

Práctica 5: Probabilidad. Distribuciones continuas.

1. R como calculadora

Se escribe en la ventana de ordenes (script) la operación y se ejecuta.

- Operaciones elementales: $2 + 3$, $2 - 3$, $2 * 3$, $2/3$
- Potencias: $2^3 = 2^3$
- Raíces cuadradas: $\sqrt{3} = \text{sqrt}(3)$
- Otras raíces (como potencias): $\sqrt[5]{3} = 3^{(1/5)}$
- Exponencial: $e^2 = \text{exp}(2)$
- Numero combinatorio: $\binom{5}{3} = \text{choose}(5,3)$

Integrales:

Se hace en dos pasos.

- En el primero se crea la función que se va a integrar, construyendo un elemento de R .
Esto se hace ejecutando la orden **integral=function(x) { ... }**
- En el segundo paso se integra el elemento que hemos creado entre a y b con la orden **integrate(integral, lower=a, upper=b)**.

Ejemplo: $\int_0^{\infty} e^{-x} dx$

integral=function(x){exp(-x)} integrate(integral, lower=0, upper=Inf)

Como ves el límite superior puede ser ∞ , en ese caso se escribe *upper=Inf* y también podría ser el límite inferior $-\infty$, en ese caso se escribe *lower=-Inf*

Solución del ejemplo: 1

Ejercicio: A lo largo del tiempo en que las compuertas de un pantano están cerradas, el caudal de un canal de riego (m^3 por segundo) es una variable aleatoria con densidad:

$$f(x) = \begin{cases} \frac{1}{16}x^2e^{-x/2}, & x > 0 \\ 0, & \text{en otro caso} \end{cases}$$

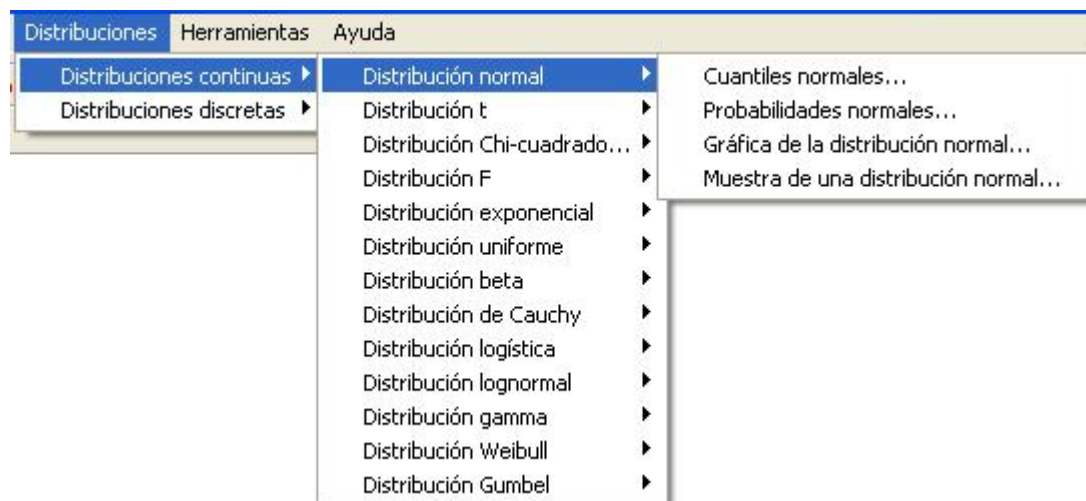
Se pide calcular:

- En qué porcentaje del tiempo el caudal no llegará a 5 m^3 por segundo.
- ¿Cuál es el caudal medio? ¿Con cuánta desviación?
- El 50 % de los días el caudal se mantendrá por encima de, ¿cuántos m^3 por segundo?

2. El menú *Distribuciones*

Como se comentó en la práctica anterior, R-commander jugará un papel similar al de una calculadora en la resolución de problemas con distribuciones de probabilidad.

El menú *Distribuciones* ya se estudió en la práctica anterior para distribuciones discretas. En la siguiente figura observamos que, para distribuciones continuas, incluye más modelos que los estudiados en clase y que las opciones para cada modelo son similares a las del caso discreto.



