Master en Big Data. Fundamentos matemáticos del análisis de datos.

Sesión 3. Poblaciones, muestras y probabilidad.

Fernando San Segundo

Curso 2019-20. Última actualización: 2019-09-01

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI
DEPARTAMENTO DE MATEMÁTICA APLICADA

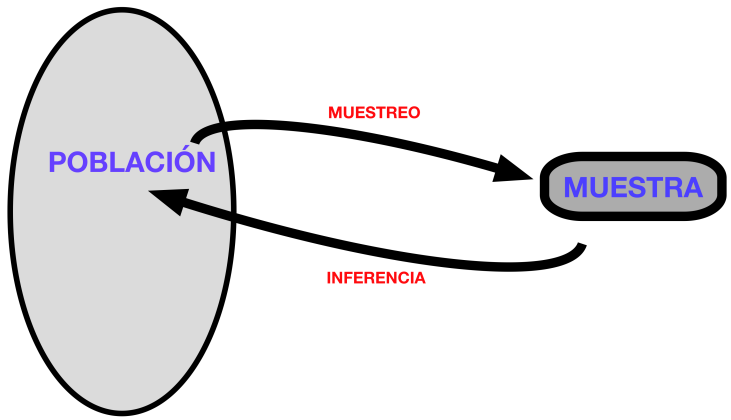
1. Población y muestra.
2. Probabilidad básica.
3. Axiomas de la Probabilidad.
4. Probabilidad condicionada e independencia.
5. Regla de Bayes.
6. Tablas de Contingencia.

Sección 1

Población y muestra.

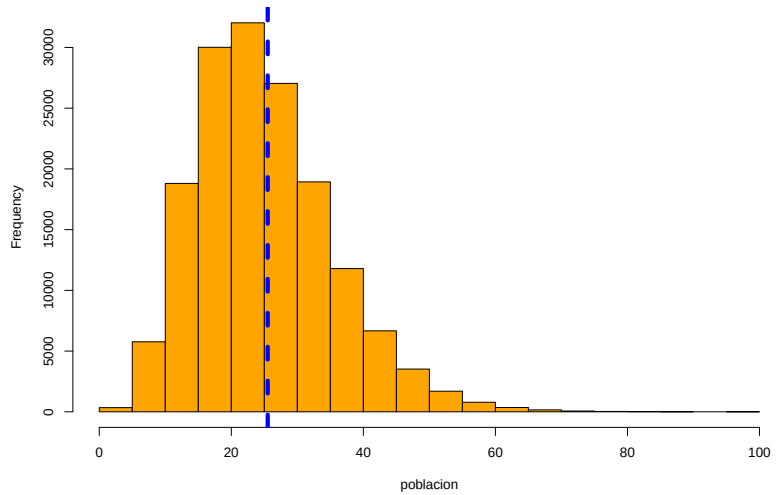
Inferencia Estadística.

- El objetivo central de la Estadística es obtener información fiable sobre las características de una **población** a partir de **muestras**. Ese término significa aquí un conjunto de entidades individuales (individuos), no necesariamente seres vivos. La poblacion pueden ser los vehículos matriculados en 2015 o las órdenes de compra recibidas por una empresa cierto mes o las especies de ave que visitan un comedero en Costa Rica en los últimos 10 años, etc.
- Muchas veces estudiar toda la población es demasiado difícil, indeseable o imposible. Entonces surge la pregunta de si podemos usar las muestras para *inferir*, o *predecir* las características de la población. ¿Hasta qué punto los datos de la muestra son *representativos*de la población?
- La *Inferencia Estadística* es el núcleo de la Estadística porque da sentido a estas preguntas, las formaliza y responde.



Poblaciones y muestras aleatorias simples con vectores con R.

- Al estudiar una población nos interesan determinadas características individuaes, que pueden cambiar de un individuo a otro y que constituyen las *variables de interés*. Cuando tomamos una muestra obtenemos los valores de esas variables en algunos individuos de la población.
- Para que la muestra sea representativa lo mejor es que sea una **muestra aleatoria simple**: elegimos a los individuos al azar y con remplazamiento (podemos incluri al mismo individuo más de una vez en la muestra).
- Para entenderlo mejor haremos un experimento con R. En este caso vamos a suponer una población de $N = 158000$ individuos. Por ejemplo, los viajeros que pasan por un aeropuerto en un día y sea la variable de interés su edad. El código de esta sesión construye un vector `poblacion` con las edades de los viajeros. Vamos a hacer una pequeña trampa y mostraremos el histograma de las edades. La línea de puntos indica *la media poblacional de la edad.* ¿Cuál crees que es?



