Esenciais USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

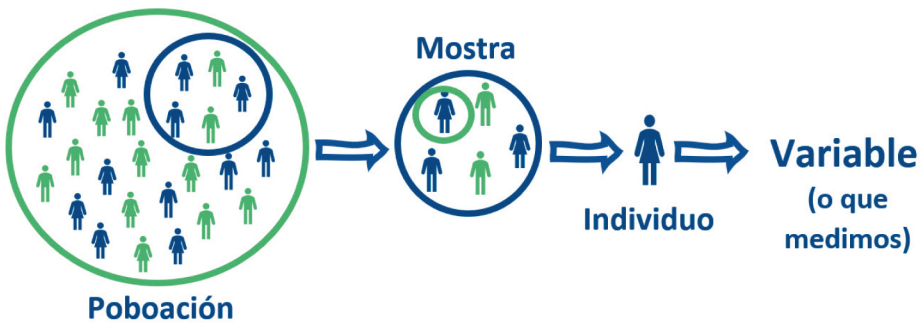
ESTATÍSTICA DESCRITIVA
resumos numéricos e representacións gráficas

María Isabel Borrajo García, Mercedes Conde Amboage e Rosa María Crujeiras Casais (Departamento de Estatística, Análise Matemática e Optimización)

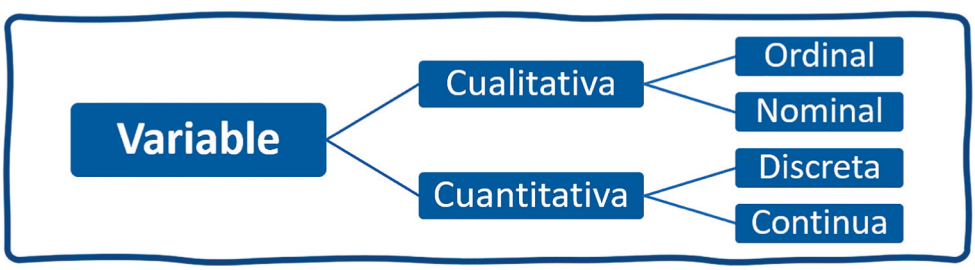
A estatística é, segundo a Real Academia Galega, "o conxunto de técnicas que estudan os feitos para chegar a unha avaliación numérica deles". No primeiro paso do estudo duns feitos, sempre a través dun conxunto de datos, precísase da estatística descritiva, que recolle, clasifica e resume (tanto numérica como graficamente) a información presente na mostra.

Para poder aplicar correctamente estas técnicas, precisamos coñecer algúns conceptos fundamentais:

- **Poboación**: universo de individuos ao que se refire o estudo que se pretende realizar.
- **Mostra**: subconxunto da poboación sobre a que coñecemos os valores da variable de interese; o número de individuos que compoñen a mostra denomínase **tamaño de mostra** e denótase con n.
- **Variable**: característica dos elementos da poboación que se pretende analizar.



Ademais, a análise descritiva dunha variable dependerá da súa tipoloxía, e polo tanto precisamos ter presente a clasificación "clásica" das variables aleatorias:



As variables **cualitativas** son aquelas que toman valores que son atributos ou categorías. Dentro delas atopamos dous grandes subgrupos que son as **ordinais** (aquelas nas que os valores teñen unha relación de orde intrínseca, como por exemplo, "Nivel de estudos" ou "Medallas (ouro, prata e bronce) nunha proba deportiva"); e as **nominais** (as categorías non teñen ningunha orde preestablecida, como por exemplo "Sexo" ou "Cor de ollos").

As variables **cuantitativas** son aquelas que toman valores numéricos. Se eses posibles valores son unha cantidade finita ou numerable diremos que a variable é **discreta** (como por exemplo "Número de fillos" ou "Número de chamadas diarias ao 112"); mentres que se hai una cantidade infinita de posibles valores estaremos ante unha variable **continua** (como o "Peso" ou o "Nivel de glicosa en sangue").