Es importante observar que R-commander proporciona las probabilidades de las colas (tanto de la derecha como de la izquierda), así, el cálculo de la probabilidad de que una variable tome valores en un intervalo acotado requiere expresarlo en función de la diferencia de las colas: $P(a < X < b) = P(X < b) - P(X \leq a)$

2.1. Ejercicios

1. El tiempo que cuesta rellenar un formulario electrónico se distribuye uniformemente entre 1.5 y 2.2 minutos. Representa sobre la gráfica de densidad y calcula la probabilidad de que cueste menos de dos minutos rellenar el formulario. El 95 % de los casos el tiempo mínimo será de, ¿ Cuántos minutos? Dibuja la distribución de probabilidad acumulada y la de densidad en una misma pantalla.
2. Con un sistema de irrigación automático la altura de las plantas, dos semanas después de la germinación, se distribuye normalmente con media 2.5 centímetros y desviación de 0.5 centímetros. ¿Cuál es la probabilidad de que la altura de la planta sea mayor

de 2.25 centímetros? ¿Y de que se encuentre entre 2 y 3 centímetros? ¿Qué altura sobrepasará el 90 % de las plantas?

Si con otro sistema de irrigación el crecimiento se distribuye como una exponencial de media 0.5, ¿con cuál de los dos sistemas hay más probabilidad de que una planta alcance en dos semanas los 3.5 centímetros?

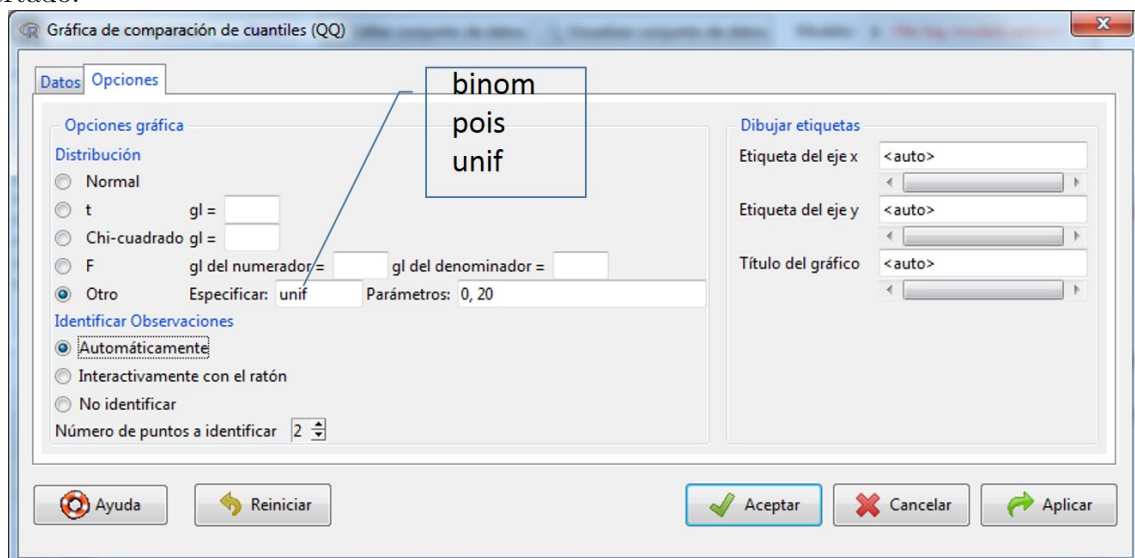
3. Ajuste de distribuciones

Desconocemos la distribución de probabilidad que sigue una variable aleatoria X . Con objeto de establecer la distribución, se observa la variable X en una muestra aleatoria simple extraída de la población.

Los valores proporcionados por la muestra son la única evidencia que nos da la realidad para establecer la distribución de la variable. Veremos dos procedimientos para estudiar este problema.

