación conjunto de 0,15.

4.– En un estudio sobre la efectividad de los métodos para dejar de fumar se quiere saber cuándo la reducción media en el número de cigarrillos diarios difiere de un método a otro entre hombres fumadores. Para ello se hace un experimento con 12 fumadores que consumían 60 cigarrillos diarios. Se aplica cada uno de los métodos a 4 de ellos seleccionados aleatoriamente. El número de cigarrillos que deja de fumar cada individuo es:

Método I	Método II	Método III
50	41	49
51	40	47
51	39	45
52	40	47

a) Contrastar mediante el análisis de la varianza si la reducción media en el número de cigarrillos es similar para los tres métodos con un nivel de significación $\alpha = 0{,}05$.

b) Construir los intervalos de confianza para la diferencia entre las medias con un nivel de confianza conjunto de 0,95. ¿Qué conclusiones se pueden obtener?

5.– A continuación se muestran los datos recogidos por las inspecciones en cuatro gasolineras elegidas aleatoriamente. Los valores de la tabla reflejan los mililitros que faltan para completar un litro en distintas mediciones sobre el mismo surtidor de cada gasolinera.

Gasol. C	17.80	18.00	17.98	18.20	18.00	17.99	18.10	17.90
Gasol. R	18.01	17.75	18.00	17.77	18.01	18.01	18.12	18.20
Gasol. S	18.10	17.92	18.01	17.88	18.30	18.22	18.56	18.10
Gasol. V	18.05	18.01	17.94	18.23	18.20	18.00	17.84	18.11

Contrastar la hipótesis de que la cantidad media de gasolina que se sirve por litro no depende de la gasolinera (tomar $\alpha = 0{,}05$).

6.– En un estudio sobre la incidencia del cáncer de garganta se van a analizar los resultados obtenidos en 10 ciudades de más de 100 000 habitantes, en otras 10 ciudades de menos de 100 000 habitantes, y en 20 pueblos. En cada población se registra la variable $Y =$ «Número de casos detectados por cada 100 000 habitantes». Los resultados obtenidos se resumen a continuación:

	$\bar{y}$	s^2
Ciudades grandes	120	105'2
Ciudades pequeñas	110	95'3
Pueblos	70	101'8

a) ¿Aportan estos datos evidencia estadística de que el hábitat influye en la incidencia del cáncer de garganta? Dar una respuesta razonada, con una confianza del 90 %, indicando el modelo y la metodología estadística empleada.

b) Suponiendo que la varianza es la misma en cada uno de los tipos de población, dar una estimación insesgada de dicha varianza.

c) Hallar intervalos de confianza para estimar simultáneamente las diferencias entre incidencias medias de la enfermedad en los tres tipos de población, con una confianza conjunta del 70 %. ¿Qué conclusiones estadísticas se pueden obtener sobre la incidencia media en los tres tipos de población?

7.– Se desea estudiar la influencia del nivel medio de renta en la tasa de mortalidad. Se consideran 60 países con diferente desarrollo económico, y se dividen en tres niveles de renta (bajo, medio y alto). Obtenemos los siguientes resultados:

Nivel			
Bajo	$n_1 = 4$	$\bar{y}_{1\cdot} = 866{,}243$	$s_1^2 = 3\,394{,}596$
Medio	$n_2 = 40$	$\bar{y}_{2\cdot} = 935{,}785$	$s_2^2 = 3\,024{,}844$
Alto	$n_3 = 16$	$\bar{y}_{3\cdot} = 970{,}321$	$s_3^2 = 4\,198{,}795$

a) Describir el modelo a usar en el análisis de los datos y las hipótesis asumidas.

b) Construir la tabla ANOVA y decidir, a nivel $\alpha = 0{,}01$, si el nivel de renta influye en la tasa de mortalidad.

c) A nivel conjunto $\alpha = 0{,}06$, ¿entre qué niveles de renta se detectan diferencias significativas?

II. Modelos de diseño de experimentos de varios factores

1.– Se quiere estudiar la producción de fresa que se obtiene con diferentes variedades. La producción obtenida con 3 variedades y en 4 tipos de suelo diferentes, se ofrece a continuación:

		Suelos			
		1	2	3	4
Variedades	1	6'3	6'9	5'3	6'2
	2	10'1	10'8	9'8	10'5
	3	8'4	9'4	9	9'2

Hacer un estudio completo.

2.– En un estudio sobre el consumo de gasolina de distintos coches se realiza el siguiente experimento: se toman cuatro coches al azar de un fabricante español, cuatro de un francés, cuatro de un alemán y cuatro de un japonés. Se prueba un coche de cada fabricante en una gran ciudad durante la hora punta y otro fuera de la hora punta, otro se prueba en carretera de montaña y el otro en una carretera llana. El consumo en litros de gasolina por cada 100 kilómetros es:

	Hora punta	Hora normal	Montaña	Carretera llana
Español	14.7	9.4	7.2	6.8
Francés	11.6	7.7	6.8	6.0
Alemán	10.8	7.2	7.2	6.4
Japonés	16.0	10.0	9.3	7.7

a) Proponer un modelo para estudiar el consumo de gasolina.

b) ¿Qué modelo de coche parece que consume más y qué modelo de coche parece que consume menos? ¿En qué condiciones parece que se consume más y en qué condiciones parece que se consume menos?

c) Obtener la tabla de análisis de la varianza y decidir si el modelo de coche tiene una influencia significativa sobre el consumo (al nivel de significación 0,05).

d) Comparar de dos en dos el consumo medio de los cuatro modelos de coche, con un nivel de confianza conjunto del 95 %. ¿Conclusiones?

e) Analizar los datos sin tener en cuenta las condiciones en que se conducen los coches, realizando un análisis de la varianza con un solo factor. Con este modelo, ¿influye el modelo de coche en el consumo de gasolina? Comparar los resultados con los obtenidos anteriormente.