Medias muestrales

- Ese es justo el tipo de preguntas que esperamos que responda la Estadística. Aunque en este caso disponemos del vector completo de edades debes tener claro que en los problemas reales no será así. Así que recurrimos a las muestras aleatorias (con remplazamiento), en inglés *random sample (with replacement)*. Por ejemplo, de tamaño 20. En R construimos una de esas muestras así:

```
n = 20
(muestra = sample(poblacion, n, replace = TRUE))
```

```
##  [1] 20 10 18 39 36 29 55 25 30 40 18 44 12 30 18 15 12 22 10 19
```

Esas son las 20 edades $x_1, \ldots, x_{20}$ de los viajeros de la muestra. Para *estimar* la edad media de *todos los viajeros* a partir de estos valores calcularíamos la **media muestral**.

$$\bar{X} = \frac{x_1 + \cdots + x_{20}}{n} = \frac{20 + 10 + \cdots + 19}{20} \approx 25.1 = \text{mean(muestra) en R}$$

- Naturalmente, si tomas otra muestra, su media muestral puede ser otra:

```
(muestra2 = sample(poblacion, n, replace = TRUE))
```

```
##  [1] 16 28 38 18 28 46 18 32 27 16 15 23 18 30 48 23 30 14 23 31
```

```
mean(muestra2)
```

```
## [1] 26.1
```

Muestras buenas y malas.

- Hemos visto que cada muestra produce una media muestral y que esas medias muestrales pueden ser distintas. ¿Cuántas muestras distintas hay? Hay una cantidad inimaginablemente grande:

$$158000^{20} = 9.4003005 \times 10^{103}$$

  Para ponerlo en perspectiva, se estima que en el universo hay menos de 10^{40} estrellas. Esta cantidad enorme de muestras, de las que solo hemos visto 2, forman lo que se llama el **espacio muestral** (de tamaño $n = 20$) de este problema.

- Entre esas muestras hay muestras *buenas* y muestras *malas*. ¿Qué queremos decir con esto? Para seguir con nuestro experimento vamos a ordenar *la población completa* por edad y tomemos los 20 primeros valores:

```
(muestra3 = sort(poblacion)[1:20])
```

```
##  [1] 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
```

  Hemos llamado `muestra3` a ese vector porque es una más de las muchísimas muestras posibles que podríamos haber obtenido al elegir al azar 20 viajeros. Y si usáramos esta muestra para estimar la media de la población obtendríamos

```
mean(muestra3)
```

```
## [1] 2.5
```

  Eso es lo que llamamos una *muestra mala*, poco representativa.

La distribución de las medias muestrales.

