Práctica 6 R-Commander

NOMBRE: *Jaime Osés Azcona*

1- Distribuciones en el muestreo

1. (En clase de TEORÍA) El peso por unidad en una variedad de mandarina tiene una distribución normal de media 50 y varianza 5. Para una muestra de tamaño 20:

a) ¿Cuál es la probabilidad de que la media muestral se encuentre entre 45 y 50?

Como X sigue una distribución Normal diremos que X~N(50, $\sqrt{5}$).

Por tanto, podremos afirmar que la media también seguirá una distribución normal, pro en este caso $\bar{X}_{20}$ ~ N(50, $\sqrt{\frac{5}{20}}$) o lo que es lo mismo $\bar{X}_{20}$ ~ N(50, 0.5)

En este caso, queremos hallar P(45 ≤ $\bar{X}_{20}$ ≤ 50) = P ($\bar{X}_{20}$ ≤ 50) - P ($\bar{X}_{20}$ ≤ 45). Por tanto, iremos al menú ***"Distribuciones-...continuas-…normal-Probabilidades nomales acumuladas"*** e introduciremos la media (50) y la desviación típica (0.5) y el valor acumulado que queremos obtener, en este caso hasta el 50 y hasta el 45.

```
> pnorm(c(50), mean=50, sd=0.5, lower.tail=TRUE)-pnorm(c(45), mean=50, sd=0.5, lower.tail=TRUE)
[1] 0.5
```

b) ¿Cuál es la probabilidad de que la varianza muestral sea mayor o igual que 7.44?

Como X se distribuye siguiendo una distribución normal, también podemos afirmar que la varianza se distribuye como una chi-cuadrado con 19 grados de libertad. $\frac{(n-1)\cdot S^2}{\sigma^2}$ ~ X^2_{19}

En este caso, queremos hallar P(S^2 ≥ 7,44) = P (X^2_{19} ≥ $\frac{19\cdot 7{,}44}{5}$) = 1 - P (X^2_{19} < 28,272).

Por tanto, iremos al menú ***"Distribuciones-...continuas-…Chicuadrado-Probabilidades Chicuadrado acumuladas"*** e introduciremos los grados de libertad (19) y el valor acumulado que queremos obtener (28,272).

```
> 1-pchisq(c(28.272), df=19, lower.tail=TRUE)
[1] 0.07833198
```

c) ¿Por encima de qué valor se encontrara la varianza muestral en el 90% de las muestras ?

En este caso, queremos hallar lo contrario que en el caso anterior. Por tanto, tendremos que ir al menú ***"Distribuciones-...continuas-…Chicuadrado-Cuantiles..."*** y pondremos el porcentaje que se nos pide, 1-0.9. Lo que hallaremos será lo siguiente :

P(S^2 ≥ x) = P(X^2_{19} ≥ a) = 1 – P(X^2_{19} < a) = 0,1

```
> qchisq(c(0.1), df=19, lower.tail=TRUE)
[1] 11.65091
```

Para sacar el valor tenemos que deshacer la cuenta que hemos hecho para que la varianza se distribuya como una chi-cuadrado.

