



Dpto. Matemática Aplicada y Estadística

Grado en IIAA y Grado en IHJ  
Asignatura: Estadística Aplicada. Curso 2011-2012  
Examen de prácticas de JUNIO 2012

**NOMBRE:**.....**APELLIDOS:**.....  
**ESPECIALIDAD:**.....

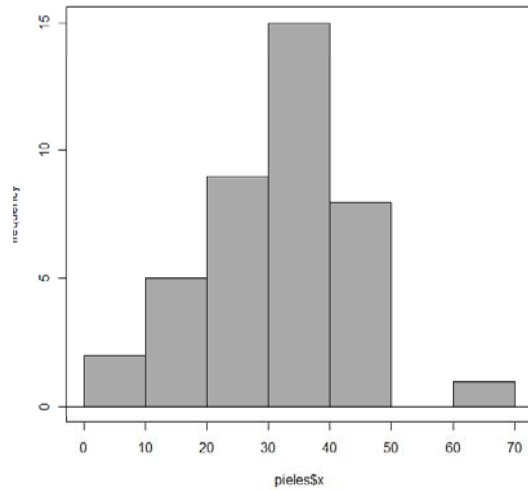
Uno de los problemas más desafiantes para el control de contaminación del agua lo presenta la industria del curtido de pieles. Los desechos de esta industria son químicamente complejos. Se caracterizan por valores elevados en la demanda de oxígeno bioquímico, los sólidos volátiles y otras mediciones de contaminación. Los datos del fichero **curtido\_pieles.txt**, son los datos experimentales que se obtuvieron de 40 muestras de desperdicios que se tratan químicamente. Para cada una de las 40 muestras se registraron las lecturas de la reducción del total de sólidos ( $X$ , en porcentajes) y de la reducción de demanda de oxígeno químico ( $Y$ , en porcentaje). Después de importar los datos, que se encuentran en la ruta habitual, se pide:

1. Realizar un histograma de ambos conjuntos de datos. Comentar las características más relevantes de ambos gráficos.
2. Realizar (en el mismo gráfico) un diagrama de caja y bigotes para cada una de las características e identificar cada una de las líneas que lo constituyen, así como los valores numéricos correspondientes.
3. ¿Existen datos atípicos? ¿Cuáles serían los valores admisibles entre los que se encontrarían los datos no atípicos para cada uno de los dos conjuntos de datos?
4. A partir de los resultados obtenidos en los apartados anteriores, ¿qué medidas de centralización y dispersión consideras más adecuadas para resumir cada uno de los conjuntos de datos? Dar el valor numérico de estos descriptivos estadísticos.
5. Supongamos que la variable  $Y$  = “Reducción de demanda de oxígeno químico” sigue un modelo normal de media  $\mu = 32.1$  y de desviación típica  $\sigma = 10.95$ , determinar la siguiente probabilidad:  $P(10.5 \leq Y \leq 40.5)$ .
6. Proporcionar un intervalo de confianza al 96% para la media de la variable “Reducción de demanda de oxígeno químico”. Indicar la distribución de probabilidad que ha utilizado para construir dicho intervalo.
7. ¿Podemos asumir que la media de la variable “Reducción de demanda de oxígeno químico” es inferior al 35.5%? Indicar el procedimiento utilizado y dar la respuesta a partir del  $p$ -valor obtenido.
8. Se quiere determinar un modelo para explicar la reducción de demanda de oxígeno químico a partir de la reducción del total de sólidos. ¿Qué modelo parece adecuado?
9. Realizar un ajuste por mínimos cuadrados con el fin de explicar la reducción de demanda de oxígeno químico a partir de la reducción del total de sólidos. Indicar la ecuación del modelo propuesto y dar una medida de la bondad del ajuste realizado.
10. Se detecta que una muestra presenta una reducción del total de sólidos igual al 33%, ¿podrías dar una estimación para la reducción de demanda de oxígeno químico? ¿Es fiable esta estimación? Razonar la respuesta.

