Claudio Bonechi (clavicordo)

EY BIGNAMI - STATISTICA E PROBABILITÀ - 1A PARTE

13 August 2020

Premessa

Tutti sappiamo che la statistica e la probabilità sono legate tra loro: ma in che modo?
Certamente non sono la stessa cosa, ma spesso vengono interscambiate, il che può creare confusione.

Allora mi sono proposto di fare una breve illustrazione dei due concetti e del loro collegamento, creando un'occasione di ripasso o di prima visione di argomenti dai quali siamo sempre più circondati. Mi sono basato soprattutto su testi anglosassoni in uso nelle scuole superiori, disponibili online, perché mi è parso che avessero il mio stesso obiettivo: mostrare gli argomenti in modo il più possibile semplice, con esempi spiegati nel dettaglio, senza tralasciare né dare niente per scontato, salvo "le quattro operazioni" o poco più.
La maggior parte degli esempi è riportata nell'originale inglese e me ne scuso con chi lo conosce poco (ma chi bazzica l'elettronica e l'elettrotecnica è difficile che non lo conosca per niente). Anche se ritengo che sia un inglese accessibile a tutti, ho spesso inserito tra () la traduzione dei termini meno immediati.

Per evitare articoli troppo lunghi ho suddiviso la presente trattazione in 3 parti:

1. Statistica

2. Teoria (elementare) della probabilità

3. Variabili casuali discrete e continue

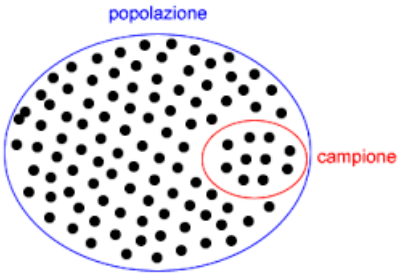
Introduzione: Statistica e Teoria della probabilità

Che cos'è la Statistica

- La statistica è la **scienza dei dati**. Comprende strumenti usati per organizzarli in modi significativi e per analizzarli. La statistica è *una metodologia e non una teoria*: lo scopo non è quello di spiegare ma di descrivere, analizzare ed eventualmente fornire elementi per assumere decisioni. L'**obiettivo** è quello di comprendere le **caratteristiche** di **raccolte di dati**; ciascuna caratteristica è depositaria di un'informazione che si rapporta all'intera raccolta in esame.

- I **dati** sono **simboli** che rappresentano le **informazioni** raccolte. I dati possono essere **qualitativi** (per esempio i colori) o **quantitativi**, descritti tramite **variabili** (dette anche "**caratteri**"), che assumono **valori qualitativi o quantitativi.**

- I dati, sia qualitativi che quantitativi, possono essere **discreti** o **continui** e descritti con variabili a loro volta discrete o continue.
- L'**insieme (set) dei dati** da analizzare ha due definizioni:
    - **Popolazione** (o insieme ***Universo***), se l'insieme comprende **tutte le unità statistiche** omogenee rispetto ad una caratteristica comune: se ad esempio la caratteristica comune è l'età, le unità statistiche sono tutte le persone di un insieme predefinito, ad esempio quelle che abitano una provincia. Ma il termine "popolazione" si riferisce a oggetti di qualunque tipo, non solo alle persone.
    - **Campione**: se il collettivo in esame costituisce **un sottoinsieme della popolazione** di riferimento. Tipicamente il campionamento statistico è **aleatorio** e segue criteri definiti.