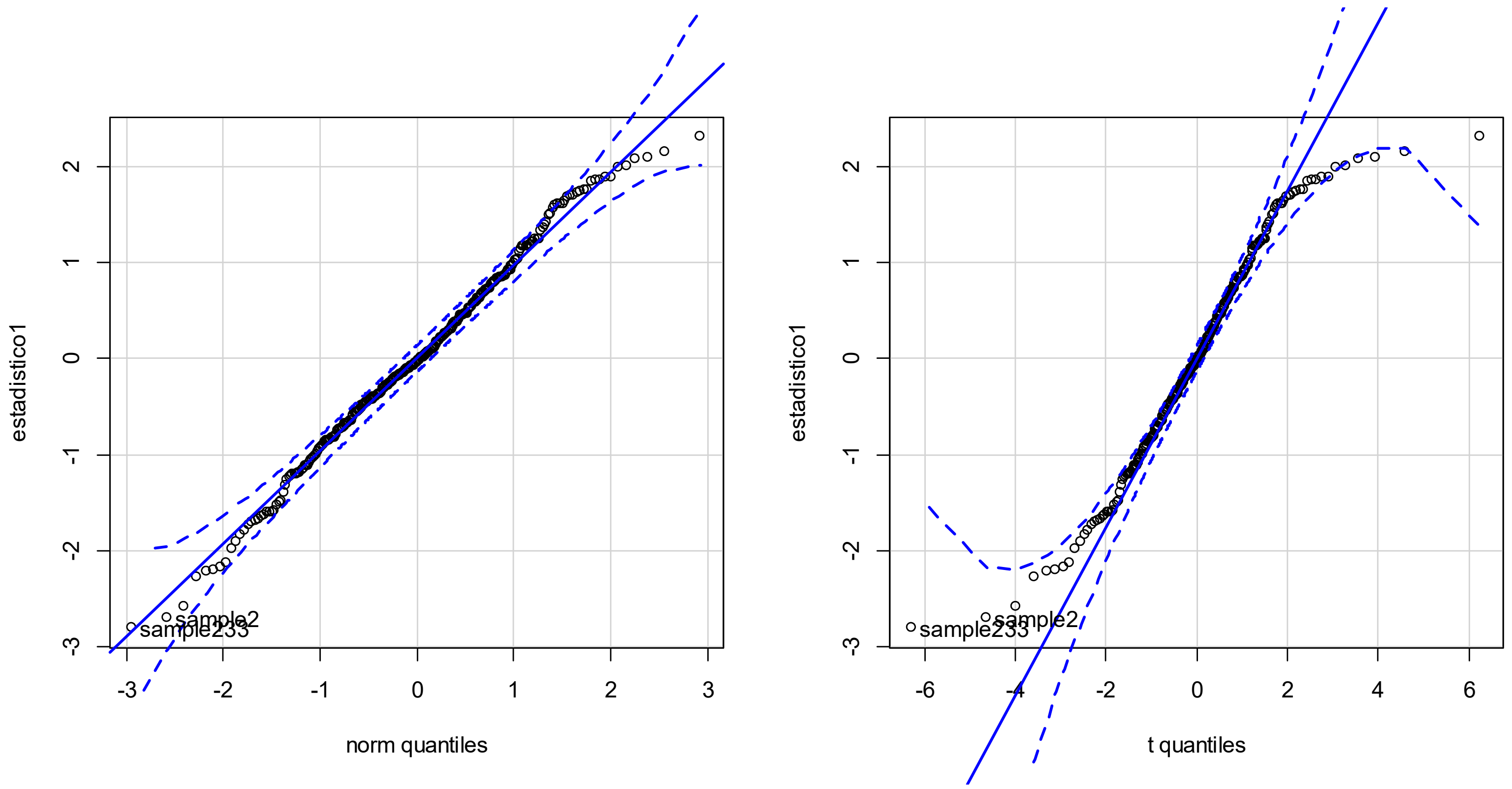
Por tanto el valor real será x = $\frac{\sigma^2\cdot a}{n-1}$ = $\frac{5\cdot 11{,}65091}{19}$ = 3,0658

2. Genera 300 muestras de tamaño 5 de una poblaciónnormal de media 8 y desviación típica 3. Para cada muestra guarda la media y la desviación típica .
Para generar una muestra iremos al menú **"Distribuciones-...*continuas*-...*normal*-Muestras de una distribución normal"** y pondremos la características que se nos pidan en el problema. En este caso nos piden la media y desviación típica de cada muestra, por tanto tendremos que seleccionar dichas casillas al realizar la muestra.

a) Comprueba que el estadístico $\frac{\bar{X}-8}{\sigma_x/\sqrt{5}}$ se comporta como una distribución normal y no como una t-student con 4 grados de libertad.

Para este apartado crearemos una nueva variable desde el menú **"Datos-Modificar variables…-Calcular una nueva variable..."** y la llamaremos estadistico1. Con esto estaremos creando el estadístico que se nos pide en el enunciado para todas las muestras realizadas anteriormente.

