Introduction

Pattern Recognition

Gamaliel Moreno Chávez

MCPI

Enero-Julio
2021

Introduction

IS PATTERN RECOGNITION IMPORTANT?

Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. We will refer to these objects using the generic term patterns

- Machine vision
- Character (letter or number) recognition
- Computer-aided diagnosis
- Speech recognition

Introduction

What is machine learning?

Machine Learning or ML

A computer program is said to learn from experience E with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

we treat all unknown quantities (e.g., predictions about the future value of some quantity of interest, such as tomorrow's temperature, or the parameters of some model) as random variables, that are endowed with probability distributions which describe a weighted set of possible values the variable may have.

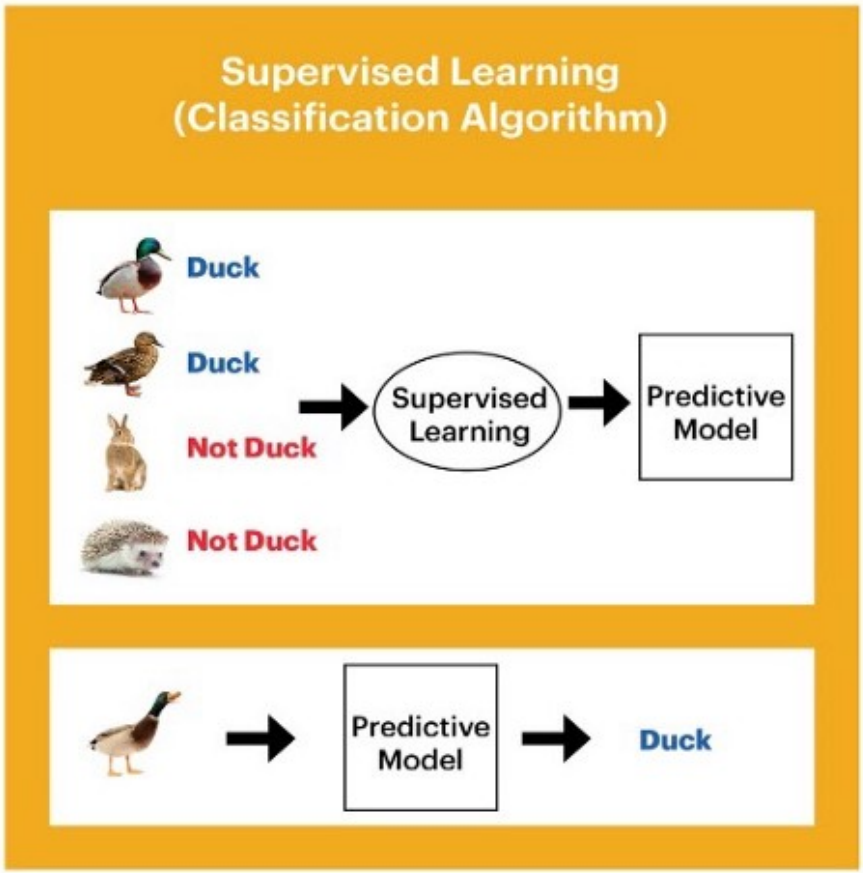
Introduction

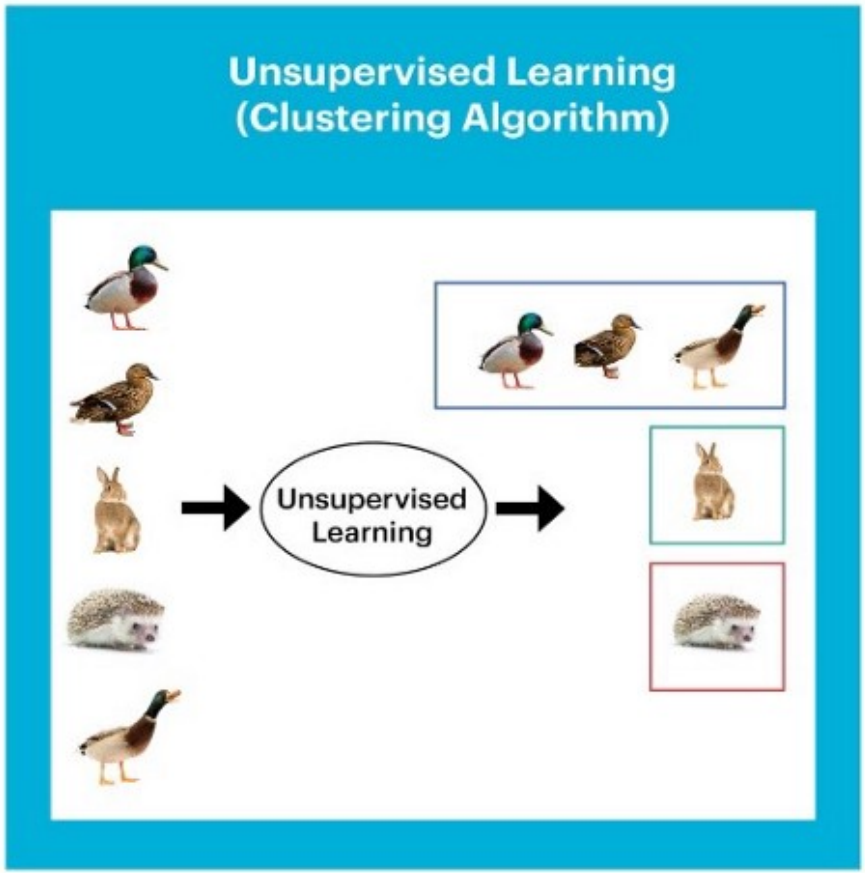
Probabilistic modeling is the language used by most other areas of science and engineering

Almost all of machine learning can be viewed in probabilistic terms, making probabilistic thinking fundamental. It is, of course, not the only view. But it is through this view that we can connect what we do in machine learning to every other computational science, whether that be in stochastic optimization, control theory, operations research, econometrics, information theory, statistical physics or bio-statistics. For this reason alone, mastery of probabilistic thinking is essential.

Introduction

Learning types: Supervised and Unsupervised learning





Introduction

Supervised learning

- the task T is to learn a mapping f from inputs $x \in X$ to outputs $y \in Y$.
- The inputs x are also called the features, covariates, or predictors. This is often a fixed dimensional vector of numbers, such as the height and weight of a person, or the pixels in an image.
- $X = \mathbb{R}^D$, where D is the dimensionality of the vector.
- The output y is also known as the label, target, or response.
- The experience E is given in the form of a set of N input-output pairs $\mathcal{D} = \{(xn, yn)\}_{n=1}^{N}$ known as the training set.
- N is called the sample size.

Classification

In classification problems, the output space is a set of C unordered and mutually exclusive labels known as classes, $Y = \{1, 2, \ldots, C\}$. The problem of predicting the class label given an input is also called pattern recognition. If there are just two classes, often denoted by $y \in \{0, 1\}$ or $y \in \{0, 1\}$, it is called binary classification.

Classification

Example: classifying Iris flowers. Consider the problem of classifying iris flowers into their 3 subspecies, setosa, versicolor and virginica.