En calquera análise estatística, un primeiro paso comprenderá sempre a identificación da variable de interese (característica da poboación que desexamos analizar e da que dispoñemos de valores nunha mostra) e a súa clasificación, que nos permitirá, mediante as indicacións que se ven a continuación, facer un estudo correcto.

Análise dunha variable cualitativa

Supoñamos que dispoñemos dunha mostra de n=150 individuos dunha **variable cualitativa ordinal** que toma k posibles valores (categoría 1, categoría 2, ..., categoría i, ..., categoría k), no noso exemplo k=3.

A primeira forma de recoller e resumir a información contida nunha mostra é efectuar un reconto do número de veces que se observou cada un dos distintos valores que pode tomar a variable. Estas cantidades denomínanse frecuencias.

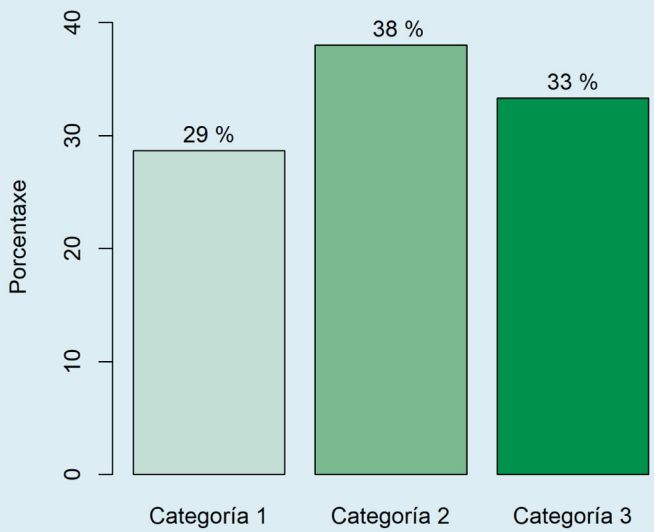
Existen distintos tipos de recontos que dan lugar a varias medidas:

- *Frecuencia absoluta (n_i):* é o número de veces que ocorre cada un dos posibles valores da variable.
- *Frecuencia relativa (f_i):* representa a proporción de datos en cada unha das categorías e obtense como n_i/n.
- *Frecuencia absoluta acumulada (N_i):* é o número de veces que se observou a categoría i-ésima e das anteriores*.
- *Frecuencia relativa acumulada (F_i):* é a proporción de datos da categoría i-ésima e das anteriores*.

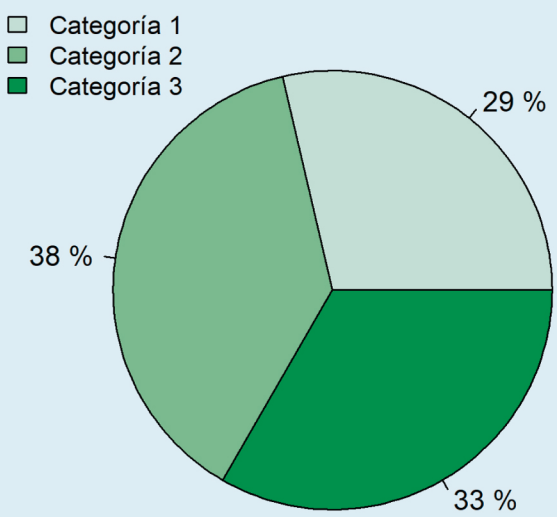
As frecuencias poden presentarse de xeito ordenado nunha **táboa de frecuencias**; este sería o resultado para a nosa mostra:

	Frec. absoluta (n_i)	Frec. relativa (f_i)	Frec. absoluta acumulada (N_i)	Frec. relativa acumulada (F_i)
Categoría 1	43	0.29	43	0.29
Categoría 2	57	0.38	100	0.67
Categoría 3	50	0.33	150	1

O **diagrama de barras** permite representar as frecuencias absolutas ou relativas. No eixe de abscisas sitúanse as distintas categorías (respectando a orde se a houbese) e debúxanse barras verticais sobre elas. As alturas das barras representan as frecuencias absolutas, relativas ou as porcentaxes.



