Dpto. Matemática Aplicada y Estadística

Grado en IIAA y Grado en IHJ
Asignatura: **Estadística Aplicada. Curso 2012-2013**
Examen de prácticas de FEBRERO 2013

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

Los niveles de contaminantes en productos alimenticios deben mantenerse bajo estrictos límites para asegurar la salud de la población. Por ello, los controles de calidad son exhaustivos en distintos tipos de alimentos. En el archivo **cacahuetes.txt** se presentan los datos correspondientes a 34 lotes de 50 Kg de cacahuetes, donde se presenta el nivel del contaminante aflatoxin (partes por billón) (X) y el porcentaje de cacahuetes no contaminados en cada lote (Y). Tras importar los datos, responder a las siguientes cuestiones.

1. Realizar un histograma de ambos conjuntos de datos. Comentar las características más relevantes de ambos gráficos. `(Copiar los gráficos obtenidos)`. (1.5 puntos)

2. Realizar, en gráficos separados, un diagrama de caja y bigotes para cada variable e identificar cada una de las líneas que lo constituyen, así como los valores numéricos correspondientes. `(Copiar los gráficos obtenidos)`. ¿Existen datos atípicos? ¿Cuáles serían los valores admisibles entre los que se encontrarían los datos no atípicos para cada una de las dos variables? (2 puntos)

3. A partir de los resultados obtenidos en los apartados anteriores, ¿qué medidas de centralización y dispersión consideras más adecuadas para resumir cada conjunto de datos? (1.5 puntos)

4. ¿Podemos asumir que la dispersión es similar en ambas variables? Comentar los resultados que consideres más relevantes así como los estadísticos que ha utilizado para dar respuesta a la cuestión anterior. (1.5 puntos)

5. Basándote en las características más relevantes del histograma correspondiente a la variable X, ¿qué modelo de distribución de probabilidad continua te parece más adecuado para describir el comportamiento de la v.a. X="Nivel del contaminante aflatoxin"? Razonar la respuesta. ¿Cuánto valdría el parámetro de la distribución? Independientemente del resultado que hayas obtenido, si asumimos que la variable X="Nivel del contaminante aflatoxin" sigue una distribución exponencial de parámetro 0.03, ¿cuál es la probabilidad de que un lote contenga entre 10 y 20 partes por billón de contaminante? ¿Y de que tenga más de 50 partes por billón de contaminante? (2 puntos)

6. Proporcionar un intervalo de confianza al 98% para la media de la variable "Nivel del contaminante aflatoxin". Indicar la distribución de probabilidad que se ha utilizado para construir dicho intervalo. (1 punto)

7. La legislación vigente dicta que el valor promedio del contaminante aflatin debe ser inferior a 46 partes por billón. ¿Podemos asumir que se cumple dicha ley al 95% de confianza? Plantear el contraste correspondiente y dar la respuesta a partir del $p-valor$ obtenido. (1 punto)

8. Representar los datos observados por ambas características mediante el diagrama de dispersión correspondiente (`copiar el gráfico obtenido`). ¿Parece adecuado un modelo lineal para explicar la variable "Porcentaje de cacahuetes no contaminados" a partir de los valores de la variable "Nivel del contaminante aflatoxin"? Razonar la respuesta. (1 punto)

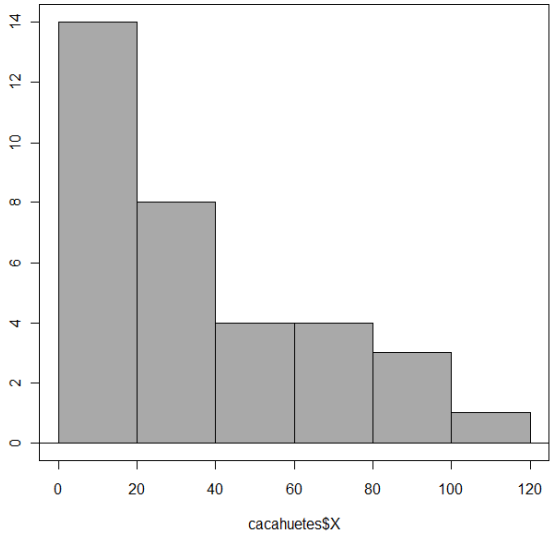
9. Realizar un ajuste por mínimos cuadrados con el fin de predecir el comportamiento de la variable "Porcentaje de cacahuetes no contaminados" a partir de los valores de la variable "Nivel del contaminante aflatoxin". Indicar la ecuación del modelo propuesto y dar una medida de la bondad del ajuste. Comentar la bondad del ajuste realizado. (2 puntos)