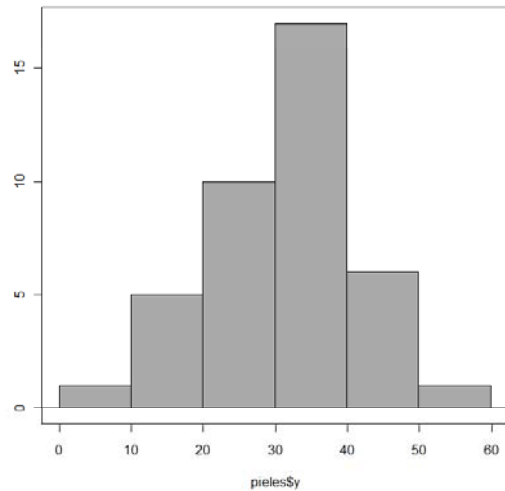
# Grado en IIAA y Grado en IHJ:

## SOLUCIÓN DEL EXAMEN DE PRÁCTICAS DE JUNIO 2012

1.-



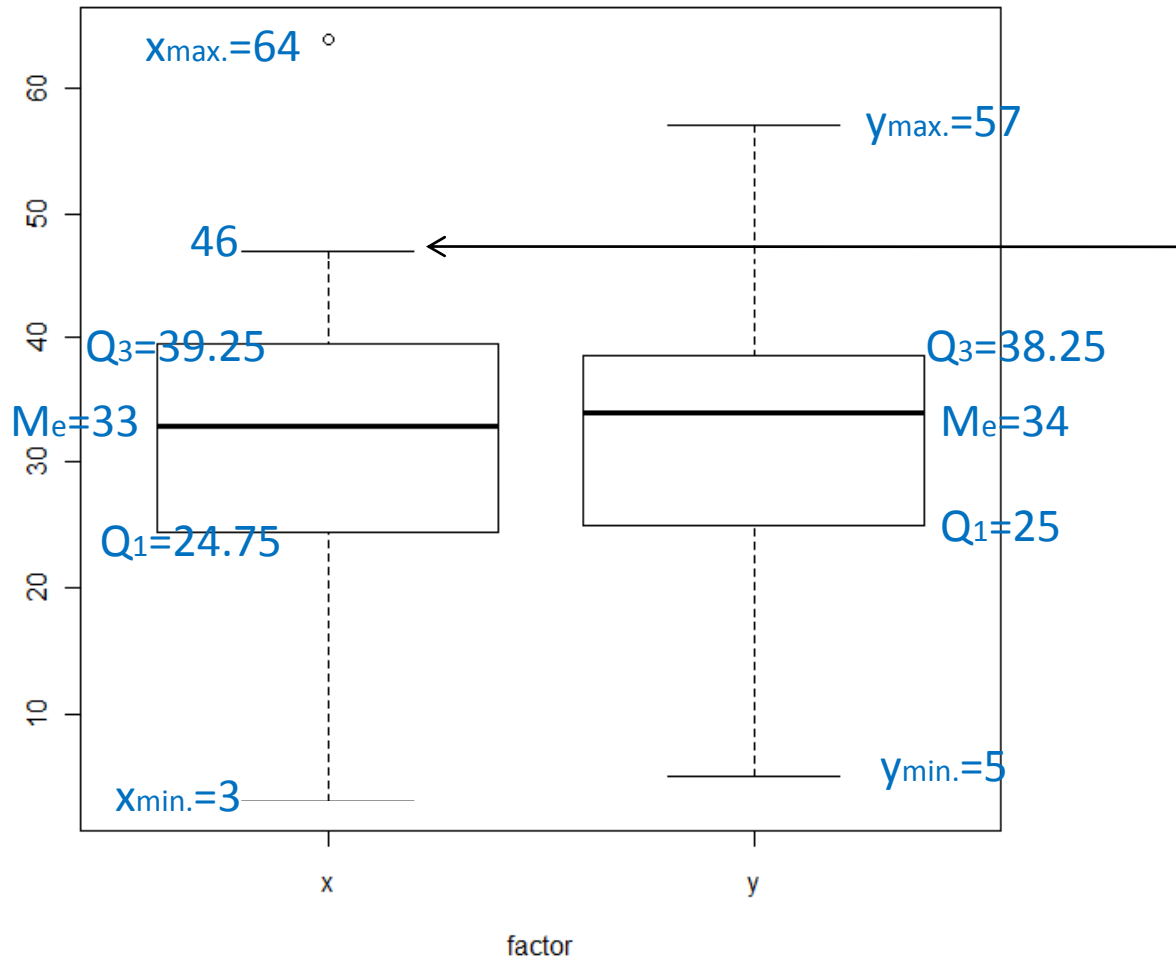
Variable X: Unimodal , existencia de datos atípicos y cierta asimetría a la izda.