O **diagrama de sectores** representa a mesma información que o diagrama de barras pero cun aspecto diferente. A súa elaboración require dividir o círculo en tantos sectores como categorías teña a variable, de xeito que a amplitude de cada sector sexa proporcional á frecuencia correspondente. Unha pequena complicación desta representación é que o ollo humano está máis acostumado a comparar lonxitudes que áreas, polo que pode resultar nalgúns casos menos intuitivo.



**Nótese que o concepto de "anteriores" ten sentido posto que estamos traballando cunha variable cualitativa ordinal; no caso dunha variable cualitativa nominal as frecuencias acumuladas non están definidas e polo tanto non se calcularían.*

Análise dunha variable cuantitativa discreta

Supoñamos que dispoñemos dunha mostra de n=150 individuos sobre os que medimos unha certa variable discreta que neste caso ten seis posibles valores: valor 1, valor 2, valor 3, valor 4, valor 5 e valor 6 (este número de valores é só ilustrativo, unha variable discreta como xa vimos pode tomar calquera cantidade finita ou numerable** de valores).

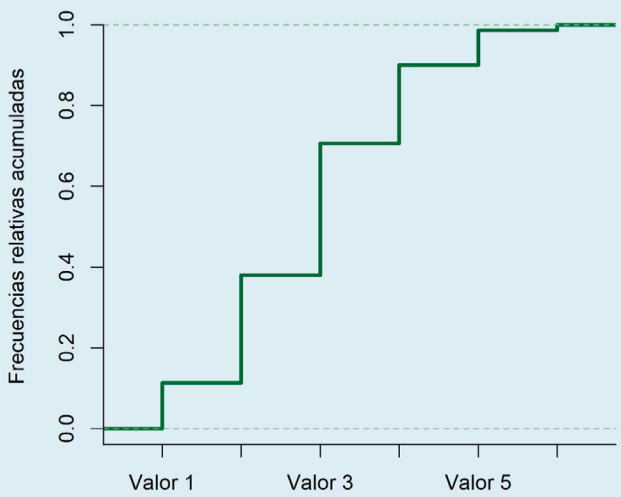
Neste caso podemos duplicar a mesma análise que fixemos para as variables de tipo cualitativo, tanto a táboa de frecuencias como as representacións gráficas. Vexamos por exemplo a táboa de frecuencias resultante para a nosa variable discreta:

	Frec. absoluta (n_i)	Frec. relativa (f_i)	Frec. absoluta acumulada (N_i)	Frec. relativa acumulada (F_i)
Valor 1	17	0.11	17	0.11
Valor 2	40	0.27	57	0.38
Valor 3	49	0.33	106	0.71
Valor 4	29	0.19	135	0.90
Valor 5	13	0.09	148	0.99
Valor 6	3	0.01	150	1

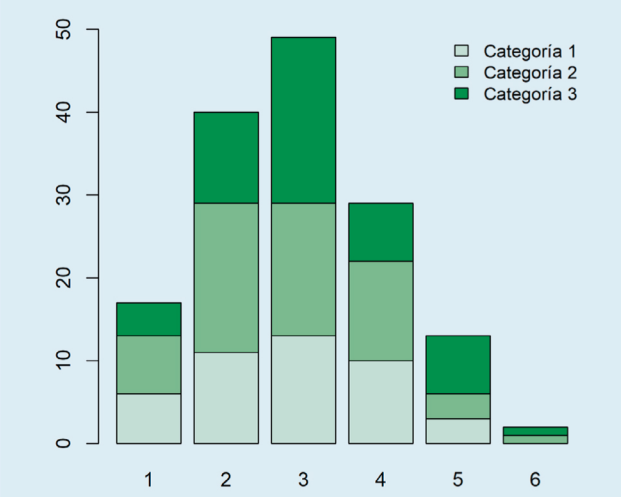
Ademais das representacións xa mencionadas para variables cualitativas, en variables discretas empregamos tamén algunhas outras que detallamos a continuación.