3.– Se quiere hacer un estudio de comparación pluviométrica entre 5 ciudades de una misma región. Para esto, se mide la lluvia recogida en esas 5 ciudades en 4 meses diferentes:

	Enero	Abril	Julio	Octubre	$\bar{y}_{i\cdot}$
Ciudad A	11	16	10	17	13.50
Ciudad B	9	9	8	12	9.50
Ciudad C	12	9	9	10	10.00
Ciudad D	11	10	10	12	10.75
Ciudad E	19	18	20	14	17.75
$\bar{y}_{\cdot j}$	12.4	12.4	11.4	13.0	$\bar{y}_{\cdot\cdot} = 12.3$

Además $\sum_{i=1}^{5}\sum_{j=1}^{4}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = 262{,}2$.

a) ¿Qué diseño se ha seguido en el experimento? Escribir el modelo.

b) Construir la tabla ANOVA y contrastar con nivel de significación 0.05 la hipótesis de que no hay diferencias pluviométricas entre las cinco ciudades.

c) Construir un intervalo de confianza al 95 % para la diferencia media de lluvia recogida entre las ciudades A y E. Con nivel de significación 0.05, ¿existe evidencia para rechazar que estas dos ciudades son iguales?

4.– En 12 grandes ciudades se hace un estudio sobre la tasa de contaminación (con cierto contaminante atmosférico). Se piensa que puede estar influida por dos factores:

- «Índice de pluviosidad», que actúa a 2 niveles (baja o alta pluviosidad).
- «Densidad de industria contaminante», que actúa a 3 niveles (baja, media o alta densidad).

Se obtienen 2 réplicas de cada posible combinación de los niveles de los 2 factores. Las tasas medias de contaminación de las dos réplicas se ofrecen a continuación:

	Densidad baja	Densidad media	Densidad alta
Baja pluviosidad	837,5	868	887
Alta pluviosidad	420,5	437	526,5

La variabilidad total viene dada por $\mathrm{SCT} = 513632,25$

Especificar modelo, hipótesis y obtener la tabla de análisis de la varianza adecuada para contestar, razonadamente, a las siguientes preguntas: ¿Existe interaccion significativa entre los 2 factores? ¿Influye apreciablemente el índice de pluviosidad sobre la tasa de contaminación? ¿Influye la densidad de industria contaminante? Dar respuestas al nivel de significación $0,05$.

5.– Si en el ejercicio de las gasolineras del capítulo anterior, los cuatro primeros datos para cada gasolinera se tomaron inmediatamente después de adquirir los surtidores, y los cuatro últimos 6 meses más tarde, plantear el modelo adecuado para investigar si hay diferencias entre las gasolineras.

a) Para este nuevo modelo, obtener la tabla ANOVA y decidir, para $\alpha = 0{,}05$, qué factores influyen en la cantidad de gasolina que se suministra por litro.

b) A partir de los resultados del apartado anterior, simplificar el modelo (si es preciso), obtener la tabla ANOVA para el modelo simplificado, y comentar los resultados.

6.– Una gran empresa desea saber si el absentismo laboral está relacionado con el tamaño del departamento y la antigüedad. Para el estudio se dispone de una muestra aleatoria de 60 empleados, de la que se conoce el número de días que no acudieron al puesto de trabajo en los últimos tres años. El tamaño del departamento se clasifica en *pequeño*, *mediano* y *grande*, y la antigüedad en *más de 5 años* y *menos de 5 años*. Los datos son:

		TAMAÑO DEL DEPARTAMENTO					
ANTIGÜEDAD		**Pequeño**		**Mediano**		**Grande**	
Más de 5 años		0	2	2	4	15	16
		2	0	4	3	10	7
		1	5	7	1	8	30
		3	6	12	5	5	3
		0	8	15	20	25	27
	Media	2.7		7.3		14.6	
Menos de 5 años		0	2	5	1	10	15
		1	7	3	3	8	4
		1	4	2	6	12	9
		0	0	0	7	3	6
		4	3	1	9	7	1
	Media	2.2		3.7		7.5	

a) Plantear el modelo modelo adecuado para analizar estos datos y estimar los parámetros.

b) Obtener la tabla de análisis de la varianza.

c) Para un nivel de significación del 5 %, ¿los contrastes F para el efecto de la antigüedad y el tamaño del departamento indican que hay que rechazar las hipótesis nulas correspondientes? ¿Qué podemos decir sobre la interacción?

d) Calcular un intervalo de confianza al 95 % para la diferencia entre los efectos de los grupos *más de 5 años* y *menos de 5 años* de antigüedad.

7.– Para estudiar el efecto de la iluminación (A=natural, B=muy fuerte, C=escasa) en la velocidad de lectura se realiza un experimento. Se mide el número de palabras leídas en un minuto para distintos tipos de papel y tamaño de letra. Los resultados que se obtienen son los siguientes:

	Papel satinado	**Papel blanco**	**Papel color**
Letra grande	258 **A**	230 **C**	240 **B**
Letra normal	235 **B**	270 **A**	240 **C**
Letra pequeña	220 **C**	225 **B**	260 **A**

¿Cuántos factores se consideran en el experimento? Construir con SPSS la tabla de análisis de la varianza y contrastar con un nivel de significación $\alpha = 0{,}05$ si los factores afectan a la velocidad de lectura.

