Problemas de Estadística Descriptiva. Resumen.

1. Un aeropuerto importante contrató recientemente al consultor Manuel López para estudiar el problema de los retrasos en el tráfico aéreo. En la siguiente tabla se puede observar la cantidad de minutos que los aviones llegaron tarde en una muestra de vuelos:

Minutos de retraso	0<10	10<20	20<30	30<40	40<50	50<60
Número de vuelos	30	25	13	6	5	4

a. Estime el número medio de minutos de retraso.
b. Estime la varianza y la desviación estándar de la muestra.

Solución

a. Para calcular el número medio de minutos de retraso, construimos la tabla siguiente donde la última fila en rojo hay las sumas las frecuencias y del producto de las frecuencias por los valores centrales:

Minutos de retraso	Frecuencia f_i (número de vuelos)	Valores centrales m_i	$f_i \cdot m_i$
0<10	30	5	150
10<20	25	15	375
20<30	13	25	325
30<40	6	35	210
40<50	5	45	225
50<60	4	55	220
Sumas	**83**		**1505**

El número medio de minutos de retraso será:

$$\overline{x} = \frac{1505}{83} = 18.133.$$

b. Para hallar la varianza y desviación típica, añadimos las columnas siguientes a la tabla anterior:

Minutos	f_i	m_i	$f_i \cdot m_i$	$m_i - \overline{x}$	$(m_i - \overline{x})^2$	$f_i \cdot (m_i - \overline{x})^2$
0<10	30	5	150	-13.133	172.475689	5173.9
10<20	25	15	375	-3.133	9.815689	245.319
20<30	13	25	325	6.867	47.155689	613.108
30<40	6	35	210	16.867	284.495689	1707.069
40<50	5	45	225	26.867	721.835689	3609.305
50<60	4	55	220	36.867	1359.175689	5436.841
Sumas	**83**		**1505**			**16785.542**

La varianza será:

$$s_x^2 = \frac{16785.542}{83-1} = \frac{16785.542}{83-1} = 204.702.$$

La desviación típica será:

$$s_x = \sqrt{s_x^2} = \sqrt{204.701734} = 14.307.$$

2. Considere las siguientes cuatro poblaciones:

- $1, 2, 3, 4, 5, 6, 7, 8$
- $1, 1, 1, 1, 8, 8, 8, 8$
- $1, 1, 4, 4, 5, 5, 8, 8$
- $-6, -3, 0, 3, 6, 9, 12, 15$

Todas estas poblaciones tienen la misma media. Sin hacer los cálculos, organice las poblaciones de acuerdo con las magnitudes de sus varianzas, de menor a mayor. Luego calcule cada una de las variaciones manualmente.

Solución

La media común de las 4 poblaciones vale 4.5.

Observamos que están ordenadas. La población con menos varianza estará entre la primera y la tercera que es dónde las diferencias con la media es más pequeña. Sin embargo, en la tercera es donde habrá diferencias más acentuadas ya que los valores extremos 1 y 8 se repiten más. En resumen las poblaciones con menos varianza serían: la primera y seguidamente la tercera.

Luego quedan la segunda y la cuarta. Es claro que el rango de valores en la cuarta es mucho mayor que en la segunda. Por tanto, las poblaciones que vienen en orden de menos a más varianza serían la segunda y por último la cuarta.

Calculemos seguidamente las varianzas a mano:

- Primera población:

x_i	$x_i - 4.5$	$(x_i - 4.5)^2$
1	-3.5	12.25
2	-2.5	6.25
3	-1.5	2.25
4	-0.5	0.25
5	0.5	0.25
6	1.5	2.25
7	2.5	6.25
8	3.5	12.25
Sumas		**42**

  La varianza de la primera población será: $s^2_{x,1} = \frac{42}{7} = 6$.

- Segunda población:

x_i	$x_i - 4.5$	$(x_i - 4.5)^2$
1	-3.5	12.25
1	-3.5	12.25
1	-3.5	12.25
1	-3.5	12.25
8	3.5	12.25
8	3.5	12.25
8	3.5	12.25
8	3.5	12.25
Sumas		**98**

  La varianza de la primera población será: $s^2_{x,2} = \frac{98}{7} = 14$.