3.1. Procedimiento gráfico: gráfico qq-plot

Este gráfico representa para cada valor de la muestra el percentil que le correspondería si hubiese sido generado por la distribución en contraste. En estos gráficos, los datos centrales suelen quedar bien ajustados a la recta pero no así las colas, salvo si el ajuste es acertado.



3.2. Inferencia estadística: Contraste de ajuste de distribuciones

A través del siguiente ejemplo se presenta un procedimiento clásico para estudiar el ajuste de distribuciones. Es un procedimiento de inferencia estadística, puesto que de las observaciones reales (valores de la variable) queremos obtener (inferir) una conclusión sobre la población (distribución de la variable).

Nos proporcionan los resultados obtenidos al lanzar un dado 120 veces. Nos planteamos si estos datos pueden ser reflejo de que el dado esté trucado. Datos:

Resultado	1	2	3	4	5	6
Frecuencia	20	14	23	12	26	25

Pasos a seguir:

a) **Planteamiento del contraste**

Escribimos la hipótesis nula (hipótesis que se va a mantener como cierta a no ser que los datos muestrales evidencien lo contrario) sobre la variable de interés que es X: “valor obtenido al lanzar un dado”

$$\left\{ \begin{array}{l} H_0: \text{ el dado está equilibrado- } X \text{ sigue una distribución discreta uniforme.} \\ H_1: \text{ el dado no está equilibrado o } X \text{ no sigue una distribución discreta uniforme.} \end{array} \right.$$

b) **Estadístico del contraste.** El estadístico de contraste es una expresión matemática en la que se mide la distancia entre la hipótesis nula y la realidad proporcionada por la muestra. Esta expresión siempre responde a una lógica.

Resultado	1	2	3	4	5	6
Frecuencia observada: O	20	14	23	12	26	25
H_0 cierta [frecuencia esperada]: E	20	20	20	20	20	20

$$\chi^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i}$$

Si H_0 es cierta, el numerador presentará muchos valores cercanos a cero, y la suma total será pequeña por lo que la distancia entre la realidad y la hipótesis nula es pequeña. Pero...¿hasta qué valor podemos decir que la distancia es pequeña?

Es importante observar que χ^2 es una v.a. y tiene, por tanto, una distribución de probabilidad. Cuando H_0 es cierta esta distribución está perfectamente determinada, y es χ -cuadrado con 5 grados de libertad, puesto que conocido el valor de 5 casillas de la Frecuencia Observada (o esperada) la sexta queda determinada (no está sujeta al azar).

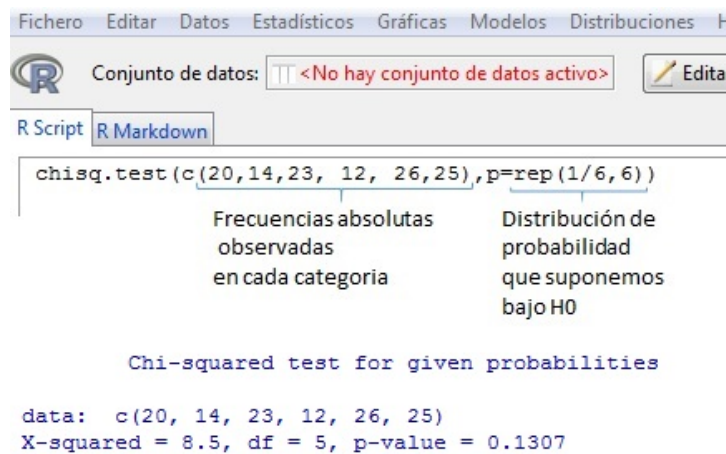
c) Establecer la región de rechazo

Necesitamos conocer el nivel de significación $\alpha = P(e_I)$ que fija la zona de valores tan grandes como para rechazar la hipótesis nula. Técnicamente, representa la proporción de muestras que nos llevarán a cometer un error de tipo uno, es decir, muestras que aun proviniendo de una distribución como indica la hipótesis nula, nos llevarán a considerarla falsa

- Error tipo I (e_I): rechazar la hipótesis nula cuando realmente es cierta.
- Error tipo II (e_{II}): aceptar la hipótesis nula cuando realmente es falsa.

