Práctica 7.Distribuciones de probabilidad

Jesús Esteban Hernández & Jesús Martín Fernández

Contenidos

1. Funciones de probabilidad y distribución en R

Las funciones de probabilidad son herramientas matemáticas que describen cómo se distribuyen las probabilidades de los posibles resultados de una variable aleatoria. . La función de densidad se utiliza para variables continuas y describe la probabilidad de que una variable tome un valor dentro de un intervalo. Por otro lado, la función de masa de probabilidad se aplica a variables discretas, asignando probabilidades a valores específicos. La función de distribución acumulada mide la probabilidad de que una variable aleatoria sea menor o igual a un valor dado, mientras que los valores inversos o *cuantiles* nos permiten encontrar el valor correspondiente a una probabilidad acumulada específica. También es común en el análisis estadístico generar números aleatorios que sigan distribuciones determinadas, lo que es útil para simulaciones y modelado de datos.

R, como lenguaje estadístico, facilita el abordaje de estos problemas mediante funciones integradas para trabajar con diferentes distribuciones de probabilidad, como la normal, binomial, Poisson, uniforme, entre otras. Estas funciones permiten calcular la densidad o masa de probabilidad, la distribución acumulada, obtener cuantiles y generar números aleatorios de manera sencilla. Gracias a estas herramientas, los usuarios pueden realizar análisis probabilísticos y simulaciones de datos de forma eficiente, aplicando conceptos clave de probabilidad en diferentes contextos analíticos.

2. Cómo trabaja R

En R, las funciones para trabajar con distribuciones de probabilidad siguen un esquema simple y flexible. Las funciones comienzan con una letra que indica la operación a realizar:

- p para calcular la probabilidad acumulada o función de distribución.
- d para obtener la densidad de probabilidad (o masa de probabilidad en el caso de distribuciones discretas).
- q para calcular el cuantil, es decir, el valor que corresponde a una probabilidad acumulada específica.
- r para generar números aleatorios que sigan una distribución determinada.

Luego de esta letra, se añade el nombre de la distribución con la que se quiere trabajar. Por ejemplo, para la distribución binomial, las funciones serían `pbinom` (para la probabilidad acumulada), `dbinom` (para la densidad o masa de probabilidad), `qbinom` (para los cuantiles), y `rbinom` (para generar números aleatorios). Lo mismo ocurre con otras distribuciones como la Poisson (`ppois`, `dpois`, etc.) o la normal (`pnorm`, `dnorm`, etc.).

Este enfoque modular y consistente facilita la realización de análisis probabilísticos y simulaciones en R.

Vamos a revisar y describir las principales distribuciones de probabilidad en R. Las discretas son la distribución binomial y la de Poisson. Son distribuciones de probabilidad contínuas la normal, la T de Student, la Chi cuadrado y la F de Schnedecor

3. Distribución binomial

La **distribución binomial** es una distribución de probabilidad discreta que describe el número de éxitos en una serie de ensayos independientes y con dos posibles resultados (éxito o fracaso) en un experimento. Cada ensayo tiene una probabilidad fija de éxito, p, y la probabilidad de fracaso es 1-p. La distribución binomial es utilizada principalmente en situaciones donde se desea modelar el conteo de éxitos en un conjunto de eventos repetidos, como lanzar una moneda un número determinado de veces, realizar encuestas con respuestas de sí/no, o en el control de calidad para contar defectos en lotes de productos. En R, esta distribución se maneja mediante funciones como `dbinom` (para calcular la probabilidad de un número exacto de éxitos), `pbinom` (para la probabilidad acumulada), `qbinom` (para obtener cuantiles), y `rbinom` (para generar números aleatorios siguiendo esta distribución).