(a) (b) (c)

Figure 1.1: Three types of iris flowers: setosa, versicolor and virginica. Used with kind permission of Dennis Kramb and SIGNA.

Classification

Example: classifying Iris flowers. Consider the problem of classifying iris flowers into their 3 subspecies, setosa, versicolor and virginica.

(a) (b) (c)

Figure 1.1: Three types of iris flowers: setosa, versicolor and virginica. Used with kind permission of Dennis Kramb and SIGNA.

Classification

In image classification, the input space X is the set of images, which is a very high dimensional space: for a color image with $C = 3$ channels (e.g., RGB) and $D_1 \times D_2$ pixels, we have $X = \mathbb{R}^D$, where $D = C \times D_1 \times D_2$.
Some botanists have already identified 4 simple, but highly informative, numeric feature: sepal length, sepal width, petal length, petal width
We will use this much lower dimensional input space,$X = \mathbb{R}^D$, for simplicity. The iris dataset is a collection of 150 labeled examples of iris flowers, 50 of each type, described by these 4 features.

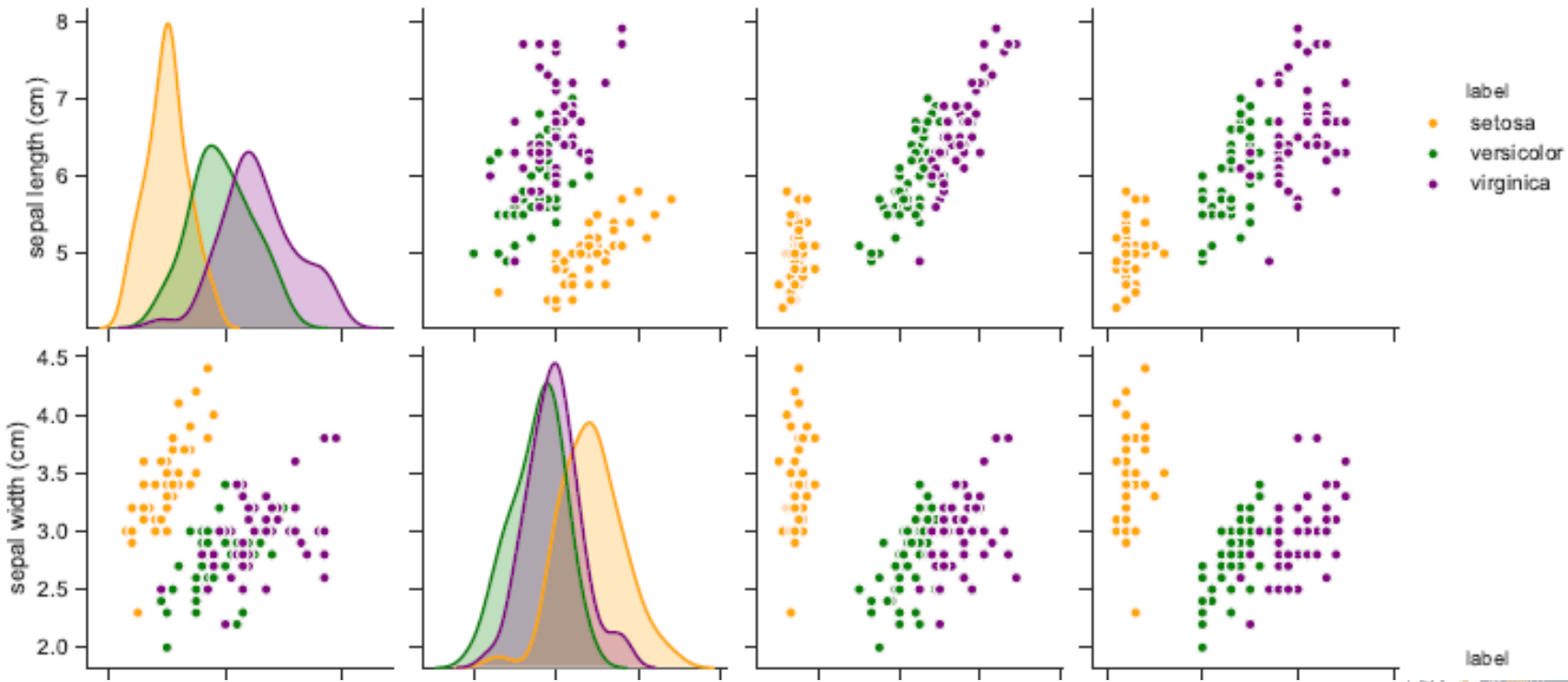
Classification

It is common to store them in an $N \times D$ matrix, in which each row represents an example, and each column represents a feature. This is known as a design matrix;

index	sl	sw	pl	pw	label
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
	...				
50	7.0	3.2	4.7	1.4	versicolor
	...				
149	5.9	3.0	5.1	1.8	virginica