Buscamos el percentil $1 - \alpha$ de la distribución χ^2 . Si se toma $\alpha = 0,05$, el percentil es 11.07, por tanto este valor establece el límite a partir del cual comienza la región de rechazo.

d) **Decisión o resolución del contraste** Calculamos el valor del estadístico con los datos observados, en este caso $\chi^2_{OBS} = 8,5$, obsérvese que el subíndice *OBS* significa observado, es decir, que hemos sustituido los datos muestrales en la fórmula. Como el valor no cae en la región de rechazo, no hay evidencias para rechazar la hipótesis nula. En R se resuelve como sigue:



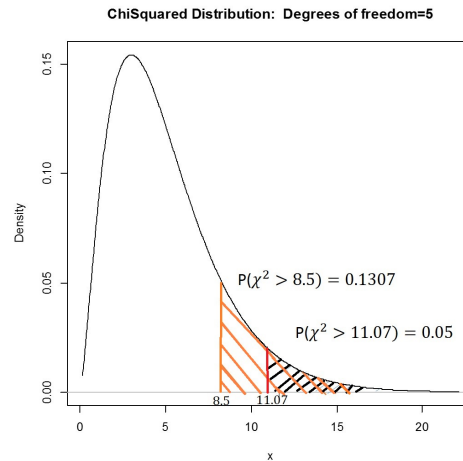
```
Fichero  Editar  Datos  Estadísticos  Gráficas  Modelos  Distribuciones  H
Conjunto de datos: <No hay conjunto de datos activo> Editar
R Script  R Markdown
chisq.test(c(20,14,23, 12, 26,25),p=rep(1/6,6))
           Frecuencias absolutas      Distribución de
           observadas                 probabilidad
           en cada categoría           que suponemos
                                   bajo H0

Chi-squared test for given probabilities

data:  c(20, 14, 23, 12, 26, 25)
X-squared = 8.5, df = 5, p-value = 0.1307
```

En la figura que sigue se pueden identificar e interpretar todos los valores que hemos utilizado en los pasos anteriores. Conviene observar con detenimiento el gráfico y entender el significado de los valores que ahí aparecen. Es importante entender el significado del p-value que nos proporciona la probabilidad de que haya valores más alejados de cero que 8.5, y esa probabilidad es 0.13. Como 0.13 es mayor que 0.05, ya tenemos la seguridad de no estar en la región de rechazo.

Realidad	Decisión	
	H_0	H_1
H_0	✓	e_I
H_1	e_{II}	✓



3.3. Ejercicios

1.- Nos dicen que un programa de ordenador genera observaciones de una distribución normal estándar. Como no estamos seguros de ello, obtenemos una muestra aleatoria de 450 observaciones mediante dicho programa y se obtienen los siguientes resultados:

Resultado	< -2	$(-2, -1)$	$(-1, 0)$	$(0, 1)$	$(1, 2)$	> 2
Frecuencia observada: O	30	80	140	110	60	30
H_0 cierta [frecuencia esperada]: E						

¿Se puede aceptar con un nivel de significación α que el programa funciona correctamente?

2.- El Rector de una Universidad opina que el 60 % de los estudiantes consideran los cursos que realizan como muy útiles, el 20 % como algo útiles y el 20 % como nada útiles. Se hace una encuesta a 100 estudiantes y 68 consideran que son muy útiles, 18 poco útiles y 14 nada útiles. ¿Es aceptable la opinión del rector?

3.- En una tómbola se desarrolla el siguiente sorteo. De un gran bombo opaco se extrae una bola al azar. En el bombo se nos dice que hay 5 bolas con lunares, 45 bolas blancas, 50 azules, 50 rojas y 50 amarillas. Una bola con lunares supone un premio importante, mientras que una blanca es un premio de consolación, el resto no tienen premio. Se observan 600 extracciones y se observa que se reparten 6 premios importantes y 160 de consolación. Se puede afirmar, con una significación del 5 % que el bombo realmente contiene la distribución de bolas que se anuncia?

