2 Distribuciones discretas de probabilidad

2.1 Distribución discreta uniforme

Seleccionamos en el menú desplegable *discrete uniform* (no se debe confundir con *uniform* que veremos más adelante). En este caso los posibles valores de la variable aleatoria son igualmente probables. El parámetro a cambiar es el número de valores diferentes que puede tomar la variable; por ejemplo este valor es 6 si estamos en el caso sencillo del lanzamiento de un dado. Si escribimos 6 en la caja etiquetada como *number* la representación de la *pdf* muestra igual probabilidad con densidad $0.1667 = 1/6$ para $1 \leq x \leq 6$ y 0 para el resto. Si en la caja correspondiente al valor x en el que se calcula la función escribimos 2, obtenemos la misma densidad 1/6. Si escribimos 2.5 obtenemos densidad nula ya que la distribución que estamos utilizando es discreta.

Pasemos a la gráfica de la probabilidad acumulada *cdf*. Si para el mismo ejemplo escribimos 3.0, obtenemos una probabilidad 0.5: la probabilidad de sacar un número menor o igual que 3 al lanzar el dado es de 1/2. Si calculamos ésta para $x = 3.5$ el resultado es idéntico por ser una distribución discreta como ya indica el aspecto escalonado de la *cdf*. ¿Cual es la probabilidad de que este número sea 6 o menor? efectivamente es la unidad puesto que ya sabíamos que los valores que puede tomar la variable aleatoria de nuestro ejemplo son $\{1, 2, 3, 4, 5, 6\}$ (espacio muestral).

2.2 Distribución binomial

Recordemos que una variable aleatoria binomial es la función que da el número de éxitos en un proceso de Bernouilli. Es decir que tenemos n ensayos repetidos e independientes cada uno de los cuales puede ser clasificado de éxito o fracaso, y cuya probabilidad de éxito es un valor p constante en todos los ensayos. La variable binomial podrá tener valores en el rango $X = \{0, 1, 2, \ldots, n\}$.

Al seleccionar en el menú correspondiente la función binomial (en inglés se escribe igual) nos aparecen como parámetros variables el número de ensayos (en inglés *trials*) y la probabilidad de cada uno de ellos. Utilicemos el gráfico de la *pdf* para resolver un ejemplo sencillo.

Ejemplo 1: *Supongamos que la probabilidad de que un componente pase una prueba de impacto sea* 3/4. *¿ Cual es la probabilidad de que exactamente 2 de los siguientes* 4 *componentes que se prueben pasen la prueba?.*
Escribimos 4 en la casilla de ensayos, 0.75 (ó 3/4) en la casilla de probabilidad y 2 en la de éxitos y obtenemos densidad de 0.2109 (21.09% de probabilidad). Sin MATLAB hubiéramos operado de la siguiente forma:

$$b(2; 4, 0.75) = \begin{pmatrix} 4 \\ 2 \end{pmatrix} \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^{4-2}$$

Ejemplo 2: *La probabilidad de que un paciente se recupere de una cierta enfermedad es* 0.4. *Si* 15 *personas contraen la enfermedad ¿cual es la probabilidad de que*

a) *al menos* 10 *sobrevivan ?* Sea X el número de supervivientes.

$$P(X \geq 10) = 1 - P(X < 10) = 1 - \sum_{x=0}^{9} b(x; 15, 0.4) = 1 - 0.9662 = 0.0338$$

Con MATLAB usando DISTTOOL utilizariamos la opción *cdf* con 15 ensayos de probabilidad 0.4 y calculamos la probabilidad acumulada hasta 9, que resulta ser 0.9662. La diferencia con 1 es el resultado esperado de 3.38%.

b) *sobrevivan entre* 3 *y* 8 *?*

$$P(3 \leq X \leq 8) = \sum_{x=3}^{8} b(x; 15, 0.4) = \sum_{x=0}^{8} b(x; 15, 0.4) - \sum_{x=0}^{2} b(x; 15, 0.4) = 0.9050 - 0.0271 = 0.8779$$

que es muy fácil de realizar con DISTTOOL.

c) *sobrevivan exactamente* 5 *personas ?*

$$P(X = 5) = \sum_{x=0}^{5} b(x; 15, 0.4) - \sum_{x=0}^{4} b(x; 15, 0.4) = 0.4032 - 0.2173 = 0.1859$$

si lo hacemos usando la *cdf* o más fácil en este caso empleando la *pdf*

$$P(X = 5) = b(5; 15, 0.4) = 0.1859$$

Como vemos usando esta herramienta llegamos pronto a los resultados, más rápido que si emplearamos las tablas. Sin embargo no siempre dispondremos de un ordenador con MATLAB a mano y por ello es importante aprender a usar las tablas (*se debe prestar atención ya que* MATLAB *proporciona el área que queda a la izquierda en la distribución, mientras que las tablas dan el área de la cola a la derecha*). Por otra parte es fundamental el planteamiento del problema, es decir establecer correctamente en qué situación nos encontramos y qué distribución debemos aplicar.