8.– Se realiza un seguimiento para estudiar la posible influencia de dos factores sobre el número de visitantes a los parques nacionales. Los factores considerados son: el clima (seco o húmedo) y el departamento encargado de la conservación (A, B ó C). Los datos que se obtienen son:

	A	B	C
Seco	60	73	85
Húmedo	63	69	88

a) Plantear el modelo y las hipótesis asumidas para hacer el estudio con los datos disponibles, razonando la elección del modelo. Estimar la influencia adicional del clima húmedo sobre el número medio global de visitantes.

b) ¿Influye el clima sobre el número de visitantes? ¿Influye el departamento encargado de la conservación? Dar respuestas razonadas con un nivel de significación del 5 %.

9.– Un laboratorio de medición atmosférica ha adquirido un nuevo equipo (que llamaremos B) para medir ozono. Para evaluar si el nuevo equipo está calibrado, se realiza un pequeño experimento en 5 observatorios diferentes. En cada observatorio se toma una medida con el nuevo equipo y otra con el equipo antiguo (que llamaremos A), obteniéndose los siguientes resultados:

Observatorio	1	2	3	4	5	Media
Equipo A	215	305	247	221	286	254.8
Equipo B	224	312	251	232	295	262.8

Teniendo en cuenta que la suma de cuadrados totales (SCT) es 12491.6, proponer un modelo adecuado para explicar los niveles de ozono que se han observado en el experimento y contrastar si existen diferencias significativas entre los dos equipos, con nivel de confianza 0.95.

10.– En una investigación de laboratorio se emplean cámaras de crecimiento para estudiar el desarrollo de ciertos microorganismos cuando se varían las concentraciones de CO_2 (baja y alta), y la temperatura (baja, media y alta). En distintas cámaras se cruzan todos los niveles de los dos factores y se obtienen tres réplicas completas del experimento.

La siguiente tabla muestra los crecimientos medios que se obtienen para cada combinación de los dos factores:

	T. baja	T. media	T. alta
Conc. baja	51	46	42
Conc. alta	59	54	48

Además, sabemos que SCT=600. Nos planteamos son las siguientes cuestiones: ¿Influye la concentración de CO_2 sobre el crecimiento? ¿Influye la temperatura sobre el crecimiento? ¿Se produce alguna interacción apreciable entre la concentración de CO_2 y la temperatura?

Proponer un modelo adecuado y hacer el estudio para dar una respuesta razonada a estas tres preguntas a un nivel de significación 0,10.

11.– Se hace un estudio para ver de qué manera influyen el tipo de población (HABITAT) y el tipo de vivienda (TIPOVIV) sobre la cantidad de papel y cartón reciclados. Para esto, se toman datos del «número de Kg. reciclados por vivienda en un mes» en 9 viviendas pequeñas (3 en ciudades pequeñas, 3 en ciudades medianas y 3 en ciudades grandes) y en 9 viviendas grandes (3 en ciudades pequeñas, 3 en ciudades medianas y 3 en ciudades grandes). Se analizan los resultados con el SPSS, obteniéndose los siguientes resultados:

Fuente	Suma cuadrados	G.l.	Media cuadrática	F	Significación
HABITAT	0,333	2	0,167	0,008	0,992
TIPOVIV	410,889	1	410,889	19,210	0,001
HABITAT*TIPOVIV	4,111	2	2,056	0,096	0,909
Error	256,667	12	21,389		
Total	672,000	17			

a) ¿Influye el tipo de población (HABITAT) sobre la cantidad reciclada? ¿Influye el tipo de vivienda (TIPOVIV) sobre la cantidad reciclada? ¿Existe interacción significativa entre los dos factores? Dar respuestas razonadas al nivel de significación $0,05$ e indicar el modelo estadístico utilizado.

b) Con los mismos datos, consideramos ahora un modelo de diseño de experimentos con un sólo factor (el tipo de vivienda). Construir la tabla ANOVA para este diseño y tomar una decisión razonada (al nivel 0,05) sobre si el tipo de vivienda influye o no sobre la cantidad reciclada.

12.– En 12 grandes ciudades se hace un estudio sobre la tasa de contaminación (con cierto contaminante atmosférico). Se cree que puede estar influida por el factor «densidad de industria contaminante». Cuatro de las ciudades tienen una baja densidad de industria contaminante, otras 4 ciudades tiene una densidad media y, finalmente, otras 4 ciudades poseen una densidad alta.

Se adjunta la salida del estudio estadístico efectuado.

ANOVA

	Suma de cuadrados	g. l.	Media cuadrática	F	Significación
Inter-grupos	$12\,720,500$	2	$6\,360,250$	0,114	0,893
Intra-grupos	$500\,911,750$	9	$55\,656,861$		
Total	$513\,632,250$	11			

Comparaciones múltiples (intervalos de confianza al 90 %)

Nivel densidad (I)	Nivel densidad (J)	Diferencia (I-J)	Error típico	Lim. inf.	Lim. sup.
Densidad baja	Densidad media	-23,50	166,81856	-442,1457	395,1457
	Densidad alta	-77,75	166,81856	-496,3957	340,8957
Densidad media	Densidad baja	23,50	166,81856	-395,1457	442,1457
	Densidad alta	-54,25	166,81856	-472,8957	364,3957
Densidad alta	Densidad baja	77,75	166,81856	-340,8957	496,3957
	Densidad media	54,25	166,81856	-364,3957	472,8957

a) Especificar claramente el modelo y las hipótesis que se han considerado.

b) Estimar, con una confianza del 90 %, la varianza común de las observaciones.

c) ¿Podemos afirmar que la densidad de industria contaminante influye sobre la tasa de contaminación? Dar una respuesta razonada al nivel de significación 0,10.

d) Deseamos hacer una comparación simultánea de los 3 grupos de ciudades (al nivel de significación conjunto de 0,10) para decidir entre qué grupos se observan diferencias significativas de la tasa de contaminación. Obtener conclusiones razonadas.

