**Estadística descriptiva**

**Estadística descriptiva**

* Describe una muestra sin extrapolar datos al resto de la población.
* **Población:** conjunto de individuos que constituyen el objetivo del estudio
* **Variable:** rasgo medible de los elementos de la población.
  + Variable cualitativas: Valores no numéricos
    - Nominales: Los valores no tienen orden (sexo, color de ojos)
    - Ordinales: Orde subyacente entre categorías (nivel de estudios)
  + Variables cuantitativas: Toman valores numéricos
    - Discretas: Valores enteros y separados
    - Continuas: Valores reales (incluye racionales)
* **Muestra:** subconjunto de la población para el que se conocen los valores de las variables a analizar. Debe ser representativa.
* **Individuo:** cada uno de los elementos de la muestra

**Extracción aleatoria**

* Permite pasar de población a muestra. Consiste en el cálculo de probabilidades.

**Inferencia estadística**

* Permite pasar de muestra a población
* Consiste en obtener conclusiones sobre la población a partir de la muestra.

**Tablas de frecuencias**

* Utilizadas para representar la información de una muestra tamaño n
* **Clase (ci):** Cada uno de los valores que puede tomar una variable.
* **Frecuencia absoluta (ni):** Número de individuos en la clase ci. 0<=ni<=n
* **Frecuencia relativa (fi):** = 0<=fi<=1
* **Frecuencia absoluta acumulada (Ni):** Número de individuos en ci o en valores anteriores. Nk = n
* **Frecuencia relativa acumulada (Fi)**: . Fk = 1
* Para unha variable cualitativa nominal non se inclúen valores absolutos.
* Para unha variable cuantitativa continua defínense **intervalos**.
  + Denominados intervalos de clase: [ei, ei+1)[[1]](#footnote-0) sendo o punto medio a marca de clase (ci)
  + Se suelen tomar como número de intervalos aproximado.
  + Suelen ser todos de igual longitud y contiguos. La longitud se calcula aproximando hacia arriba para evitar excluir datos.

**Ejemplos tabla de frecuencia**

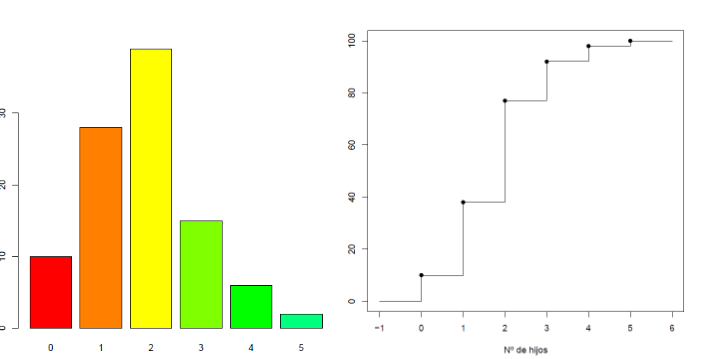
| ei | ni | fi | Ni | Fi |
| --- | --- | --- | --- | --- |
| [2.5, 3.1) | 1 | 0.1 | 1 | 0.1 |
| [3.1, 3.7) | 4 | 0.4 | 5 | 0.5 |
| [3.7, 4.3) | 5 | 0.5 | 10 | 1.0 |

Variable cualitativa ordinal Variable cuantitativa continua

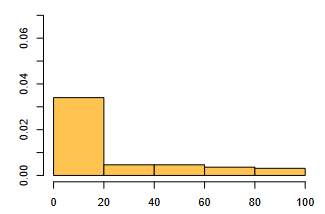
| ci | ni | fi | Ni | Fi |
| --- | --- | --- | --- | --- |
| Baixa | 2 | 0.2 | 2 | 0.2 |
| Media | 5 | 0.5 | 7 | 0.7 |
| Alta | 3 | 0.3 | 10 | 1.0 |

**Representacións gráficas**

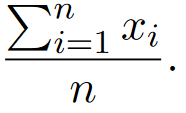
* Variables cualitativas o cuantitativas discretas: **diagrama de barras** o sectores, y diagrama de frecuencias acumuladdas



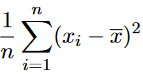
* Variables cuantitativas continuas: **histograma** o diagrama de caja



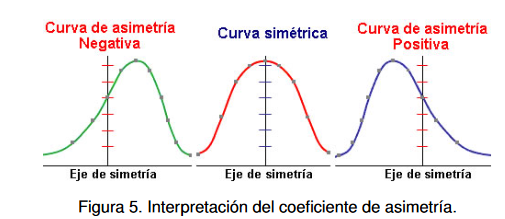
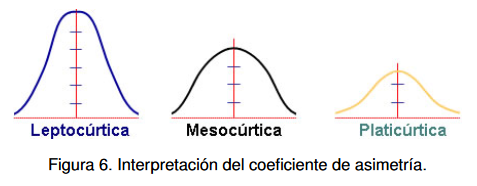
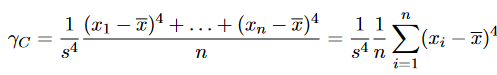
**Medidas de posición**

* Medidas de posición: indican a posición que ocupa a mostra. Poden ser de tendencia central (indican o centro da mostra, media, moda) ou non central. 
  + Media: En caso de intervalos se puede realizar elemento por elemento o intervalo por intervalo.
  + Media truncada: elimínanse un porcentaxe dos datos máis extremos
    - Media recortada: os datos extremos son reemprazados polo punto de corte
  + Cuantiles: Medidas non centrais. Exemplo: O cuantil q0.45 será o dato que deixa á súa esquerda, como mínimo, o 45% dos datos.
    - En un conxunto de 10 datos, q0.45 e q0.4 serán ambos o 5º dato.
    - Os cuartiles (Q1, Q2, Q3) son cuantiles de orden (0.25, 0.5, 0.75). Existen tamén deciles e percentiles.
    - A mediana é o segundo cuartil. Se hai un dato impar de datos, é a media dos dous centrais.