O **diagrama de frecuencias acumuladas**, tamén coñecido como **diagrama en escaleira**, permite visualizar mediante unha liña as frecuencias acumuladas dunha variable discreta.

No eixe de abscisas sitúanse os valores da variable e no eixe de ordenadas as frecuencias acumuladas. Pódese realizar tanto coas frecuencias absolutas como coas relativas: no primeiro caso o eixe de abscisas tomaría valores entre 0 e o tamaño mostral (n); e no segundo entre 0 e 1. A altura dos chanzos corresponderase, respectivamente, coa frecuencia absoluta ou relativa de cada valor da variable.



Se a maiores da variable discreta dispoñemos, por exemplo, dunha variable cualitativa con k categorías (neste caso k=3), podemos elaborar un **diagrama de barras aniñadas.** É de estrutura similar ao diagrama de barras clásico, pero permite coñecer a distribución por categorías (dadas pola variable cualitativa) para cada un dos valores da variable discreta.



*** Xeralmente, cando toma unha cantidade moi grande de valores, o tratamento de tal variable como discreta deixa de ter sentido, e traballarase con ela como se fose unha variable cuantitativa continua.*

Análise dunha variable cuantitativa continua

As variables cuantitativas continuas tamén se poden resumir empregando **táboas de frecuencias**, aínda que neste caso non será posible considerar todos os seus valores un a un, posto que son infinitos. Entón, será necesario construír modalidades artificiais de maneira que se agrupen valores en intervalos. Estas novas modalidades denomínanse **intervalos de clase**.

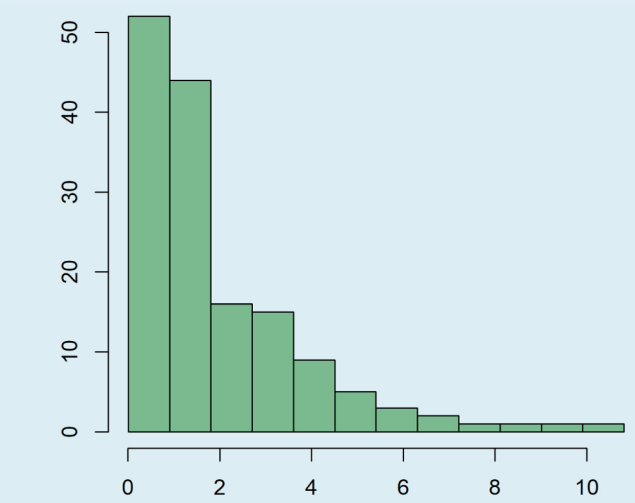
A partir dunha mostra $\{x_1, x_2, \ldots, x_{150}\}$ de valores dunha variable continua, os intervalos de clase constrúense do seguinte xeito:

- Denotamos por $\text{mín}\{x_i\} = e_0 < e_1 < \ldots < e_k = \text{máx}\{x_i\}$ os extremos dos k intervalos de clase. Cada intervalo será da forma $[e_{i-1}, e_i)$ con habitualmente $k = \sqrt{n}$ intervalos.
- A amplitude de cada intervalo será $a_i = e_i - e_{i-1}$, en moitas ocasións a mesma para todos os intervalos de clase (é dicir, a amplitude sería $\left(máx\{x_i\} - mín\{x_i\}\right) / \sqrt{n}$).
- Os puntos $c_i = \dfrac{e_{i-1} + e_i}{2}$, que son os centros dos intervalos, coñécense como **marcas de clase**.

