TEOREMA DEL LÍMITE CENTRAL

Selección de actividades resueltas

© Jose Fco. Martínez Boscá, Arnau Mir Torres, Lluís M. Pla Aragonés, Àngel J. Gil Estallo (Autors) & Ángel A. Juan (Editor)

UOC Universitat Oberta de Catalunya

Introducción

La importancia del TCL radica en que **sea cuál sea** la distribución de la población original (v.a. *X*), conforme el tamaño de las muestras (*n*) aumenta, la distribución de las medias se va aproximando a la de una normal (de la cual conocemos muchas propiedades).

Por ejemplo, un portal de reservas de vuelos por Internet ha observado que el número de reservas semanales que al final no se formalizan sigue una distribución de media 92, y de desviación estándar 17,2. Tenemos que la variable X es la distribución del número semanal de reservas que finalmente no se formalizan y que no sabemos si es normal o no. Si se coge una muestra aleatoria de tamaño muestral 40 (semanas), por el TCL, la distribución de la media se aproximará a una normal independientemente de la variable X.

Otro ejemplo seria: Un programador ha realizado un programa informático que lo quiere hacer correr en dos ordenadores con configuraciones diferentes y con unos determinados tiempos de ejecución. Tenemos que la variable *Xa* es la distribución del tiempo que tarda en hacer los cálculos de una aplicación un PC con una configuración A y que no sabemos si es normal o no. Si se coge una muestra aleatoria de tamaño muestral 35, por el TCL, la distribución de la media se aproximará a una normal independientemente de la distribución de la variable *Xa*. Es decir, como n=35 > 30, podemos utilizar el TCL para afirmar que la distribución de medias muestrales $\overline{Xa}$ se podrá aproximar por una normal ($\overline{Xa} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$) con media 1,8943 y desviación estándar 0,0364.

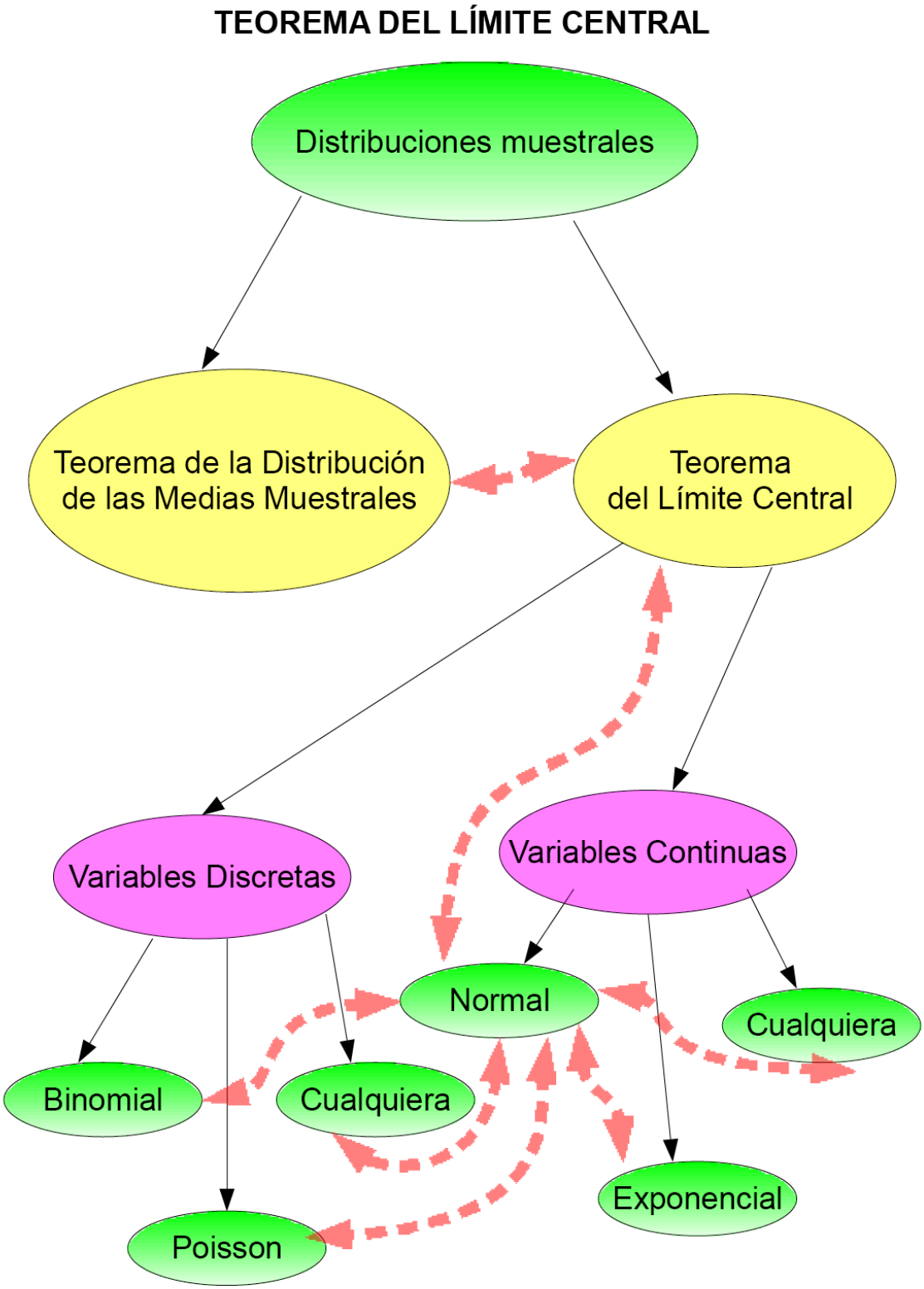
De forma parecida razonamos para la segunda muestra y tenemos que la distribución de medias muestrales $\overline{Xb}$ se podrá aproximar por una normal ($\overline{Xb} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$) con media 2,0387 y desviación estándar 0,0291.
A partir de aquí ya podemos calcular probabilidades, intervalos, hacer contrastes de hipótesis para las medias, etc. Eso lo veremos en los módulos siguientes.