10. Obtener la predicción para el porcentaje de cacahuetes sin contaminar, si se observa un nivel de contaminante de 60 partes por billón. ¿Es fiable esta predicción? Razonar la respuesta. (1.5 puntos)

Solución del examen de prácticas de FEBRERO 2013: Grado en IIAA e IHJ

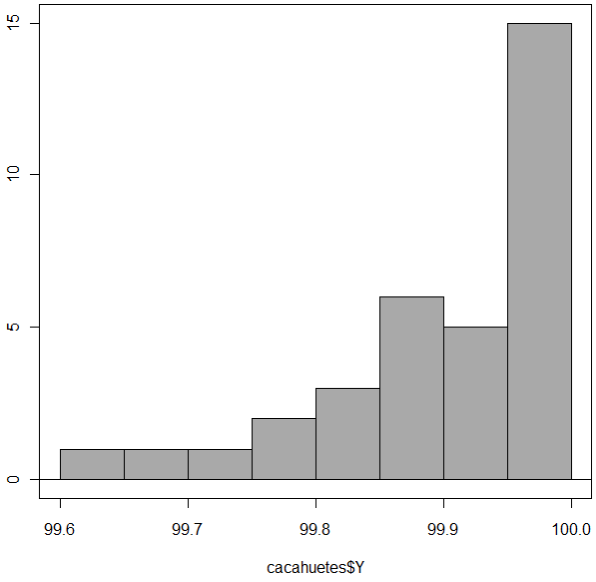
1.-

Histograma X



Unimodal, asimétrico a la derecha y parece no presentar datos atípicos.

