Análisis de datos

Apellidos, nombre: *Grupo de prácticas:*
DNI:

Notas:

1) Esta práctica se realizará los días 3, 4, 5 y 6 de octubre de 2023.
2) La fecha límite de entrega de esta práctica, que debe presentarse en el Campus Virtual UVa, son las 14 horas del miércoles **11 de octubre** de 2023.

En esta práctica se utilizará el fichero de datos **Datos-Valladolid.sgd**.

El fichero **Datos-Valladolid.sgd** es un fichero en formato *Statgraphics* que contiene datos de los 41 municipios de la provincia de Valladolid con más de 1000 habitantes en una determinada fecha F_1. La información que aparece recogida en el fichero proviene de diferentes fuentes oficiales.

Para cada municipio, la información aparece organizada en dos apartados. El primero recoge datos relativos al territorio y la población. El segundo ofrece indicadores sobre el empleo y la contratación laboral (datos referidos a fecha F_2, posterior a F_1).

Así, las variables que aparecen en este fichero son las siguientes:

TERRITORIO Y POBLACIÓN

Municipio	Denominación del municipio
Partido	Partido judicial al que pertenece el municipio
Población	Población oficial, a fecha F_1

ESTRUCTURA PRODUCTIVA

Tasa Paro	Tasa de paro, a fecha F_2
Demandantes Empleo	Número de personas demandantes de empleo, a fecha F_2
Paro	Paro registrado, datos referidos a fecha F_2
CAg	Número de contratos registrados en el sector de la Agricultura, a fecha F_2
CIn	Número de contratos registrados en el sector de la Industria, a fecha F_2
CCo	Número de contratos registrados en el sector de la Construcción, a fecha F_2
CSe	Número de contratos registrados en el sector Servicios, a fecha F_2

En el anexo se presentan los datos que aparecen en el fichero.

Este fichero de datos puede ser obtenido:

1) En \\sanson\Estadistica de la red del laboratorio de Informática.
2) En Campus Virtual UVa.

Importar el fichero **Datos-Valladolid.sgd** al programa *Statgraphics* y almacenarlo en memoria.

EJERCICIO 1

1.1 Clasificar las variables que aparecen en el fichero. Justificar la respuesta.

1.2 Dibujar un diagrama de sectores para la variable *Partido*. ¿Cuál es el porcentaje de municipios vallisoletanos, entre los que tienen más de 1000 habitantes, que pertenecen al partido judicial de Valladolid?

1.3 Insertar una nueva variable al fichero, *Sector*, para clasificar los municipios en cuatro grupos: (i) Industrial (si más del 14.8 % de los contratos registrados se ha realizado en el sector industrial), (ii) Agrícola (si no es industrial y más del 14.8 % de los contratos registrados se ha realizado en el sector de la agricultura), (iii) Construcción (si no es ni industrial ni agrícola y más del 14.8 % de los contratos registrados se ha realizado en el sector de la construcción) y (iv) Servicios, en caso contrario. Construir la tabla de contingencia de las variables *Partido* y *Sector*. ¿Qué porcentaje de los municipios del partido judicial de Valladolid aparecen clasificados como industriales? ¿Y del partido judicial de Medina del Campo? Razonar la respuesta.

EJERCICIO 2

2.1 ¿Qué municipio tiene una tasa de paro superada por el 69 % de los municipios? ¿Cuál es su población? ¿En qué sector aparece clasificado?

2.2 Calcular los tres cuartiles de la variable *Paro*, ¿a qué municipios corresponden?

2.3 Calcular la media de la variable *Paro* e indicar qué municipio tiene la tasa más cercana a ese valor.

2.4 Con el fin de comparar la población de los municipios distintos de la capital en función del sector en que fueron clasificados en el ejercicio anterior, construir los correspondientes diagramas de caja. Identificar, si existen, los outliers y los posibles outliers.

2.5 Realizar ahora un diagrama de caja para la variable *Población* en los municipios distintos de la capital. Identificar, si existen, los outliers y los posibles outliers.

2.6 Calcular los indicadores de centralización, dispersión y forma más adecuados para la variable *Población* en los municipios distintos de la capital, justificando la elección. ¿Qué se puede decir acerca de la estructura de esta variable?

2.7 Con el fin de eliminar los outliers de la variable anterior (*Población* en los municipios distintos de la capital), se propone realizar la transformación $Ln(Población - 500)$. Dibujar el diagrama de caja correspondiente, indicando, si existen, los valores "raros". Describir la estructura de esta nueva variable.

2.8 Calcular los cuartiles y los sextiles para la variable definida en el apartado anterior (es decir, $Ln(Población - 500)$ en los municipios distintos de la capital). Identificar los municipios a los que corresponden. Calcular, a partir de los valores anteriores, los correspondientes a la variable *Población*. ¿Coinciden estos valores con los obtenidos directamente para la variable *Población*? Justificar la respuesta.

EJERCICIO 3

3.1 Realizar el diagrama de tallo-hojas para la variable *Tasa Paro*.

3.2 Construir una tabla de frecuencias adecuada para la variable *Tasa Paro*.

3.3 Dibujar el histograma con el agrupamiento realizado en el apartado anterior.

3.4 Realizar un diagrama de caja para la variable *Tasa Paro*. Identificar, si existen, los outliers y posibles outliers. ¿Qué se puede decir acerca de la estructura de esta variable?