13.– Unos laboratorios desean analizar la eficacia de cuatro analgésicos infantiles ante las cefaleas. Para ello, realizan un experimento en el que se proporcionan estos analgésicos a niños con cinco diferentes tipos de cefalea. Se combinan de todas las formas posibles los analgésicos con los tipos de

cefalea, obteniéndose, en total, 20 datos sobre el tiempo de remisión de la cefalea. Cuando se analizan los datos con SPSS se obtienen los resultados que se adjuntan al final.

a) Describe detalladamente el modelo empleado. A la vista del estudio, ¿qué podemos decir sobre las hipótesis del modelo? Dar respuestas razonadas.

b) ¿Influye significativamente el tipo de analgésico sobre el tiempo de remisión? ¿y el tipo de cefalea? Dar respuestas al nivel $0,05$.

c) El laboratorio comercializa los analgésicos A y D y está especialmente interesado en ellos. ¿Cuál de ellos parece ser el más efectivo? ¿Existe diferencia significativa entre ellos con un nivel de significación $0,05$?

d) Con los mismos datos, consideramos un modelo de diseño de experimentos con un sólo factor: el tipo de analgésico empleado. Construir la nueva tabla ANOVA. ¿Influye significativamente el tipo de analgésico sobre el tiempo de remisión, al nivel $0,05$? Si la conclusión es diferente de la obtenida en el apartado b), ¿con qué conclusión nos quedamos? ¿por qué?

Tabla ANOVA[a]

Variable dependiente: Tiempo de remisión de la cefalea

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
analgésico	175,000	3	58,333	9,459	,002
cefalea	785,200	4	196,300	31,832	,000
Error	74,000	12	6,167		
Total	1034,200	19			

a. R cuadrado = ,995 (R cuadrado corregida = ,992)

1. Media global

Variable dependiente: Tiempo de remisión de la cefalea

Media	Error típ.	Intervalo de confianza al 95%.	
		Límite inferior	Límite superior
26,700	,555	25,490	27,910

2. Tipo de analgésico empleado

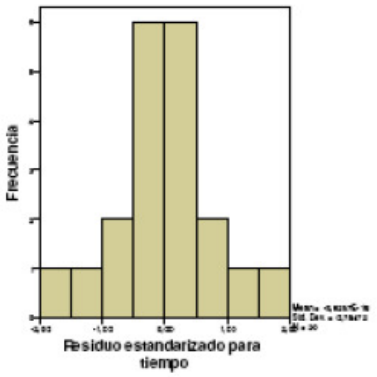
Variable dependiente: Tiempo de remisión de la cefalea

Tipo de analgésico empleado	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
Analgésico A	26,400	1,111	23,980	28,820
Analgésico B	24,000	1,111	21,580	26,420
Analgésico C	24,800	1,111	22,380	27,220
Analgésico D	31,600	1,111	29,180	34,020

3. Tipo de cefalea

Variable dependiente: Tiempo de remisión de la cefalea

Tipo de cefalea	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
Primer tipo de cefalea	29,500	1,242	26,795	32,205
Segundo tipo de cefalea	17,500	1,242	14,795	20,205
Tercer tipo de cefalea	24,000	1,242	21,295	26,705
Cuarto tipo de cefalea	36,500	1,242	33,795	39,205
Quinto tipo de cefalea	26,000	1,242	23,295	28,705



III. Modelo de regresión simple

1.– Se quiere estudiar la posible relación lineal entre Y = «Porcentaje de asfalteno» y X = «Porcentaje de resina» en asfaltos utilizados para la fabricación de telas asfálticas. Se dispone de datos de 22 tipos diferentes de asfaltos:

X	22	21	25	23	29	26	25	27	25	21	24	26	23	24	22	27	29	24	24	27	24	34
Y	35	35	29	32	27	29	28	31	30	36	39	33	31	31	36	26	32	31	29	27	27	23

a) Plantear modelo e hipótesis. Mediante el análisis de los residuos, ¿qué se puede decir sobre dichas hipótesis?

b) Obtener la recta de regresión y el coeficiente de correlación lineal r. ¿Qué indica el valor del coeficiente de correlación obtenido?

c) ¿Influye el porcentaje de resina sobre el porcentaje de asfalteno? Obtener una conclusión al nivel de significación $0,01$.

d) Estimar, con una confianza del 95 %, el valor medio del porcentaje de asfalteno para aquellos asfaltos que tienen un 30 % de resina.

2.– Ajustar un modelo de regresión adecuado, de Y sobre X, a los siguientes pares de datos:

x	1	2	3	4	5
y	1	9	90	900	12000

3.– Se estudia la influencia en el nivel de contaminación por nitratos (Y) del % de población conectada a sistemas de tratamiento de residuos (X) en 20 áreas de la UE. Los datos obtenidos son los siguientes

$$\sum x_i = 692 \quad \sum \log x_i = 67{,}34 \quad \sum y_i = 82{,}7 \quad \sum x_i y_i = 2332{,}9$$
$$\sum x_i^2 = 30430 \quad \sum (\log x_i)^2 = 235{,}01 \quad \sum y_i^2 = 432{,}49 \quad \sum y_i \log x_i = 262{,}37$$

Ajustar un modelo de regresión logarítmico $Y = \beta_0 + \beta_1 \log X$. ¿Es bueno este ajuste?