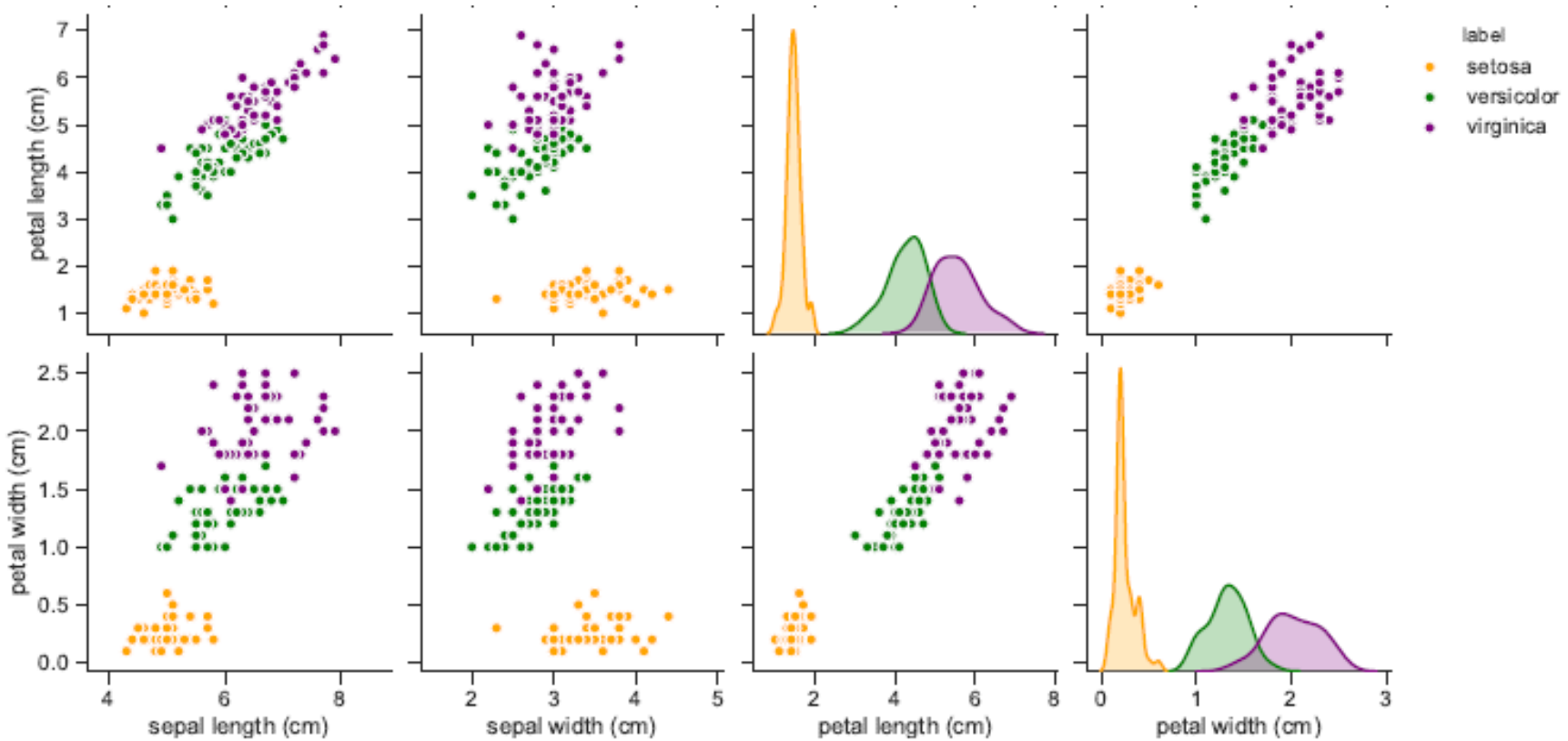
Classification

Exploratory data analysis. Before tackling a problem with ML, it is usually a good idea to perform exploratory data analysis.



Classification

Exploratory data analysis. Before tackling a problem with ML, it is usually a good idea to perform exploratory data analysis.



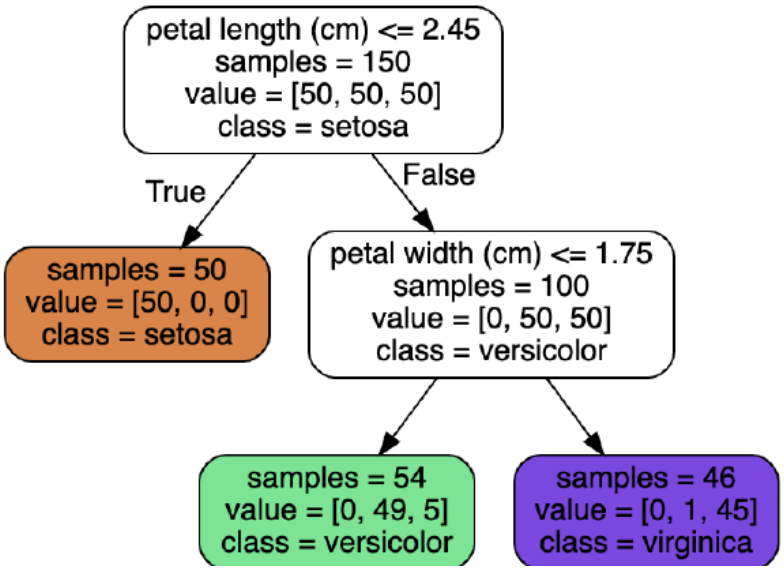
Classification

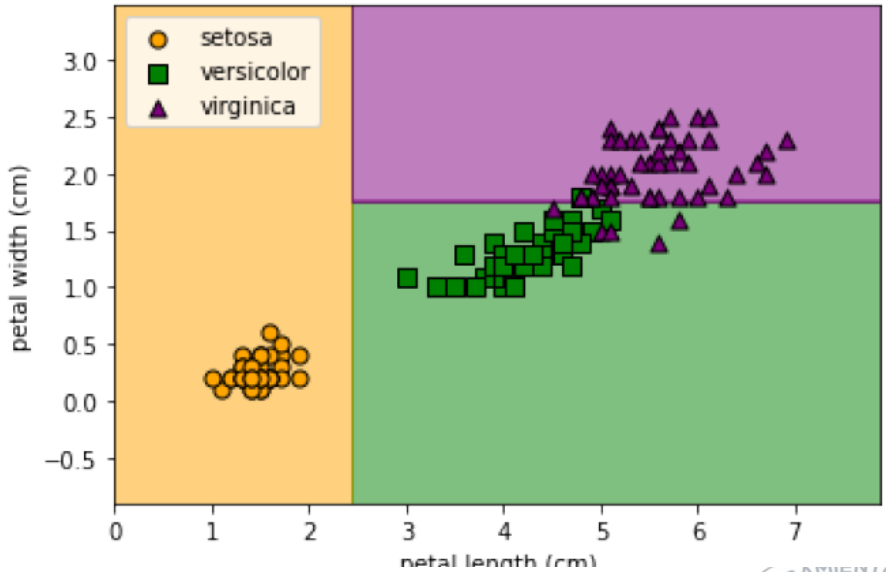
We can see that the setosa class is easy to distinguish from the other two classes. For example, suppose we create the following decision rule:

$$f(\mathbf{x};\boldsymbol{\theta}) = \begin{cases} \text{Setosa if petal length} < 2.45 \\ \text{Versicolor or Virginica otherwise} \end{cases}$$

Classification

We can arrange these nested rules in to a tree structure, **called a decision tree**.





Model fitting

The goal of supervised learning is to automatically come up with classification models. A common way to measure performance on this task is in terms of the misclassification rate on the training set:

$$\mathcal{L}(\theta) \triangleq \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(y_n \neq f(x_n; \theta))$$

where $\mathbb{I}(e)$ is the binary indicator function,

$$\mathbb{I}(e) = \begin{cases} 1 & \text{if } e \text{ is true} \\ 0 & \text{if } e \text{ is false} \end{cases}$$

Model fitting

For example, suppose we are foraging in the wilderness and we find some iris flowers. Furthermore, suppose that setosa and versicolor are tasty, but virginica is poisonous. In this case, we might use the asymmetric loss function

		Estimate		
		Setosa	Versicolor	Virginica
Truth	Setosa	0	1	10
	Versicolor	1	0	10
	Virginica	1	1	0

Model fitting

We can then define empirical risk to be the average loss of the predictor on the training set:

$$\mathcal{L}(\theta) \triangleq \frac{1}{N} \sum_{n=1}^{N} l(y_n, f(x_n; \theta))$$

the empirical risk is equal to misclassification rate when we use zero-one loss for comparing the true label with the prediction

$$l_{01}(y, \hat{y}) = \mathbb{I}(y_n \neq \hat{y})$$

Model fitting

One way to define the problem of model fitting or training is to find a setting of the parameters that minimizes the empirical risk on the training set:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta) = \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} l(y_n, f(x_n; \theta))$$