- Tercera población:

x_i	$x_i - 4.5$	$(x_i - 4.5)^2$
1	-3.5	12.25
1	-3.5	12.25
4	-0.5	0.25
4	-0.5	0.25
5	0.5	0.25
5	0.5	0.25
8	3.5	12.25
8	3.5	12.25
Sumas		**50**

  La varianza de la primera población será: $s^2_{x,3} = \frac{50}{7} = 7.1428571$.

- Cuarta población:

x_i	$x_i - 4.5$	$(x_i - 4.5)^2$
-6	-10.5	110.25
-3	-7.5	56.25
0	-4.5	20.25
3	-1.5	2.25
6	1.5	2.25
9	4.5	20.25
12	7.5	56.25
15	10.5	110.25
Sumas		**378**

  La varianza de la primera población será: $s^2_{x,4} = \frac{378}{7} = 54$.

Por tanto, se cumple, tal como vaticinamos:

$$s^2_{x,1} < s^2_{x,3} < s^2_{x,2} < s^2_{x,4}.$$

3. Los tiempos en minutos que están 50 clientes en un supermercado local para realizar la compra son los siguientes:

26.88	28.60	20.73	34.00	35.87	25.99	20.94	26.45	29.54	26.27
26.51	29.70	29.55	33.52	30.49	31.49	21.28	23.57	22.47	23.15
19.51	23.85	30.98	9.81	26.59	29.68	30.48	25.38	23.49	25.11
19.35	33.80	23.14	13.56	24.63	24.26	37.18	22.20	21.37	28.30
11.02	25.59	24.38	25.29	29.17	25.55	26.94	27.24	19.10	27.44

a. Calcular el tiempo medio que tardan los 50 clientes.
b. Calcular la varianza y la desviación típica de los tiempos.
c. Calcular el percentil 90.
d. Calcular los 5 números resumen.
e. Calcular el coeficiente de variación.

Solución:

a. El tiempo medio será:

$$\overline{x} = \frac{26.88 + \cdots + 27.44}{50} = \frac{1281.39}{50} = 25.6278.$$

b. La varianza de los tiempos será:

$$s_x^2 = \left(\frac{26.88^2 + \cdots + 27.44^2}{50} - 25.6278^2\right) \cdot \frac{50}{49} = \left(\frac{34353.7941}{50} - 25.6278^2\right) \cdot \frac{50}{49} = 30.9099481.$$

c. Para calcular el percentil 90, primero ordenamos los tiempos:

9.81	11.02	13.56	19.10	19.35	19.51	20.73	20.94	21.28	21.37
22.20	22.47	23.14	23.15	23.49	23.57	23.85	24.26	24.38	24.63
25.11	25.29	25.38	25.55	25.59	25.99	26.27	26.45	26.51	26.59
26.88	26.94	27.24	27.44	28.30	28.60	29.17	29.54	29.55	29.68
29.70	30.48	30.49	30.98	31.49	33.52	33.80	34.00	35.87	37.18

El percentil 90 es el que deja a su "izquierda" el 90% de los valores. Por tanto, deja a su izquierda $0.9 \cdot 50 = 45$ valores. El percentil 90 será el valor que ocupa el lugar 45 en los tiempos ordenados. Dicho valor vale 31.49.

d. Los cinco números resumen son el mínimo, el primer cuartil, la mediana, el segundo cuartil y el máximo:

- el mínimo vale: 9.81.
- el primer cuartil es el que deja a su izquierda el 25% de los datos, es decir, es la mediana de los 25 primeros datos que correspondería al valor que ocupa el lugar 13. Dicho valor vale: $Q_1 = 23.14$.
- la mediana es el que deja a su izquierda el 50% de los datos. Correspondería a la semisuma de los dos valores centrales que ocupan los lugares 25 y 26. Dichos valores son 25.59 y 25.99. La mediana será, pues, $Q_2 = \frac{25.59+25.99}{2} = 25.79$.
- el tercer cuartil es el que deja a su izquierda el 75% de los datos, es decir, es la mediana de los 25 últimos datos que correspondería al valor que ocupa el lugar $25+13 = 38$. Dicho valor vale: $Q_3 = 29.54$.
- el máximo vale: 37.18.