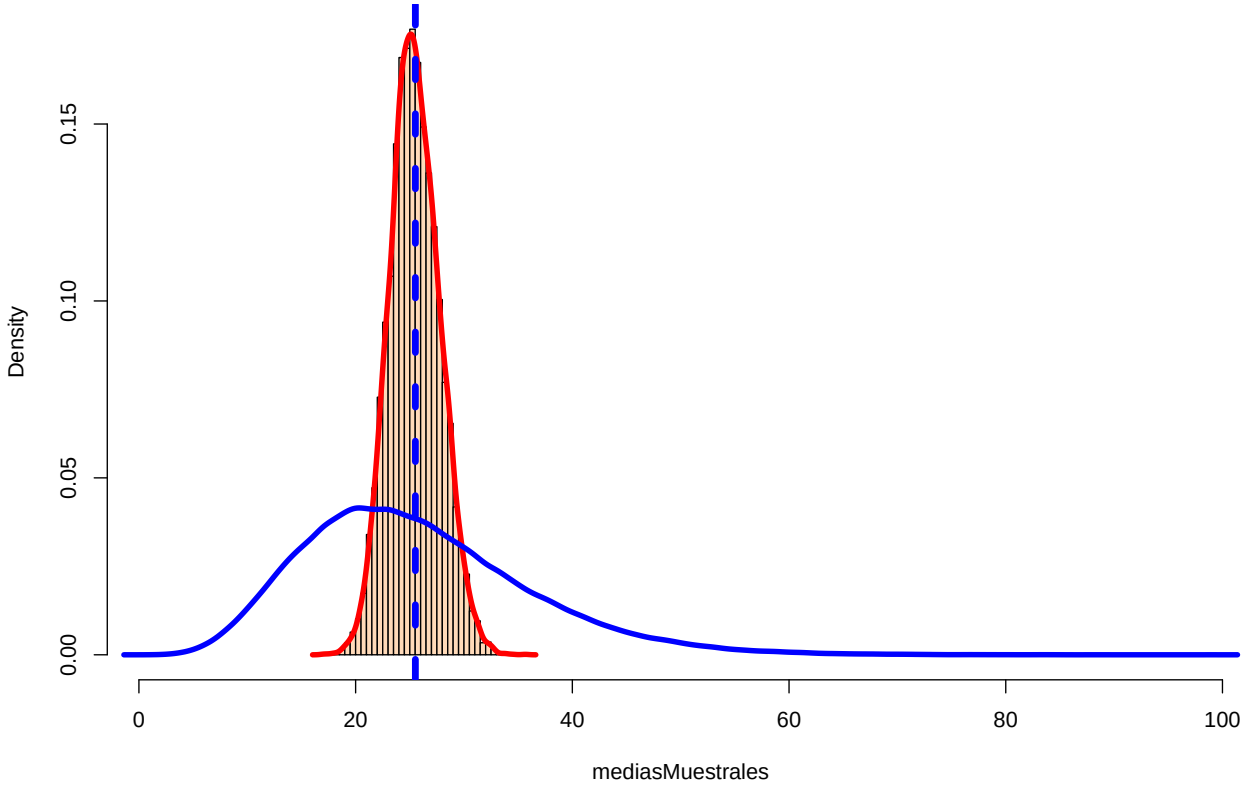
- La última muestra que hemos examinado era muy poco representativa. Pero la pregunta esencial para la estadística es ¿cuál es la relación entre muestras buenas y malas? Al elegir una muestra al azar, ¿cómo de probable es que nos toque una muestra tan mala en lugar de una buena?
- Podemos hacer otro pequeño experimento para explorar el espacio muestral. No podemos repasar todas las muestras una por una para clasificarlas en buenas o malas (eso sería demasiado incluso para R) pero podemos tomar *muchas* muestras aleatorias (pongamos $k = 10000$) y ver como de buenas o malas son (hacemos una *muestra de muestras*). En R es muy fácil hacer esto usando la función `replicate`:

```
k = 10000
# replicate repite k veces los comandos entre llaves y guarda el resultado
# del último comando en el vector mediasMuestrales
mediasMuestrales = replicate(k, {
  muestra = sample(poblacion, n, replace = TRUE)
  mean(muestra)
})
head(mediasMuestrales, 10)
```

```
##  [1] 25.00 28.70 24.85 26.05 25.75 27.15 28.05 25.15 28.40 28.40
```

Se muestran las primeras 10 de las 10000 medias muestrales que hemos obtenido.

- En lugar de examinar una a una esas 10000 medias muestrales vamos a representarlas en un histograma y una curva de densidad. Además, aprovechándonos de que en este caso tenemos acceso a la población completa hemos añadido su curva de densidad:



- Este es posiblemente **el gráfico más importante del curso**. Fíjate en tres cosas:
  - La media de las medias muestrales es la media de la población.
  - Prácticamente no hay *muestras malas*. Es *extremadamente improbable* que una muestra elegida al azar sea muy mala.
  - La distribución de las medias muestrales tiene forma de campana (y es muy estrecha).

  Para entender bien estas ideas *necesitaremos aprender más sobre Probabilidad*.

Otra población, mismos resultados.