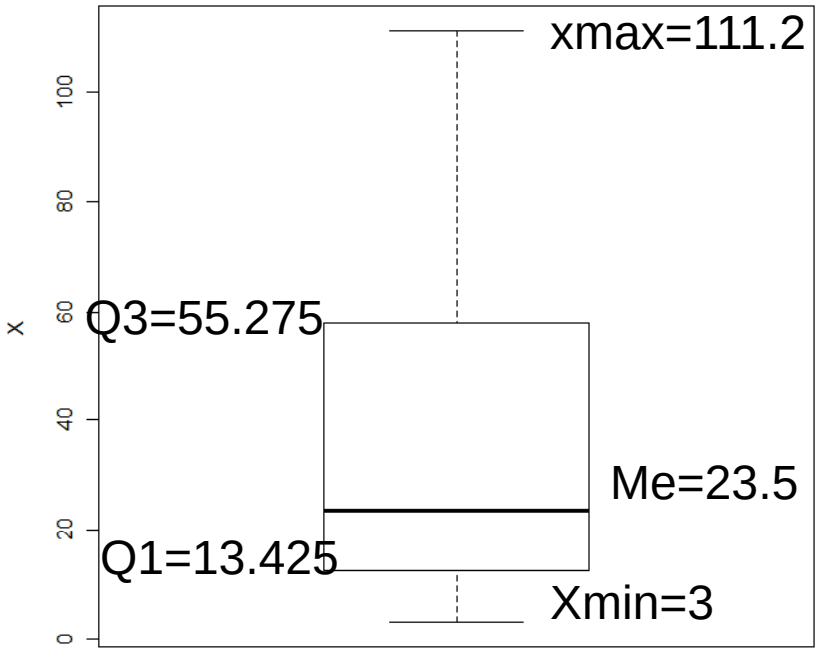
Histograma Y



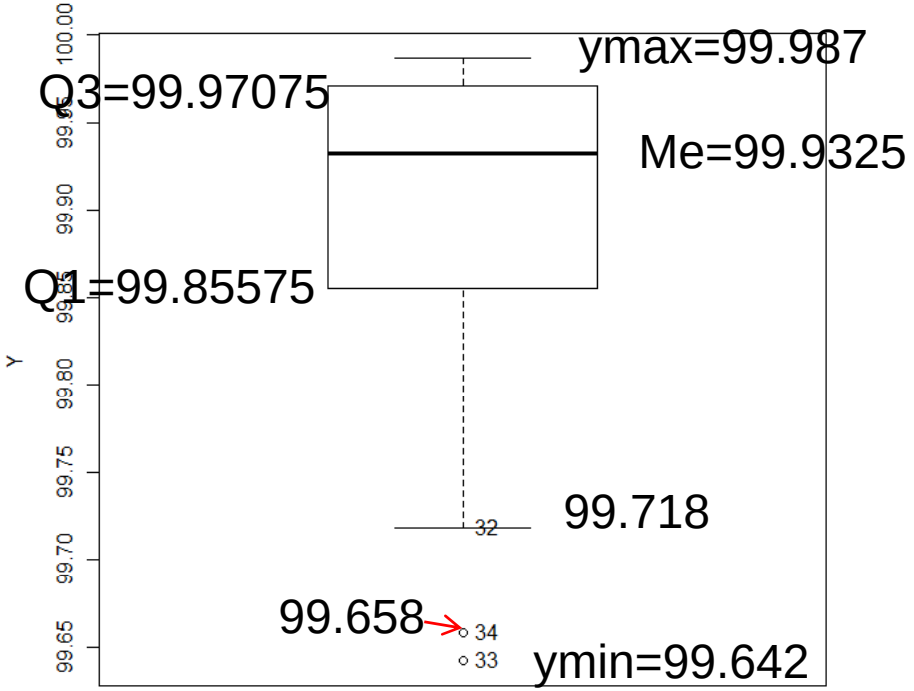
Unimodal, asimétrico a la izquierda y no parece que presente datos atípicos.

2.-

Variable X



Variable Y



La variable Y presenta dos datos atípicos:
caso 34 de valor 99.658 y caso 33 de valor 99.642 que es el valor mínimo de la muestra.

Descriptivos numéricos:

```
> numSummary(cacahuetes[,c("X", "Y")], statistics=c("mean", "sd", "quanti
      mean           sd     0%      25%     50%      75%    100%  n
X 36.60294 29.31939564  3.000 13.42500 23.5000 55.27500 111.200 34
Y 99.89582  0.09352442 99.642 99.85575 99.9325 99.97075  99.987 34
```

RIQ_X=55.275-13.425=41.85

RIQ_Y=99.97075-99.85575=0.115

Valores **No atípicos** para la variable X:
[Q1-1.5*RIQ,Q3+1.5*RIQ]=[13.425-1.5*41.85,55.275+1.5*41.85]=
=[-49.35,118.05]
Entonces la variable X no presenta datos atípicos

Valores **No atípicos** para la variable Y:
[Q1-1.5*RIQ,Q3+1.5*RIQ]=[99.85575-1.5*0.115,99.97075+1.5*0.115]=
=[99.68325,100.14325]
Entonces la variable Y presenta dos datos atípicos: 99.642 y 99.658

3.- **Variable X:**

Medida de centro: mediana=23.55
Medida de dispersión: RIQ=41.85
Debido a la gran asimetría a pesar de que no presenta datos atípicos.

Variable Y:

Medida de centro: mediana=99.9325
Medida de dispersión: RIQ=0.115
Debido a la gran asimetría y a que sí presenta dos datos atípicos.