	Marca de clase (c_i)	Frec. absoluta (n_i)	Frec. relativa (f_i)	Densidade (f_i/a_i)
Intervalo 1: $[e_0, e_1)$	$c_1 = \dfrac{e_0 + e_1}{2}$	n_1	f_1	f_1 / a_1
Intervalo 2: $[e_1, e_2)$	$c_2 = \dfrac{e_1 + e_2}{2}$	n_2	f_2	f_2 / a_2
Intervalo 3: $[e_2, e_3)$	$c_3 = \dfrac{e_2 + e_3}{2}$	n_3	f_3	f_3 / a_3
...	...	...	...	...
Intervalo k: $[e_{k-1}, e_k]$	$c_k = \dfrac{e_{k-1} + e_k}{2}$	n_k	f_k	f_k / a_k

O **histograma** permite representar as frecuencias recollidas na táboa anterior. No eixe de abscisas sitúanse os distintos intervalos de clase sobre os que elevamos barras verticais. As alturas das barras poden representar as frecuencias absolutas, relativas ou as densidades (a suma de todas as áreas das barras é 1).

Nótese que, a diferenza do que ocorre no caso das variables cualitativas ou cuantitativas discretas, neste caso as barras son contiguas.



O **diagrama de talo e follas** é unha representación que permite observar os datos e que á vez dá unha idea da súa distribución. Primeiro selecciónase o número de cifras significativas (talo) que se colocan á esquerda, trázase unha liña vertical e inclúense ao lado as cifras seguintes de cada dato observado (follas). Pódese observar una forma moi similar á do histograma xirándoo 90º en sentido contrario ás agullas do reloxo.

```
The decimal point is at the |

  0 | 000011111222222233334444444445555566666677777888999
  1 | 000000001111112222222333333445555555666677888899
  2 | 11122233345677899
  3 | 00123444458899
  4 | 0022259
  5 | 2338
  6 | 00
  7 | 114
  8 | 2
  9 | 8
 10 | 8
```

Análise dunha variable cuantitativa continua

Dada unha variable cuantitativa***, as **medidas características** permiten describir a información contida nunha mostra de n=150 datos na que os valores serán, de xeito xenérico, $\{x_1, x_2, \ldots, x_{150}\}$. Estas medidas utilízanse para resumir a información atendendo a dous aspectos principais: arredor de que valores se encontran os datos (medidas de posición) e canto se dispersan (medidas de dispersión).

Medidas de posición

- A **media** é o valor promedio dos datos da mostra: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- A **mediana** é o valor que deixa a metade das observacións por debaixo e a outra metade por enriba.
- Os **cuartís** Q_1, Q_2, e Q_3 dividen a mostra en catro partes iguais, de maneira que por debaixo de Q_1 teremos o 25% dos datos, entre Q_1 e Q_2 encóntrase outro 25% Q_1, entre Q_2 e Q_3 outro 25% e por enriba de Q_3 outro 25%. Esta idea de dividir a mostra en partes iguais pódese xeneralizar dando lugar á definición de **cuantil** (sendo habituais os decís, en 10 partes iguais, ou os percentís, en 100 partes iguais).
- A **moda** é o valor ou valores que máis se repiten na mostra.

Medidas de dispersión

- A **varianza** mide a distancia cadrática de cada punto da mostra á súa media, e vén dada por $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. Habitualmente trabállase coa desviación padrón da mostra $s = +\sqrt{s^2}$ correspondéndolle as mesmas unidades que as dos datos.
- O **rango** representa a diferenza entre os valores máximo e mínimo da mostra: $máx\left(x_i\right) - \min\left(x_i\right)$.
- O **rango intercuartil** representa a diferenza entre os cuartís Q_3 e Q_1: RIC =Q_3 - Q_1.
- O **coeficiente de variación** defínese como $CV = \frac{s}{\bar{x}}$ cando $\bar{x} > 0$ e permítenos comparar variables aínda que estean rexistradas en distintas unidades de medida.