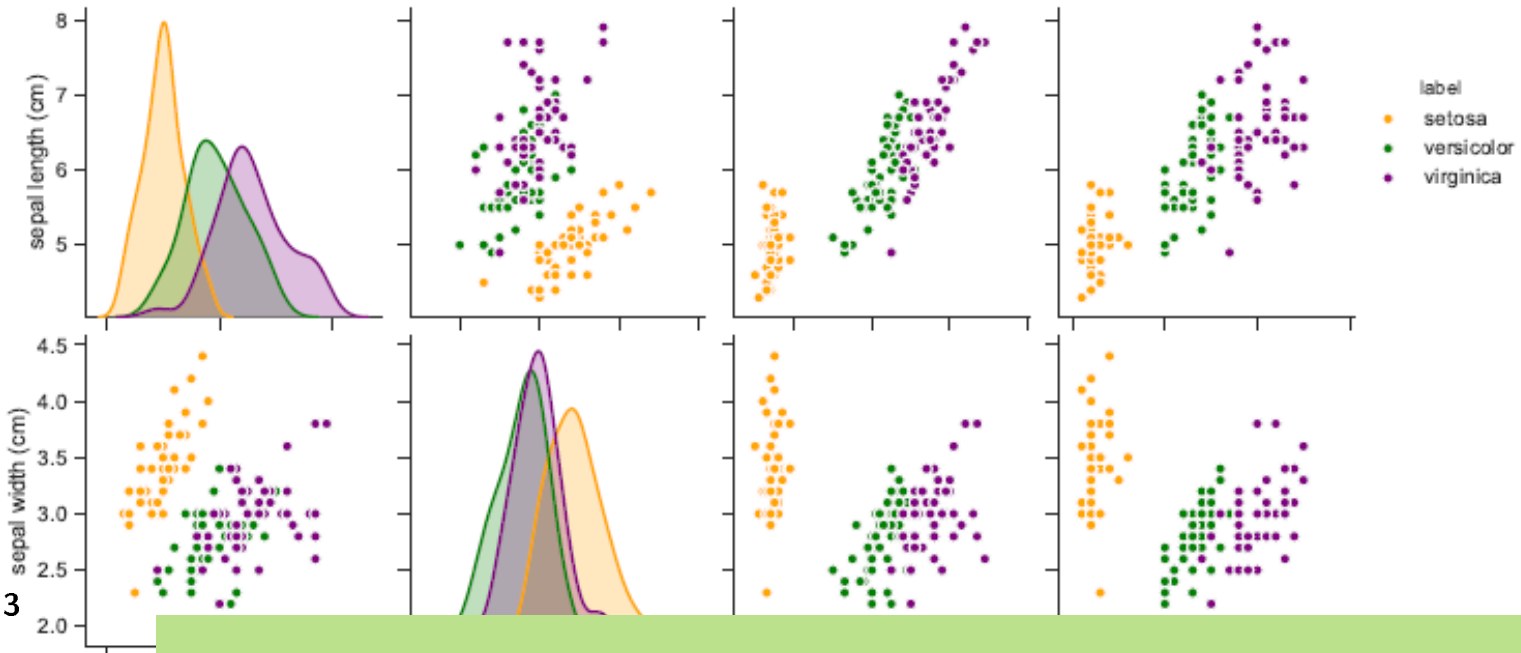
This is called empirical risk minimization.

Rango

Definición

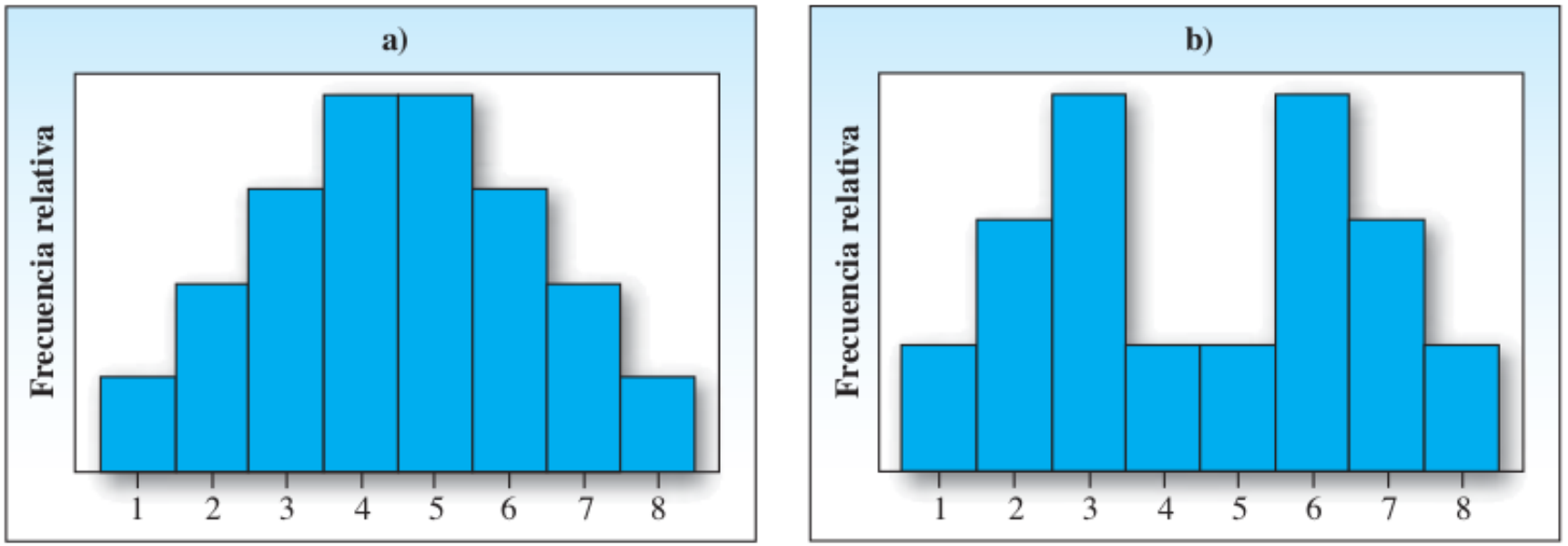
El rango, R, de un conjunto de n mediciones se define como la diferencia entre la medición más grande y la más pequeña.

Para los datos de peso al nacer de la tabla, las mediciones varían de 5.6 a 9.4. Por tanto, el rango es $9.4 - 5.6 = 3.8$.



Rango

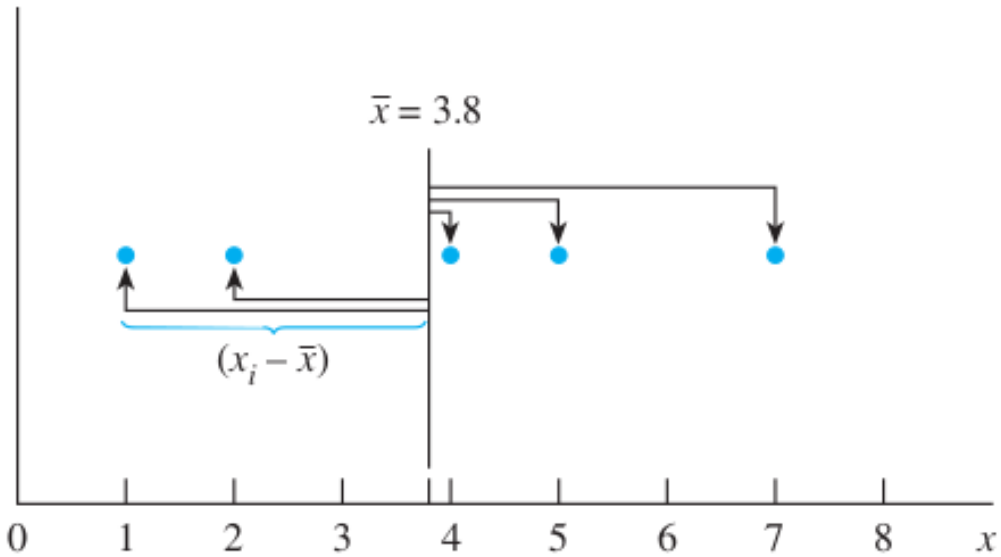
El rango es fácil de calcular, fácil de interpretar y es una medida adecuada de variación para conjuntos pequeños de datos. Pero, para conjuntos grandes, el rango no es una medida adecuada de variabilidad. Por ejemplo, las dos distribuciones de frecuencia relativa de la figura tienen el mismo rango pero muy diferentes formas y variabilidad.