Popolazione-campione

- Si parla di *variabili semplici* il cui oggetto è **una sola caratteristica** specifica e di *variabili multiple* — doppie, triple, ecc. — i cui oggetti sono più caratteristiche. Esempi di **caratteri** su un gruppo o su un campione di persone sono: altezza, età, colore degli occhi, genere, segno zodiacale, credo religioso ecc.
- Una variabile quantitativa **discreta** ha **valori numerici esatti**
- Una variabile quantitativa **continua** può assumere valori numerici la cui **accuratezza** e **precisione** dipendono dagli strumenti che li generano.
- Una popolazione (o un campione) viene suddivisa in "**unità statistiche**", ossia individui o enti sui quali viene effettuata una rilevazione statistica, raccogliendone dati significativi.
- La statistica ha due accezioni:

- "**descrittiva**", che "concerne il trattamento e l'esposizione razionalmente ordinata dei dati relativi a un fenomeno e la loro analisi"; si applica a una popolazione o a un suo campione.
- "**inferenziale**" (o "induttiva"), che parte da un campione aleatorio (ovvero casuale) per descriverne le proprietà statistiche oppure **risalire** o inferire al modello probabilistico applicato a una popolazione; questa parte della statistica è collegata più direttamente alla teoria della probabilità.

Che cos'è la Teoria della probabilità

La teoria della probabilità si occupa di fornire **modelli** matematici ovvero "**distribuzioni di probabilità**" adattabili ai vari **fenomeni aleatori** (o *casuali*) reali, definendo i parametri della variabile aleatoria in questione.

La **probabilità** è un **numero** espresso in **percentuale** (o **compreso tra 0 e 1**) che viene associato a un **carattere** collegato a una variabile aleatoria.

Statistica

Statistica descrittiva

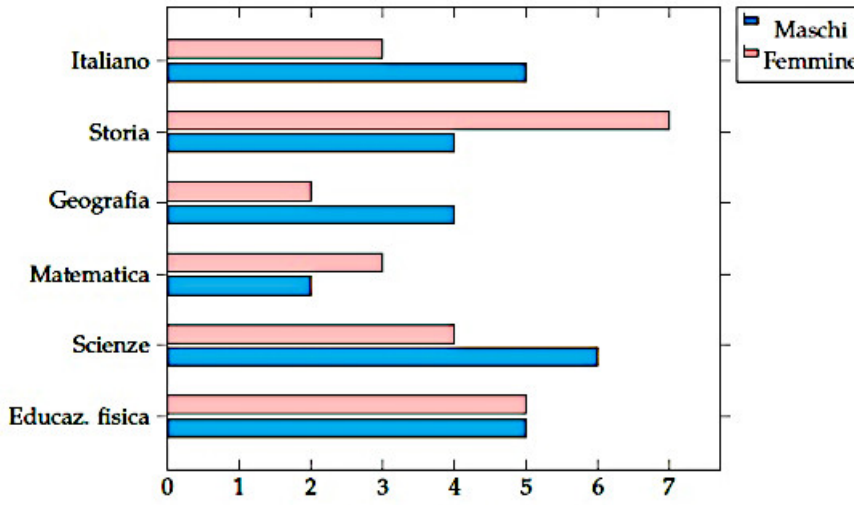
La statistica Lo **scopo** della statistica descrittiva è **sintetizzare i dati** attraverso strumenti grafici (diagrammi a barre, a torta, istogrammi, etc.) e **indicatori statistici** che sono di **posizione** (come la *media*, la *mediana*, la *moda*, etc.), di **dispersione** (come la *varianza* e lo *scarto interquartile*), di **concentrazione**, di **correlazione**, di **forma**.

Altri indicatori si usano nella pratica per descrivere vari contesti come i **tassi** (di natalità, di mortalità, di immigrazione, di inflazione, etc.).

I dati vengono raggruppati in **variabili statistiche** dalle quali vengono calcolati gli **indicatori**.

Strumenti grafici e **indicatori** descrivono gli aspetti salienti dei dati osservati, formando così il **contenuto statistico**.