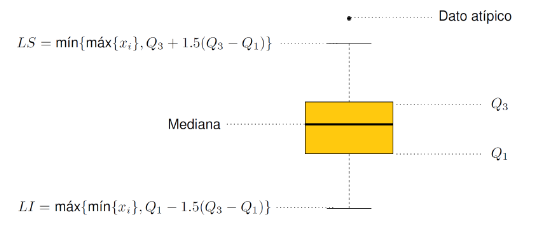
**Medidas de dispersión**

* Recorrido: diferencia entre o máximo e mínimo
  + Recorrido intercuartílico: diferenza entre Q3 e Q1
* Varianza (**s2**):  . A desviación típica é a raíz da varianza, s
  + Siempre positiva.
  + Si se modifican los datos haciendo yi=axi+b, se cumplirá que sy2=a2\*sx2
* Cuasivarianza(**sn-1**): Igual, pero tomando n-1. (ainda asi con todos os datos)
* Coeficiente de variación: CV: s / media. Util porque a magnitude da desviación típica é maior en datos de media maior.

**Medidas de forma**

* Coeficiente de asimetría: 0 se é simétrica respecto á media. Valores positivos se hai máis datos por encima da media, negativos se hai máis por debaixo.
  + 
* Coeficiente de curtosis: mide o grao de apuntamento da distribución 
  + Nunha curva normal é de 3. Maior de 3 é leptocúrtica, menor é platicúrtica.
  + 

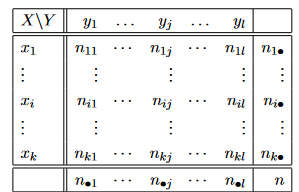
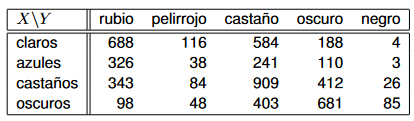
**Diagrama de caixa**

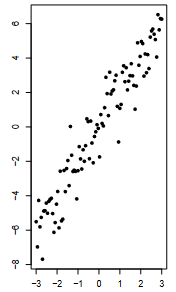
* Representa variables cuantitativas continuas.
* A caixa está delimitada polos cuartiles, sendo a raia do medio o segundo cuartil (a mediana). A altura da caixa é o rango intercuartílico, e contén un 50% dos datos.
* Os bigotes (LS e LI) calcúlanse coa fórmula do gráfico.
* Se quedan datos fóra dos bigotes, considérense atípicos:
  + Defínense outras cotas inferiores e superiores que sexan Q1-3RI e Q3+3RI
  + Se están entre bigotes e a cota especificada, son moderados (\*)
  + Se están fóra das cotas, son atípicos extremos (º)
* 

**Estadística descriptiva bivariante**

* Permite analizar simultáneamente 2 o máis variables
* Se son categóricas ou discretas emprégase a tabla de continxencia
* Con variables continuas realizase unha recta de regresión
* Se temos variables de distinto tipo, por exemplo unha categórica e unha continua, creamos distintos grupos de estudo segundo a variable categórica

**Tablas de continxencia**

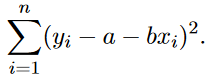
* 
* A suma dos datos dunha columna ou fila concreta denomínanse distribucións marxinais (n**.**1)



**Diagrama de dispersión**

* Represéntase o conxunto de individuos como puntos nun plano onde os eixos x e y son as dúas variables a medir
  + Permite describir a relación entre variables
* **Covarianza:** 
  + O signo describe a relación entre as variables: se é directa Sxy>0, se é inversa Sxy<0, se non hai relacion é proxima a 0
* Para evitar que a covarianza esté influida pola media dos datos, calcúlase o **coeficiente de correlación lineal:** 
  + Dise que a relación é significativa se |rxy| ≥ 0.7, e que existe algunha se é maior de 0.3

**Recta de regresión**

* Consiste en calcular a recta que mellor representa a mostra. A recta será do tipo Y = a + bX + ε, sendo ε o erro cometido.
* **Método de mínimos cuadrados:** Consiste en minimizar a suma dos cadrados dos residuos.
  + Os residuos son a diferenza entre o valor y de cada punto é o da recta.
  + . A partir de esto: 
  + Os valores deben ser **homcedasticos:** A variabilidade dos residuos debe ser constante.
* **Coeficiente de regresión:** Parámetro **b** de la recta de regresión.
  + Si b>0, aumentar los valores de X también aumenta Y.
* **Coeficiente de determinación (R2):** Mide la proporción de variabilidad de Y que explica X. Entre 0 y 1.
  + Si R es próximo a 1, la recta es un buen ajuste de la muestra.
  + 

1. el último intervalo, como excepción, puede ser cerrado por la derecha: [ek-1, ek] [↑](#footnote-ref-0)