Una propiedad importante de la distribución binomial es que será simétrica en el caso de $p = q$ y presentará asimetría a la derecha (serán más probables los valores bajos de x) cuando $p < q$ (y al contrario), como es lógico esperar. Esto podemos verlo con DISTTOOL sin más que seleccionar en *pdf* 20 ensayos de probabilidad 0.1, 0.5 y 0.8 por ejemplo. Para el caso de probabilidad 0.5, que es el clásico ejemplo de lanzamiento de monedas al aire, observamos que es simétrica y que el máximo de la probabilidad ocurre en $X = 10$, es decir obtención de 10 caras o 10 cruces en 20 lanzamientos independientes.

2.3 Distribución de Poisson

Decimos que un experimento aleatorio es un proceso de Poisson cuando los sucesos aparecen aleatoriamente de forma independiente (sin memoria), con una probabilidad por intervalo que es proporcional a la longitud de dicho intervalo, si el proceso es además estable (probabilidad constante) y la probabilidad de que ocurra más de un resultado en un intervalo suficientemente pequeño es despreciable. La distribución de Poisson es un caso límite de la binomial cuando el número de observaciones en esta última es muy grande y la probabilidad de que en una observación se dé el suceso es muy pequeña.

La variable aleatoria de Poisson es el número de resultados que aparecen en un experimento que sigue el proceso de Poisson. La distribución de probabilidad asociada con esta variable se denomina distribución de Poisson y dependerá fundamentalmente del número medio de resultados por intervalo, que denotaremos por λ. De esta forma, la distribución de Poisson se escribe:

$$f(x) = P(X = x) = p(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Ejemplo 6: El número promedio de partículas radiactivas que pasan a través de un contador durante 1 *ms en un experimento de laboratorio es* 4*, ¿cual es la probabilidad de que entren 6 partículas al contador en un milisegundo determinado?.*
En este caso $x = 6$, $\lambda = 4$ luego debemos calcular $p(6; 4)$,

```
>> p=poisspdf(6,4)            % distribucion de Poisson
p =
     0.1042
```

Si nos piden hallar la probabilidad de que entren 5 ó 6 partículas,

$$P(5 \leq X \leq 6) = \sum_{x=0}^{6} p(x; 4) - \sum_{x=0}^{4} p(x; 4)$$

y por lo tanto hay que emplear las probabilidades acumuladas o función de distribución *cdf*,

```
>> p6=poisscdf(6,4);          % probabilidad de que entren hasta 6 particulas
>> p4=poisscdf(4,4);          % idem 4 particulas
p=p6-p4                       % idem 5 o 6
     0.2605
```

Ya hemos visto los comandos que se emplean en MATLAB para estos cálculos, veamos ahora cómo ayudarse de la herramienta DISTTOOL. Arrancamos la aplicación con el comando `disttool`, y seleccionamos la distribución de Poisson. Para resolver el ejemplo nos colocamos en la opción *pdf* e introducimos en las casillas correspondientes $\lambda = 4$, $x = 6$, resultando una probabilidad $p = 0.1042$ como obtuvimos antes. La segunda parte necesita la opción *cdf*; para $x \leq 6$ obtenemos $p = 0.8893$ y para $x \leq 4$, $p = 0.6288$. La diferencia es 0.2605 ó 26%.

Ejemplo 7: Supongamos que en exploraciones del cielo se encuentran con una cierta técnica un promedio de 3.2 *galaxias por grado cuadrado. Se pide encontrar el area que debemos explorar para tener la seguridad en un* 95% *de encontar más de* 100 *galaxias.*
Con el diagrama *cdf*, colocamos en la casilla correspondiente al número de casos el valor 100. Vamos probando valores en lambda y leyendo la probabilidad hasta que valga 0.05. Resulta que $\lambda = 118$ proporciona probabilidad 0.05074 de encontrar $x \leq 100$, o lo que es lo mismo la probabilidad de encontrar $x > 100$ es 0.95. Traducido a nuestro problema: esperamos 118 galaxias en $118/3.2 = 37$ grados cuadrados. Si exploramos este área hay una probabilidad de 0.95 de encontrar más de 100 galaxias.

