PRACTICA 6: VARIABLES ALEATORIAS CONTINUAS.
Curso 2022/23
Grado en Biología. Universidad de Alcalá

1. En los siguientes ejercicios sea X una variable aleatoria de tipo $N(5, 3)$. Calcula las probabilidades y valores que se indican.

   (a) Calcula $P(X < 4)$, $P(X < 3)$ y, finalmente, $P(X < 0)$. ¿Qué observas?

   (b) Calcula $P(X > 6)$, $P(X > 7)$ y, finalmente, $P(X > 10)$. ¿Ves alguna relación con los valores del anterior apartado? Para entenderlo, puedes hacer un dibujo de la normal con la que estás trabanjando

   (c) $P(4.5 < x < 5.5)$, $P(2 < X < 8)$ y, finalmente, $P(0 < X < 10)$.

   (d) Los valores k_1 y k_2 tales que $P(X < k_1) = 0.90$ y $P(X < k_2) = 0.95$.

   (e) Los valores k_1 y k_2 tales que $P(X > k_1) = 0.1$ y $P(X > k_2) = 0.05$. ¿Ves alguna relación con los valores del anterior apartado?

   (f) $P(X > 2|X \leq 6)$

   Ahora, con la misma variable X, responde a estas preguntas sin usar el ordenador:

   (g) El valor $P(X < 7)$ ¿es mayor o menor que $\frac{1}{2}$?

   (h) El valor $P(X > 8)$ ¿es mayor o menor que $\frac{1}{2}$?

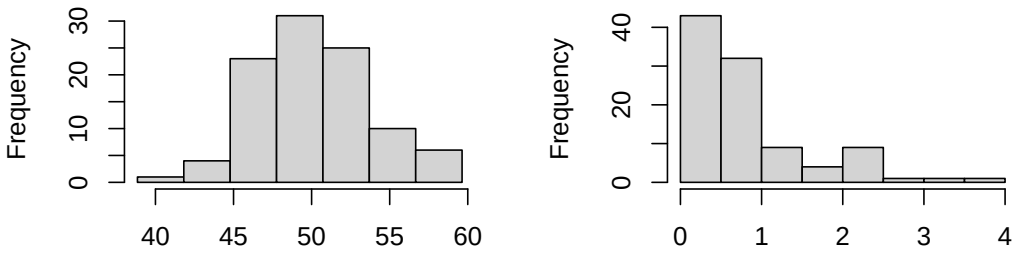
   (i) ¿Cuál de estos dos valores es más grande, $P(X > 4)$ o $P(X < 5.5)$?

   (j) El valor k tal que $P(X > k) = 0.6$ ¿es mayor o menor que 5?

   (k) El valor k tal que $P(X < k) = 0.1$ ¿es mayor o menor que 5?

   Soluciones en la página 3.

2. Estás trabajando con dos variables aleatorias continuas. Tomas una muestra de cada una de ellas y haces un histograma para observar la distribución de cada una de las muestras. ¿A qué tipo de variable aleatoria discreta que conozcas podría corresponder cada uno de ellos?



3. Para comprobar si unos datos provienen o no de una distribuón normal se suelen usar, en primera instancisa, métodos gráficos. Copia, pega y ejecuta varias veces este código en un script de R

```
par(mfrow = c(2,3))
muestra = rnorm(200)
hist(muestra, main = "Muestra normal", cex.main = .8)
boxplot(muestra, main = "Muestra normal", cex.main = .8)
qqnorm(muestra, main = "Muestra normal", cex.main = .8)
qqline(muestra)
muestra2 = rexp(200, rate = .4)
hist(muestra2, main = "Muestra NO normal", cex.main = .8)
boxplot(muestra2, main = "Muestra NO normal", cex.main = .8)
qqnorm(muestra2, main = "Muestra NO normal", cex.main = .8)
qqline(muestra2)
par(mfrow = c(1,1))
```

La tercera columna muestra los qqplots:

- en el eje horizontal, los cuantiles teóricos de una normal
- en el eje horizontal, los cuantiles de la muestra

A medida en que los puntos se alejan de la linea se deja de creer que los datos provienen de una población normalmente distribuida.

Comprueba ahora que para muestras pequeñas, aunque los datos provengan de una población normal, puede ser difícil detectarlo con boxplots, histogramas o qqplots. Copia este código en un script de R y ejecutalo varias veces

```
par(mfrow = c(2,3))
muestra = rnorm(15)
hist(muestra, main = "Muestra normal", cex.main = .8)
boxplot(muestra, main = "Muestra normal", cex.main = .8)
qqnorm(muestra, main = "Muestra normal", cex.main = .8)
qqline(muestra)
muestra2 = rnorm(20)
hist(muestra2, main = "Muestra normal", cex.main = .8)
boxplot(muestra2, main = "Muestra normal", cex.main = .8)
qqnorm(muestra2, main = "Muestra normal", cex.main = .8)
qqline(muestra2)
par(mfrow = c(1,1))
```

4. La concentración (en ppm) de cierta sustancia en el aire se modeliza con una distribución normal de media 55 y desviación típica 10.