Varianza

¿Hay una medida de variabilidad que sea más sensible que el rango?

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{19}{5} = 3.8$$



Cálculo de $\Sigma(x_i - \bar{x})^2$

x	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
5	1.2	1.44
7	3.2	10.24
1	−2.8	7.84
2	−1.8	3.24
4	.2	.04
19	0.0	22.80

Varianza

De la suma de desviaciones cuadradas, se calcula una sola medida llamada varianza. Para la varianza de una muestra usamos el símbolo s^2 y la varianza de una población σ^2. La varianza será relativamente grande para datos muy variables y relativamente pequeña para datos menos variables.

Definición

La **varianza de una población** de N mediciones es el promedio de los cuadrados de las desviaciones de las mediciones alrededor de su media μ. La varianza poblacional se denota con σ^2 y está dada por la fórmula

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Varianza

Definición

La **varianza de una muestra** de n mediciones es la suma de las desviaciones cuadradas de las mediciones alrededor la media $\bar{x}$ dividida entre $(n-1)$. La varianza muestral se denota con s^2 y está dada por la fórmula

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

Varianza

Para el conjunto de $n = 5$ mediciones muestrales presentadas en la tabla

suma

$$\sum (x_i - \bar{x})^2 = 22.80$$

varianza muestral

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{22.80}{4} = 5.70$$

Cálculo de $\Sigma(x_i - \bar{x})^2$

x	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
5	1.2	1.44
7	3.2	10.24
1	−2.8	7.84
2	−1.8	3.24
4	.2	.04
19	0.0	22.80

Desviación estándar

La varianza se mide en términos del cuadrado de las unidades originales de medición. Tomando la raíz cuadrada de la varianza, obtenemos la desviación estándar, que regresa la medida de variabilidad a las unidades originales de medición.

Definición

La **desviación estándar** de un conjunto de mediciones es igual a la raíz cuadrada positiva de la varianza.

Notación

NOTACIÓN

n: número de mediciones en la muestra

s^2: varianza muestral

$s = \sqrt{s^2}$: desviación muestral estándar

N: número de mediciones en la población

σ^2: varianza poblacional

$\sigma = \sqrt{\sigma^2}$: desviación poblacional estándar

Resumen

- El valor de s es siempre mayor o igual a cero.
- Cuanto mayor sea el valor de s^2 o de s, mayor es la variabilidad del conjunto de datos.
- Si s^2 o s es igual a cero, todas las mediciones deben tener el mismo valor.
- Para medir la variabilidad en las mismas unidades que las observaciones originales, calculamos la desviación estándar $s = \sqrt{s^2}$.

Coeficiente de variación

El coeficiente de variación es la relación entre la desviación típica de una muestra y su media.

$$CV = \left(\frac{\sigma}{\bar{x}}\right) 100$$

El coeficiente de variación permite comparar las dispersiones de dos distribuciones distintas, siempre que sus medias sean positivas.
Mayor dispersión el valor del coeficiente de variación será mayor.

Coeficiente de variación

Una distribución tiene $\bar{x} = 140$ y $\sigma = 28.28$ y otra $\bar{x} = 150$ y $\sigma = 24$. ¿Cuál de las dos presenta mayor dispersión?

$$CV_1 = \frac{28,28}{140} \cdot 100 = 20,2\,\%$$

$$CV_2 = \frac{24}{150} \cdot 100 = 16\,\%$$

La primera distribución presenta mayor dispersión.