4. La tabla siguiente nos da unos indicadores socio-económicos para cada una de las 47 provincias de habla francesa de Suiza en 1888:

	Fertilidad	Agricultura	Examen	Educación	Católicos	Mortalidad infantil
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6
Glane	92.4	67.8	14	8	97.16	24.9
Gruyere	82.4	53.3	12	7	97.67	21.0
Sarine	82.9	45.2	16	13	91.38	24.4
Veveyse	87.1	64.5	14	6	98.61	24.5
Aigle	64.1	62.0	21	12	8.52	16.5
Aubonne	66.9	67.5	14	7	2.27	19.1
Avenches	68.9	60.7	19	12	4.43	22.7
Cossonay	61.7	69.3	22	5	2.82	18.7
Echallens	68.3	72.6	18	2	24.20	21.2
Grandson	71.7	34.0	17	8	3.30	20.0
Lausanne	55.7	19.4	26	28	12.11	20.2
La Vallee	54.3	15.2	31	20	2.15	10.8
Lavaux	65.1	73.0	19	9	2.84	20.0
Morges	65.5	59.8	22	10	5.23	18.0
Moudon	65.0	55.1	14	3	4.52	22.4
Nyone	56.6	50.9	22	12	15.14	16.7
Orbe	57.4	54.1	20	6	4.20	15.3
Oron	72.5	71.2	12	1	2.40	21.0
Payerne	74.2	58.1	14	8	5.23	23.8
Paysd'enhaut	72.0	63.5	6	3	2.56	18.0
Rolle	60.5	60.8	16	10	7.72	16.3
Vevey	58.3	26.8	25	19	18.46	20.9
Yverdon	65.4	49.5	15	8	6.10	22.5
Conthey	75.5	85.9	3	2	99.71	15.1
Entremont	69.3	84.9	7	6	99.68	19.8
Herens	77.3	89.7	5	2	100.00	18.3
Martigwy	70.5	78.2	12	6	98.96	19.4
Monthey	79.4	64.9	7	3	98.22	20.2
St Maurice	65.0	75.9	9	9	99.06	17.8
Sierre	92.2	84.6	3	3	99.46	16.3
Sion	79.3	63.1	13	13	96.83	18.1
Boudry	70.4	38.4	26	12	5.62	20.3
La Chauxdfnd	65.7	7.7	29	11	13.79	20.5
Le Locle	72.7	16.7	22	13	11.22	18.9
Neuchatel	64.4	17.6	35	32	16.92	23.0
Val de Ruz	77.6	37.6	15	7	4.97	20.0
ValdeTravers	67.6	18.7	25	7	8.65	19.5
V. De Geneve	35.0	1.2	37	53	42.34	18.0
Rive Droite	44.7	46.6	16	29	50.43	18.2
Rive Gauche	42.8	27.7	22	29	58.33	19.3

donde:

- Fertilidad: indica el índice de fertilidad de la provincia,
- Agricultura: indica el porcentaje de hombres que se dedican a la agricultura,
- Examen: indica el porcentaje de reclutas que reciben la calificación más alta en el examen del ejército,
- Educación: indica el porcentaje de reclutas que tienen una educación superior a la primaria,
- Católicos: indica el porcentaje de católicos,
- Mortalidad infantil: indica el porcentaje de bebés que viven menos de un año.

a. Dar la tabla de frecuencias de la variable Educación.
b. Calcula la media y la varianza de la variable Educación como datos agrupados.
c. Calcular los cuartiles de la variable Educación. Realizar un diagrama de caja de la variable Fertilidad según el cuartil dónde esté la provincia correspondiente. Es decir, dibujar 4 diagramas de caja para la variable Fertilidad, uno para las provincias que estén en el primer cuartil de la variable Educación, otro para las provincias que estén en el segundo cuartil de la variable Educación y lo mismo para los cuartiles tercero y cuarto. Comentar los resultados observados.