Ejemplo 8: En un proceso de fabricación de componentes electrónicos, a veces se producen defectos que los hacen inservibles. Supongamos que 2 *de cada* 1000 *piezas salen defectuosas en promedio. ¿cuál es la probabilidad de que en una muestra aleatoria de* 6000 *piezas menos de* 9 *sean inservibles?.*
Este ejemplo es de distribución binomial y se resuelve muy fácilmente como,

$$P(X < 9) = \sum_{x=0}^{8} b(x; 6000, 0.002) = 0.1548$$

o con DISTTOOL, seleccionando binomial y *cdf* con ensayos (*trials* $n = 6000$, de probabilidad p=0.002 y número de casos 8, obtenemos el mismo resultado. Como la probabilidad es muy pequeña ($p \simeq 0$) y n es bastante grande, podemos hacer la aproximación con la distribución de Poisson utilizando $\lambda = 6000 \times 0.002 = 12$. Volvemos a seleccionar la distribución de Poisson con $\lambda = 12$ y número de casos 8, y obtenemos una probabilidad de 0.155.

> *Relléne ahora el cuestionario 3-A.*
> *Si dispone de tiempo libre continúe la práctica 3 hasta el final de la clase.*
> *Esta práctica se terminará en la próxima sesión (cuestionario 3-B).*

3 Distribuciones continuas de probabilidad

3.1 Distribución continua uniforme

Ya vimos la distribución uniforme discreta. A diferencia de ella, la función de densidad *pdf* de la distribución uniforme continua (*uniform* en DISTTOOL) toma valores constantes en un intervalo. Están claras las analogías entre las dos distribuciones. Como ejercicio se seleccionará la distribución uniforme y se calculará $P(X \geq 4)$ y $P(3 < X < 5.5)$ para una distribución uniforme continua de límites $a = 2$ y $b = 7$

3.2 Distribución normal

Una variable aleatoria continua X sigue una distribución normal de media μ y desviación típica σ si su función de densidad es:

$$f(x) = N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

De esta forma, una vez que se especifican μ y σ la distribución queda determinada completamente. La distribución de probabilidad normal tiene forma de campana (llamada campana de Gauss, o curva normal), simétrica (por depender de x a través del término $(x - \mu)^2$), centrada en μ y con anchura proporcional a σ.

Sabemos que la curva de cualquier distribución continua de probabilidad o función de densidad está construída de forma que el área bajo la curva limitada por los puntos $x = x_1$ y $x = x_2$ es igual a la probabilidad de que la variable aleatoria X asuma un valor entre $x = x_1$ y $x = x_2$. Como la resolución de las integrales para cada curva normal no es fácil, es aconsejable utilizar tablas. Para no tener que presentar estas tablas para todos los posibles valores de μ y σ se utiliza una variable normal tipificada Z definida como $Z = (X - \mu)/\sigma$ con lo que sustituyendo nos queda la función de densidad de X:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = N(0, 1)$$

La variable tipificada sigue una distribución normal con media 0 y desviación típica 1, llamada función de densidad tipificada, o estándar. Es claro que esta distribución no depende de ningún parámetro y su representación gráfica es una campana simétrica respecto al eje $z = 0$, en el que alcanza el máximo valor. El problema de calcular la probabilidad de que X se encuentre en un intervalo (x_1, x_2) se reduce entonces a calcular la probabilidad de que Z esté en un intervalo equivalente (z_1, z_2) con $z_1 = (x_1 - \mu)/\sigma$ y $z_2 = (x_2 - \mu)/\sigma$. Por ello que sólo se tabula esta función de densidad.

Ejemplo 9: Cierto tipo de batería dura un promedio de 3 *años, con una desviación estándar de 0.5 años. Suponiendo que la duración de las baterías están normalmente distribuídas, encuentrese la probabilidad de que una determinada batería dure menos de* 2.3 *años.*
Con DISTTOOL en la opción de la función de distribución *cdf* y con $\mu = 3$, $\sigma = 0.5$ y número 2.3, obtenemos $P(X < 2.3) = 0.08076$. Para comprobar que estamos haciéndolo bien podemos utilizar las tablas. En este caso $P(X < 2.3) = P(Z < -1.4) = 0.0808$ ya que $z = (2.3 - 3)/0.5 = -1.4$

Ejemplo 10: En un examen la calificación promedio fue 82 *y la desviación estándar* 5*. El número total de estudiantes con nota entre* 88 *y* 94 *inclusive fue* 8*. Determínese el número de alumnos presentados.* Si suponemos que las notas se dan como números enteros, debemos establecer los límites de manera que calculemos la probabilidad $P(87.5 < X < 94.5) = P(X < 94.5) - P(X < 87.5))$ que con DISTTOOL es fácil de calcular, obteniendose 0.1295. Para comprobar el resultado transformamos a $z_1 = 1.1$ y $z_2 = 2.5$ y comprobamos en las tablas $P(1.1 < Z < 2.5) = P(Z < 2.5) - P(Z < 1.1) = 0.1295$. Si 8 son el 12.95% de los alumnos, el total serán 62. Si en este ejemplo nos piden hallar el sexto decil D_6 introducimos 0.6 en la casilla de probabilidad acumulada y obtenemos 83.27, es decir que el 60% de las calificaciones son iguales o inferiores a a 83.