3.5 Calcular los indicadores de centralización, dispersión y forma más adecuados, justificando la elección.

3.6 Se clasifican los municipios en tres grupos: *Po*, *No* y *Mu*. El grupo *Po* está constituido por el 15 % de los municipios con menor tasa de paro; el grupo *Mu* por el 15 % de los municipios con mayor tasa de paro y el grupo *No* por el resto de municipios. Identificar, razonando la respuesta, los componentes de los grupos *Po* y *Mu*.

EJERCICIO 4

4.1 Realizar el diagrama de dispersión de las variables *CIn* y *Demandantes Empleo*. ¿Qué tipo de correlación se tiene? ¿Es adecuado realizar un ajuste lineal? Justificar la respuesta.

4.2 Construir una nueva variable, *Auxiliar Paro*, igual al producto de las variables *Tasa Paro* y *Población*. Repetir el apartado anterior con las variables *Paro* y *Auxiliar Paro*.

4.3 Calcular la recta de regresión de *Demandantes Empleo* sobre *Auxiliar Paro*.

4.4 ¿Qué número de demandantes de empleo se espera en un municipio con 1650 habitantes y una tasa de paro del 20 %? ¿Es adecuado el modelo? Justificar la respuesta.

4.5 Realizar el diagrama de dispersión de las variables ln(*Poblacion-500*) y *Demandantes Empleo*. ¿Qué tipo de relación entre las dos variables sugiere el gráfico?

4.6 Calcular la recta de regresión de ln(*DemandantesEmpleo*) sobre ln(*Poblacion-500*). ¿Qué número de demandantes de empleo se espera en un municipio con 2100 habitantes? ¿Y en un municipio de 1200 habitantes? Justificar la respuesta.

4.7 Guardar en la variable *Residuos* el valor de los errores que se cometen cuando se sustituyen los valores reales de la variable ln(*DemandantesEmpleo*) por los correspondientes a la recta de regresión calculada en el apartado anterior. Realizar un diagrama de caja y un histograma con esa variable. Identificar, si existen, los outliers e interpretar los resultados obtenidos.

Anexo

Municipio	Partido	Población	Tasa Paro	Demandantes Empleo	Paro	CAg	CIn	CCo	CSe
Alaejos	Medina del Campo	1429	30.81	145	126	9	0	1	21
Aldeamayor de San Martín	Valladolid	4891	16.49	366	300	2	47	9	64
Arroyo de la Encomienda	Valladolid	17572	18.43	1366	1083	0	39	21	647
Boecillo	Valladolid	3989	5.42	245	204	42	27	4	278
Cabezón de Pisuerga	Valladolid	3622	24.91	355	288	4	12	2	52
Campaspero	Valladolid	1174	14.24	55	45	176	2	0	27
Carpio	Medina del Campo	1068	25.58	117	99	30	1	0	16
Cigales	Valladolid	5008	36.22	497	410	3	7	23	82
Cistérniga	Valladolid	8734	16.67	749	606	2	46	18	179
Fuensaldaña	Valladolid	1468	16.83	126	104	0	4	1	24
Íscar	Valladolid	6678	19.53	718	558	53	49	7	126
Laguna de Duero	Valladolid	22555	29.99	2202	1787	12	30	43	220
Matapozuelos	Medina del Campo	1007	19.82	79	66	13	4	0	80
Mayorga	Medina de Rioseco	1687	15.98	116	89	4	0	3	34
Medina del Campo	Medina del Campo	21274	27.97	3115	2588	41	176	43	408
Medina de Rioseco	Medina de Rioseco	4906	14.26	441	351	6	23	2	137
Mojados	Medina del Campo	3184	14.85	252	198	11	10	8	92
Nava del Rey	Medina del Campo	2091	24.12	188	164	5	2	3	23
Olmedo	Medina del Campo	3759	13.52	352	264	13	281	8	206
Pedraja de Portillo, La	Valladolid	1134	25.8	86	73	1	0	0	6
Pedrajas de San Esteban	Valladolid	3503	22.41	316	264	17	29	3	182
Peñafiel	Valladolid	5428	16.83	508	414	34	61	5	68
Portillo	Valladolid	2409	17.26	162	130	170	9	4	31
Quintanilla de Onésimo	Valladolid	1109	9.39	56	45	3	5	0	32
Renedo de Esgueva	Valladolid	3507	27.07	267	206	0	0	21	48
Rueda	Medina del Campo	1332	15.31	131	113	14	10	2	10
Santovenia de Pisuerga	Valladolid	4155	21.55	439	359	0	30	14	100
Seca, La	Medina del Campo	1127	17.86	94	80	6	18	0	17
Serrada	Medina del Campo	1184	22.32	126	100	22	12	5	21
Simancas	Valladolid	5331	17.24	368	289	2	3	1	115
Tordesillas	Valladolid	8973	24.12	1045	832	131	12	34	253
Traspinedo	Valladolid	1126	27.08	90	78	8	0	2	38
Tudela de Duero	Valladolid	8717	31.37	831	710	19	8	27	153
Valdestillas	Medina del Campo	1742	28.37	164	143	1	2	4	117
Valladolid	Valladolid	306830	15.73	30755	24926	322	2103	674	8935
Viana de Cega	Valladolid	2031	27.52	142	120	2	1	1	36
Villabrágima	Medina de Rioseco	1054	18.21	68	61	1	0	1	10
Villalón de Campos	Medina de Rioseco	1733	18.71	141	107	4	0	3	20
Villanubla	Valladolid	2484	9.53	197	161	2	10	16	121
Villanueva de Duero	Medina del Campo	1202	34.67	114	95	18	0	1	4
Zaratán	Valladolid	6029	17.46	479	399	3	7	8	210