Esempio di grafico statistico per la rappresentazione dei dati raccolti: voti per materie e sesso di una classe di studenti:



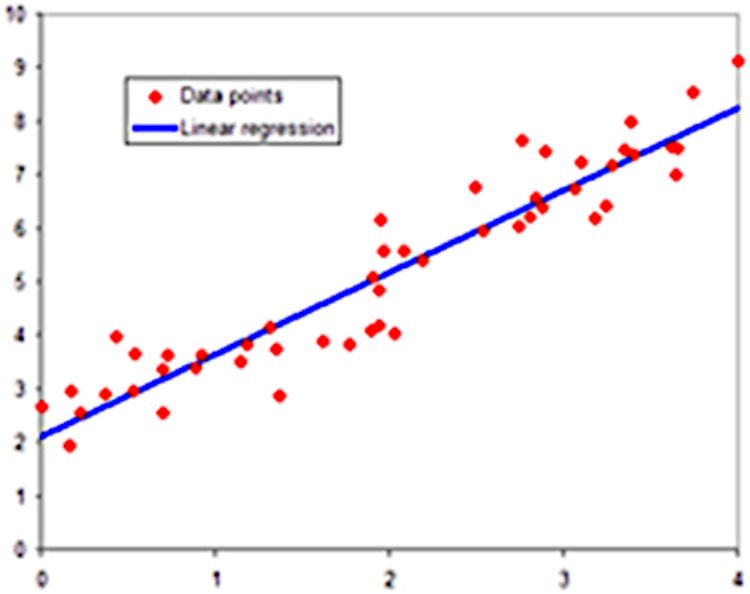
Statistica inferenziale

La statistica inferenziale (o "induttiva") utilizza i dati statistici, anche opportunamente sintetizzati dalla statistica descrittiva, per fare **previsioni di tipo probabilistico** su situazioni **future** o comunque **incerte**. Ad esempio esaminando un piccolo campione estratto da una grande popolazione cerca di valutare la frazione della popolazione che possiede una certa caratteristica, ha un certo reddito o voterà per un certo candidato. Le inferenze possono riguardare la **natura teorica** (la legge probabilistica) del fenomeno che si osserva.

La conoscenza di questa natura permetterà poi di fare una **previsione** (si pensi, ad esempio, che quando si dice che "l'inflazione il prossimo anno avrà una certa entità" deriva dal fatto che esiste un modello dell'andamento dell'inflazione derivato da tecniche inferenziali). La statistica inferenziale è quindi **fortemente legata alla teoria della probabilità.**

Sotto questo punto di vista descrivere in termini probabilistici o statistici un **fenomeno aleatorio nel tempo**, caratterizzabile dunque da una variabile aleatoria, vuol dire descriverlo in termini di **densità di distribuzione di probabilità** e dei suoi parametri come **media** e **varianza**, argomenti sui quali torneremo.

Esempio di inferenza (regressione lineare) dedotta a partire da un insieme di dati.



La retta blu è ottenuta **interpolando** i dati con un metodo matematico definito, il più conosciuto dei quali è detto "**dei minimi quadrati**". Esistono anche altre funzioni di interpolazione, la cui scelta dipende dalla forma della distribuzione dei dati e dagli obiettivi di impiego.

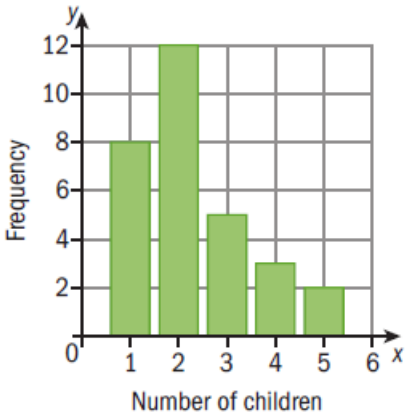
Aspetti generali

Frequenza statistica

Il punto di passaggio tra statistica e Teoria della Probabilità è il concetto di **frequenza**, che è un concetto **diverso** da quello a cui siamo abituati in elettrotecnica e in elettronica. Il concetto di frequenza in statistica si applica sia alla statistica descrittiva che a quella inferenziale.