Uno de los casos más habituales en los que podemos aplicar el teorema del límite central es a la hora de hacer un proceso de control de calidad. Entenderemos por control de calidad el seguimiento de cierta variable aleatoria en un proceso de producción a partir de la media de muestras sucesivas.

Estableceremos un intervalo, de manera que las medias que caigan fuera de este intervalo nos indicarán que existe alguna anomalía en el proceso de producción en aquel instante y habrá que revisarlo. Los límites de este intervalo se denominan límites de control.

A pesar que la estadística es muy útil en el control de calidad, muchas personas que las ven por primera vez sienten cierto rechazo. Sin embargo, cuando se comprenden las ideas que hay detrás de los métodos estadísticos, su utilización en la práctica es muy sencilla; todo lo que se necesita son conocimientos elementales de aritmética (sumar, restar, multiplicar y dividir).

Mapa conceptual



Actividad 1: Simulación del Teorema de Distribución de las Medias Muestrales.

Simulación. Media Muestral. Normal. Tamaño muestral.

A fin de visualizar el *Teorema de Distribución de las Medias Muestrales*, vamos a "simular" la extracción de *k=100* muestras de una variable normal con media 80 y desviación típica 5. Tomaremos como *n=9* el tamaño de cada muestra.

Para realizar la simulación de 100 muestras de una distribución normal de media 80 min. y desviación estándar 5, debemos generar una matriz de 9 columnas y 100 filas. Cada componente de esta matriz es una observación aleatoria proveniente de una distribución normal de media 80 y desviación estándar 5.

```
> x<-rnorm(900,80,5)
> dim(x)<-c(100,9)
> x
```

Habremos generado así una matriz de 9 columnas y 100 filas. Cada componente de esta matriz es una observación aleatoria proveniente de una distribución normal de media 80 y desviación estándar 5.

Observación: en los apartados de simulación cada uno obtendrá un resultado ligeramente diferente.