4.– En un estudio se trata de explicar la supervivencia de cierta especie animal en función de las temperaturas máximas alcanzadas en los hábitats naturales en los que se desarrolla. Se seleccionan aleatoriamente 20 reservas naturales de esta especie y se mide el porcentaje de supervivientes al final del año, Y, y la temperatura máxima registrada en ^oF, X. Los resultados que se obtienen son:

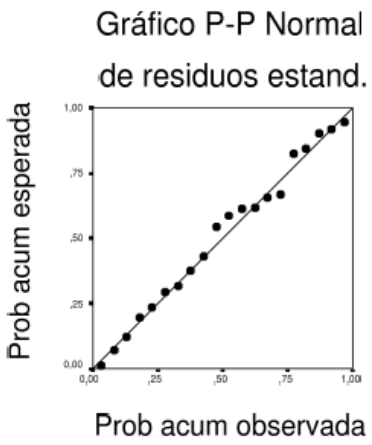
$$\sum_{i=1}^{n} x_i = 1379 \quad \sum_{i=1}^{n} y_i = 823 \quad \sum_{i=1}^{n} x_i^2 = 100055 \quad \sum_{i=1}^{n} y_i^2 = 42063 \quad \sum_{i=1}^{n} x_i y_i = 62103$$

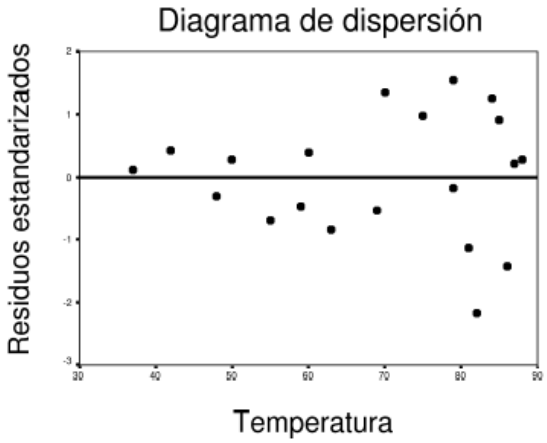
a) Calcular la recta de regresión.

b) Calcular la varianza residual S_R^2.

c) Realizar el contraste de la regresión. A nivel $\alpha = 0{,}05$, ¿podemos rechazar la hipótesis nula de que la temperatura no afecta a la supervivencia?

d) A continuación se presentan algunos gráficos de residuos estandarizados. Analizar gráficamente si se cumplen las hipótesis de normalidad, homocedasticidad y linealidad.





5.– Se obtuvieron los siguientes datos de la latitud del espacio natural de cría (x) y la duración del periodo de cría en días (y) de $n = 11$ especies de patos buceadores:

$$\sum_{i=1}^{n} x_i = 548 \qquad \sum_{i=1}^{n} x_i^2 = 28230 \qquad \sum_{i=1}^{n} y_i = 620 \qquad \sum_{i=1}^{n} y_i^2 = 42626 \qquad \sum_{i=1}^{n} x_i y_i = 28895$$

a) Obtener la recta de regresión de Y sobre X.

b) Teniendo en cuenta que

$$S_R^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = 379,$$

decidir si la latitud tiene alguna influencia sobre la duración del período de cría (al nivel de significación $0,05$) y calcular el porcentaje de variabilidad explicado por la regresión.

c) Calcular un estimador de la duración media del periodo de cría para aves cuyo espacio natural de cría está a una latitud de 35 grados. Hallar un intervalo de confianza del 95 % para este parámetro.

d) Calcular un estimador de la longitud del periodo de cría para un ave cuyo espacio natural de cría está a una latitud de 35 grados. Hallar un intervalo de confianza del 95 % para el valor predicho.

6.– Un estudio sobre el efecto de la temperatura en el rendimiento de un proceso químico proporciona los siguientes resultados:

Temperatura (x)	-5	-4	-3	-2	-1	0	1	2	3	4	5
Rendimiento (y)	1	5	4	7	10	8	9	13	14	13	18

a) Asumiendo el modelo $y_i = \beta_0 + \beta_1 x_i + u_i$, calcular los estimadores mínimo cuadráticos de β_0 y β_1. ¿Cuál es la recta de regresión estimada?

b) Construir la tabla ANOVA y contrastar la hipótesis. $H_0 : \beta_1 = 0$ con un nivel de significación $\alpha = 0{,}05$

c) Construir un intervalo de confianza al 95 % para β_1.

d) Construir un intervalo de confianza al 95 % para la verdadera media de la distribución de y cuando $x = 3$.

e) Construir un intervalo de confianza al 95 % para la predicción del rendimiento de un nuevo proceso que se realice a temperatura $x = 3$.

7.– El siguiente conjunto de datos corresponde a una muestra de las aguas de 20 lagos en Estados Unidos. Para cada lago se calculan el número de factorías por kilómetro que hay situadas en la orilla (x) y el porcentaje de impurezas en el agua (y):

x	y	x	y	x	y	x	y
5.00	0.005	8.20	0.910	1.14	0.003	1.47	0.009
8.90	0.430	1.01	0.008	1.18	0.010	3.00	0.150
7.15	0.009	0.61	0.030	0.73	0.001	2.40	0.250
6.10	0.240	0.72	0.010	0.40	0.950	4.10	0.200
7.70	0.007	0.68	0.005	0.87	0.200	4.00	0.010

a) Plantear un modelo de regresión para explicar el porcentaje de impurezas en el agua en función del número de factorías. Hallar la recta de regresión y decidir si el número de factorías influye sobre el porcentaje de impurezas (al nivel $0,05$).

b) Repetir el estudio eliminando el dato atípico $x = 8{,}20$, $y = 0{,}910$. Comentar las diferencias que se observan.

8.– En algunas reservas naturales se controla el número Y de ejemplares de cierta especie al final del año, y la temperatura media anual X.