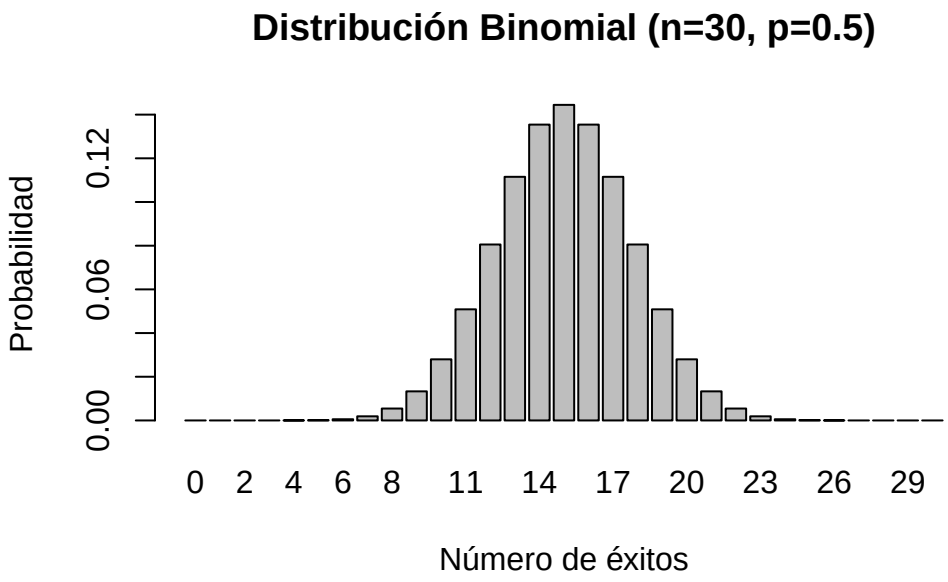
Vamos a dibujar la representación gráfica de una distribución binomial dibujando un diagrama de barras que represente la probabilidad de obtener determinado número de caras en el lanzamiento de una moneda 30 veces

```r
n<-30
p<-0.5

valores <- 0:n  # desde 0 hasta n

# Calcular las probabilidades de cada número de éxitos
probabilidades <- dbinom(valores, n, p)

# Graficar las probabilidades con un gráfico de barras
barplot(probabilidades, names.arg = valores,
        main = "Distribución Binomial (n=30, p=0.5)",
        xlab = "Número de éxitos", ylab = "Probabilidad")
```



Ahora vamos a calcular la probabilidad de obener, exactamentes 10 caras. La probabilidad de obtener exactamente k éxitos en n ensayos, donde la probabilidad de éxito en cada ensayo es p, se calcula con la función de masa de probabilidad de la distribución binomial. acumulada)

```r
# Parámetros: 30 lanzamientos, probabilidad de éxito p = 0.5
n <- 30
p <- 0.5
k <-10

# Calcular la probabilidad de obtener exactamente 10 caras
```

```r
prob_10_caras <- dbinom(k, n, p)
prob_10_caras
```

```
[1] 0.0279816
```

También podemos preguntarnos por a probabilidad de obtener, al menos 10 caras. Al preguntar por la probabilidad de obtener al menos un número de éxitos, estamos calculando la probabilidad acumulada desde ese número .

```r
k <- 6   # número mínimo de éxitos

# Probabilidad de obtener al menos 10 caras
prob_al_menos_10 <- 1 - pbinom(k-1, n, p)
prob_al_menos_10
```

```
[1] 0.9998375
```

Una cuestión con diferente enfoque sería preguntarse cuál es número mínimo de caras que podríamos sacar con determinada probabilidad, por ejemplo cuál es el número mínimo de caras que podríamos sacar con una probabilidad mínima del 80 % en el experimento anterior.

```r
cuantil_80 <- qbinom(0.8, n, p)
cuantil_80
```

```
[1] 17
```

Con una probabilidad del 80% sacaremos, el menos, 17 caras.

La ultima utilidad de la distribución es la de simular un experimento y ver qué resultados tendríamos. Vamos a simular el experimento descrito, lanzar una moneda al aire 30 veces, y a ver los resultados aleatorios de repetirlo 10 veces

```r
set.seed(123) # Fijar la semilla para reproducibilidad
experimentos <- rbinom(10, n, p)
experimentos
```

```
 [1] 13 17 14 18 19 10 15 18 15 15
```

La salida de R indica el número de caras que tendríamos en cada uno de los experimentos.

4. Distribución de Poisson

La distribución de Poisson es una distribución discreta que modela el número de veces que ocurre un evento en un intervalo de tiempo o espacio, bajo la condición de que los eventos sean independientes y ocurran a una tasa promedio constante (lambda). Es especialmente útil para describir eventos raros o poco frecuentes, como el número de llamadas recibidas en una central telefónica por hora, la cantidad de accidentes en una carretera durante un mes o errores en un sistema de producción. Se aplica en diversas áreas como la teoría de colas, biología, medicina y telecomunicaciones, facilitando el análisis de la frecuencia de eventos aleatorios en un tiempo o espacio determinado.