Variable Y: Unimodal , aparentemente simétrica y parece no existir datos atípicos.

2.-

Previamente tenemos que apilar las variables:



Notar que este dato es la observación más grande de la variable X que no es clasificado como atípico

Resúmenes numéricos de las características:

	mean	sd	0%	25%	50%	75%	100%	n
x	31.85	12.01612	3	24.75	33	39.25	64	40
y	32.10	10.94696	5	25.00	34	38.25	57	40

|

3.- Existe un dato atípico para la X.

El rango de valores para la X entre los que se encuentran las observaciones NO atípicas es:

$$L_{\text{inf.}} = Q_1 - 1.5 * RIQ = 24.75 - 1.5 * (39.25 - 24.75) = 3$$

$$L_{\text{sup.}} = Q_3 + 1.5 * RIQ = 39.25 + 1.5 * (39.25 - 24.75) = 61$$

➡ [3,61] ➡ Por eso, el dato  $x=64$  es dato atípico.

Y para la característica Y sería:

$$L_{\text{inf.}} = Q_1 - 1.5 * RIQ = 25 - 1.5 * (38.25 - 25) = 5.125$$

$$L_{\text{sup.}} = Q_3 + 1.5 * RIQ = 38.25 + 1.5 * (38.25 - 25) = 58.125$$

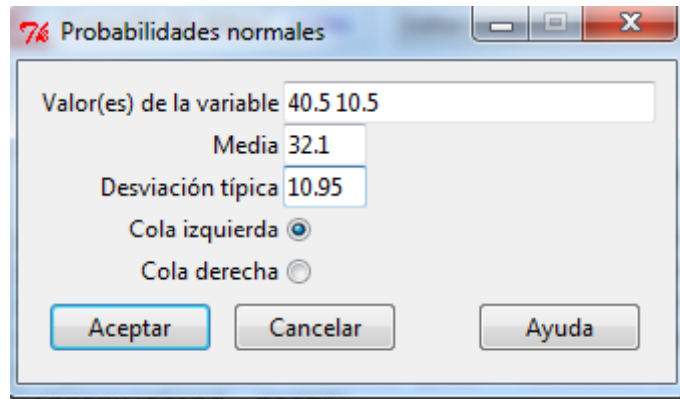
➡ [5.125,58.125]

4.-

Para la característica X tomaremos la  $Me=33$  como medida de centro y el  $RIQ=14.5$  Como medida de dispersión ya que hemos visto que presenta un dato atípico.

Para la característica Y tomaremos la  $media =33$  como medida de centro y la desviación típica  $s=32.10$  como medida de dispersión ya que hemos visto que la distribución es aparentemente simétrica sin datos atípicos.

## 5.- Instrucciones:



## Resultados:

```
> pnorm(c(40.5,10.5), mean=32.1, sd=10.95, lower.tail=TRUE)
[1] 0.77849589 0.02427042
```

De donde, la probabilidad que nos piden quedaría:

$$P(10.5 \leq Y \leq 40.5) = 0.7785 - 0.0243 = 0.7542$$

6.- El intervalo de confianza al 96% para la media de Y quedaría como:

```
96 percent confidence interval:
| 28.42236 35.77764
```

 $\rightarrow$  (28.4224,35.7776)

Trabajamos con la distribución  $T = \frac{\bar{Y} - \mu_Y}{S_Y / \sqrt{n_Y}} \approx t_{n_Y-1}$