In statistica la **frequenza assoluta f** è il **numero delle volte che un certo "carattere" di fenomeno osservato si presenta**. Ad esempio, data una classe di 30 studenti, rappresentiamo il carattere "numero di figli" per famiglia:

Children	f
1	8
2	12
3	5
4	3
5	2



I **caratteri** si possono rappresentare in una **tabella** oppure in un "**istogramma**" (diagramma a barre), che è molto intuitivo.

E' chiaro che, in questo caso, la somma delle frequenze deve essere 30, pari al numero degli studenti.

Quella appena considerata è una **popolazione intera** (detta anche **Universo**) e non ha nessuna casualità, nessuna incertezza, nessuna possibilità di errore.

Ma quando la popolazione è molto numerosa e la massa dei dati aumenta considerevolmente, diventa conveniente utilizzare un **campione della popolazione**, campione che deve essere scelto con criteri tali che lo rendano **rappresentativo** della popolazione. Ci si avvale allora della statistica inferenziale. Ossia, dall'esame di un campione deduciamo, sia pure con un certo margine di errore, il comportamento dell'intera popolazione. **Non trattiamo però dei criteri di scelta del campione**, di quanto deve essere ampio per fornire informazioni utili: questo è un capitolo della statistica piuttosto complesso, di cui non è il caso di parlare qui.

Oltre alla frequenza assoluta **f**, si definisce anche la **frequenza relativa**, come **rapporto** tra il numero dei casi favorevoli (associati ai relativi valori di un carattere) e il numero totale dei casi osservati**.**

Nell'esempio appena visto, le frequenze relative sono, per ogni valore del carattere, le seguenti:

N. figli	Frequenza assoluta	Frequenza relativa	decimale	percentuale
1	8	8/30	0,266 (periodico)	26,6%
2	12	12/30	0,400	40%
3	5	5/30	0,166 (periodico)	16,6%
4	3	3/30	0,100	10%
5	2	2/30	0,066 (periodico)	6%
totale	30	30/30	1	100%

Se si passa da una popolazione a un campione, la frequenza relativa diventa probabilità.

Ma prima cominciamo a parlare delle raccolte di dati e alla loro analisi un po' più in dettaglio.

L'analisi dei dati detta "**univariata**" coinvolge una singola variabile, ad esempio l'altezza

di tutti gli studenti di una classe. L'analisi **bivariata** e **multivariata**, coinvolge due o più variabili statistiche

Cominciamo con l'analisi **univariata**

Indicatori di posizione (central tendency).

I più usati sono: **moda** (mode), **media** (mean) e **mediana** (median)

Moda: in un insieme di dati, è il dato che compare **il maggior numero di volte**.

Find the mode of: 9, 3, 9, 41, 17, 17, 44, 15, 15, 15, 27, 40, 13

Answer

The mode is 15 (15 occurs the most at 3 times).

La moda può non essere unica:

3, 5, 5, 7, 9, 9, 2, 1, 8

qui la moda ha 2 valori: 5 e 9.

Quando un carattere è in forma di "**fascia di valori**" (class) la moda diventa "**classe modale**" (modal class) come nel caso *Times* dell'esempio

Find the mode or modal class for these frequency tables.

a

Goals	Frequency
0	4
1	7
2	3
3	3
4	1

b

Times	Frequency
$0 \le t < 5$	1
$5 \le t < 10$	5
$10 \le t < 15$	6
$15 \le t < 20$	7
$20 \le t < 25$	6

Answers

a The mode is 1 goal.

b The modal class is $15 \le t < 20$.

Common errors:

1 *The mode is 7.*
No. *The biggest frequency is 7.*

2 *The mode is 3.*
No. *The most common frequency is 3.*

The mode from a grouped frequency table is called the modal class.

Mode or modal class.png

La moda è costituita da quel/quei carattere/i (o classe/i modale/i) **la cui frequenza è più alta di tutte**.