Sobre la significancia práctica de la desviación estándar

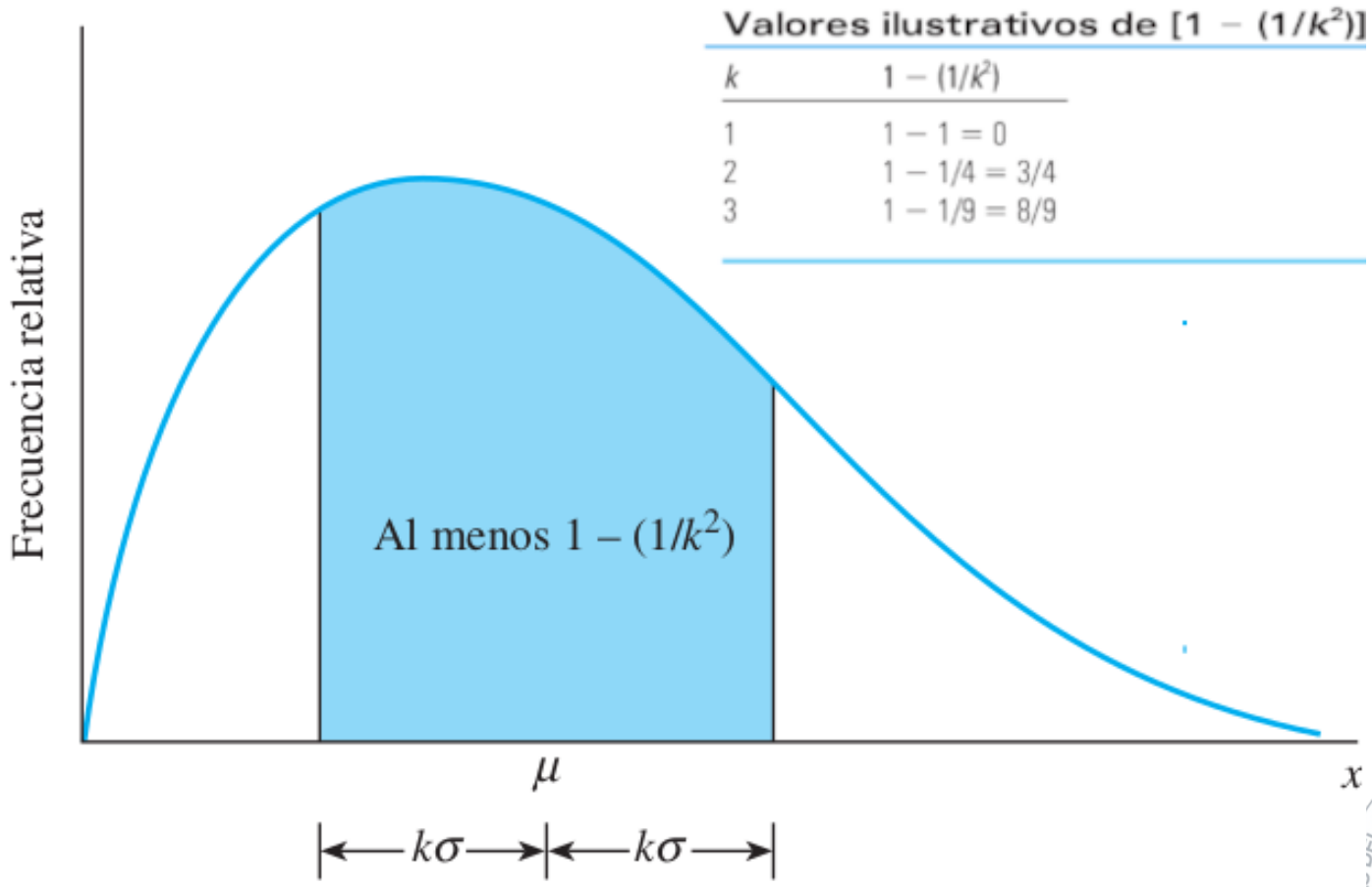
Teorema de Chebyshev

Dado un número k mayor o igual a 1 y un conjunto de n mediciones, al menos $[1 - (1/k^2)]$ de las mediciones estarán dentro de k desviaciones estándar de su media.

Se construye un intervalo al medir una distancia $k\sigma$ a cualquier lado de la media m. El número k puede ser cualquier número mientras sea mayor o igual a 1. Entonces el teorema de Chebyshev expresa que al menos $[1 - (1/k^2)]$ del número total n de mediciones está en el intervalo construido.

Teorema de Chebyshev

Ilustración del teorema de Chebyshev



Valores ilustrativos de $[1 - (1/k^2)]$

k	$1 - (1/k^2)$
1	$1 - 1 = 0$
2	$1 - 1/4 = 3/4$
3	$1 - 1/9 = 8/9$

Teorema de Chebyshev

Ilustración del teorema de Chebyshev

Valores ilustrativos de $[1 - (1/k^2)]$

k	$1 - (1/k^2)$
1	$1 - 1 = 0$
2	$1 - 1/4 = 3/4$
3	$1 - 1/9 = 8/9$

- Al menos ninguna de las mediciones está en el intervalo $\mu - \sigma$ a $\mu + \sigma$.
- Al menos 3/4 de las mediciones están en el intervalo $\mu - 2\sigma$ a $\mu + 2\sigma$.
- Al menos 8/9 de las mediciones están en el intervalo $\mu - 3\sigma$ a $\mu + 3\sigma$.

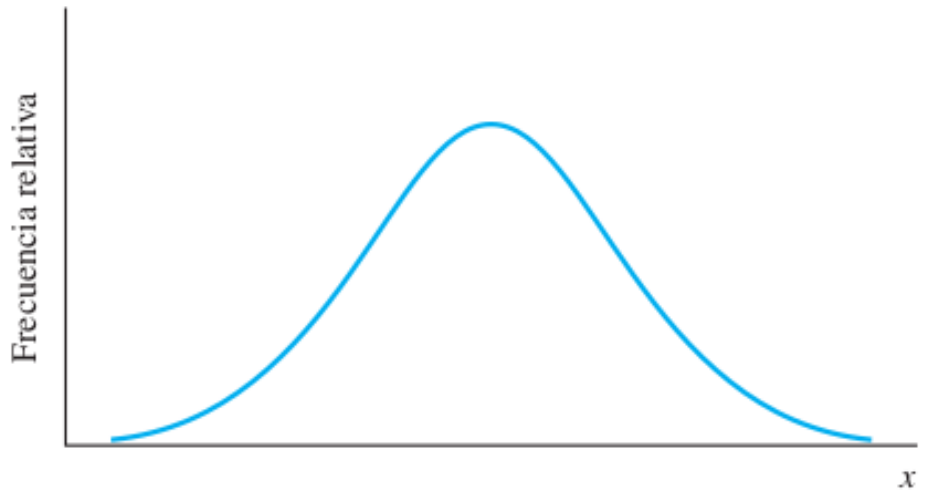
Teorema de Chebyshev

Ejemplo. La media y varianza de una muestra de $n = 25$ mediciones son 75 y 100, respectivamente. Use el teorema de Chebyshev para describir la distribución de mediciones.
Solución Nos dan $\bar{x} = 75$ y $s^2 = 100$. La desviación estándar es $s = \sqrt{100} = 10$. La distribución de mediciones está centrada alrededor de x = 75, y el teorema de Chebyshev establece que:

- Al menos 3/4 de las 25 mediciones están en el intervalo $\bar{x} \pm 2s = 75 \pm 2(10)$, esto es, 55 a 95.
- Al menos 8/9 de las mediciones están en el intervalo $\bar{x} \pm 3s = 75 \pm 3(10)$, esto es, 45 a 105.

Regla empírica

Como el teorema de Chebyshev se aplica a cualquier distribución, es muy conservador. Ésta es la razón por la que hacemos hincapié en "al menos $1 - (1/k^2)$" en este teorema. Otra regla para describir la variabilidad de un conjunto de datos no funciona para todos los conjuntos de datos, pero funciona muy bien para datos que "se apilan" en la conocida forma de montículo de la figura



Regla empírica

Regla empírica Dada una distribución de mediciones que tiene forma aproximada de montículo:

El intervalo $(\mu \pm \sigma)$ contiene aproximadamente 68% de las mediciones.
El intervalo $(\mu \pm 2\sigma)$ contiene aproximadamente 95% de las mediciones.
El intervalo $(\mu \pm 3\sigma)$ contiene aproximadamente 99.7% de las mediciones.

Regla empírica

Ejercicio

En un estudio de tiempo efectuado en una planta manufacturera, el tiempo para completar una operación especificada se mide para cada uno de los $n = 40$ trabajadores. Se encuentra que la media y la desviación estándar son 12.8 y 1.7, respectivamente. Describa los datos muestrales usando la Regla empírica.

Regla empírica

Solución Para describir los datos, calcule estos intervalos:

$$(\bar{x} \pm s) = 12.8 \pm 1.7 \quad \text{o} \quad 11.1 \text{ a } 14.5$$
$$(\bar{x} \pm 2s) = 12.8 \pm 2(1.7) \quad \text{o} \quad 9.4 \text{ a } 16.2$$
$$(\bar{x} \pm 3s) = 12.8 \pm 3(1.7) \quad \text{o} \quad 7.7 \text{ a } 17.9$$

De acuerdo con la Regla empírica, se espera que aproximadamente 68 % de las mediciones caigan en el intervalo de 11.1 a 14.5, aproximadamente 95 % caigan en el intervalo de 9.4 a 16.2, y aproximadamente 99.7 % caigan en el intervalo de 7.7 a 17.9.