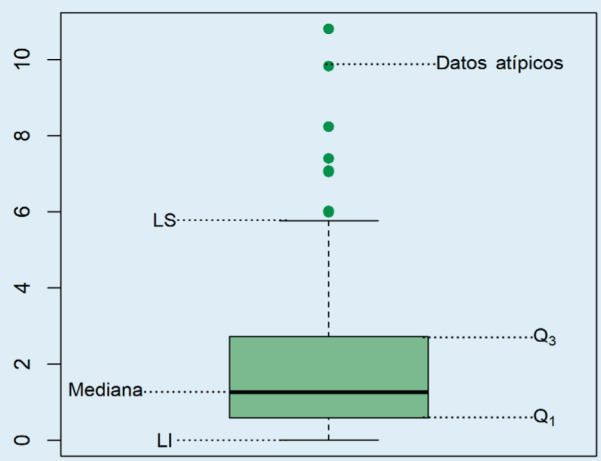
O diagrama de caixa obtense utilizando as seguintes medidas:

- Cuartís Q_1 e Q_3 como límites da caixa central. A altura da caixa vén dada polo RIC.
- Os límites inferior e superior obtéñense como:

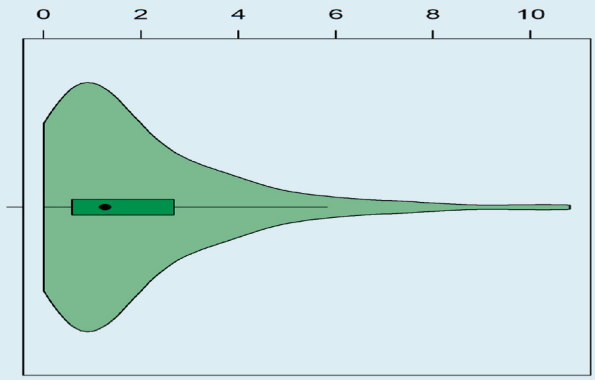
$$LI = máx\left\{\text{mín}\left\{x_i\right\}, Q_1 - 1.5 * RIC\right\}$$

$$LS = mín\left\{\text{máx}\left\{x_i\right\}, Q_3 + 1.5 * RIC\right\}$$

- A mediana debúxase na caixa cunha liña máis grosa.
- Os datos fóra dos límites son atípicos.



O **gráfico de violín** é unha representación que combina nun mesmo gráfico un diagrama de caixas e un histograma "suavizado".



**** Todas estas medidas características poden calcularse tanto para variables cuantitativas discretas como para variables cuantitativas continuas. Pola contra, para variables cualitativas so podemos calcular a moda.*

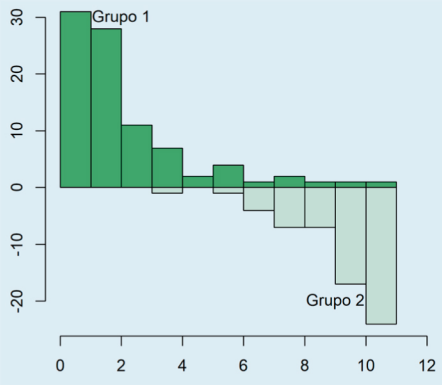
Outros resumos gráficos

Nesta última sección abórdanse algunhas representacións gráficas máis elaboradas que permiten resumir a información dos datos tendo en conta máis características ou información sobre eles. Pode ocorrer que dispoñamos doutras variables, ademais da inicial, que establezan por exemplo agrupamentos, ou máis dunha variable cuantitativa que se pretenda representar conxuntamente.

Existen tamén datos con características moi particulares, como poden ser datos espaciais, espazo-temporais, direccionais ou circulares... que precisan de representacións específicas que recollan toda a información que conteñen.

