

## Solución al boletín de problemas 2

### Parte A.

#### Capítulo 3.

1. 66 y 71,999

2. Véase adjunto.

(a) edad 1 :  $8/5 = 1.6\%$

edad 11 :  $13/9 = 1,444\%$

Por lo tanto, hay más niños con la edad de 1.

(b) edad 21 :  $10/7 = 1,4\%$

edad 31 :  $9/5 = 1,8\%$

Por lo tanto, hay más con la edad de 31.

(c) más edad de 35-44

(d) alrededor de 50%

4. (a)  $1,8(5) + 1(5) + 0,8(10) + 0,3(10) = 9+5+8+3 = 25$ . Respuesta: 25%

(b) 99%

(c) 135-140mm :  $1\% \times 5 = 5\%$

140-150mm :  $0,8\% \times 10 = 8\%$

más mujeres en el intervalo 140-150mm.

(d) intervalo 135-140mm.

(e)  $5 \times 2,1\% = 10,5\%$

(f) 102-103mm.

(g) 115-120mm.

8.(a) Verdadero.

$7,3\%/5 = 1,46\%$  (10-15),  $15,6\%/10 = 1,56\%$  (15-25),  $15\%/10 = 1,5\%$  (25-35)

las familias que ganan entre 10.000\$ y 35.000\$ están distribuidas de manera equitativa.

(b) Falso.

$$19,2\%/15 = 1,38\% (35-50), 19,6\%/25 = 0,78\% (50-75)$$

el porcentaje de familias que ganan entre 35.000\$ y 50.000\$ es mayor que el de familias que ganan entre 50.000\$ y 75.000\$.

(c) Falso. En el historiograma, la altura de un bloque representa el porcentaje por unidad horizontal. Aquí, el porcentaje está representado como área. Además, el x-axis no está dividido por la unidad determinada.

9.

(a) Verdadero.

(b) Verdadero.

(c) Se puede analizar de dos maneras. Quizás hay muchos estudiantes que han estudiado lo suficiente para aprobar la clase. Por lo tanto, muchos estudiantes apenas pasaron y recibieron un nota media (GPA) de 2. De manera alternativa, quizás el profesor solía ser más "blando" (a diferencia del asistente (TA) con los estudiantes más débiles animándolos a conseguir una "C."

10.

(a) Véase adjunto.

(b) Podría tratarse de datos erróneos que provienen de la persona entrevistada o del entrevistador. Hay que tener en cuenta el bajo nivel de formación existente y la carencia en el pasado tanto de un sistema preciso de registro de partidas de nacimiento como de herramientas fiables para la elaboración del censo. Es posible que el hecho de que la gente no supiera con exactitud su fecha de su nacimiento se debiera a que no siempre se extendían partidas de nacimiento y a que no todo el mundo sabía leer. Las personas entrevistadas pueden haber respondido a la pregunta sobre su edad en función de si se hizo la pregunta a mitad de década o al final, lo que es una respuesta más sencilla (por ejemplo: "unos 40 años"). Debido a que el censo se elabora en años que terminen en "0", puede que haya una tendencia a dar como respuesta años de nacimiento que también terminen en cero.

(c) Con el paso del tiempo, el desarrollo del sistema de seguridad social y a otros factores, la elaboración del censo y la expedición de partidas de nacimiento han adquirido un mayor grado de perfeccionamiento. También los niveles de alfabetización son más altos.

(d) Incluso en ambas épocas (*Even in both times*).

#### Capítulo 4.

1.

(a)  $(41+48+50+50+54+57)/6 = 50.$

$$[\{(50-41)^2 + (50-48)^2 + (50-50)^2 + (50-54)^2 + (57-50)^2 + (50-50)^2\}/6]^{1/2} = 5$$

media = 50 & SD = 5

$$(b) 50 + 0,5(5) = 52,5$$

$$50 - 0,5(5) = 47,5$$

48, 50, 50 dentro de 0,5 SDs de media.

$$50 + 1,5(5) = 57,5$$

$$50 - 1,4(4) = 42,5$$

48, 50, 50, 54, 57 dentro de 1,5 SDs de media.