Consideraremos que cada una de las filas obtenidas es una muestra, y lo que haremos ahora será calcular la media asociada a cada una de estas 100 muestras.
En la variable media_filas calculamos la media de las 100 filas:

```
> media_filas<-apply(x,1,mean)
> media_filas
```

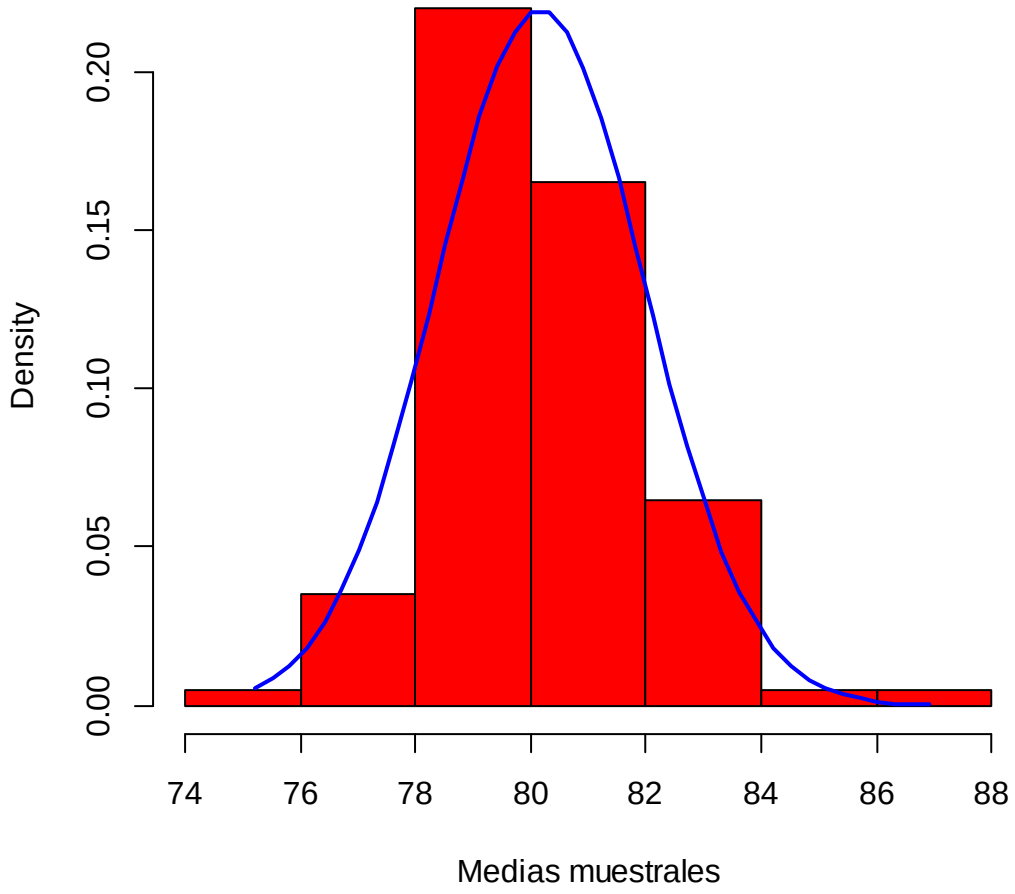
[1] 82.43585 79.82845 79.62773 82.77435 84.45286 82.44325 78.79425 80.61491 82.01091
80.26040 80.57207 80.26302 83.78623 78.91728 80.46891 80.79037 78.67052 82.27575
[19] 81.85368 78.56232 77.52486 82.44447 86.92261 81.96172 78.81054 81.58585 79.72984
81.58484 80.71797 80.82387 79.24158 78.73051 78.13033 79.58422 79.45598 81.79801
[37] 80.79622 80.82231 81.52369 77.18913 79.75902 82.73239 79.27873 81.01397 78.17444
79.41111 83.98582 80.32767 79.75012 80.89130 78.43534 79.51610 80.12659 77.42119
[55] 79.39917 78.22530 80.81746 79.34257 79.28187 82.80543 79.47672 79.13153 79.58254
79.77636 79.64871 78.61575 81.26895 78.89019 81.80122 80.71970 79.16976 77.83976
[73] 80.08356 79.81303 79.10042 78.50707 78.04565 77.47971 79.50715 82.57421 78.49880
81.31346 78.22448 75.20927 81.46834 79.04851 78.13508 77.74716 81.30095 80.29984
[91] 80.09539 82.99574 79.84758 80.27280 76.66568 81.62007 79.60086 81.08365 82.77367
79.50476

A continuación, calculamos la media, la varianza y la desviación típica de la variable media_filas:

```
> mean(media_filas)
[1] 80.16215
> var(media_filas)
[1] 3.308112
> sd(media_filas)
[1] 1.818822
```