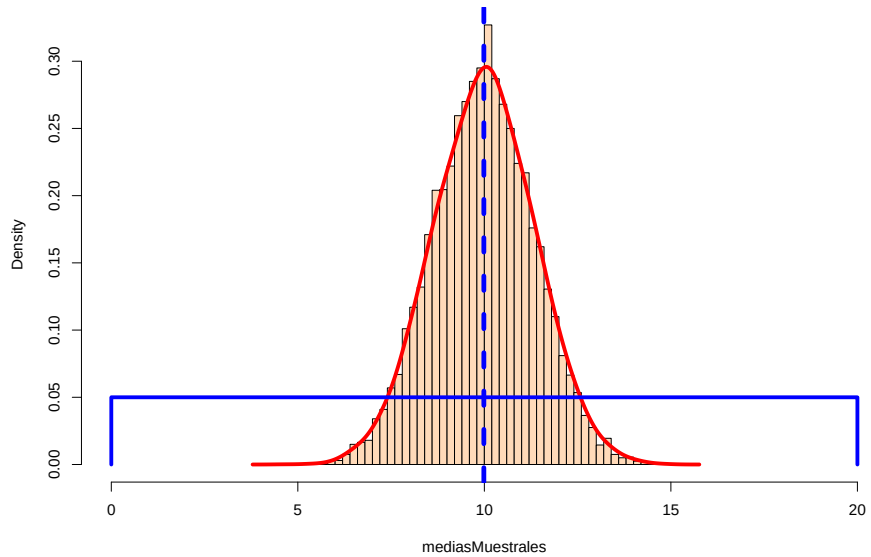
- Pero antes de lanzarnos a la probabilidad vamos a asegurarnos de algo. Puede que te preguntes si la población con la que hemos empezado tenía algo especial. Probemos con otra muy distinta. La población la forman 20000 números elegidos al azar del 0 al 20, siendo todos los valores igual de probables (su curva de densidad es horizontal).

```
poblacion = sample(0:20, 20000, replace = TRUE)
```

Y ahora repetimos el proceso de construcción de medias muestrales usando replicate

```
k = 10000
mediasMuestrales = replicate(k, {
  muestra = sample(poblacion, n, replace = TRUE)
  mean(muestra)
})
```

El gráfico del resultado muestra el mismo comportamiento de las medias muestrales, lo que se conoce como **Teorema Central del Límite**:



Sección 2

Probabilidad básica.

- Para entender resultados como el Teorema Central del Límite tenemos que aprender el mínimo vocabulario necesario para poder hablar con precisión sobre la Probabilidad.
- Lo primero de lo que hay que ser conscientes es de que nuestra intuición en materia de probabilidad suele ser muy pobre. Vamos a empezar usando ejemplos de juegos de azar (dados, naipes, etc.) para poder desarrollar el lenguaje, igual que sucedió históricamente.

Experimentos del Caballero de Méré.

- ¿Qué es más probable?

  (a) obtener al menos un seis en cuatro tiradas de un dado, o
  (b) obtener al menos un seis doble en 24 tiradas de dos dados?

- Los jugadores que en el siglo XVIII se planteaban esta pregunta pensaban así:

  (a) La probabilidad de obtener un seis en cada tirada es $\frac{1}{6}$. Por lo tanto, en cuatro tiradas es $\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{2}{3}$.

  (b) La probabilidad de un doble seis en cada tirada de dos dados es $\frac{1}{36}$, (hay 36 resultados distintos) y todos aparecen con la misma frecuencia. Por lo tanto, en veinticuatro tiradas será $\frac{24}{36}=\frac{2}{3}$.

  Así que en principio ambas apuestas parecen iguales,

- Vamos a usar R para jugar a estos dos juegos sin tener que jugarnos el dinero. Descarga este fichero de código y ejecútalo.

La paradoja del cumpleaños.

- Otro experimento que puede sevrir para afianzar la idea de que la probabilidad es poco intuitiva. Si en una sala hay 1000 personas entonces es seguro que hay dos que cumplen años el mismo día. De hecho basta con que haya 367 personas. Si hay menos de ese número, la probabilidad de que dos cumpleaños coincidan disminuye. ¿Cuál es el *menor número de personas* que nos garantiza una probabiidad mayor del 50 % de coincidencia?
- Usemos R para averiguar ese número. Repite el experimento varias veces para convencerte..

```
n = 366 # Número de personas en la sala

# Vamos a repetir el experimento N veces (N salas de n personas)
N = 10000
pruebas = replicate(N, {
  fechas = sort(sample(1:366, n, replace=TRUE))
  max(table(fechas)) # si el máximo es mayor que 1 es que 2 fechas coinciden
})
mean(pruebas > 1) # ¿qué proporción de salas tienen coincidencias?
```

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

Regla de Laplace.

Tablas de frecuencia relativas, probabilidades. Modelos, primera visita.

Sección 3

Axiomas de la Probabilidad.

Sección 4

Probabilidad condicionada e independencia.

Sección 5

Regla de Bayes.

Sección 6

Tablas de Contingencia.

Referencias para la sesión

Enlaces

- Código de esta sesión

Bibliografía