Bioestadística

Grado en Medicina USC

Apuntes

CAPÍTULO 1: ESTADÍSTICA DESCRIPTIVA

INTRODUCCIÓN

Estadística:

La estadística **descriptiva** se encarga de resumir y describir un conjunto de datos para su comprensión.
La estadística **inferencial** se encarga de extraer conclusiones sobre una población a partir del estudio de una muestra mediante técnicas probabilísticas.

Individuo: elemento bien definido que presenta la característica a estudiar. Pueden ser personas, animales, objetos o grupos. *Por ejemplo: un alumno de la clase*

Muestra: subconjunto representativo de individuos de la población extraído por algún método válido de muestreo. Intentaremos extrapolar los resultados a la población*. Por ejemplo: 20 alumnos cogidos al azar.*

Población: el conjunto de individuos sobre los que queremos hacer el estudio. *Por ejemplo: la clase entera.*

Estudios experimentales: hay una intervención durante el estudio manipulando el factor de estudio (exposición, tratamiento…) y aleatorizando los sujetos en base a ese factor.

Estudios observacionales: el estudio no involucra ninguna intervención. El investigador selecciona individuos de la población de interés, mide las variables de interés y estudia después posibles asociaciones durante el análisis de datos.

Variable: característica observable que cambia entre los elementos de una población. Característica que puede ser medida y que puede adoptar más de un valor.

Variables cualitativas o categóricas: son aquellas que recogen una característica que no se puede expresar mediante una cantidad, aunque sí con una categoría.
Por ejemplo: sexo (hombre/mujer), color de ojos (azules/marrones/verdes...), etc.

- **Nominales:** tienen un conjunto de categorías sin ningún tipo de jerarquía.
- **Ordinales:** tienen un conjunto de categorías con una jerarquía u orden.

Variables cuantitativas o numéricas: son aquellas variables que recogen como información una cantidad numérica de lo que se está observando.
Por ejemplo: edad, peso, tensión arterial, número de hijos, etc.

- **Discretas:** tienen un conjunto finito de valores por ejemplo si únicamente toman números enteros.
- **Continuas:** el conjunto de posibles valores entre dos números fijos es infinito.

PRESENTACIÓN DE DATOS: TABLAS DE FRECUENCIAS

Frecuencia absoluta (n_i): número de veces que se repite un determinado valor. Puede ser individual o acumulada (N_i) (en les variables nominales no tiene sentido acumular).

Frecuencia relativa (f_i): proporción que representa las apariciones de ese valor respecto al total $f_i = \frac{n_i}{n}$

También puede ser individual o acumulada (F_i). Si se multiplica por 100 nos dará un porcentaje.

Si la **variable es continua** o tiene muchos valores se puede dividir en $\sqrt{n}$ intervalos. El **rango** (r) es la diferencia entre el valor superior e inferior del intervalo y la **amplitud** (a) es $r/\sqrt{n}$. La **marca de clase** (c) es punto central del intervalo.

Ejemplo: clase a la que va el alumno B, B, A, C, B, B, B, C, B, A.

x_i	n_i	f_i	P_i	N_i	F_i	PA_i
A						
B						
C						

Ejemplo: numérico continuo, notas de un examen: 3,4; 5,3; 6,8; 9,1; 7,2; 8,3; 1,9; 5,1; 4,4; 3,2.

x_i	n_i	N_i	f_i	F_i	r_i	c_i	a_i
[0-2)							
[2-4)							
[4-6)							
[6-8)							
[8-10)							

En R Studio:

```
counts:
Hijos
0 1 2
2 5 3
```

```
percentages:
Hijos
0  1  2
20 50 30
```

```
$breaks
0 2 4 6 8 10
$counts
1 2 3 2 2
```

MEDIDAS DE CENTRALIZACIÓN

Media aritmética:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\sum x_i \cdot n_i}{n}$$