   (a) Calcula la probabilidad de que haya más de 60ppm.

   (b) Has tomado una muestra en una zona en la que sabes que la concentración es, al menos, de 75ppm, ¿cuál es la probabilidad de en la muestra haya una concentración de entre 70ppm y 80ppm?

(c) Si se decide seleccionar, de entre todas las muestras posibles, el 10% con mayor concentración. ¿Cuál es la concentración mínima para que una muestra pueda ser seleccionada?

Soluciones en la página 5.

5. Se fumiga una plantación de zanahorias con un producto tóxico. Se sabe que la cantidad de producto que absorbe una zanahoria (en mg) es una variable aleatoria con distribución normal de media 4 y desviación típica 1.5. Se considera que una zanahoria está contaminada si ha absorbido más de 6 mg del producto tóxico:

   (a) Calcula la probabilidad de que una zanahoria seleccionada al azar haya sido contaminada en el proceso de fumigación.

   (b) Si se seleccionan al azar 5 zanahorias, ¿cuál es la probabilidad de que al menos dos de ellas estén contaminadas?

   (c) Si se seleccionan al azar 5 zanahorias, ¿cuál es la probabilidad de que al menos dos de ellas estén contaminadas, sabiendo que hay al menos 1 contaminada?

Soluciones en la página 7.

6. Se considera que una muestra de agua está contamidad por PCBs cuando la concetración está por encima de 10ng/L (umbral inventado). En la zona A de un gran lago la concetración de PCB sigue una distribución $N(8, 2)$ y en la zona B se puede modelizar con una curva exponencial negativa de parámetro $\lambda = 0.15$. Si el 30% de las muestras provienen de la zona A y el 70% de la zona B,

   (a) ¿Cuál es la probabilidad de que, tomada una muestra al azar, esté contaminada? Solución en la página 7.

   (b) Si tienes en las manos una muestra contaminada, ¿cuál es la probabilidad de que provenga de la zona B del lago?

7. Siete mil corredores participan en una carrera local en la que se pueden clasificar para correr la maratón de la ciudad de New York si completan los 42 km en un tiempo inferior a 3 horas y 10 minutos. Solamente 6350 corredores han terminado la carrera. Si los tiempos empleados en completar los 42 km se distribuyen normalmente con una media de 3 horas 40 minutos y una desviación típica de 28 minutos, se pide:

   (a) ¿Cuál es la probabilidad de que un corredor elegido al azar haya empleado menos de 3 horas en completar la carrera?

   (b) ¿Cuántos corredores se han clasificado para la maratón de Nueva York?

   (c) Si se clasifican más de 800 corredores, la organización aplicará como criterio de selección para acudir a la maratón de Nueva York haber conseguido un tiempo que esté incluido dentro del 5% de los mejores tiempos. ¿Qué tiempo máximo se debe emplear en la carrera local para ser seleccionado para la maratón de Nueva York con este criterio?

Soluciones en la página 8.

SOLUCIONES

1. Ejercicio 1, pág. 1

   Todas las respuestas son aproximadas

(a) $P(X < 4)$

```
pnorm(4, mean= 5, sd= 3)
```

```
[1] 0.3694413
```

También sepuede hacer el cálculo "directo": observa el uso del argumento `lower.tail` (su valor por defecto es `FALSE`)

```
pnorm(3, mean= 5, sd= 3)
```

```
[1] 0.2524925
```

```
pnorm(0, mean= 5, sd= 3)
```

```
[1] 0.04779035
```

A medida que nos desplazamos hacia la derecha los valores van siendo más y más pequeños.

(b) En este caso:

```
1 - pnorm(c(6, 7, 10), mean= 5, sd= 3)
```

```
[1] 0.36944134 0.25249254 0.04779035
```

Las respuestas son las mismas del anterior apartado, por la simetría de la curva normal respecto de la media, que está en $\mu = 5$. Por ejemplo, los valores 3 y 7 están a la misma distancia, a izquierda y derecha de μ, respectivamente, y por eso

$$P(X < 3) = P(X > 7) = 0.2525$$

Observa que los siguientes cálculos producen el mismo resultado

```
1 - pnorm(c(4, 3, 0), mean= 5, sd= 3)
```

```
[1] 0.6305587 0.7475075 0.9522096
```

```
pnorm(c(4, 3, 0), mean= 5, sd= 3, lower.tail = TRUE)
```

```
[1] 0.36944134 0.25249254 0.04779035
```