Se ajusta un modelo de regresión lineal $y_i = \beta_0 + \beta_1 x_i + u_i$ con los datos de seis reservas, para explicar la dependencia entre el tamaño de la población y la temperatura media anual. A continuación se da un resumen de la salida obtenida con el SPSS:

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	237,729	1	237,729	860,741	,000[a]
	Residual	1,105	4	,276		
	Total	238,833	5			

[a] Variables predictoras: (Constante), temperatura
[b] Variable dependiente: Y

Coeficientes[a]

		Coeficientes no estandarizados		Coeficientes estandarizados		
Modelo		B	Error típ.	Beta	t	Sig.
1	(Constante)	100,933	,489		206,302	,000
	temperatura	3,686	,126	,998	29,338	,000

[a] Variable dependiente: Y

a) Obtener estimaciones puntuales de los tres parámetros del modelo.

b) Estimar, mediante un intervalo de confianza al 95 %, el parámetro β_0 del modelo de regresión. ¿Qué interpretación tiene esta estimación?

c) ¿Tiene influencia la temperatura media anual sobre el tamaño de la población? Dar una respuesta al nivel de significación 0.01. Evaluar la fuerza de la relación mediante el coeficiente adecuado.

9.– Una empresa quiere estudiar cómo influye la inversión en publicidad local sobre el nivel de ventas de una nueva agenda electrónica. Hace un estudio en varias ciudades con el objeto de expresar la variable respuesta Y="Número de unidades vendidas en una semana" en función de X="Miles de euros en publicidad local". A continuación, se ofrecen los resultados obtenidos con SPSS en una regresión lineal de Y sobre el logaritmo neperiano de X.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.974	.948	.922	2.04

ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	152.438	1	152.438	36.680	.026
	Residual	8.312	2	4.156		
	Total	160.750	3			

Coeficientes

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	38.683	3.704		10.444	.009
	Log (X)	18.019	2.975	.974	6.056	.026

a) Utilizar estos resultado para hallar una regresión logarítmica de Y sobre X.

b) Hallar el coeficiente de correlación lineal entre Y y $\log X$. ¿Qué significa?

c) ¿Cuál sería el número esperado de unidades vendidas en una semana con una inversión de 3500 euros en publicidad local?

10.– Se lleva a cabo un estudio del nivel de glucosa (Y) en 20 pacientes diabéticos en función de la duración de la enfermedad (X). Se dispone de los siguientes resultados:

$\sum x_i = 3714 \qquad \sum x_i^2 = 793444$

$\sum y_i = 1682 \qquad \sum y_i^2 = 143294$

$\sum x_i y_i = 312532$

a) Calcular la recta de regresión de Y sobre X.

b) Calcular la varianza residual.

c) Obtener la tabla de análisis de la varianza. ¿Se puede concluir, al nivel de significación 0,05, que la duración de la enfermedad influye sobre el nivel de glucosa?

IV. Modelo de regresión múltiple

1.– Los datos de la siguiente tabla corresponden a un estudio sobre la contaminación acústica realizado en distintas zonas de la misma ciudad. La variable y mide la contaminación acústica en decibelios y las variables x_1 y x_2 la hora del día y el tráfico de vehículos por minuto, respectivamente.

Decibelios	0.9	1.6	4.7	2.8	5.6	2.4	1.0	1.5
Hora	14	15	16	13	17	18	19	20
Vehículos	1	2	5	2	6	4	3	4

Analizar un modelo de regresión múltiple para explicar el número de decibelios en función de las otras variables.

2.– En el Ayuntamiento de Madrid se hizo un estudio hace varios años sobre la conveniencia de instalar mamparas de protección acústica en una zona de la M-30. Un técnico del Ayuntamiento piensa que si el ruido afecta mucho a los habitantes de la zona esto debe reflejarse en los precios de las viviendas. Su idea es que el precio de una casa, en miles de pesetas, en esa zona (y) depende del número de metros cuadrados (x_1), del número de habitaciones (x_2) y de la contaminación acústica, medida en decibelios, (x_3). Para una muestra de 20 casas vendidas en los últimos tres meses, se estima el siguiente modelo:

$$\hat{y}_i = 5970 + \underset{(2{,}55)}{22{,}35\, x_{1i}} + \underset{(1820)}{2701{,}1\, x_{2i}} - \underset{(15{,}4)}{67{,}6730\, x_{3i}} \qquad R^2 = 0{,}9843,$$

donde los errores típicos de las estimaciones de los coeficientes aparecen entre paréntesis.

a) ¿Qué hipótesis hemos tenido que asumir sobre el modelo que estima el técnico?

b) Calcular el efecto que tendría sobre el precio un descenso de 10 decibelios.

c) Contrastar con $\alpha = 0{,}05$ la hipótesis nula de que el número de habitaciones no influye en el precio.

d) Con $\alpha = 0{,}05$, ¿se puede afirmar que el efecto de la contaminación acústica es reducir el precio?

e) Contrastar con $\alpha = 0{,}05$ la hipótesis nula de que las tres variables no influyen conjuntamente en el precio.

f) Estimar el precio medio de las casas (no incluidas en la muestra) que tienen 100 metros cuadrados, dos habitaciones y una contaminación acústica de 40 decibelios.

3.– Se dispone de datos sobre la «duración de la estancia (en horas) en la UVI» de 17 pacientes, de su «índice de gravedad» y del «tamaño del hospital» (codificado con 0 si es grande y con 1 si es pequeño):

duración	24	48	37	81	48	2	24	64	12	8	4	36	12	8	4	88	54
gravedad	22	80	55	140	90	10	40	120	30	25	10	77	32	14	15	200	120
tamaño	0	0	0	1	0	1	1	1	1	1	1	0	0	1	1	0	0

Se lleva a cabo un análisis de regresión lineal con el SPSS para tratar de explicar la duración de la estancia en función de las otras dos variables.

a) Plantear el modelo utilizado.

b) ¿Es razonable asumir normalidad, linealidad y homocedasticidad?

c) Aceptando la validez del modelo de regresión, ¿cuál es el aumento estimado en la duración de la estancia si el índice de gravedad aumenta 5 unidades? ¿Cuál es la duración media estimada de estancia en la UVI de hospitales pequeños para los enfermos con un índice de gravedad de 50?

d) Con una confianza del 95 %, ¿se puede aceptar que el modelo es explicativo? ¿Influye el índice de gravedad en la duración de la estancia? ¿Influye el tamaño del hospital en la duración de la estancia?