Supongamos que en promedio se producen 7 avisos a la semana en una Unidad de emergencias. Queremos saber cuál es la probabilidad de que se produzcan exactamente 6 avisos en una semana (masa de probabilidad)

```
lambda <- 7  # Tasa promedio de eventos (avisos/semana)
k <- 6       # Número exacto de avisos

# Probabilidad de recibir exactamente 6 avisos
prob_6_avisos <- dpois(k, lambda)
prob_6_avisos
```

```
[1] 0.1490028
```

Y de que se reciban al menos 3 avisos en una semana (probabilidad acumulada).

```
prob_al_menos_3_avisos <- 1 - ppois(2, lambda)
# 1 - probabilidad acumulada hasta 2 avisos
prob_al_menos_3_avisos
```

```
[1] 0.9703638
```

También podrríamos preguntarnos cuál es el número m,aximo de avisos que ocurren con una probabilidad del 90%

```
num_avisos <- qpois(0.9, lambda)
num_avisos
```

```
[1] 10
```

Con una probabilidad del 90% ocurren hasta 10 avisos

5 Distribución normal

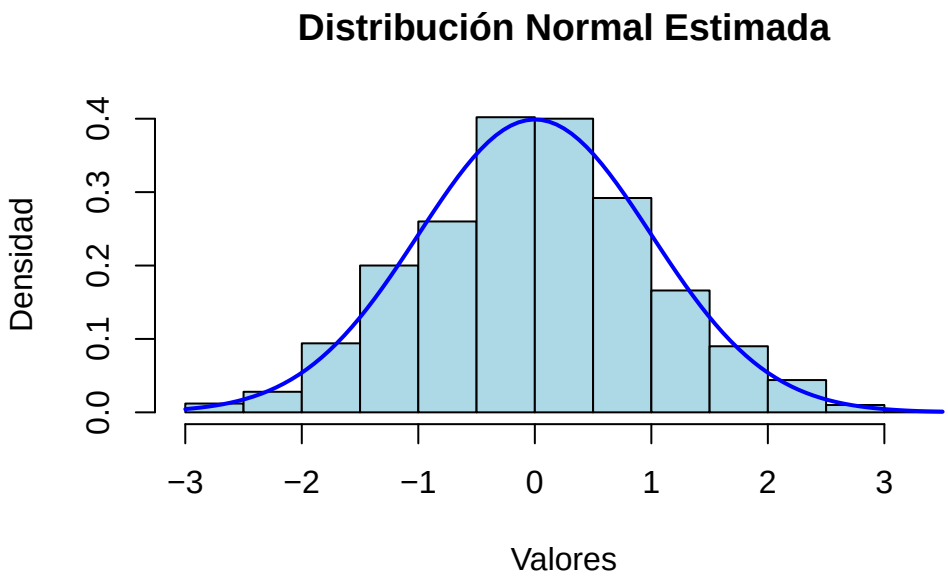
Vamos a describir ahora las distribuciones de probabilidad para variables continuas. Empezaremos por la distribución normal. La distribución de probabilidad normal, o distribución de Gauss, es una de las más utilizadas en estadística debido a su relevancia en fenómenos naturales, sociales y científicos. Se caracteriza por una curva simétrica en forma de campana, donde la mayoría de los valores se agrupan alrededor de la media (mu) y se dispersan conforme se alejan de ella. En esta distribución, la media, la mediana y la moda coinciden, y la desviación estándar (sigma) determina el grado de dispersión de los datos. Su área total bajo la curva es 1, lo que representa la probabilidad total. Las colas de la curva se extienden indefinidamente, aunque con probabilidades muy bajas para valores lejanos a la media. Esta distribución es especialmente útil porque, según el teorema central del límite, la suma de un gran número de variables aleatorias independientes tiende a seguir una distribución normal, independientemente de la distribución original de las variables, lo que la convierte en una herramienta esencial en numerosas disciplinas para modelar datos y tomar decisiones bajo incertidumbre.