Media (mean): con questo termine si intende la *media aritmetica*dei dati, ossia la loro somma divisa per la loro numerosità. Si indica con la lettera greca **μ**. La media indica il "punto centrale" (mid point) dell'insieme di dati.

Find the mean of

a 89, 73, 84, 91, 87, 77, 94

b 2, 3, 3, 4, 6, 7

Answers

a $\mu = \frac{\sum x}{N} = \frac{89+73+84+91+87+77+94}{7}$

$= \frac{595}{7} = 85$

b $\mu = \frac{\sum x}{N} = \frac{2+3+3+4+6+7}{6} = \frac{25}{6} = 4.1\dot{6}$

In questo esempio, ogni dato compare una volta (in b il 3 compare due volte, ma viene conteggiato separatamente), ossia la sua frequenza assoluta è considerata = 1.

Se i dati invece vengono **suddivisi per frequenza**, la formula si "accorcia", moltiplicando prima ogni valore del dato per la relativa frequenza.

Se ogni suddivisione è **una fascia** di valori**, si usa la media interna alla fascia:**

Find the mean of each set of data displayed below.

a

Grade (g)	Frequency
0	11
1	10
2	19
3	10

b

Age (a)	Frequency
$10 \le t < 12$	4
$12 \le t < 14$	8
$14 \le t < 16$	5
$16 \le t < 18$	3

Answers

a

Grade (g)	Frequency	fg
0	11	0
1	10	10
2	19	38
3	10	30
Total	50	78

Add a 3rd column; fg means $f \times g$.

The total of the fg column is the sum of all of the grades.
The total of the f column is the number of grades.

$$\text{Mean} = \frac{\sum fg}{\sum f} = \frac{78}{50} = 1.56$$

b

Age (a)	f	Mid point (m)	fm
$10 \le t < 12$	4	11	44
$12 \le t < 14$	8	13	104
$14 \le t < 16$	5	15	75
$16 \le t < 18$	3	17	51
Total	20		274

When the data are grouped, we can calculate the mean by assuming that all of the data values are equally spread around the midpoint.

$$\text{Mean} = \frac{\sum fm}{\sum f} = \frac{274}{20} = 13.7$$

Si tratta quindi sempre di una media "pesata" e il peso è costituito dalla frequenza.

Mediana: la mediana è il valore del dato che sta in **posizione centrale**, *una volta che i dati siano stati ordinati in una successione di valori crescenti*. Si indica con "m".

Find the median of 2, 13, 7, 5, 19, 23, 39, 23, 42, 23, 14, 12, 55, 23, 29.	
Answer 2, 5, 7, 12, 13, 14, 19, 23, 23, 23, 23, 29, 39, 42, 55 The median value of this set of numbers is **23**.	*Write the numbers in order.* *There are **15** numbers. Our middle number will be the **8**th number.*

Se il numero n dell'insieme di dati $\{a_n\}$ è pari, si suddivide l'insieme in due metà $a_1 - a_k$ e $a_{k+1} - a_n$, dove k=n/2 e la mediana è $(a_k + a_{k+1})/2$. Esempio:

$\{a_n\} = 1, 3, 8, 7, 4, 9 \rightarrow 1, 3, 4, 7, 8, 9$

n=6, k=3, m = (4+7)/2 = 5,5

Se il numero n è grande, conviene calcolare la posizione della mediana con n(n+1)/2

Nota: mentre la media è influenzata dai valori estremi della successione, la moda e la mediana non lo sono.

Indicatori di dispersione (Measures of dispersion)

Misurano lo **scostamento** dagli indicatori centrali

I più usati sono: **range**, **quartili** (quartiles), **Frequenza cumulativa** (cumulative frequency), **varianza e deviazione standard** (variance and standard deviation).

Eccettuata la varianza-deviazione standard, questi indicatori si calcolano **dopo** aver ordinato i dati in successione crescente.