Usando las instrucciones siguientes de R hacer el histograma de la curva normal:

```
> xfit<-seq(min(media_filas),max(media_filas),length=40)
> yfit<-dnorm(xfit,mean=mean(media_filas),sd=sd(media_filas))
> lines(xfit, yfit, col="blue", lwd=2)
```



Como podemos observar el histograma de las medias se parecen bastante a la curva normal. La media y la desviación típica de las medias son:

```
> mean(media_filas)
[1] 80.16215
> var(media_filas)
[1] 3.308112
> sd(media_filas)
[1] 1.818822
```

1. La distribución de la v.a. inicial X era normal, parece que también la distribución de la v.a. media_filas) es normal, de media muy similar y desviación estándar menor (los puntos de la media_filas están menos “dispersos” que los de la X).

2. Más concretamente, la media de los 100 valores contenidos en Columna Media _filas (y que es una aproximación a la media de la v.a. media_filas) es de 80.16215, valor muy similar a la media de X (que era de 80). Esto es coherente con lo que la teoría nos indica:

$$\mu_{\bar{X}} = \mu$$

3. La desviación estándar de los 100 valores en Columna Media_filas (que será una aproximación a la desviación estándar de media_filas) es de 1.818822. Si tomamos la desviación estándar de X (que era de 5) y la dividimos por 3 (raíz de 9, el tamaño de la muestra), obtenemos el valor 1.667. Ambos valores son muy parecidos, tal y como la teoría predice:

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

Es interesante notar que aun no habiendo tomado inicialmente una variable normalmente distribuida, las conclusiones obtenidas serían semejantes siempre que el tamaño muestral *n* sea lo suficientemente grande (tal y como predice el *Teorema Central del Límite*). El proceso a seguir es análogo al anterior, por lo que se deja como práctica especialmente recomendada.

Actividad 2: Estudio sobre la formalización de reservas hechas en un portal de vuelos por Internet.

TCL. Distribución de probabilidad. Función de distribución. Media muestral.

Un portal de reservas de vuelos por Internet ha observado que el número de reservas semanales que al final no se formalizan sigue una distribución de media 92, y de desviación estándar 17,2.
Si se coge como muestra los datos de las 40 últimas semanas, ¿Cuál es la distribución de la media muestral?
En el supuesto del apartado de antes, ¿cuál es la probabilidad que la media semanal de reservas de vuelo que al final se cancelan esté entre 86 y 100?
Si queremos conseguir sólo una probabilidad del 50%, ¿qué número de vuelos exactamente deberán ser cancelados?

Solución

1. Tenemos que la variable X es la distribución del número semanal de reservas que finalmente no se formalizan y que no sabemos si es normal o no. Si se coge una muestra aleatoria de tamaño muestral 40, por el TCL, la distribución de la media se aproximará a una normal independientemente de la variable X.
Es decir, como n=40 > 30, podemos utilizar el TCL para afirmar que la distribución de medias muestrales $\bar{X}$ se podrá aproximar por una normal ($\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$) con media 92 y desviación estándar 2.72.

2. La Probabilidad que tenemos que obtener es P(86< $\bar{X}$ <100).
Utilizaremos los comandos de R:

```
> pnorm(100,92,2.72)
[1] 0.9983652

> pnorm(86,92,2.72)
[1] 0.01369612
```

P(86< $\bar{X}$ <100)=P($\bar{X}$ <100)-P($\bar{X}$ <86)=0.9984 -0.0137=0.9847

En conclusión, podemos decir que existe una probabilidad del 98.47% que el número medio semanal de reservas canceladas de vuelos de este portal de Internet esté entre 86 y 100.

3. Queremos saber cuánto vale "c" para que P($\bar{X}$ <c)=0.50. Mediante las instrucciones:

```
> qnorm(c(0.50),mean=92,sd=2.72,lower.tail=TRUE)
[1] 92
```

Por lo tanto, el número de vuelos cancelados debería ser 92.

Actividad 3: Validación experimental del Teorema del Límite Central a partir de una variable exponencial.

Simulación. Exponencial. Histograma. Error estándard.