Vamos a generar una serie de 1000 números aleatorios y a representarlos gràficamente para ver la forma de la distribución normal

```
valores <- rnorm(1000, mean = 0, sd = 1)

# Graficar el histograma de los valores generados
hist(valores, probability = TRUE, main = "Distribución Normal Estimada",
     xlab = "Valores", ylab = "Densidad", col = "lightblue", border = "black")

# Superponer la curva teórica de la distribución normal
curve(dnorm(x, mean = 0, sd = 1), col = "blue", lwd = 2, add = TRUE)
```



Ahora vamos a suponer que la altura de una poblción dada tiene una distribución normal de media 178,5 y ds 8,25 cms. ¿Cuál es la probabilidad de encontrar a un sujeto con altura <= 180?

```
pnorm(180, mean = 178.5, sd = 8.25)
```

```
[1] 0.5721373
```

Este valor, 0,596 indica dicha probabilidad.

Si quisiésemos conocer la probabilidad de que la altura sea mayor de 180 podemos hacerlo de dos formas que dan el mismo resultado:

```
1- pnorm(180, mean = 178.5, sd = 8.25)
```

```
[1] 0.4278627
```

```
pnorm(180, mean = 178.5, sd = 8.25, lower.tail = FALSE)
```

```
[1] 0.4278627
```

La primera expresión tiene sentido porque la probabilidad total bajo la normal es 1 y e opuesto de una probabilidad se pude calcular restando a 1 dicha probabilidad.

Para saber la probabilidad de que un sujeto esté en un rango de la distribución basta con restar la probabilidad acumulada del valor menor, de la probabilidad acumulada del valor mayor, Por ejemplo para saber cuál es la probabilidad de que un sujeto mida entre 170 y 180 cms en esa población usaríamos la siguiente expresión:

```
pnorm(180, mean = 178.5, sd = 8.25)- pnorm(160, mean = 178.5, sd = 8.25)
```

```
[1] 0.5596703
```

También podemos preguntar cuál es el valor en la población por debajo del cuál está el 95% de los sujetos (que será el mismo por encima del cuál sólo están el 5% de las personas de dicha población).

```
qnorm(0.95, mean = 178.5, sd = 8.25)
```

```
[1] 192.07
```

6. T de Student

La distribución t de Student es una distribución de probabilidad utilizada principalmente en estadística inferencial, especialmente en situaciones donde el tamaño de la muestra es pequeño y la varianza poblacional es desconocida. Esta distribución se asemeja a la distribución normal, pero con colas más largas, lo que la hace más adecuada para muestras reducidas ya que contempla una mayor variabilidad en los datos. A medida que aumenta el tamaño de la muestra, la distribución t se aproxima a la distribución normal estándar. Se emplea comúnmente en pruebas de hipótesis, como el **test t**, y para la construcción de intervalos de confianza, proporcionando una herramienta robusta cuando no se puede asumir que los datos sigan estrictamente una distribución normal.

La función `pt()` calcula la probabilidad acumulada para la distribución t de Student. Por ejemplo la Probabilidad P(T<= -2) con 10 grados de libertad

```
pt(-2, df = 10)
```

```
[1] 0.03669402
```

Este código calcula la probabilidad de que un valor en una distribución t con 10 grados de libertad sea menor o igual a -2.

La función `qt()` te da el valor crítico (o cuantil) de la distribución t para un nivel de probabilidad dado. Por ejemplo, calculamos cuál es el cuantil para un nivel de confianza del 95% (0.975 para dos colas) con 15 grados de libertad

```
qt(0.975, df = 15)
```

```
[1] 2.13145
```

Este código devuelve el valor t crítico para una distribución t con 15 grados de libertad en un intervalo de confianza del 95%. Este valor es útil para pruebas de hipótesis y para construir intervalos de confianza.

Si queremos calcular la probabilidad complementaria, es decir, la probabilidad de que TTT sea mayor que un valor dado en una distribución t, puedes usar la opción `lower.tail = FALSE`. La probabilidad P(T>1.96)con 20 grados de libertad, se calcula con la siguiente expresión:

```
pt(1.96, df = 20, lower.tail = FALSE)
```

```
[1] 0.03203913
```

Esto calcula la probabilidad de que un valor en una distribución t con 20 grados de libertad sea mayor que 1.96.