(c) El cálculo es

```
pnorm(5.5, mean= 5, sd= 3)-pnorm(4.5, mean= 5, sd= 3)
```

```
[1] 0.1323677
```

Observa que también se puede calcular así

```
(pnorm(8, mean= 5, sd= 3)-0.5)*2
```

```
[1] 0.6826895
```

```
(pnorm(10, mean= 5, sd= 3)-0.5)*2
```

```
[1] 0.9044193
```

(d) Se trata de los percentiles 90 y 95, y la sintaxis vertorial permite calcularlos de una vez

```
qnorm(c(0.9, 0.95), mean= 5, sd= 3)
```

```
[1] 8.844655 9.934561
```

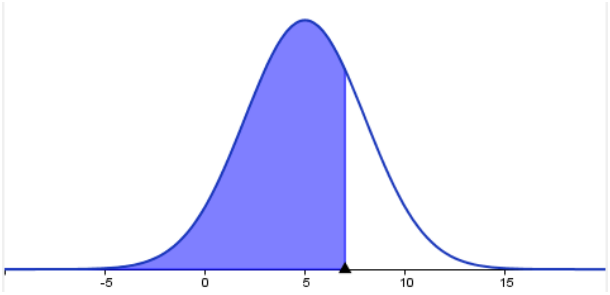
(e) Los mismos valores de k_1 y k_2, de nuevo por la simetría de la curva normal.

(f) $P(X > 2|X \leq 6) = \frac{P((X>2)\cap(X\leq 6))}{p(x\leq 6)}$

```
numerador2 = pnorm(6, mean= 5, sd= 3)-
             pnorm(2, mean= 5, sd= 3)
denominador2 = pnorm(6, mean= 5, sd= 3)
numerador2/denominador2
```

```
[1] 0.7483894
```

(g) Es mayor que 1/2. Se tiene $P(X < 7) = 0.7475$. Este apartado y el siguiente se resuelven teniendo en cuenta que el área de cada una de las dos mitades de la curva es $\frac{1}{2}$, observando si el punto que hemos tomado está a la derecha o la izquierda de μ, y si la probabilidad que calculamos incluye todos los valores mayores o menores. Por ejemplo, para la primera pregunta tenemos que pensar en un dibujo como este:



con el que resulta evidente que la respuesta es mayor que $\frac{1}{2}$.

(h) Es menor que 1/2. Se tiene $P(X > 8) = 0.1587$.

(i) El valor $P(X > 4)$ es más grande. Se tiene $P(X > 4) = 0.6306$, mientras que $P(X < 5.5) = 0.5662$. En este caso la respuesta es fácil de ver porque el valor 4 está más lejos de $\mu = 5$ que 5.5.

(j) Las ideas para este apartado y el siguiente son las mismas, pero ahora tenemos que pensar en los valores del eje x, en vez de pensar en las probabilidades (áreas) que definen esos valores. El valor tiene que ser menor que 5 (de otra manera, la probabilidad sería menor que $\frac{1}{2}$). Se obtiene $k = 4.24$

(k) El valor tiene que ser menor que 5. Se obtiene $k = 1.156$.

2. La de la izquierda podría ser una normal y la de la derecha una exponencial negativa.

3. Ejercicio 4, pág. 2 La cantidad de contaminante X sigue una $N(\mu = 55, \sigma = 10)$

   (a) La probabilidad pedida es $P(X > 60)$, que se puede calcular de varias formas diferentes con la función `pnorm`. Como se trata de la cola derecha de la distribución y `pnorm` calcula probabilidades con al cola izquierda (es decir, cosas como $P(X < 60)$) se puede calcular la probabilidad del complementario: $P(X > 60) = 1 - P(X > 60)$

```r
1-pnorm(60, mean = 55, sd = 10)
```

```
[1] 0.3085375
```

si quieres hacer saber a `qnorm` que en realidad te interesa la cola derecha, puedes utilizar directamente

```r
pnorm(60, mean = 55, sd = 10, lower.tail = FALSE)
```

```
[1] 0.3085375
```

donde `lower.tail = FALSE` le indica a la función `qnorm` que te interesan las probabilidades a la derecha de $X = 60$, y no a su izquierda.