2.

(a) (ii) tiene un SD menor. Ya que no hay diferencia debido a los tres 50 adicionales, y está dividido por 10 en vez de por 7, genera un SD menor que (i).

(b) (i) tiene un SD menor. Dos elementos adicionales (1 y 99) alargarán las distancias de la media 50 que excederá el denominador incrementado.

3.

(a) 5

(b) Considerando que la media es 5, su SD debería ser de 3, ya que la media más o menos 2SD debería abarcar el 95% de los datos.

5. Asumiendo que tiene una distribución normal, el límite inferior es 96 ( $124 - 2 \times 14$ ) y el superior es 152 ( $124 + 2 \times 14$ ). Es decir, 80mm y 210mm son cantidades bastante bajas y altas respectivamente si las comparamos con la media.

6. (a) (i) media 60

(ii) media 50

(iii) media 40

(b) (i) media < mediana

(ii) media = mediana

(iii) media > mediana

(c) 15

(d) Falso. (i) parece ser más disperso, al tener una varianza mayor que (iii) en la gráfica.

7.

(a) Media masculina = 66,  $SD_m = 9$

Media femenina = 55,  $SD = 9_f$

	Media	SD
Hombres	145,2	19,8
Mujeres	121	19,8

(b)  $66 - 9 = 57.66 + 9 = 75$ . Por tanto, 1 SD de la media que abarca al 68% de los hombres.

(c) Más de 9 Kg.

Pregunta qué ocurre si se combinan las dos variables. Las dos variables tienen el mismo SD, pero diferentes medias. Imagine, al contrario de lo que indican los datos, que las dos variables tienen la misma media. A continuación, divida la muestra moviendo la mitad a la izquierda y la mitad a la derecha en cantidades iguales conservando el mismo SD de la muestra dividida. El SD de la muestra completa se incrementará. (O, al menos, es el desarrollo intuitivo que yo he empleado).

10.

(a) La mejor estimación es 163.

(b) 8 dólares.

### Capítulo 8.

1.

	Media IQ	SD
Esposos	100	15
Esposas	100	15

$r = 0,6$

rangos de x e y : 70-130 ( $15 \times 2 = 30$ )

(a) Las medias están fuera del rango.

(b) El rango es demasiado pequeño para x e y.

(c) El rango es demasiado amplio para x e y.

(d) Diagrama de dispersión correcto.

2.

(a) Negativo. Conforme un automóvil se vuelve más antiguo aumenta el consumo de gasolina y el ahorro de gasolina disminuye. Además, los vehículos nuevos tienen que ajustarse a unos estándares mínimos de consumo de combustible. Ambos factores se añaden para producir una correlación negativa entre la edad del automóvil y el ahorro de combustible.

(c) Las personas con mayores ingresos pueden permitirse comprar automóviles nuevos que gastan menos combustible que los viejos o usados.

3. El coeficiente de correlación es 1 porque hay una relación lineal perfecta.

6. Falso. No hay relación directa entre dos coeficientes de correlación distintos, ya que los coeficientes de correlación son cifras estandarizadas.

7. Tal y como muestra el orden...

8. 0,62 -1

-0,85 0,97

0,06 -0,38

11. **respuesta : -1**

media correcta: 6,4  $SD_r = 2$

media incorrecta: 3,6  $SD_w = 2$

correcta = 10 - incorrecta

$Corr(r,w) = Cov(r,w) / (SD_r \times SD_w)$

$Cov(r,w) = \sum (r_i - 6,4)(w_i - 3,6) / n = \sum (r_i w_i) / n - (6,4)(3,6)$

$= \sum r_i(10 - r_i) / n - (6,4)(3,6) = \sum (10 r_i - r_i^2) / n - (6,4)(3,6) = 10 \sum r_i / n - \sum r_i^2 / n - 23,04$

$= 10(6,4) - \sum r_i^2 / n - 23,04$

$\sum r_i^2 / n = Var(r) + media(r)^2 = 4 + (6,4)^2$   $Var(r) = \sum r_i^2 / n - media(r)^2$

$\sum Cov(r,w) = 10(6,4) - (4 + 6,4^2) - 23,04$

$= -4$

$\sum Corr(r,w) = -4 / (2 \times 2) = -1$ .