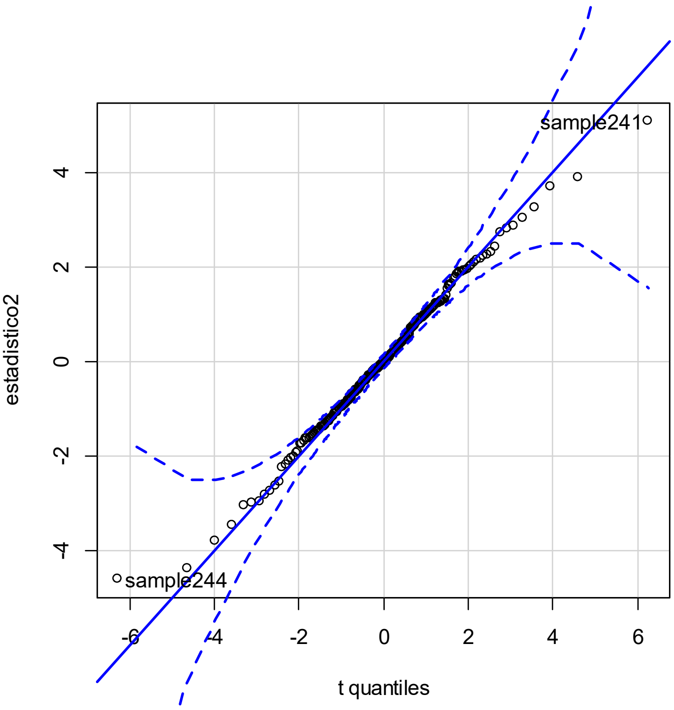
Para comprobar si se comporta como una normal o una t-student iremos al menú **"Gráficas-Gráfica de comparación de cuantiles"** y seleccionaremos el estadístico creado y el tipo de distribución con el que queremos comprobarlo. Lo haremos para las dos que se nos pide.

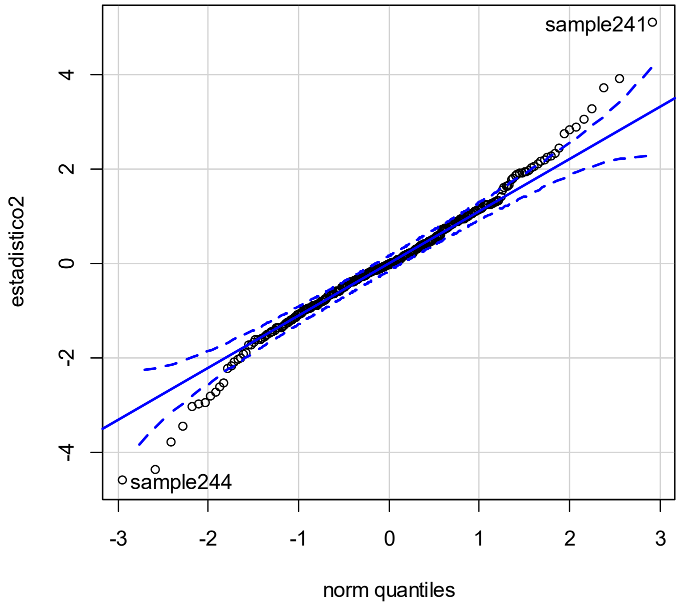


Viendo las 2 gráficas vemos que los puntos se ajustan más a una distribución normal ya que en dicha gráfica los puntos están casi todos pegados a la recta, mientras que en la gráfica de la distribución t-student las colas tienen los puntos muy separados de la línea.

b) Comprueba que el estadístico $\frac{\bar{X}-8}{S/\sqrt{5}}$ se comporta como una distribución t-student con 4 grados de libertad y no como una normal.

En este apartado realizaremos lo mismo que en el apartado anterior, pero esta vez llamando a la variable creada estadistico2 y en ella crearemos el estadístico que se nos pide en el enunciado para todas las muestras realizadas anteriormente.





Viendo las 2 gráficas vemos que los puntos se ajustan más a una distribución t-student ya que en dicha gráfica los puntos están casi todos pegados a la recta, mientras que en la gráfica de la distribución normal las colas tienen los puntos muy separados de la línea.

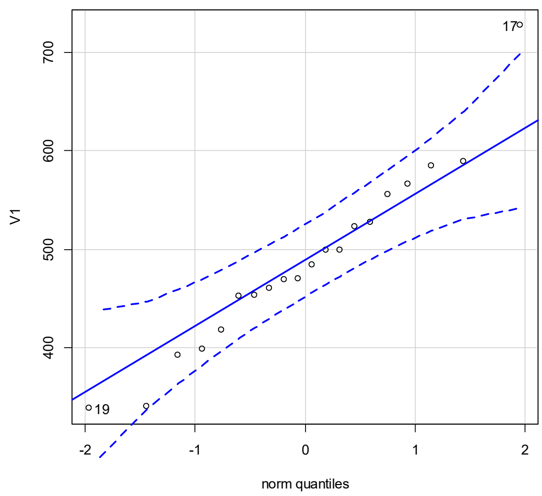
2. Intervalos de confianza y contrastes para la media con varianza desconocida y la proporción.