Solución

a. La tabla de frecuencias de la variable Educación es la siguiente:

Educación	Frecuencia f_i
1	1
2	3
3	4
5	2
6	4
7	7
8	4
9	3
10	2
11	1
12	5
13	3
15	1
19	1
20	1
28	1
29	2
32	1
53	1

b. Para calcular la media de la variable Educación, construimos la tabla siguiente:

Valores educación x_i	f_i	$f_i \cdot x_i$
1	1	1
2	3	6
3	4	12
5	2	10
6	4	24
7	7	49
8	4	32
9	3	27
10	2	20
11	1	11
12	5	60
13	3	39
15	1	15
19	1	19
20	1	20
28	1	28
29	2	58
32	1	32
53	1	53
Sumas	**47**	**516**

La media será:

$$\overline{x} = \frac{516}{47} = 10.978723.$$

Para calcular la varianza, hacemos la tabla siguiente:

Valores educación x_i	f_i	$f_i \cdot x_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_i \cdot (x_i - \overline{x})^2$
1	1	1	-9.979	99.575	99.575
2	3	6	-8.979	80.617	241.852
3	4	12	-7.979	63.66	254.64
5	2	10	-5.979	35.745	71.49
6	4	24	-4.979	24.788	99.151
7	7	49	-3.979	15.83	110.812
8	4	32	-2.979	8.873	35.491
9	3	27	-1.979	3.915	11.746
10	2	20	-0.979	0.958	1.916
11	1	11	0.021	0	0
12	5	60	1.021	1.043	5.215
13	3	39	2.021	4.086	12.257
15	1	15	4.021	16.171	16.171
19	1	19	8.021	64.341	64.341
20	1	20	9.021	81.383	81.383
28	1	28	17.021	289.724	289.724
29	2	58	18.021	324.766	649.533
32	1	32	21.021	441.894	441.894
53	1	53	42.021	1765.788	1765.788
Sumas	**47**	**516**			**4252.979**

La varianza y la desviación típica serán, pues,

$$s_x^2 = \frac{4252.979}{47-1} = 92.456059, \quad s_x = \sqrt{92.456059} = 9.615407.$$

c. Para hallar los cuartiles de la variable Educación, primero los ordenamos:

1	2	2	2	3	3	3	3	5	5	6	6	6	6	7	7	7	7	7	7
7	8	8	8	8	9	9	9	10	10	11	12	12	12	12	12	13	13	13	15
19	20	28	29	29	32	53													

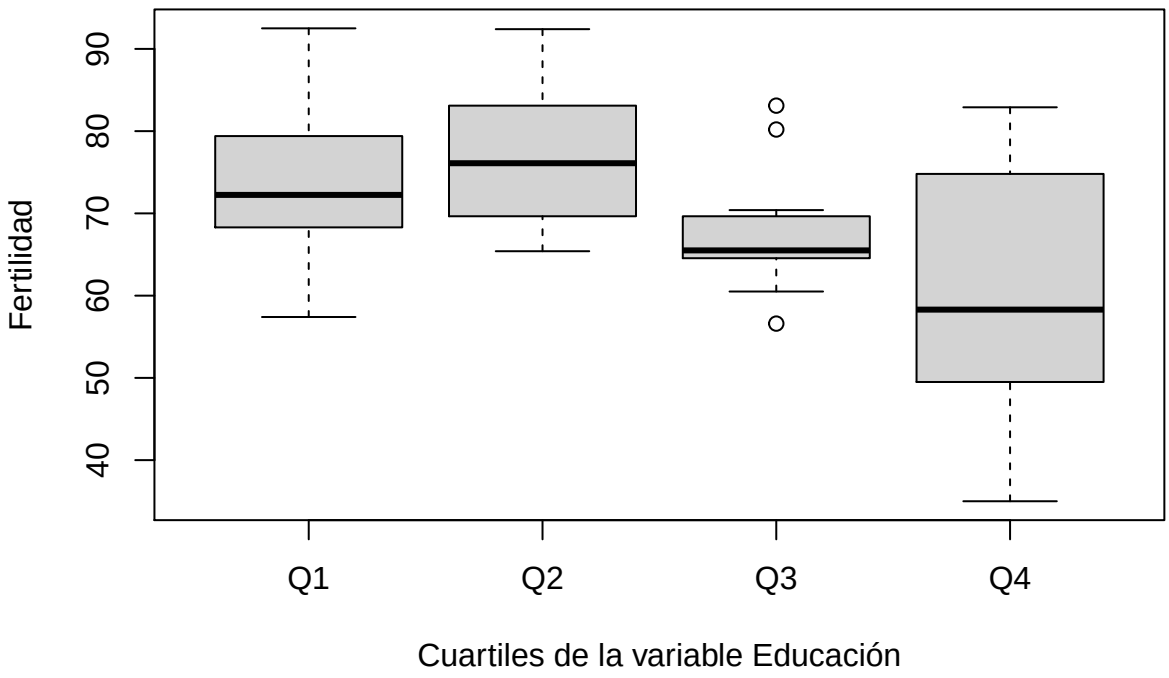
Los cuartiles de la variable Educación serían los siguientes:

- el primer cuartil es el que deja a su izquierda el 25% de los datos, es decir, es la mediana de los 23 primeros datos que correspondería al valor que ocupa el lugar 12. Dicho valor vale: $Q_1 = 6$. (en rojo)
- la mediana es el que deja a su izquierda el 50% de los datos. Correspondería al valor que ocupa el lugar 24. Dichos valor es 8. (en verde)
- el tercer cuartil es el que deja a su izquierda el 75% de los datos, es decir, es la mediana de los 23 últimos datos que correspondería al valor que ocupa el lugar $24 + 12 = 36$. Dicho valor vale: $Q_3 = 12$. (en magenta)

A continuación mostramos la tabla con las variables Educación, Fertilidad y una variable nueva que nos indica en qué cuartil se encuentra la provincia dependiendo de su valor respecto de la variable Educación.

	Educación	Fertilidad	Cuartil respecto educación
Courtelary	12	80.2	Q3
Delemont	9	83.1	Q3
Franches-Mnt	5	92.5	Q1
Moutier	7	85.8	Q2
Neuveville	15	76.9	Q4
Porrentruy	7	76.1	Q2
Broye	7	83.8	Q2
Glane	8	92.4	Q2
Gruyere	7	82.4	Q2
Sarine	13	82.9	Q4
Veveyse	6	87.1	Q1
Aigle	12	64.1	Q3
Aubonne	7	66.9	Q2
Avenches	12	68.9	Q3
Cossonay	5	61.7	Q1
Echallens	2	68.3	Q1
Grandson	8	71.7	Q2
Lausanne	28	55.7	Q4
La Vallee	20	54.3	Q4
Lavaux	9	65.1	Q3
Morges	10	65.5	Q3
Moudon	3	65.0	Q1
Nyone	12	56.6	Q3
Orbe	6	57.4	Q1
Oron	1	72.5	Q1
Payerne	8	74.2	Q2
Paysd'enhaut	3	72.0	Q1
Rolle	10	60.5	Q3
Vevey	19	58.3	Q4
Yverdon	8	65.4	Q2
Conthey	2	75.5	Q1
Entremont	6	69.3	Q1
Herens	2	77.3	Q1
Martigwy	6	70.5	Q1
Monthey	3	79.4	Q1
St Maurice	9	65.0	Q3
Sierre	3	92.2	Q1
Sion	13	79.3	Q4
Boudry	12	70.4	Q3
La Chauxdfnd	11	65.7	Q3
Le Locle	13	72.7	Q4
Neuchatel	32	64.4	Q4
Val de Ruz	7	77.6	Q2
ValdeTravers	7	67.6	Q2
V. De Geneve	53	35.0	Q4
Rive Droite	29	44.7	Q4
Rive Gauche	29	42.8	Q4

Los diagramas de caja pedidos de la variable Fertilidad dependiendo del cuartil en dónde están las provincias respecto de la variable Educación son los siguientes:



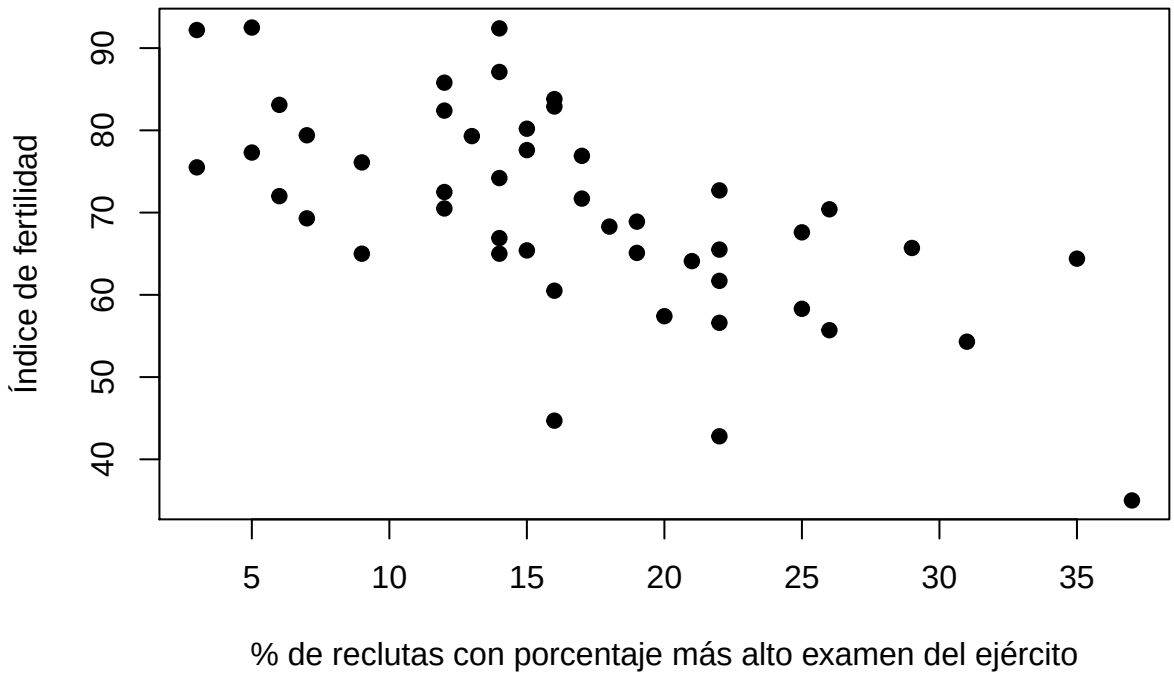
Observamos un índice de fertilidad mayor en las provincias con un nivel de educación de los reclutas menor pero también hay más dispersión de índice de fertilidad en aquellas provincias con un porcentaje mayor de reclutas con educación superior a la primaria.

5. Usando los datos de la tabla anterior, queremos estudiar la posible relación entre las variables Fertilidad y Examen.

- Realiza un gráfico de puntos de las variables anteriores, indicando en el eje X o abscisas la variable Examen y en el Y, la variable $Fertilidad$.
- Calcula la covarianza y la correlación entre las variables anteriores. ¿A qué conclusión llegas?

Solución

- El gráfico pedido es el siguiente:



- Para calcular la covarianza, primero hay que calcular las medias.

La media de la variable Examen vale:

$$\overline{X} = \frac{775}{47} = 16.489.$$

La media de la variable Fertilidad vale:

$$\overline{Y} = \frac{3296.7}{47} = 70.143.$$

Para calcular la Para ello, realizamos la tabla siguiente:

X	Y	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	$Y_i - \overline{Y}$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
15	80.2	-1.489	2.218	10.057	101.152	-14.979
6	83.1	-10.489	110.027	12.957	167.895	-135.915
5	92.5	-11.489	132.005	22.357	499.855	-256.873
12	85.8	-4.489	20.154	15.657	245.156	-70.292
17	76.9	0.511	0.261	6.757	45.663	3.451
9	76.1	-7.489	56.091	5.957	35.491	-44.617
16	83.8	-0.489	0.239	13.657	186.526	-6.683
14	92.4	-2.489	6.197	22.257	495.394	-55.407
12	82.4	-4.489	20.154	12.257	150.245	-55.028
16	82.9	-0.489	0.239	12.757	162.752	-6.243
14	87.1	-2.489	6.197	16.957	287.555	-42.213
21	64.1	4.511	20.346	-6.043	36.512	-27.256
14	66.9	-2.489	6.197	-3.243	10.514	8.072
19	68.9	2.511	6.303	-1.243	1.544	-3.120
22	61.7	5.511	30.367	-8.443	71.277	-46.524
18	68.3	1.511	2.282	-1.843	3.395	-2.783
17	71.7	0.511	0.261	1.557	2.426	0.795
26	55.7	9.511	90.452	-14.443	208.587	-137.358
31	54.3	14.511	210.559	-15.843	250.986	-229.886
19	65.1	2.511	6.303	-5.043	25.427	-12.660
22	65.5	5.511	30.367	-4.643	21.553	-25.583
14	65.0	-2.489	6.197	-5.143	26.446	12.802
22	56.6	5.511	30.367	-13.543	183.401	-74.628
20	57.4	3.511	12.325	-12.743	162.373	-44.734
12	72.5	-4.489	20.154	2.357	5.558	-10.583
14	74.2	-2.489	6.197	4.057	16.463	-10.100
6	72.0	-10.489	110.027	1.857	3.450	-19.483
16	60.5	-0.489	0.239	-9.643	92.979	4.719
25	58.3	8.511	72.431	-11.843	140.246	-100.788
15	65.4	-1.489	2.218	-4.743	22.492	7.063
3	75.5	-13.489	181.963	5.357	28.702	-72.269
7	69.3	-9.489	90.048	-0.843	0.710	7.995
5	77.3	-11.489	132.005	7.157	51.229	-82.234
12	70.5	-4.489	20.154	0.357	0.128	-1.605
7	79.4	-9.489	90.048	9.257	85.700	-87.847
9	65.0	-7.489	56.091	-5.143	26.446	38.514
3	92.2	-13.489	181.963	22.057	486.531	-297.541
13	79.3	-3.489	12.176	9.157	83.859	-31.954
26	70.4	9.511	90.452	0.257	0.066	2.448
29	65.7	12.511	156.516	-4.443	19.736	-55.579
22	72.7	5.511	30.367	2.557	6.541	14.093
35	64.4	18.511	342.644	-5.743	32.977	-106.298
15	77.6	-1.489	2.218	7.457	55.614	-11.107
25	67.6	8.511	72.431	-2.543	6.465	-21.639
37	35.0	20.511	420.686	-35.143	1234.999	-720.796
16	44.7	-0.489	0.239	-25.443	647.324	12.451
22	42.8	5.511	30.367	-27.343	747.615	-150.675
775	**3296.7**	**0.000**	**2927.745**	**0.000**	**7177.955**	**-2960.879**

La covarianza será, pues:

$$\text{cov}(X, Y) = \frac{-2960.879}{47 - 1} = -64.367.$$

Para calcular el coeficiente de correlación r entre X e Y, necesitamos calcular primeramente las varianzas correspondientes:

$$s_X^2 = \frac{2927.7447}{46} = 63.6466, \quad s_Y^2 = \frac{7177.9549}{46} = 156.0425.$$

El coeficiente de correlación será:

$$r = \frac{\text{cov}(X,Y)}{s_X \cdot s_Y} = \frac{-64.3669}{\sqrt{63.6466} \cdot \sqrt{156.0425}} = -0.6459.$$

Vemos que hay una correlación negativa entre las variables. Esto significa que a medida que el porcentaje de reclutas que reciben la calificación más alta en el examen del ejército aumenta, se espera un índice de fertilidad menor aunque hay que decir también que dicha correlación indica que la relación lineal entre las variables anteriores no es muy acentuada.