Este ejercicio está dedicado a validar experimentalmente el Teorema del Límite Central a partir de una distribución exponencial. Seguid las indicaciones siguientes y responded a las preguntas que se plantean.

a) Generaremos 200 muestras de tamaño 100 de una variable aleatoria exponencial de esperanza 0,25. Para hacerlo con el *R-Commander -> Distribuciones -> Escoged si discreta o continúa -> Escoged la distribución -> Muestra de una distribución ...* Guardaremos cada muestra de las 200 en 200 filas diferentes. Interpretad el fichero obtenido.
b) A continuación calcularemos las medias de todas las filas rellenando el cuadro de diálogo. Marcando las opciones: Media de cada muestra
A esta columna la denominaremos "Mean". Interpretad los datos de esta columna.
c))Haced un histograma de esta columna –histograma dónde aparezca la normal que mejor se ajuste.
d) Relacionad los cálculos anteriores con el Teorema del Límite Central.
e) Contad cuántos valores de "Mean" son más grandes que 0.26.
f) La parte de CPU que utiliza un determinado programa es del 0.25 y se puede modelizar con una distribución exponencial. Ejecutamos 100 veces el programa. Utilizando el Teorema del Límite Central, ¿cuál es la probabilidad que la media de CPU utilizado sea superior a 0,26?
g) Relacionad el valor obtenido en f) con el resultado obtenido en el apartado e).

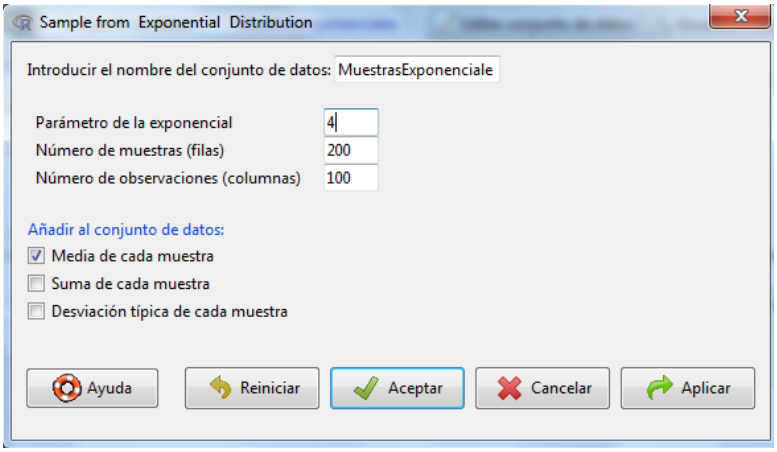
Observación: en los apartados de simulación cada uno obtendrá un resultado ligeramente diferente

Solución

a) Es un fichero con 200 muestras de tamaño 100 de una distribución exponencial de esperanza 0,25. Cada muestra está en una fila diferente.

Con R-*Commander* hacemos: *Distribuciones -> Distribuciones continuas -> Distribución exponencial -> Muestra de una distribución exponencial*
Marcad las opciones: Media de cada muestra. El parámetro de la exponencial es 4 ya que como indica el enunciado la esperanza 0.25,