Ejercicio

Los maestros-estudiantes son capacitados para desarrollar planes de lecciones, en la suposición de que el plan escrito les ayudará a trabajar de manera satisfactoria en el salón de clases. En un estudio para evaluar la relación entre planes de lección escritos y su implementación en el salón de clases, se calificaron 25 planes de lección en una escala de 0 a 34 de acuerdo a una Lista de verificación de Plan de lección. Las 25 calificaciones se muestran en la tabla. Use el teorema de Chebyshev y la Regla empírica (si es aplicable) para describir la distribución de estas calificaciones de evaluación.

Calificaciones para evaluación de Plan de lección

26.1	26.0	14.5	29.3	19.7
22.1	21.2	26.6	31.9	25.0
15.9	20.8	20.2	17.8	13.3
25.6	26.5	15.7	22.1	13.8
29.0	21.3	23.5	22.1	10.2

Mediciones de posición relativa

A veces es necesario conocer la posición de una observación respecto a otras de un conjunto de datos.

Definición

El **puntaje z muestral** es una medida de posición relativa definida por

$$\text{puntaje } z = \frac{x - \bar{x}}{s}$$

Un puntaje z mide la distancia entre una observación y la media, medidas en unidades de desviación estándar.

Mediciones de posición relativa

Por ejemplo, suponga que la media y desviación estándar de los puntajes de examen (basados en un total de 35 puntos) son 25 y 4, respectivamente. El puntaje z para una calificación de 30 se calcula como sigue:

$$\text{puntaje } z = \frac{x - \bar{x}}{s} = \frac{30 - 25}{4} = 1.25$$

El puntaje de 30 está a 1.25 desviaciones estándar arriba de la media ($30 = \bar{x} + 1.25s$).

Mediciones de posición relativa

De acuerdo con el teorema de Chebyshev y la Regla empírica,

- al menos 75% y más probablemente 95% de las observaciones están a no más de dos desviaciones estándar de su media: sus puntajes z están entre -2 y $+2$. *Las observaciones con puntajes z mayores a 2 en valor absoluto se presentan menos del 5% del tiempo y son consideradas un tanto improbables.*
- al menos 89% y más probablemente 99.7% de las observaciones están a no más de tres desviaciones estándar de su media: sus puntajes z están entre -3 y $+3$. *Las observaciones con puntajes z mayores a 3 en valor absoluto se presentan menos del 1% del tiempo y son consideradas muy poco probables.*

Percentil

Un percentil es otra medida de posición relativa y se usa con más frecuencia para conjuntos grandes de datos. (Los percentiles no son muy útiles para conjuntos pequeños de datos).

Definición

Un conjunto de n mediciones de la variable x se ha reacomodado en orden de magnitud. El p-ésimo percentil es el valor de x que es mayor a $p\,\%$ de las mediciones y es menor que el restante $(100 - p)\,\%$.

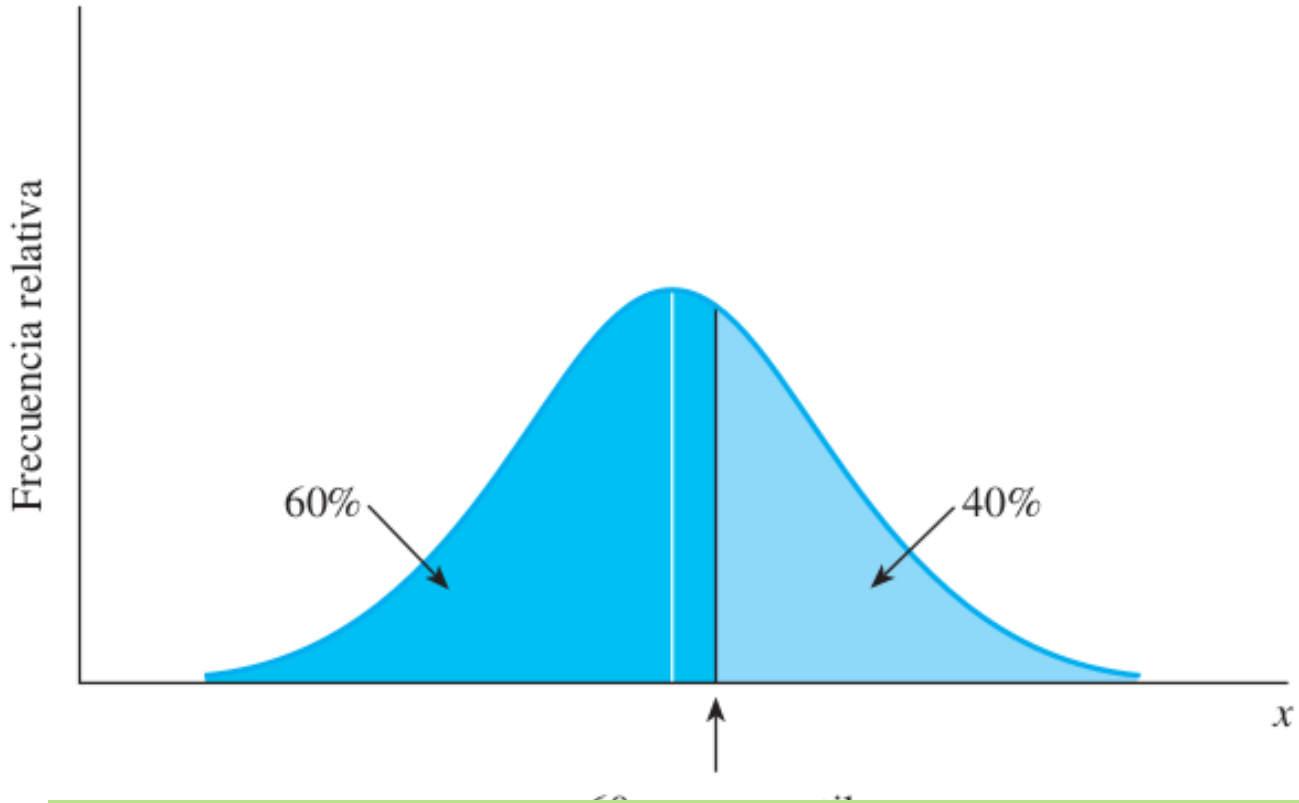
Percentil

Supongamos que usted ha sido notificado que su calificación de 610, en el Examen verbal de graduación, lo ha colocado en el 60avo percentil en la distribución de calificaciones. ¿Dónde está su calificación de 610 en relación a las calificaciones de los otros que tomaron el examen?

Solución Calificar en el 60avo percentil significa que 60 % de todas las calificaciones de examen fueron más bajas que la calificación de usted y 40 % fueron más altas.