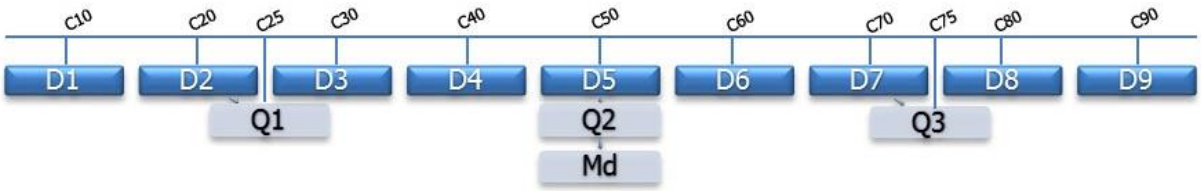
Mediana:

$$Pos_{Md} = \frac{n+1}{2}$$

Moda:

MEDIDAS DE POSICIÓN

$$Pos_{IP} = \frac{x}{100} \cdot (n + 1)$$



Ejemplo R Studio:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
1.000   6.000   7.000   7.311   9.000  10.000      10
```

MEDIDAS DE DISPERSIÓN

Rango, Recorrido o Amplitud:

$$R = V_{máx} - V_{mín}$$

Rango intercuartílico (Rq):

$$R_{IQ} = Q_3 - Q_1$$

Se considera un dato atípico si supera en 1,5 veces el Rq por encima del Q3 o por debajo del Q1

Varianza:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum(x_i - \bar{x})^2 \cdot n_i}{n-1}$$

La desviación estándar/típica:

$$s = \sqrt{s^2}$$

Coeficiente de Variación de Pearson:

$$CV = \frac{s}{\bar{x}} \cdot 100$$

MEDIDAS DE FORMA

Asimetría:

As>0 asimetría positiva

As=0 simetría

As<0 asimetría negativa

Curtosis:

C>0 leptocúrtica

C=0 mesocúrtica

C<0 platicúrtica

Ejemplo R Studio:

var	sd	IQR	cv	skew	kurtosis
4.000	2.00	3.00	0.274	-0,854	0.021

REPRESENTACIONES GRÁFICAS

Diagrama de sectores

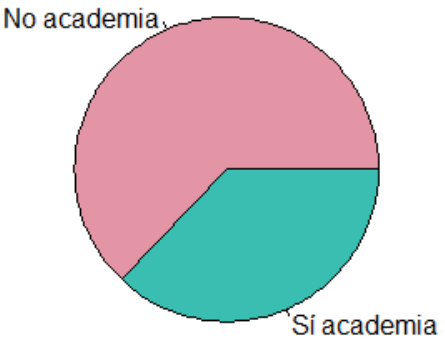
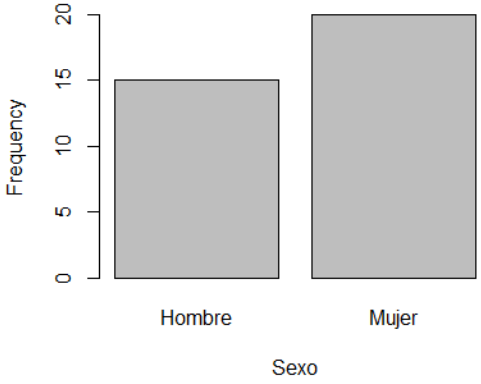
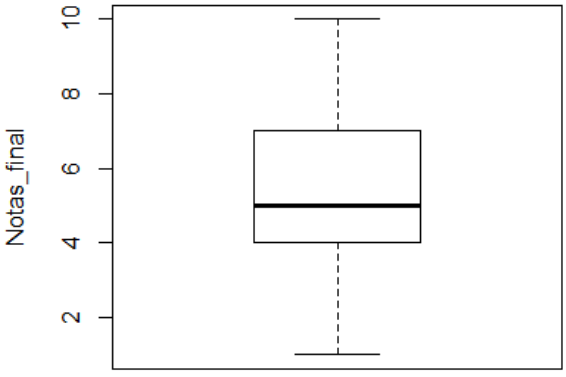


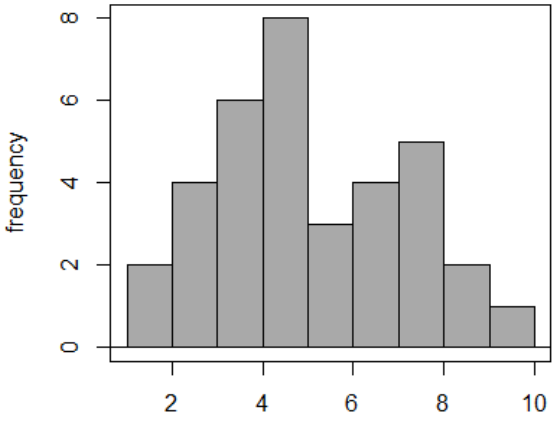
Diagrama de barras



Boxplot



Histograma



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA, ANÁLISE MATEMÁTICA E OPTIMIZACIÓN

Capítulo 1: Estadística descriptiva

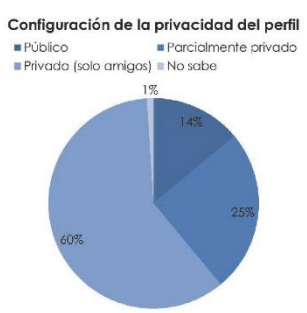
1. El Ministerio de Sanidad, Servicios Sociales e Igualdad ha publicado recientemente el Barómetro Sanitario nacional, correspondiente a 2011.

   (a) Según los resultados del estudio, de los 7757 encuestados, 2487 afirman ser fumadores en la actualidad, 5262 afirman ser no fumadores y el resto no contesta. Construye la tabla de frecuencias correspondiente y realiza una gráfica representativa de la distribución de la muestra según el hábito de fumar.

   (b) A aquellos que habían declarado ser fumadores se les preguntó a continuación como habían influido en su consumo las medidas de la nueva Ley del tabaco. A continuación se muestra una tabla incompleta que resume las respuestas. Completa la tabla y realiza una gráfica representativa.