$E(X)=\frac{1}{\lambda}$, luego el valor de $\lambda=\frac{1}{0.25}=4$



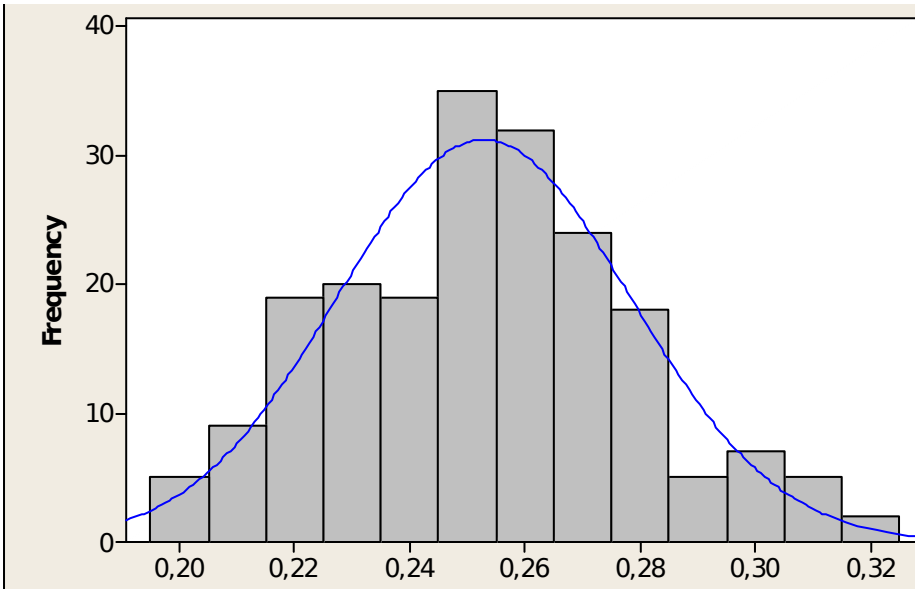
Data: MuestrasExponenciales

	obs96	obs97	obs98	obs99	obs100	mean
1	0.3727886	0.4209906	0.03491093	0.7889122	0.05735168	0.2778568
2	0.266566	0.2719667	0.1801708	0.284844	0.6432555	0.276052
3	0.2334981	0.07536193	0.06212011	0.5705603	0.02824117	0.2272467
4	0.1273145	0.07759368	0.1433404	0.02419104	0.2985677	0.2646351
5	0.09500207	0.04584819	1.014787	0.07998576	0.03499963	0.2207421
6	0.3884456	0.149224	0.05420697	0.07152278	0.9886407	0.2386673
7	0.3326909	0.06948523	0.6740734	0.07578369	0.5146055	0.2375708
8	0.07971302	0.2948031	0.018944	0.3846072	0.01210837	0.2618798
9	0.2419751	0.09957056	0.03880823	0.6079544	0.00806375	0.2383112
10	0.08306276	0.3149919	0.2420862	0.3170033	0.04767987	0.2872308
11	0.6825676	0.6709814	0.04036441	0.1245768	0.027935	0.2052775
12	0.6618572	0.217704	0.3154593	0.1434567	0.5396954	0.2660988
13	0.01448	0.2368504	0.03084488	0.2562708	0.317191	0.2706364
14	0.5751064	0.02671716	0.2822083	0.7668809	0.06556665	0.2455709
15	0.6839913	0.4598182	0.1511014	0.1133986	0.2462045	0.2131633
16	0.3326255	0.1071024	0.2285939	0.07939901	0.2329609	0.2967742
17	0.07501664	0.1931366	0.03722257	0.05567966	0.3084858	0.2539074
18	0.07243991	0.1605531	0.1655727	0.3174418	0.09949058	0.2117407
19	0.1752433	0.01365897	0.2630948	0.208192	0.0059266>	0.2417808

b) En esta columna tenemos la media de cada muestra. Por lo tanto tenemos una muestra de tamaño 200 de medias de muestras.

c) Realizamos el histograma

```
> xfit<-seq(min(mean),max(mean),length=0.5)
> yfit<-dnorm(xfit,mean=mean(mean),sd=sd(mean))
> lines(xfit, yfit, col="blue", lwd=2)
```



d) Obtenemos los estadísticos descriptivos

```
> numSummary(MuestrasExponenciales[,"mean"], statistics=c("mean", "sd", "IQR", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
```

mean	sd	IQR	0%	25%	50%	75%	100%	n

```
0.2474433 0.0250831 0.03505702 0.1896266 0.2276837 0.2487452 0.2627408 0.3150455 200
```

Vemos que los datos de la columna “Mean” se ajustan bien a una normal, tal y como lo dice el Teorema del Límite Central.

e) Hay 76 valores

.

f) Por el Teorema del Límite Central la media muestral sigue una distribución normal con media 0.25 y desviación típica el error estándar, es decir, 0.25/10=0.025.
Por lo tanto la probabilidad que la media de CPU utilizado sea superior a 0,26, con R tenemos:

```
> pnorm(0.26,0.25,0.025)
[1] 0.6554217
```

Por lo tanto la probabilidad es de 1-0,65=0,35

g) Observamos que 76/200=0,38 que por el TLC debería ser la probabilidad calculada en f).

Actividad 4: Validación experimental del Teorema del Límite Central a partir de una variable Poisson.

Simulación. Estandarizar una variable. Ajuste. Histograma. TCL.