Intuitivamente, podemos imaginar que las respuestas incorrectas y las correcta tendrán relación lineal negativa exacta.

**Parte B.**

1. y 2. deberían tener la misma respuesta que en la Parte C
3. Hay un GRAN valor atípico (1), que no es tan obvio en la versión *logged*. La versión *logged* se puede mejorar.

**Parte C.**

1. -0,566
  2. -0,503
  3. Gráficas adjuntas, véase también el archivo *log* adjunto debajo.
- Ahora está más claro que la correlación *log* es mejor, tal y como se describe en B3.

## Problema C Log

```
. corr rating enrollment
(obs=26)
| rating enroll~t
-----+-----
rating | 1.0000
enrollment | -0.5655 1.0000
. // -0.566
. gen logenrol = log(enrollment)
. corr rating logenrol
(obs=26)
| rating logenrol
-----+-----
rating | 1.0000
logenrol | -0.5032 1.0000
. \\ -0.503
. graph rating enrollment
. graph rating logenrol
. graph rating enrollment, xlog
. summ rating enrollment logenrol
Variable | Obs Mean Std. Dev. Min Max
-----+-----
rating | 26 5.626923 .7102437 3.8 6.6
enrollment | 26 48.96154 95.02609 7 501
logenrol | 26 3.334779 .8608651 1.94591 6.216606
. log close
```

**Parte D.**

1. Véanse los archivos *do* y *log* de más abajo. La correlación correcta es 0,475. La relación no está bien descrita ya que la asociación no es claramente lineal. Esto resulta evidente al representar los en la gráfica.

3. Dependiendo de si lo ha hecho como yo, tal y como se indica debajo, o usando el coeficiente de ponderación habría obtenido los resultados correctos. De cualquier manera, queda claro que el sistema de escaneo central produce una GRAN diferencia en los votos no contados.

Ponderado: 0,005. 0,057

No ponderado: 0,006. 0,061

Archivo Do File Pset2-D

```

use "E:\My Documents\17871\fla_precinct_subset.dta", clear
gen resid= undervote + overvote
gen resrate= resid/ total_ball
corr blackrv resrate
sort county
save fla_precinct_subset, replace
use "E:\My Documents\17871\fla_county_subset.dta", clear
sort county
save fla_county_subset, replace
merge county using fla_precinct_subset
save fla_merged, replace
table centraltab, c(mean resrate)
Problem D Log
. do Pset2-D
. use "E:\My Documents\17871\fla_precinct_subset.dta", clear
. gen resid= undervote + overvote
. gen resrate= resid/ total_ball
(70 missing values generated)
. corr blackrv resrate
(obs=5816)
| blackrv resrate
-----+-----
blackrv | 1.0000
resrate | 0.4748 1.0000
. sort county
. save fla_precinct_subset, replace
file fla_precinct_subset.dta saved
. use "E:\My Documents\17871\fla_county_subset.dta", clear
. sort county
. save fla_county_subset, replace
file fla_county_subset.dta saved
. merge county using fla_precinct_subset
. save fla_merged, replace
file fla_merged.dta saved
. table centraltab, c(mean resrate)
-----
CENTRAL |
OR |
PRECINCT |

```

```
TAB | mean(resrate)
-----+-----
1 | .0063025
2 | .0613287
9 | .0420813
-----
.
end of do-file
```