(b) Se trata de una probabilidad condicionada: $P(70 < X < 80|X \geq 75)$. Con la definición de probabilidad condicionada

$$P(70 < X < 80|X \geq 75) = \frac{P((70 < X < 80) \cap (X > 75))}{P(X \geq 75)}$$

los números que están entre 70 y 80 y, a la vez, son mayores que 75 son los números que están entre 75 y 80. Por tanto, la probabilidad pedida es

```r
numerador = pnorm(80, mean = 55, sd = 10) -
            pnorm(75, mean = 55, sd = 10)
denominador = 1-pnorm(75, mean = 55, sd = 10)
numerador/denominador
```

```
[1] 0.7270493
```

Se usa esa sintaxis (numerador y denominador) para que quepa todo en la página.

(c) En este caso queremos calcular el valor de la concentración de contaminante que deja por encima de sí el 10% de las muestras (y, por debajo, el 90%). Esa cantidad es el percentil 90 y se calcula así:

```r
qnorm(0.9, mean = 55, sd = 10)
```

```
[1] 67.81552
```

Observa que, de nuevo, existe la posibilidad de usar la cola derecha de la distribución:

```r
qnorm(0.1, mean = 55, sd = 10, lower.tail = FALSE)
```

```
[1] 67.81552
```

4. Ejercicio , 6.

(a) X = cantidad de contaminante en una zanahoria. $X \sim N(4, 1.5)$; hay que caluclar $P(X > 6)$

```r
(p = pnorm(6, mean = 4, sd = 1.5, lower.tail = FALSE))
```

```
## [1] 0.09121122
```

(b) Ahora la variable Y = "n zahanorias contaminadas de las 5 inspeccionadas" es una variable binomial, $Y \sim B(n = 5, prob = 0.0912)$. Se pide $P(Y \geq 1)$

```
pbinom(1, size = 5, prob = p, lower.tail = FALSE)
```

```
## [1] 0.06903121
```

(c) Se pide $P(Y \geq 2|Y \geq 1)$

```
sum(dbinom(2:5, size = 5, prob = p))/sum(dbinom(1:5, size = 5, prob = p))
```

```
## [1] 0.1816086
```

.

5. Pág. 3

(a) La probabilidad de que la muestra esté contaminada si viene de la zona A es

```
(contamina_A = pnorm(10, mean = 8, sd = 2, lower.tail = F))
```

```
[1] 0.1586553
```

y, siviene de la zona B es

```
(contamina_B = pexp(10, rate = 0.15, lower.tail = F))
```

```
[1] 0.2231302
```

de donde, por el teorema de la Probabilidad Total se tiene

$$\begin{aligned} P(contaminada) &= P((contaminada \cap A) \cup (contaminada \cap B)) \\ &= P(contaminada \cap A) + P(contaminada \cap B) \\ &= P(A)P(contaminada|A) + P(B)P(contaminada|B) \end{aligned}$$

```
(resultado1 = 0.3 * contamina_A + 0.7*contamina_B)
```

```
[1] 0.2037877
```

redondeando a cuatro cifras significativas da

```
signif(resultado1, digits = 4)
```

```
[1] 0.2038
```

(b) Ahora nos pide $P(B|contaminada)$, es decir,

$$P(B|contaminada) = \frac{P(B \cap contaminada)}{P(contaminada)} = \frac{P(B)P(contaminada|B)}{P(contaminada)}$$

lo que es una apliación del Teorema de Bayes, y que podemos resolver con la información obtenida en el apartado anterior:

```
resultado2 = 0.7*contamina_B / (0.3 * contamina_A + 0.7*contamina_B)

signif(resultado2, digits = 4)

[1] 0.7664
```

6. pág. 3 La variable X = "tiempo empleado en completar el recorrido" es una normal con media 3 horas y 40 minutos y desviación típica 28 minutos. Lo primero es traducir todo a minutos (o a horas) para que las unidades sean las mismas; usaremos minutos: $X \sim N(220, 28)$

   (a) $P(X < 180)$

```
pnorm(180, mean = 220, sd = 28)

[1] 0.07656373
```

   (b) cada corredor debe emplear menos de 190 minutos, por lo que la probabilidad de que necesite, como mucho, ese tiempo es $P(X \leq 190)$

```
(prob190 = pnorm(190, mean = 220, sd = 28))

[1] 0.1419884
```

   Nos interesa la variable Y = "número de corredores que termina la maratón en menos de 190 minutos de entre los 6350 que acabaron la carrera". Si lo piensas, Y es una variable binomial (cada corredor sólo puede llegar a tiempo o no y corren de forma independiente) de parámetros $n = 6350$ y $p = P(X \leq 190)$. El número esperado de éxitos es np, es decir

```
6350 * prob190

[1] 901.6263
```

   es decir,

```
floor(6350 * prob190)

[1] 901
```

   (c) Los mejores tiempos son los más rápidos, es decir, se trata del pecentil 5 (en minutos):

```
qnorm(0.05, mean = 220, sd= 28)

[1] 173.9441
```