4.– Se desea hacer una regresión lineal que explique el contenido en sales minerales de la parte dominante del húmero en función de X_1=“Contenido en sales minerales de la parte dominante del

radio" y de X_2="Contenido en sales minerales de la parte no dominante del radio". Se toman datos en 5 personas y se analizan con el SPSS (los resultados se ofrecen a continuación).

¿Son explicativas las variables X_1 y X_2? ¿Es explicativo el modelo en su conjunto? Dar respuestas razonadas y claras, al nivel de significación 0,05. A partir de las respuestas anteriores obtener conclusiones razonadas sobre cuál sería el procedimiento a seguir para quedarnos con un modelo de regresión adecuado.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.989	.977	.955	3.4990E-02

ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	.106	2	5.289E-02	43.196	.023
	Residual	2.449E-03	2	1.224E-03		
	Total	.108	4			

Coeficientes

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	.619	.136		4.544	.045
	Radio (parte dominante)	.471	.326	.345	1.444	.286
	Radio (parte no dominante)	.956	.342	.668	2.798	.107

5.– En un conjunto de 60 ciudades, se estudia la tasa de mortalidad (por cada cien mil habitantes) en función del tamaño medio de las viviendas, del tanto por ciento de los empleados en trabajos no manuales, y del tanto por ciento de las familias con ingresos bajos. Los resultados obtenidos son los siguientes:

ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	57083.775	3	19027.925	6.223	.001
	Residual	171223.869	56	3057.569		
	Total	228307.644	59			

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	628.288	237.174		2.649	.010
	Tamaño casas	99.033	60.081	.215	1.648	.105
	% trab. no manuales	-1.780	1.731	-0.132	-1.028	.308
	% fam. ing. bajos	4.936	1.799	.330	2.744	.008

a) Indica el modelo usado en el análisis y las hipótesis asumidas.

b) A la vista de los gráficos de la figura 1, ¿es razonable asumir estas hipótesis? Explicar detalladamente las conclusiones.

c) A nivel de significación 0,01, decidir si las variables consideradas son explicativas y si el modelo es conjuntamente explicativo. Indicar en qué se fundamenta cada respuesta.

d) Dar una estimación puntual de la tasa de mortalidad de una ciudad en la que el tamaño medio de las viviendas fuera 3,2, el % de trabajadores no manuales fuera del 51 %, y el % de familias con ingresos bajos fuera del 11 %. Dar una estimación para la tasa de mortalidad media para los mismos valores de las variables regresoras.

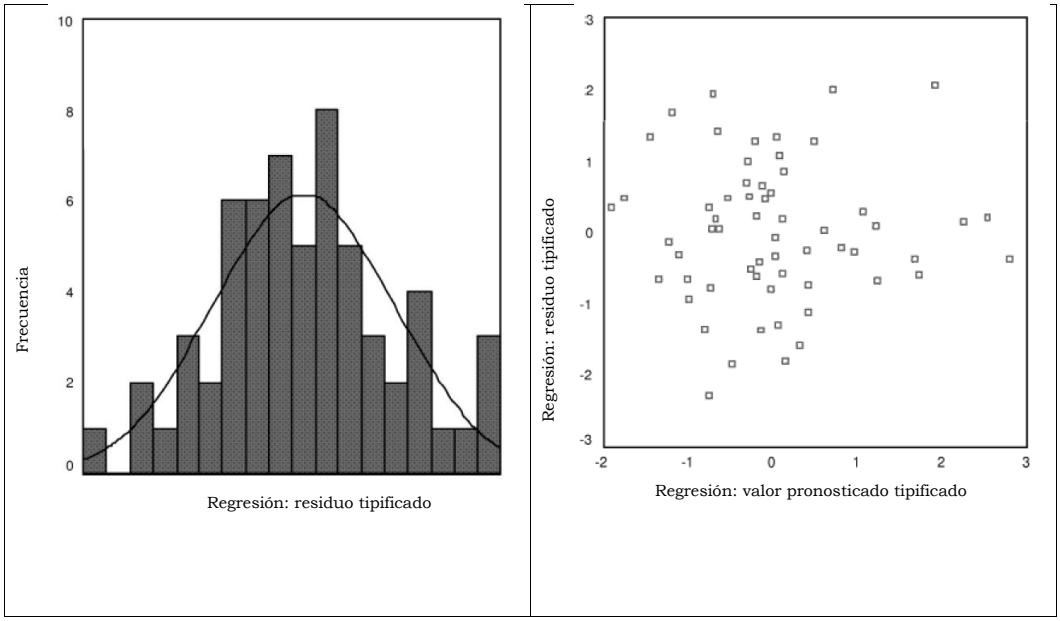


FIGURA 1.

e) A nivel de significación 0,05, ¿podemos concluir que la variable «% de familias con ingresos bajos» influye positivamente en la tasa de mortalidad?

6.– Se hace un estudio para ver de qué manera influyen el «número de metros cuadrados» y el «número de moradores» de la vivienda sobre la cantidad de papel y cartón reciclados. Para esto, se toman datos del «número de Kg. reciclados por vivienda en un mes» en 18 viviendas de diferentes tamaños. Se analizan los datos con el SPSS y se obtienen los resultados que se adjuntan en la figura 2 de la página 16.