4.-

Para comparar la dispersión de ambos conjuntos de datos vamos a calcular el coeficiente de variación de cada uno de ellos:

Variable X:

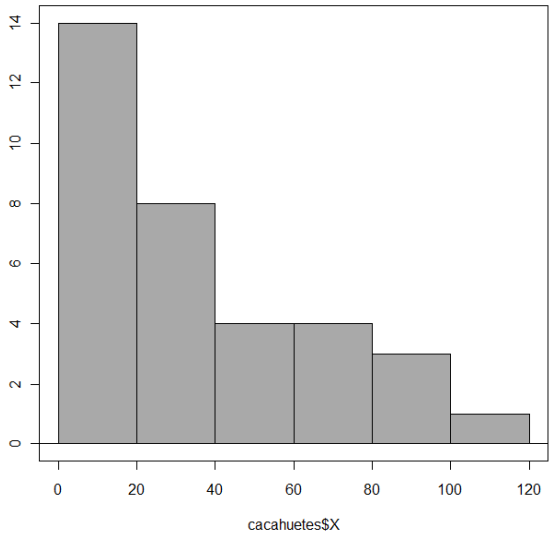
$$CV_X = \frac{s_x}{\overline{x}} = \frac{29.3194}{36.6029} = 0.8010$$

Variable Y:

$$CV_Y = \frac{s_y}{\overline{y}} = \frac{0.0935}{99.8958} = 0.0009$$

Entonces la variable Y presenta menor dispersión relativa .

5.- Como el histograma de X es:



Podemos asumir que el modelo de probabilidad más adecuado para modelizar la variable X es un **modelo exponencial** de media:

$$\lambda = \frac{1}{mean(X)} = \frac{1}{36.6029} = 0.02732$$

Supongamos que $X \rightarrow Exp(\lambda = 0.03) \Rightarrow$

$$P(10 < X < 20) =$$

$$= 0.4511884 - 0.2591818 = 0.1920066$$

$$P(X > 50) = 0.2231302$$

```
> pexp(c(20,10), rate=0.03, lower.tail=TRUE)
[1] 0.4511884 0.2591818

> pexp(c(50), rate=0.03, lower.tail=FALSE)
[1] 0.2231302
```

6.- Intervalo de confianza al 98% para la media de X es:

```
> t.test(cacahuetes$X, alternative='two.sided', mu=0.0, conf.level=.98)

        One Sample t-test

data:  cacahuetes$X
t = 7.2795, df = 33, p-value = 2.367e-08
alternative hypothesis: true mean is not equal to 0
98 percent confidence interval:
 24.30994 48.89594
sample estimates:
mean of x
 36.60294
```

Para construir el intervalo de confianza de la media de X utilizamos la distribución:

$$t_{33}$$

7.- Nos planteamos el contraste:

$$\left.\begin{array}{l} H_0 : \mu_X = 46 \\ H_1 : \mu_X < 46 \end{array}\right\}$$