En este ejercicio validaremos experimentalmente el Teorema del Límite Central a partir de una Poisson. Seguid las siguientes indicaciones y responded las cuestiones que se piden.

a) Generaremos 200 muestras de tamaño 100 de una variable aleatoria Poisson de parámetro 4.
Con R-*Commander* hacemos: *Distribuciones -> Distribuciones discretas -> Distribución Poisson -> Muestra de una distribución de Poisson*
Marcad las opciones: Media de cada muestra y Desviación típica de cada muestra. Cada muestra la guardamos en una fila diferente. Explicad cómo es la tabla de datos, cuántas filas y cuántas columnas tiene.

Data: MuestrasPoisson

	row.names	obs1	obs2	obs3	obs4	obs5	obs6	obs7	obs8	obs9
1	sample1	3	6	2	4	2	6	3	5	7
2	sample2	5	2	1	5	4	4	3	6	2
3	sample3	4	2	3	8	3	1	3	4	5
4	sample4	3	4	5	7	7	6	1	5	2
5	sample5	2	11	3	2	5	4	6	6	2
6	sample6	4	5	3	6	2	4	2	2	5
7	sample7	6	2	4	4	5	5	4	8	6
8	sample8	5	5	3	6	5	6	1	3	7
9	sample9	4	4	2	4	6	3	2	2	7
10	sample10	10	9	6	2	4	1	8	3	5
11	sample11	3	5	2	4	9	0	4	5	5
12	sample12	9	0	4	2	5	1	5	4	3
13	sample13	10	7	2	6	3	6	2	2	1
14	sample14	1	4	4	7	4	8	5	2	4
15	sample15	5	4	3	4	3	7	6	5	1
16	sample16	5	2	2	3	2	2	6	2	4
17	sample17	5	6	5	3	5	5	3	2	2
18	sample18	1	7	3	4	2	4	3	3	5
19	sample19	3	6	5	4	4	3	8	1	3

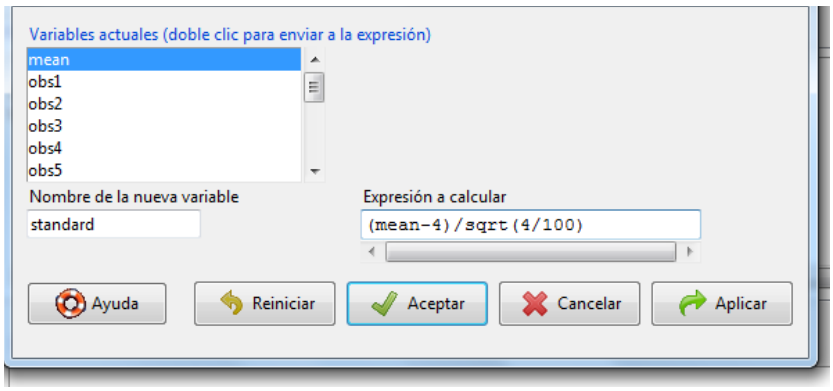
Observación: en los apartados de simulación cada uno obtendrá un resultado ligeramente diferente.

b) A continuación calcularemos las medias de todas las filas usando la opción .A esta columna la denominaremos “Mean”. Interpretad los datos de esta columna.

Data: MuestrasPoisson

	obs96	obs97	obs98	obs99	obs100	sd	mean
1	4	7	2	2	3	1.983874	3.94
2	8	6	5	4	6	1.968925	3.89
3	3	2	5	4	6	1.994031	3.94
4	4	3	2	0	3	2.021376	3.93
5	5	3	4	7	5	2.14323	3.45
6	3	4	3	10	6	2.034525	4.11
7	4	2	3	2	4	1.887626	4.15
8	7	4	4	3	3	2.091348	4.1
9	3	6	8	7	3	2.047319	4.48
10	2	3	8	5	3	2.15566	3.86
11	2	6	2	5	4	1.888936	3.74
12	5	6	2	4	4	2.302809	3.99
13	6	5	7	5	6	1.844758	4.03
14	8	3	3	5	5	2.048996	3.94
15	4	6	5	3	2	2.292257	4.09
16	2	2	6	3	3	1.816062	3.57
17	8	5	1	2	3	2.023698	4.16
18	4	1	5	2	3	1.835783	4.06
19	3	3	7	5	8	1.810491	4.07

c) Ahora calculad una nueva columna “standard” donde a cada dato de la variable “Mean” le restamos 4 y lo dividimos por $\sqrt{4/100}$.
Para hacerlo con R: Datos->Modificar variables del conjunto de datos activo->Calcular una nueva variable