4. Teorema central del límite

En esta sección ilustraremos uno de los teoremas más importantes en Estadística, el Teorema central del límite, que fundamenta el cuerpo teórico de la inferencia estadística.

El razonamiento lo presentaremos con un ejemplo. Se considera X el caudal (m^3/s) de un río se distribuye con ley exponencial de media 1 (así que $\mu = E[X] = 1$ y $\sigma^2(X) = 1$). Se van a tomar observaciones del caudal para lo que se efectúan n medidas. El experimento consiste en repetir 500 veces la toma de las n observaciones.

1. Simula el resultado del experimento para distintos valores de n , en concreto $n = 10, 20, 50, 100$. Crea, para cada n , un conjunto de datos denominado “caudal n ” que recoja las observaciones así como la media de las observaciones realizadas.

Es importante observar que cada vez que se toma las n observaciones se va a obtener una realización del vector (X_1, X_2, \dots, X_n) que se almacena en filas en el archivo de datos.

El vector (X_1, X_2, \dots, X_n) está formado por las variables X_i : caudal (m^3/s) de la observación i . Estas variables son independientes y están idénticamente distribuidas como la variable X .

Observa que \bar{X}_n es una variable aleatoria porque según las n observaciones que proporcione el azar tomará un valor u otro. De hecho, tenemos a la vista 500 observaciones (realizaciones) de la v.a. \bar{X}_n .

2. Comprueba que se cumple el siguiente resultado de la teoría de probabilidad.

Teorema central del límite: Si $\{X_i\} i = 1, \dots, n$ son variables aleatorias independientes e idénticamente distribuidas a otra variable aleatoria X entonces la distribución de la variable aleatoria \bar{X}_n se acerca a la distribución normal si n es suficientemente grande (cuanto más grande es n más se acerca a la distribución normal). En la práctica si $n \geq 25$, diremos:

$$\bar{X}_n \sim N(\mu_X, \sigma_X/\sqrt{n})$$

o bien en versión tipificada,

$$\frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0, 1) : Z$$

Nota: Si X se distribuye con la ley normal, el resultado es exacto para cualquier valor n .

Para la comprobación realizaremos un histograma de los valores tipificados de \bar{X}_n , para cada tamaño muestral, $n=10, 25, 50$ y 100 , para poder hacer comparaciones y observar que conforme n más grande es el histograma de los valores de \bar{X}_n más se parece a la campana normal. Para ver los 4 histogramas juntos dividiremos la pantalla en 4 partes.

- a) Calculamos para cada tabla de datos una nueva columna con los valores tipificados \bar{X}_n . Es decir, creamos una nueva variable que se llame por ejemplo “tipf.mean” con calcular para recoger los valores de $\frac{\bar{X}_n - \mu_X}{\sigma_X / \sqrt{n}}$.

Despues hacemos para cada tamaño muestral el histograma de la variable “tipf.mean”.

Pero para que queden en la misma ventana de gráficos primero la dividimos en cuatro partes (`par(mfrow=c(2,2))`) y a continuación hacemos uno por uno los cuatro histogramas.

- b) Observa cómo la variable \bar{X}_n tipificada tiene una distribución campaniforme que, conforme mayor se hace el tamaño de muestra (n), más se asemeja a la $Z : N(0, 1)$

Este resultado es clave para comprender la base del proceso inferencial que se estudiará en los siguientes temas teóricos.

Ejercicio propuesto

El tiempo entre llegadas de clientes a un banco se comporta como una exponencial. En el fichero *tiempos.rda* se tiene el tiempo transcurrido entre la llegada de los 100 primeros clientes en 30 días distintos. Se añade la variable media de estos tiempos.

1. Construye un histograma para los valores de la medias de las 30 muestras y a la luz del gráfico da una estimación del tiempo medio que transcurre entre dos llegadas.
2. El tiempo medio entre llegadas en un 75 % de los días es superior a, ¿qué valor?
3. Resuelve la pregunta anterior utilizando el teorema central del límite y haciendo uso de la estimación del primer apartado.