5. El instituto de meteorología de un país proporciona la media anual de precipitaciones, en mm (espesor en mm de la capa de agua acumulada sobre un suelo horizontal l/m2), durante 20 años consecutivos:

499,2	**555,2**	**398,8**	**391,9**	**453,4**	**459,8**	**483,7**	**417,6**	**469,2**	**452,4**
499,3	**340,6**	**522,8**	**469,9**	**527,2**	**565,5**	**584,1**	**727,3**	**558,6**	**338,6**

a) Comprueba si los datos siguen una ley normal.

Para comprobar si los datos siguen una distribución normal, haremos lo mismo que en el ejercicio 2, pero esta vez introduciremos nosotros los datos manualmente.



Como podemos observar en la gráfica, la mayoría de los puntos están muy próximos a la línea central, a excepción de los de las colas que están ligeramente separados. Sin embargo estos últimos entran dentro de los limites establecidos, por tanto podemos afirmar que los datos siguen una distribución normal.

b) Proporciona un intervalo de confianza al 95% para la media anual de precipitaciones.

Para calcular dicho intervalo podemos hacerlo manualmente con la siguiente fórmula:

IC = $\bar{X} \pm t_{19,\alpha/2} \cdot \frac{S}{\sqrt{n}}$. Pero también podemos hacerlo en R de una forma más sencilla y rápida. Iremos al menú **"Estadísticos-Medias-Test t..."** e introduciremos el valor de confianza (0,95).

```
        One Sample t-test

data:  V1
t = 23.725, df = 19, p-value = 1.399e-15
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 444.2686 530.2414
sample estimates:
mean of x
  487.255
```

IC (95%) = (444.2688, 530.2414)

c) Si un país se considera árido cuando la media anual de precipitaciones es inferior a 420mm; efectúa un contraste que permita aceptar o rechazar que este país sea árido.

Realizaremos el siguiente contraste:

$H_0 : \mu \geq 420$

$H_1 : \mu < 420$

Para ello iremos al menú del apartado anterior, y seleccionando el contraste unilateral adecuado introduciremos los datos que se nos piden.

```
        One Sample t-test

data:  V1
t = 3.2747, df = 19, p-value = 0.998
alternative hypothesis: true mean is less than 420
95 percent confidence interval:
     -Inf 522.7678
sample estimates:
mean of x
  487.255
```

Como el p-value es mayor que el nivel de significación no rechazamos la hipótesis nula.

Otra forma es calculando t_{obs} y comparando con $t_{19,\alpha}$:

```
> qt(c(0.05), df=19, lower.tail=TRUE)
[1] -1.729133
```

Lo que hemos dicho anteriormente también se puede deducir ya que la t_{obs} (3.2747) es mayor que la $t_{19,\alpha}$ y por tanto no se rechaza la hipótesis nula.

Por tanto, podemos rechazar que este país sea árido.

6. Una muestra aleatoria de 500 consumidores de vino participaron en una cata de vinos promovida por de la marca "Bodega". En la cata solo uno de los vinos era de esta marca. Cuando se les preguntó cuál era el vino preferido de todos los catados, 180 de ellos eligieron el de esta marca.

a) Proporciona un intervalo de confianza al 95% para que representa la probabilidad de que un consumidor elija este vino.

En este apartado, para realizar el intervalo de confianza seguiremos la siguiente fórmula:

IC = p ± $Z_{\alpha/2} \cdot \sqrt{\frac{\pi(1-\pi)}{n}}$. Para simplificar al error máximo lo llamaremos E.

Calculamos E:

```
> qnorm(c(0.975), mean=0, sd=1, lower.tail=TRUE)
[1] 1.959964
```

$$E= 1,959964 \cdot \sqrt{\frac{0,36(1-0,36)}{500}} = 0,0421$$

Por tanto IC(95%) = (0.3179, 0.4021)

b) Efectúa el contraste con hipótesis alternativa de que el porcentaje de consumidores que eligen este vino es mayor del 35 %.

Asumimos un nivel de confianza del 95%.

$H_0 : \pi \leq 0.35$

$H_1 : \pi > 0.35$

Para calcular dicho contraste tendremos que hallar la Z_{obs} y $Z_{1-\alpha}$

$$Z_{obs}= \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.36-0,35}{\sqrt{\frac{0,35(1-0,35)}{500}}} = 0,4688$$

```
> qnorm(c(0.95), mean=0, sd=1, lower.tail=TRUE)
[1] 1.644854
```

Como Z_{obs} es menor que $Z_{1-\alpha}$ no se rechaza la hipótesis nula.

c)¿Cuántos catadores deberán participar a la prueba para obtener un intervalo de confianza al 95% con un error máximo de 0,02?