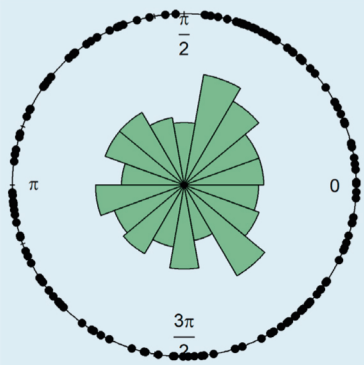
Supoñamos que dispoñemos dunha certa variable continua na que ademais están establecidos dous grupos, como pode ser calquera variable antropométrica dividida por sexo ou unha certa magnitude medida entre un grupo de casos e outro de controis.

Nun primeiro paso poderiamos representar o histograma conxunto da mostra como xa amosamos, pero tamén sería de interese representar o histograma de cada grupo por separado véndoos sobre o mesmo lenzo, para isto podemos empregar o **histograma en espello**.

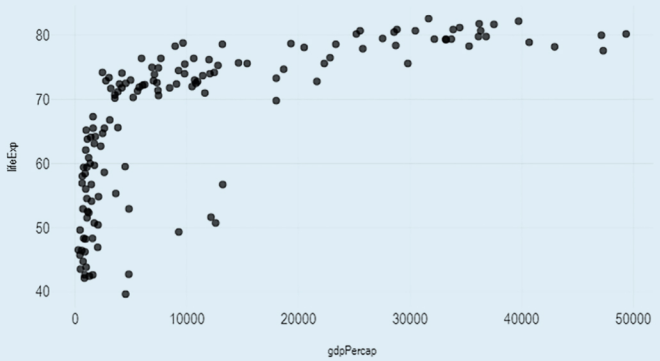


O **diagrama de rosa** é un gráfico definido exclusivamente para a representación de datos circulares ou direccionais, é dicir, datos que teñen asociada una periodicidade (como por exemplo variables que ocorren ao longo dos 12 meses do ano) ou cunha clara natureza direccional (como a dirección do vento).

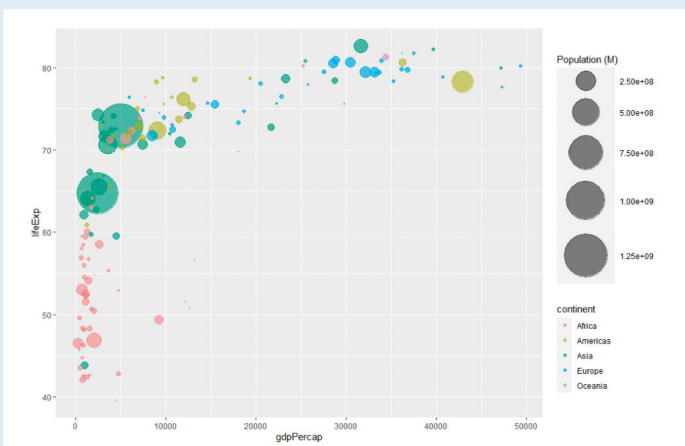
Esta representación consiste nun círculo exterior no que se sitúan os datos da mostra, unhas barras con orixe no seu centro e con alturas proporcionais á frecuencia dos datos nese sector circular.



Para representar conxuntamente dúas variables cuantitativas empregamos un **diagrama de dispersión**: no eixe de abscisas colócanse os valores dunha das variables, no eixe de ordenadas os valores da outra variable, e debúxase un punto para cada individuo na posición correspondente aos seus valores, dando lugar a unha nube de puntos.



Unha ampliación é o coñecido **diagrama de burbullas**, co que podemos representar, a maiores da información do diagrama de dispersión, unha terceira variable continua (raio das burbullas) e unha variable cualitativa ou discreta (cor das burbullas).



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
XUNTA DE GALICIA
RAGC/ REAL ACADEMIA GALEGA DE CIENCIAS