Consideramos primero el siguiente análisis de regresión lineal simple:

a) ¿Influye el número de moradores sobre la cantidad reciclada? Dar una respuesta razonada al nivel de significación 0,05.

b) Estimar la cantidad media reciclada en aquellas viviendas en las que viven 4 personas.

c) Predecir, al nivel 0,95, la cantidad reciclada en una nueva vivienda donde vivan 4 personas, sabiendo que el número medio de moradores en la muestra era de 3,11 y su varianza era de 2,10.

Para finalizar, consideramos el siguiente análisis de regresión múltiple:

d) ¿Es significativamente influyente la superficie de la vivienda? ¿Es significativamente influyente el número de moradores? ¿Es explicativo el modelo en su conjunto? Dar respuestas razonadas al nivel 0,05.

e) Comparar, razonadamente, los resultados obtenidos en los apartados a) y d), y sacar una conclusión sobre qué modelo o modelos deberíamos utilizar.

7.– Un equipo de ornitólogos está estudiando la posible relación entre el tiempo de incubación de los huevos de una especie de ave y el déficit de saturación (humedad) del aire circundante. Se hace un seguimiento de 10 huevos con diferente déficit de saturación. Los resultados del análisis de regresión lineal obtenidos con SPSS se dan en el Cuadro 1 de la página 17; la variable predictora es el déficit de saturación y la variable dependiente es el tiempo de incubación.

a) ¿Influye significativamente el déficit de saturación sobre el tiempo de incubación? Dar una respuesta al nivel 0,05.

b) Con una confianza del 95 %, ¿cuál sería el tiempo de incubación de un nuevo huevo si el déficit de saturación es de 1,1?

Regresión lineal simple

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,927(a)	,858	,850	2,44

a Variables predictoras: (Constante), Número de moradores

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	576,898	1	576,898	97,057	,000(a)
	Residual	95,102	16	5,944		
	Total	672,000	17			

a Variables predictoras: (Constante), Número de moradores
b Variable dependiente: Número de Kg. recogidos

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	11,174	1,392		8,026	,000
	Número de moradores	4,016	,408	,927	9,852	,000

a Variable dependiente: Número de Kg. recogidos

Regresión lineal múltiple

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,966(a)	,933	,924	1,74

a Variables predictoras: (Constante), Superficie de la vivienda, Número de moradores

ANOVA(b)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	626,781	2	313,390	103,958	,000(a)
	Residual	45,219	15	3,015		
	Total	672,000	17			

a Variables predictoras: (Constante), Superficie de la vivienda, Número de moradores
b Variable dependiente: Número de Kg. recogidos

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	8,692	1,164		7,466	,000
	Número de moradores	,469	,919	,108	,511	,617
	Superficie de la vivienda	,165	,041	,862	4,068	,001

a Variable dependiente: Número de Kg. recogidos

FIGURA 2. Influencia de la calidad de residencia en el reciclado de papel

c) Si quisiéramos saber, con una confianza del 95 %, el tiempo de incubación de un nuevo huevo con un déficit de saturación de 1,5, ¿el intervalo sería más grande o más pequeño que en el apartado anterior? Dar una respuesta razonada sin hacer ningún cálculo.

8.– En 12 grandes ciudades se hace un estudio sobre la tasa de contaminación (con cierto contaminante atmosférico). Se piensa que puede estar influida por tres variables:

X_1=«Índice de pluviosidad»,

X_2=«Densidad de industrias contaminantes en el término municipal»,

X_3=«Millones de habitantes».

Se adjunta en la Figura 3 la salida del estudio de regresión lineal efectuado.

a) Especificar claramente el modelo y las hipótesis que se han considerado.

ANOVA

	Suma cuadrados	g.l.	Media cuadrática	F	Significación
Regresión	48,185	1	48,185	436,307	0,000
Residual	0,884	8	0,110		
Total	49,069	9			

Coeficientes

	B	Error típico	t	Significación
Constante	16,551	0,195	84,735	0,000
Déficit de saturación	3,821	0,183	20,888	0,000

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típica
Déficit de saturación	10	0,0	1,8	0,900	0,6055
Tiempo de incubación	10	16,6	23,2	19,990	2,3350

CUADRO 1. Tiempo de incubación

b) ¿Cuál sería la tasa de contaminante estimada con el modelo de regresión para una ciudad con 3 millones de habitantes, un índice de pluviosidad de 250 y una densidad de industria contaminante de 25?

c) ¿Qué variables son indivualmente explicativas? ¿El modelo es globalmente explicativo? Dar respuestas razonadas al nivel de significación 0,05. ¿Se detecta algún problema de multicolinealidad?

d) ¿Podemos afirmar que el aumento del índice de pluviosidad disminuye la tasa de contaminación? Dar una respuesta razonada al nivel de significación $0,05$.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,999[a]	,998	,997	12,13473

a. Variables predictoras: (Constante), Millones de habitantes, Indice de pluviosidad, Densidad de industrias contaminantes

b. Variable dependiente: Tasa registrada de contaminante

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	512454,237	3	170818,079	1160,042	,000[a]
	Residual	1178,013	8	147,252		
	Total	513632,250	11			

a. Variables predictoras: (Constante), Millones de habitantes, Indice de pluviosidad, Densidad de industrias contaminantes

b. Variable dependiente: Tasa registrada de contaminante

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	1010,036	14,648		68,956	,000
	Indice de pluviosidad	-2,023	,035	-,987	-58,205	,000
	Densidad de industrias contaminantes	2,016	,184	,196	10,962	,000
	Millones de habitantes	-3,271	4,524	-,013	-,723	,490

a. Variable dependiente: Tasa registrada de contaminante

FIGURA 3. Influencia de diversas variables en la contaminación atmosférica