Data: MuestrasPoisson

	obs97	obs98	obs99	obs100	sd	mean	standard
1	7	2	2	3	1.983874	3.94	-0.3
2	6	5	4	6	1.968925	3.89	-0.55
3	2	5	4	6	1.994031	3.94	-0.3
4	3	2	0	3	2.021376	3.93	-0.35
5	3	4	7	5	2.14323	3.45	-2.75
6	4	3	10	6	2.034525	4.11	0.55
7	2	3	2	4	1.887626	4.15	0.75
8	4	4	3	3	2.091348	4.1	0.5
9	6	8	7	3	2.047319	4.48	2.4
10	3	8	5	3	2.15566	3.86	-0.7
11	6	2	5	4	1.888936	3.74	-1.3
12	6	2	4	4	2.302809	3.99	-0.05
13	5	7	5	6	1.844758	4.03	0.15
14	3	3	5	5	2.048996	3.94	-0.3
15	6	5	3	2	2.292257	4.09	0.45
16	2	6	3	3	1.816062	3.57	-2.15
17	5	1	2	3	2.023698	4.16	0.8
18	1	5	2	3	1.835783	4.06	0.3
19	3	7	5	8	1.810491	4.07	0.35

d) Haced un histograma de esta columna donde aparezca la normal que mejor se ajuste.

Relacionad los cálculos anteriores con el Teorema del Límite Central.

Solución

a) Es un fichero con 200 muestras de tamaño 100 de una distribución de Poisson de parámetro 4. Cada muestra está en una fila distinta.

b) En la columna mean tenemos la media de cada muestra. Por lo tanto tenemos una muestra de tamaño 100 de 200 de muestras.

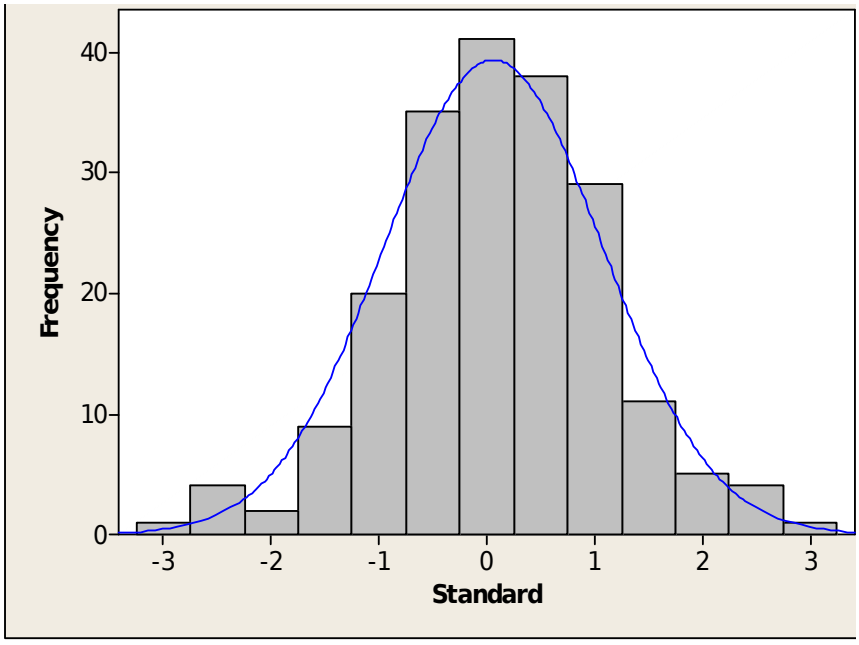
c) En esta columna calculamos para cada una de las 200 muestras, $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{\bar{x}-4}{\frac{2}{\sqrt{100}}}$.

d) Vemos que los datos de la columna “standard” se ajustan bien a una normal estándar, tal y como dice el Teorema del Límite Central.

```
> numSummary(MuestrasPoisson[,"standard"], statistics=c("mean", "sd", "IQR", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
```

```
     mean       sd    IQR    0%   25%   50%    75% 100%   n
-0.0605 0.9051859 1.2125 -2.75 -0.65 -0.15 0.5625  2.4 200
```

e)



Vemos que los datos de la columna “standard” se ajustan bien a una normal estándar, tal y como dice el Teorema del Límite Central.