Percentil

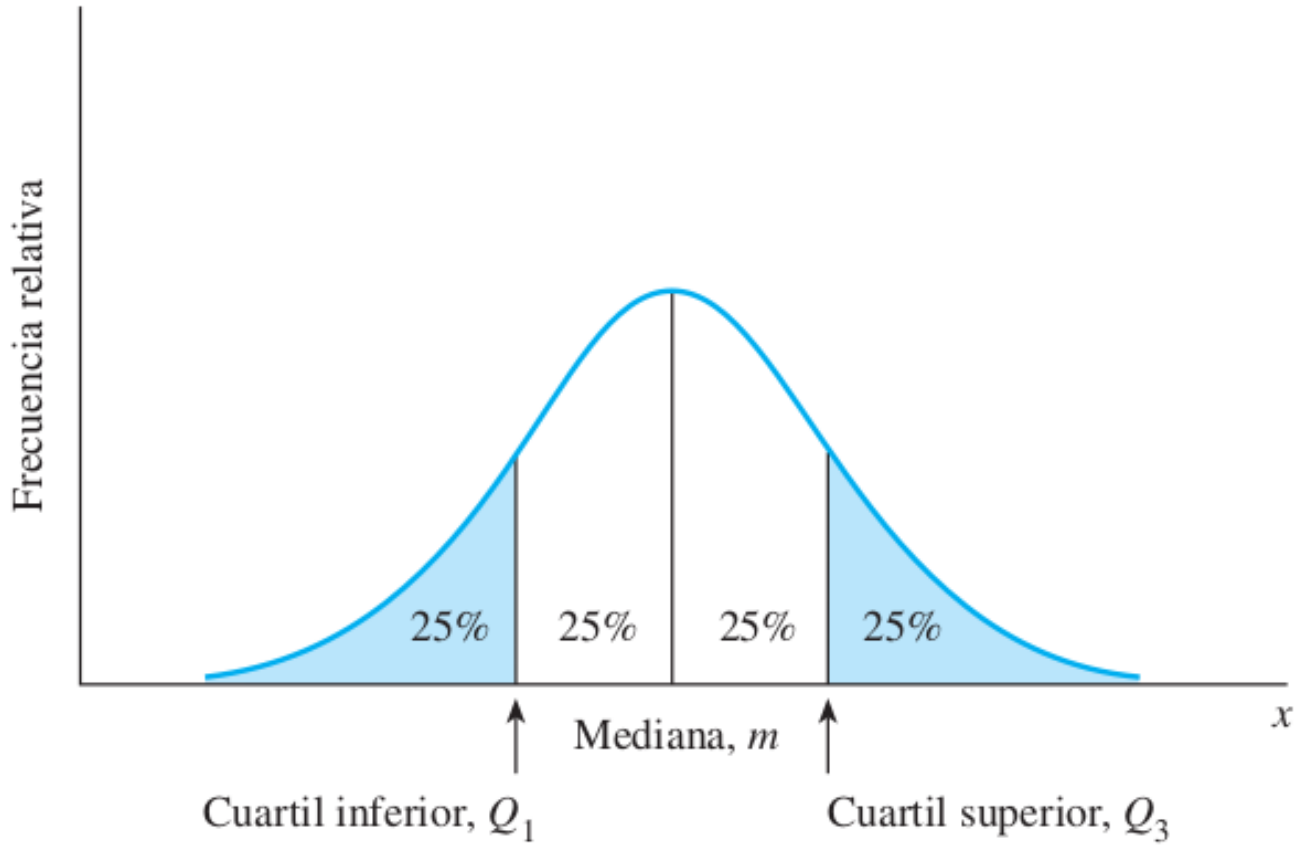
En general, el 60avo percentil para la variable x es un punto en el eje horizontal de la distribución de datos que es mayor a 60 % de las mediciones y menor que las otras.



Percentil-Cuartil

- La mediana es igual que el 50avo percentil.
- Los percentiles 25avo y 75avo, llamados cuartiles inferior y superior
- Veinticinco por ciento de las mediciones serán menores que el cuartil inferior (primero).
- 50 % serán menores que la mediana (el segundo cuartil).
- 75 % serán menores que el cuartil superior (tercero).

Cuartil


Frecuencia relativa
25%
25%
25%
25%
x
Mediana, m
Cuartil inferior, Q_1
Cuartil superior, Q_3

Cuartil

Definición

Un conjunto de n mediciones en la variable x se ha acomodado en orden de magnitud. El **cuartil inferior (primer cuartil),** Q_1 , es el valor de x que es mayor a un cuarto de las mediciones y es menor que los restantes tres cuartos. El **segundo cuartil** es la mediana. El **cuartil superior (tercer cuartil),** Q_3 , es el valor de x que es mayor a tres cuartos de las mediciones y es menor que el restante un cuarto.

Cuartil

CÁLCULO DE CUARTILES MUESTRALES

- Cuando las mediciones están dispuestas en orden de magnitud, el **cuartil inferior**, Q_1, es el valor de x en la posición $.25(n + 1)$, y el **cuartil superior**, Q_3, es el valor de x en la posición $.75(n + 1)$.
- Cuando $.25(n + 1)$ y $.75(n + 1)$ no son enteros, los cuartiles se encuentran por interpolación, usando los valores de las dos posiciones adyacentes.[†]

Cuartil

Encuentre los cuartiles inferior y superior para este conjunto de mediciones:

16, 25, 4, 18, 11, 13, 20, 8, 11, 9

Solución Ordene las $n = 10$ mediciones de menor a mayor:

4, 8, 9, 11, 11, 13, 16, 18, 20, 25

Calcule

$$\text{Posición de } Q_1 = .25(n + 1) = .25(10 + 1) = 2.75$$
$$\text{Posición de } Q_3 = .75(n + 1) = .75(10 + 1) = 8.25$$

Como estas posiciones no son enteros, el cuartil inferior se toma como el valor 3/4 de la distancia entre la segunda y tercera mediciones ordenadas, y el cuartil superior se toma como el valor 1/4 de la distancia entre la octava y novena mediciones ordenadas. Por tanto,

$$Q_1 = 8 + .75(9 - 8) = 8 + .75 = 8.75$$

y

$$Q_3 = 18 + .25(20 - 18) = 18 + .5 = 18.5$$

Intercuartil

Como la mediana y los cuartiles dividen la distribución de datos en cuatro partes, cada una de ellas conteniendo alrededor de 25 % de las mediciones, Q_1 y Q_3 son las fronteras superior e inferior para el 50 % central de la distribución. Podemos medir el rango de este "50 % central" de la distribución usando una medida numérica llamada rango intercuartil.

Definición

El rango intercuartil (IQR) para un conjunto de mediciones es la diferencia entre los cuartiles superior e inferior; esto es,
$IQR = Q_3 - Q_1$.

El resumen de cinco números y la gráfica de caja

El **resumen de cinco números** consta del número más pequeño, el cuartil inferior, le mediana, el cuartil superior, y el número más grande, presentados en orden de menor a mayor:

Min Q_1 **Mediana** Q_3 **Max**

Por definición, un cuarto de las mediciones del conjunto de datos se encuentre entre cada uno de los cuatro pares adyacentes de números.

El resumen de cinco números y la gráfica de caja

PARA CONSTRUIR UNA GRÁFICA DE CAJA

- Calcule la mediana, los cuartiles superior e inferior y el IQR para el conjunto de datos.
- Trace una recta horizontal que represente la escala de medición. Forme una caja un poco arriba de la recta horizontal con los extremos derecho e izquierdo en Q_1 y Q_3. Trace una recta vertical que pase por la caja en la ubicación de la mediana.

Una gráfica de caja se muestra en la figura 2.17.