Influencia de la Ley del tabaco	n_i	f_i
Fuma menos que antes		
Fuma más que antes	125	
No ha influido en el consumo		0.68275
No contesta	28	

2. Un estudio llevado a cabo por el Pew Research Center's Internet & American Life Project (http://www.pewinternet.org) tiene como objetivo analizar la actitud de los jóvenes en EEUU ante las redes sociales y su configuración de la privacidad. Para ello se ha llevado a cabo una encuesta entre usuarios de Facebook. A continuación se muestran dos gráficas con datos de dicho estudio.



   (a) ¿Qué gráficas aparecen representadas? ¿A qué tipo de variables hacen referencia?

   (b) ¿Cuál es el tamaño muestral? ¿Cuántos encuestados tienen su perfil parcialmente privado? ¿Qué porcentaje de usuarios comparte la misma información con sus padres y amigos?

   (c) Supongamos que les preguntamos a los encuestados cuántos amigos tienen en Facebook. ¿Qué tipo de gráfico crees que deberías utilizar para resumir esa información?

3. Se muestran a continuación los tiempos (en minutos) de 15 participantes en la última etapa de la Vuelta a España (2014), correspondiente a la contrarreloj celebrada en Santiago de Compostela el 14 de septiembre de 2014.

   12.33, 12.34, 12.43, 12.66, 13.05, 11.95, 12.56,
   12.15, 11.33, 11.96, 13.83, 13.56, 11.83, 13.08, 12.81.

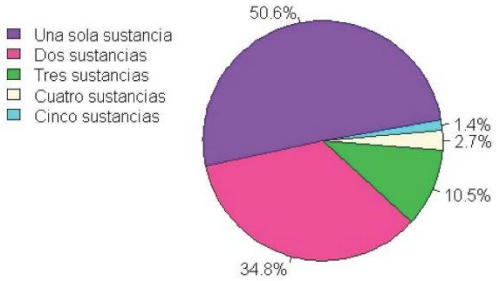
   (a) Contruye la tabla de frecuencias correspondiente.

   (b) Realiza una gráfica representativa e interprétala.

   (c) Calcula el tiempo medio de la etapa.

   (d) Representa el diagrama de cajas correspondiente.

4. De los datos recogidos en la "Encuesta domiciliaria sobre alcohol y drogas en España 2009/2010" (Delegación del Gobierno para el Plan Nacional Sobre Drogas. Ministerio De Sanidad, Política Social e Igualdad) se deduce que el policonsumo de drogas (legales e ilegales) es un patrón de consumo cada vez más prevalente en España y en Europa. Analiza el siguiente gráfico sobre el porcentaje de consumidores que han consumido una o más sustancias en el último año.



   (a) ¿Qué variable se resume en el diagrama de sectores? ¿De qué tipo es?