En este apartado solo tendremos que sustituir en la fórmula de apartado **a) :**

$$1,959964 \cdot \sqrt{\frac{0,36(1-0,36)}{n}} = 0,02$$

$$n = \frac{0,36(1-0,36)}{0,02^2/1,959964^2} = 2212,68$$

7. Se estudia la fracción de plantas de maíz de genotipo A. Un modelo genético sugiere que la proporción de plantas con genotipo A es mayor que 1/3. Se hace un análisis de genotipo en una muestra aleatoria de 300 plantas, la cual da como resultado 130 de genotipo A.

a) Construye un intervalo de confianza al 95% y establece conclusiones.

En este apartado, para realizar el intervalo de confianza seguiremos la siguiente fórmula:

$IC = p \pm Z_{\alpha/2} \cdot \sqrt{\frac{\pi(1-\pi)}{n}}$. Para simplificar al error máximo lo llamaremos E.

Calculamos E:

```
> qnorm(c(0.975), mean=0, sd=1, lower.tail=TRUE)
[1] 1.959964
```

$$E = 1{,}959964 \cdot \sqrt{\frac{0{,}43(1-0{,}43)}{300}} = 0{,}056$$

Por tanto IC(95%) = (0.374, 0.486)

Con esto podemos observar que el extremo inferior del intervalo de confianza es mayor que el valor que la proporción que se nos dice en el enunciado. Por tanto la proporción del genotipo A será mayor a 1/3, cosa que hace que la afirmación del enunciado sea cierta.

b) ¿Qué tamaño muestral hubiera sido necesario para mantener el margen de error pero aumentar el nivel de confianza al 99 %?

En este apartado solo tendremos que sustituir en la fórmula de apartado **a)** pero en este caso con diferente $Z_{\alpha/2}$, ya que el nivel de confianza es distinto.

```
> qnorm(c(0.995), mean=0, sd=1, lower.tail=TRUE)
[1] 2.575829
```

$$2{,}575829 \cdot \sqrt{\frac{0{,}43(1-0{,}43)}{n}} = 0{,}056$$

$$n = \frac{0{,}43(1-0{,}43)}{0{,}056^2/2{,}575829^2} = 518{,}56$$

c) Efectúa el contraste de hipotesis correspondiente planteando como hipótesis alternativa la sospecha del enunciado. Asume α = 0,05.

Asumimos un nivel de confianza del 95%.

$H_0 : \pi \leq 1/3$

$H_1 : \pi > 1/3$

Para calcular dicho contraste tendremos que hallar la Z_{obs} y $Z_{1-\alpha}$

$$Z_{obs} = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.43-0{,}33}{\sqrt{\frac{0{,}33(1-0{,}33)}{300}}} = 3{,}6835$$

```
> qnorm(c(0.95), mean=0, sd=1, lower.tail=TRUE)
[1] 1.644854
```

Como Z_{obs} es mayor que $Z_{1-\alpha}$ se rechaza la hipótesis nula. Podremos decir entonces que la sospecha del enunciado era cierta.

d) También se pide estimar con un intervalo al 99% de confianza la proporción de plantas que tienen el hongo cuitlacoche (Ustilago maydis). Se sabe que no superará el 20% pero no se conoce exactamente la prevalencia. En la misma muestra han aparecido 40 plantas con el hongo.

En este apartado, para realizar el intervalo de confianza seguiremos la siguiente fórmula:

IC = p ± $Z_{\alpha/2} \cdot \sqrt{\frac{\pi(1-\pi)}{n}}$. Para simplificar al error máximo lo llamaremos E.

Calculamos E:

```
> qnorm(c(0.995), mean=0, sd=1, lower.tail=TRUE)
[1] 2.575829
```

$$E = 2,575829 \cdot \sqrt{\frac{0,13(1-0,13)}{300}} = 0,05$$

Por tanto IC(99%) = (0.08 , 0.18)

e)¿Qué tamaño muestral hubiera sido necesario si en el apartado anterior hubiéramos exigido un margen de error máximo del 2 %?

En este apartado solo tendremos que sustituir en la fórmula de apartado **d)** pero con un error máximo de 0,02:

$$2,575829 \cdot \sqrt{\frac{0,13(1-0,13)}{n}} = 0,02$$

$$n = \frac{0,13(1-0,13)}{0,02^2/2,575829^2} = 1876$$