Ejemplo 11: Se utilizan medidores para rechazar todos los componentes cuyas dimensiones no se encuentren dentro de la especificación $1.50 \pm d$*. Se sabe que esta dimensión está normalmente distribuída con una media de* 1.50 *y una desviación estándar de* 0.2*. Determine el valor d para que el* 95% *de los componentes se encuentren dentro de la especidicación.*
Si el 95% de las medidas están centrados en la media, deben quedar dos colas idénticas a ambos lados que no cumplan las especificaciones. La cola de la derecha será del 2.5% ó 0.025, luego a la izquierda quedará $1 - 0.025 = 0.975$. Otra vez colocando los valores de los parámetros encontramos que probabilidad acumulada de 0.975 se obtiene para $x = 1.892$. Luego la distancia d entre la media y el valor crítico de la cola a la derecha es $d = 1.892 - 1.5 = 0.392$. Análogamente podríamos haber calculado la cola de la izquierda de 0.025 $P(X < 1.108) = 0.025$, luego $d = 1.5 - 1.108 = 0.392$ igual que antes.

3.3 Distribución χ^2 de Pearson

La utilidad de las tres distribuciones que veremos a continuación se centra en las pruebas de contraste de hipótesis que veremos en una práctica posterior. Por ello nos limitaremos de momento a estudiar su aspecto.

La función de densidad asociada a la variable,

$$\chi_n^2 = X_1^2 + X_2^2 + \ldots + X_n^2$$

donde $X_1, X_2, \ldots, X_n$ n son variables aleatorias normales con media 0 y varianza 1 independientes entre sí, se llama distribución χ^2 de Pearson.

La variable χ^2 toma únicamente valores positivos, al ser una suma de cuadrados. Además su distribución depende únicamente del parámetro n, o número de grados de libertad. Gráficamente, su función de densidad es muy asimétrica (para $n = 1$ corresponde a elevar al cuadrado una curva normal tipificada), pero se va haciendo más simétrica a medida que n aumenta.

Utilícese DISTTOOL (función *chisquare*) para comprobar la variación de la forma de la función de densidad con el aumento del número de grados de libertad (*df* del inglés *degree of freedom*.

3.4 Distribución t de Student

La función de densidad asociada a la variable,

$$t_n = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$$

donde $X_1, X_2, \ldots, X_n$ y X, son $n+1$ variables aleatorias normales con media 0 y desviación típica σ independientes entre sí, recibe el nombre de distribución t de Student. La variable se dice que tiene n grados de libertad.

El campo de variabilidad de la variable t de Student será de $-\infty$ a ∞ y su función de densidad dependerá únicamente del parámetro n (grados de libertad). Nótese que, al depender $f(t)$ de t a través de t^2, la función de densidad será simétrica alrededor de $t = 0$. Su forma será campaniforme, siendo más achatada para valores bajos de n.

Utilícese de nuevo DISTTOOL (función *T*). Compruébese la simetría alrededor del punto cero en el gráfico de la función de densidad *pdf* y que por lo tanto la probabilidad acumulada hasta el valor cero es 0.5 en el gráfico de la función de distribución *cdf*.

3.5 Distribución F de Fisher

La función de densidad asociada a la variable,

$$F_{n_1,n_2} = \frac{\frac{\chi^2_{n_1}}{n_1}}{\frac{\chi^2_{n_2}}{n_2}}$$

donde $\chi^2_{n_1}$ y $\chi^2_{n_2}$ son dos variables χ^2 de Pearson con n_1 y n_2 grados de libertad e independientes entre sí, se llama distribución F de Fisher. La variable recibe el nombre de F de Fisher con n_1 y n_2 grados de libertad.

El campo de variabilidad de la variable F es entre 0 e ∞ (al ser un cociente de cuadrados) y su función de densidad depende exclusivamente de los dos parámetros n_1 y n_2, aunque es importante el orden en el que se dan éstos. Compruébese utilizando DISTTOOL (función F).