   (b) Calcula la moda, la mediana y el cuantil 0.9.

   (c) ¿Cuál es el número medio de sustancias consumidas por individuo?

   (d) Supongamos que los datos de la gráfica han sido obtenidos a partir de una muestra formada por 1000 individuos. Si se añaden 500 consumidores más al estudio cuyo consumo medio es de 2.5 sustancias, ¿cuál es el número medio de sustancias consumidas para el total de 1500 personas?

5. Los siguientes datos corresponden a los pesos (en kg.) de 12 niñas de 5 años de edad.

15.0	17.3	18.3	21.9	13.8	20.8
17.5	19.7	15.1	26.7	20.4	16.4

   (a) Calcula el peso medio y el peso mediano. Calcula la varianza y la desviación típica.

   (b) Calcula Q_1, Q_3 y el cuantil 0.9. Representa el boxplot de los datos.

   Puedes ver los patrones de crecimiento infantil de la OMS en http://www.who.int/childgrowth/es/

6. El volumen corpuscular medio (VCM) es uno de los parámetros calculados en un examen de conteo sanguíneo completo. El VCM indica el tamaño de los glóbulos rojos y se mide en fentolitros. A continuación se muestran los valores de VCM de 19 pacientes que se han sometido a un examen de conteo sanguíneo completo.

   83, 77, 82, 84, 85, 92, 92, 93, 91, 86, 89, 109, 81, 79, 81, 88, 110, 90, 80.

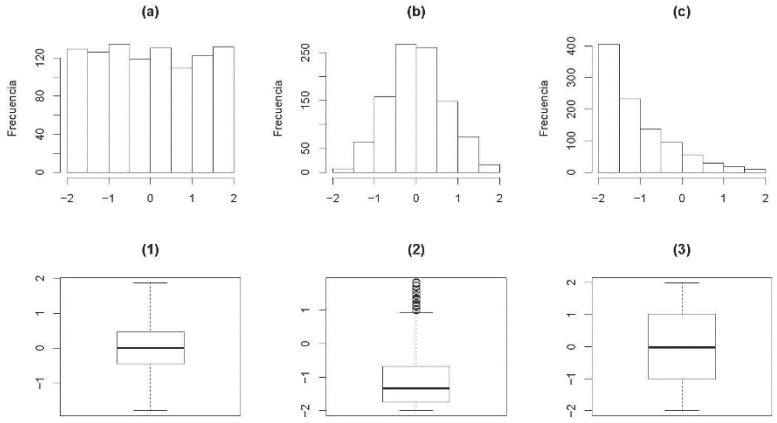
   (a) Construye la tabla de frecuencias y representa el histograma correspondiente.

   (b) Dibuja el boxplot de estos datos.

7. En un centro de salud se registra el peso (en kg.) y la altura (en cm.) de los pacientes que han acudido a consulta el último mes. La siguiente tabla presenta un resumen de las respuestas proporcionadas por dicho grupo de pacientes. ¿Qué medidas presentan mayor variabilidad?

Variable	Media	Desviación típica
Peso	65.4	12.2
Altura	170.5	9.42

8. A continuación aparecen representados los histogramas y diagramas de cajas de tres conjuntos de datos distintos. Empareja cada histograma con el diagrama de cajas que le corresponde.



9. Se ha registrado la temperatura de 9 pacientes. La temperatura mínima registrada ha sido 36ºC y la temperatura máxima registrada ha sido 39.2ºC. La mediana es 37.2ºC.