7. F de Schnedecor

La distribución F de Fisher-Snedecor, comúnmente conocida como distribución F, es una distribución de probabilidad que se utiliza principalmente en el análisis de varianza (ANOVA), en la comparación de varianzas entre dos poblaciones y en la regresión lineal. Esta distribución es útil cuando se desea evaluar si las diferencias observadas entre grupos son estadísticamente significativas. La distribución F surge como el cociente de dos varianzas muestrales independientes, lo que la hace asimétrica y con colas largas hacia la derecha. Sus dos parámetros son los grados de libertad asociados a los dos conjuntos de datos que se comparan. En estadística, la distribución F es clave para evaluar hipótesis sobre la igualdad de varias medias poblacionales y la homogeneidad de varianzas, así como para probar la calidad de los modelos de regresión ajustados.

Vamos a ver un ejemplo sencillo

Supongamos una distribución de una variable en 4 grupos de sujetos que tienen en total 112 sujetos. Si estamos estudiando las diferencias de medias y encontramos un valor de F de 2.75, tendríamos 3 grados de libertad en el numerado (4 grupos menos 1) y 109 grados de libertad en el denominador. Para ver si este valor indica que hay diferencias significativas entre grupos buscaremos la probabilidad acumulada de que una distribución F de un valor igual o inferior a 2,75. El p value que buscamos será 1 menos la probabilidad acumulada.

```
1- pf(2.75,3,109)
```

```
[1] 0.04624339
```

En este caso, como la p es pequeña , asumimos que puede haber diferencias significativas entre alguno de los grupos.

8. Chi cuadrado

La distribución chi-cuadrado es una distribución de probabilidad continua que se utiliza principalmente en el ámbito de la estadística para inferir información sobre las varianzas y las frecuencias observadas en un conjunto de datos. Esta distribución se caracteriza por su forma asimétrica, que se desplaza hacia la derecha y se define por un parámetro conocido como los grados de libertad, los cuales dependen del tamaño de la muestra y del número de parámetros estimados. La distribución chi-cuadrado es fundamental en diversas pruebas estadísticas, como el test de bondad de ajuste, que evalúa si un conjunto de datos se ajusta a una distribución teórica, y el test de independencia, que determina si dos variables categóricas están relacionadas. Su uso es común en análisis de varianza, regresión y en la construcción de intervalos de confianza.

Las opciones en R para trabajar con la distribución Chi cuadrado son las siguientes:

`dchisq(x, df)` , esta función calcula la densidad de probabilidad de la distribución chi-cuadrado para un valor específico **`x`**, **`df`**: Número de grados de libertad.

`pchisq(q, df, lower.tail = TRUE)` : esta función calcula la probabilidad acumulada (función de distribución acumulativa) de que una variable aleatoria chi-cuadrado sea menor o igual a un valor dado; `q`: valor para el cual deseas calcular la probabilidad acumulada, `df`: número de grados de libertad; `lower.tail`: Si es `TRUE`, devuelve P(X<=q), si es `FALSE`, devuelve P(X>q).

Imagina que en una prueba Chi cuadrado has obtenido un valor 3,85 para 1 grado de libertad, por ejemplo en una tabla de 2x2. ¿Cuál será el p-value de esa prueba?

```r
pchisq(3.85, 1, lower.tail = FALSE)
```

```
[1] 0.04974599
```

Diríamos que las diferencias en las proporciones no se esperan por azar.

¿Qué ocurriría si ese mismo valor proviniese de una tabla de 3x2 (dos grados de libertad)?

```r
pchisq(3.85, 2, lower.tail = FALSE)
```

```
[1] 0.1458758
```

Ya no podríamos descartar el papel del azar en esas diferencias. La forma de la distribución Chi cuadrado cambia según el número de grados de libertad,

`qchisq(p, df, lower.tail = TRUE)` : esta función calcula el cuantil de la distribución chi-cuadrado. Dado un nivel de probabilidad, devuelve el valor correspondiente; `p`: nivel de probabilidad (valor entre 0 y 1); `df`: número de grados de libertad; `lower.tail`: comportamiento similar al de `pchisq()`.

Vamos a suponer que queremos calcular el valor crítico de una distribución chi-cuadrado con un nivel de probabilidad de 0.95 y 1 grado de libertad (como en el caso de una tabla 2x2).

```r
valor_critico <- qchisq(0.95, 1, lower.tail = TRUE)
valor_critico
```

```
[1] 3.841459
```