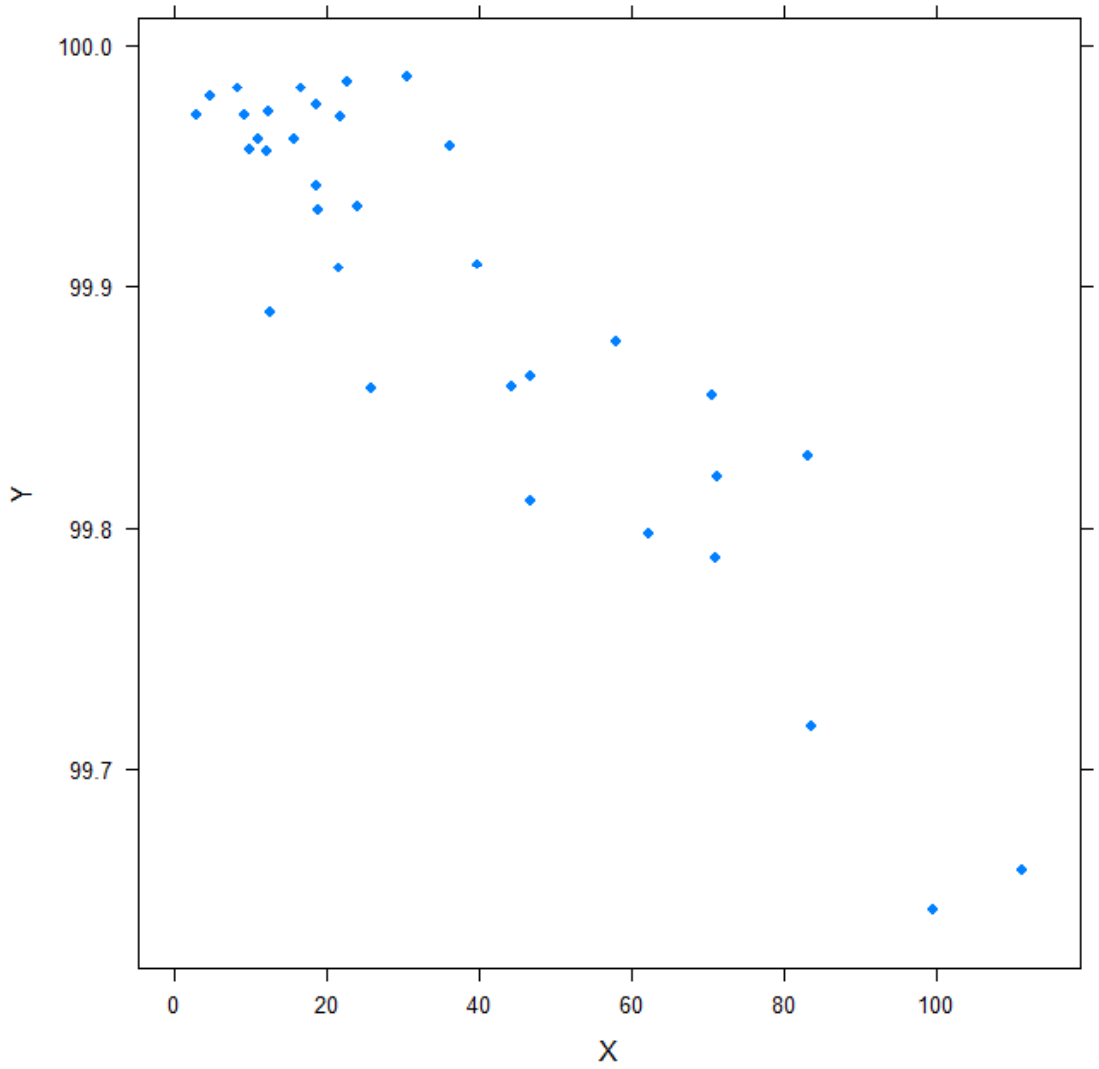
```
> t.test(cacahuetes$X, alternative='less', mu=46, conf.level=.95)

        One Sample t-test

data:  cacahuetes$X
t = -1.8689, df = 33, p-value = 0.03527
alternative hypothesis: true mean is less than 46
95 percent confidence interval:
     -Inf 45.11253
sample estimates:
mean of x
 36.60294
```

Obtenemos un:

p-valor=0.03527<0.05 $\Longrightarrow$ Rechazamos H0 con una gran confianza

$\Longrightarrow$ **El nivel medio de contaminante es significativamente MENOR a 46 partes por billón y, por lo tanto, se cumple la legislación vigente.**

8.-



Vemos que la relación entre X e Y es aparentemente lineal con dependencia negativa.

9.-

```
Call:
lm(formula = Y ~ X, data = cacahuetes)

Residuals:
      Min        1Q    Median        3Q       Max
-0.076516 -0.020012 -0.004806  0.027094  0.073747

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.000e+02  1.089e-02 9184.91  < 2e-16 ***
X           -2.903e-03  2.335e-04  -12.44 8.54e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03933 on 32 degrees of freedom
Multiple R-squared: 0.8285,  Adjusted R-squared: 0.8232
F-statistic: 154.6 on 1 and 32 DF,  p-value: 8.538e-14
```

$$\Rightarrow Y = 100.0 - 0.002903 * X$$

$$R^2 = 0.8285 \approx 0.83$$

Ajuste bueno

10.-

El valor estimado para Y cuando la variable X es igual a 60 partes por billón es:

$$\hat{y} = 100.0 - 0.002903 * 60 = 99.82582$$

Esta estimación es fiable pues el ajuste es bueno y X=60 pertenece al rango observado de las X's=[3.00,111.20]