   (a) Si eliminamos la observación correspondiente a la temperatura mínima registrada y la sustituimos por una nueva temperatura de 35.9ºC, ¿cuál es la mediana del nuevo conjunto de datos?

   (b) Si eliminamos la observación correspondiente a la temperatura máxima registrada y la sustituimos por una nueva temperatura, ¿cuál es la mediana del nuevo conjunto de datos?

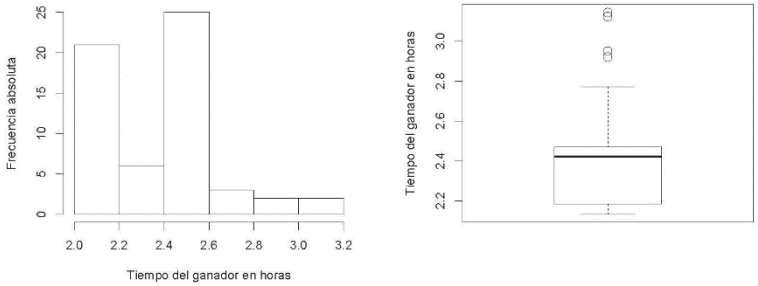
10. Para cada uno de los apartados a continuación, compara los conjuntos de datos (1) y (2) en términos de media y desviación típica. No es necesario que calcules los valores exactos, simplemente discute en qué casos son iguales o de qué manera difieren.

    (a) (1) 0, 2, 4, 6, 8, 10.
    (2) 20, 22, 24, 26, 28, 30.

    (b) (1) 100, 200, 300, 400, 500.
    (2) 0, 50, 300, 550, 600.

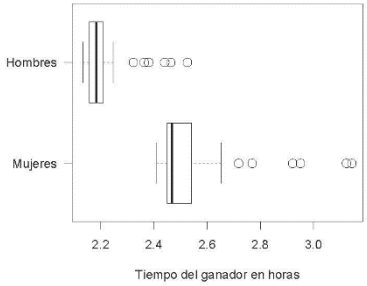
    (c) (1) 0, 2, 4, 6, 8, 10.
    (2) 0, 6, 12, 18, 24, 30.

11. El histograma y el diagrama de cajas que se muestran a continuación muestran la distribución de los tiempos de los ganadores (categoría masculina y femenina) del maratón de Nueva York entre 1970 y 1999.



    (a) ¿Qué aspectos de la distribución de los tiempos de los ganadores se ven reflejados en el histograma pero son menos evidentes en el boxplot? ¿Qué aspectos de la distribución de los tiempos de los ganadores se deducen del boxplot pero son menos evidentes en el histograma? ¿Por qué la que la distribución de los tiempos es bimodal?

    (b) A la vista del diagrama de cajas que aparece a continuación, compara la distribución de los tiempos de los ganadores en función del sexo.



En las primeras ediciones el maratón se celebraba dando varias vueltas alrededor de un circuito por Central Park, pero la carrera fue adquiriendo gran popularidad y en 1976 se cambió el trazado. Los datos atípicos corresponden a los tiempos de 1970 hasta 1975.

Este dosier está hecho para seguir la clase de prueba.

Si te apuntas al curso te enviaremos por correo el dosier entero con todos los temas que faltan, ejercicios y exámenes de años anteriores

Más información en:

www.estadistix.com

Y si tienes cualquier consulta,
escríbenos un whatsapp al 644310902