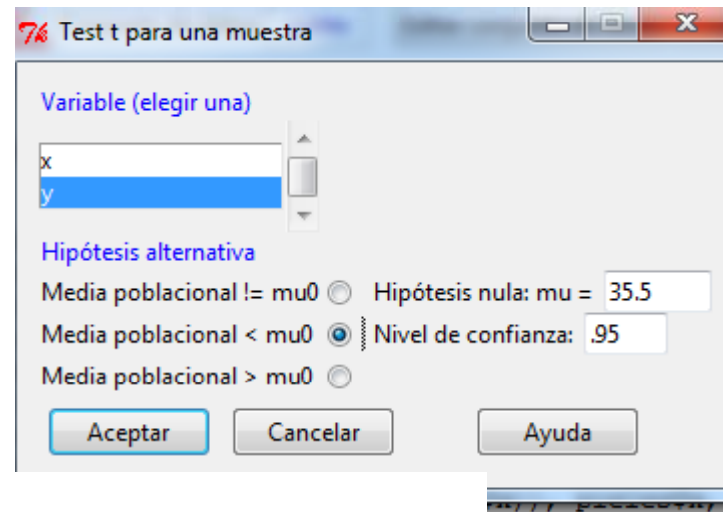
que para nuestra muestra sería la  $t_{39}$

7.- El contraste que queremos realizar es:

$$\begin{cases} H_0 : \mu_Y = 35.5 \\ H_1 : \mu_Y < 35.5 \end{cases}$$



Para realizar este contraste procedemos como sigue:



Y obtenemos:

One Sample t-test

```
data: pieles$y
t = -1.9643, df = 39, p-value = 0.02832
alternative hypothesis: true mean is less than 35.5
95 percent confidence interval:
 -Inf 35.01629
sample estimates:
mean of x
 32.1
```

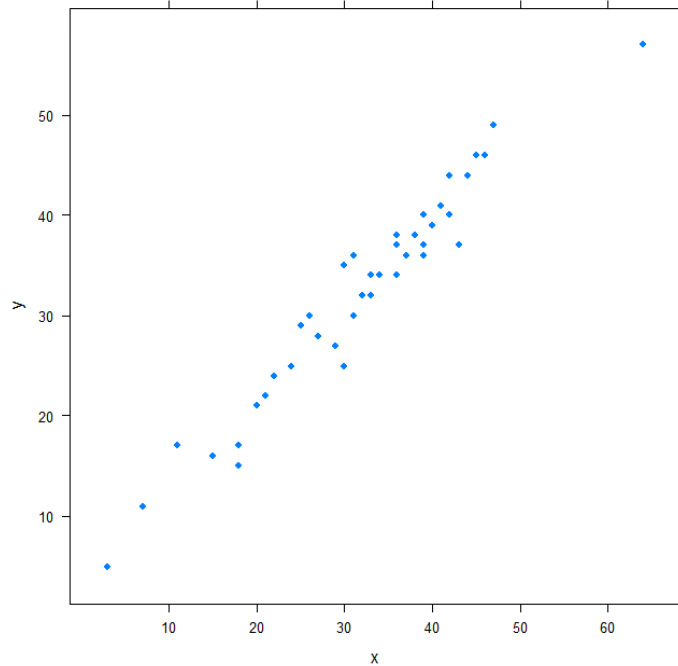


$t_0 = -1.9643$  y el p-valor = 0.02835, de donde, hay mucha confianza de que rechazar  $H_0$  es la decisión correcta.



La media poblacional de la variable Y es significativamente menor al 35.5%

8.- Empecemos realizando la nube de puntos :



Observamos que existe una dependencia lineal de y en función de x de tendencia positiva.

9.- Al ajustar la recta de mínimos cuadrados de y sobre x obtenemos que:

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.83313   1.13407    3.38  0.00169 **
x            0.88750   0.03337   26.60 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.504 on 38 degrees of freedom
Multiple R-squared: 0.949, Adjusted R-squared: 0.9477
F-statistic: 707.5 on 1 and 38 DF,  p-value: < 2.2e-16
```




Recta ajustada:  
 $Y=3.83313+0.8875 \cdot X$

$R^2=0.949$ , lo que implica un ajuste muy bueno.

10.- Para realizar la estimación de  $y$  cuando  $x=33\%$  sólo hay que sustituir en el modelo ajustado:

$$Y=3.83313+0.8875*33=33.1206\%$$

Para estudiar su fiabilidad tenemos que comprobar si  $x=33\%$  pertenece al rango observado de las  $X`s=(3,64)$ , lo que sí ocurre.  
Además el ajuste era muy bueno pues el  $R^2$  estaba muy próximo a 1

 La estimación es fiable.