Range: in una successione crescente è la differenza tra il valore più grande e quello più piccolo.

Es. in 2, 4, 5, 13, 24, 31 il **range** è 31- 2 = 29

Quartili

> in una **successione crescente** i quartili dividono i dati in 4 parti (ognuna ampia il 25%), dove i primi 2 quartili sono separati dagli altri 2 dalla mediana.

First quartile	The **first quartile** is the value one-quarter of the way into the data. One quarter of the data lies below the first quartile and three-fourths lies above. It is also called the 25th percentile and often has the symbol Q_1.
Second quartile	The **second quartile** is another name for the median of the entire set of data and is also called the 50th percentile.
Third quartile	The **third quartile** is three-quarters of the way in. Three-fourths of the data lies below the third quartile and one-fourth lies above. It is also called the 75th percentile and has the symbol Q_3.

$Q_1 = \frac{1}{4}(n+1)$th value and $Q_3 = \frac{3}{4}(n+1)$th value where n is the number of data values in the data set.

I quartili danno una sensazione sintetica della **distribuzione** dei dati.

Ad esempio, considerando le votazioni di un esame per una classe numerosa di studenti bravi:

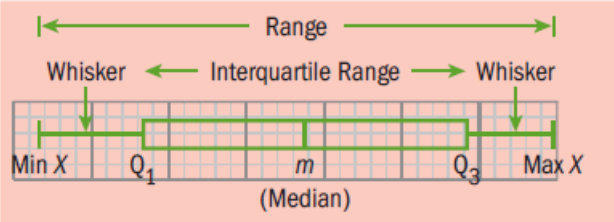
Minimum	First quartile	Median	Third quartile	Maximum
65	70	80	90	100

f13 quartili es.png

Se non si conoscono i singoli punteggi, la tabella dice che metà dei punteggi sono sotto 80 e l'altra metà sopra 80. Inoltre, dice che il 50% dei punteggi sono compresi tra 70 e 90. Il range è 35.

La differenza tra il terzo e il primo quartile è chiamata "**Range Interquartile**" (Interquartile Range o **IQR**) = Q3 – Q1

Nell'esempio IQR = 90 – 70 = 20



La rappresentazione soprastante è detta "box and whiskers" perché sembra una scatola con baffi.

I valori estremi, distanti almeno 1,5 volte IQR, si chiamano "**valori anomali**" (outliers).

a Find the range, the median, the lower quartile, the upper quartile and the interquartile range of this set of scores.
18, 27, 34, 52, 54, 59, 61, 68, 78, 82, 85, 87, 91, 93, 100
b Show the data in a box and whisker plot.
c Check if 18 is an outlier.

Answers

a Range = 100 – 18 = 82 — *Range = largest value – smallest value*

18, 27, 34, 52, 54, 59, 61, 68, 78, 82, 85, 87, 91, 93, 100 — *Write the data in order.*

Median

$= \left(\frac{n+1}{2}\right)\text{th} = \left(\frac{15+1}{2}\right)\text{th value}$ — *There are 15 numbers in the data set.* $\therefore n = 15$.

$= 8\text{th value} = 68$

$Q_1 = \frac{1}{4}(n+1)\text{th value}$

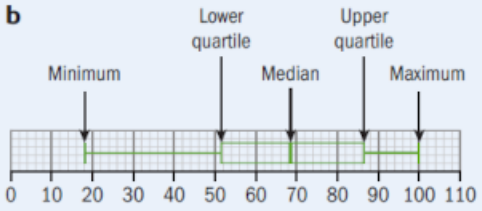
$= \frac{1}{4}(15+1)\text{th} = 4\text{th value} = 52$

$Q_3 = \frac{3}{4}(n+1)\text{th value}$

$= \frac{3}{4}(15+1)\text{th} = 12\text{th value} = 87$

$\text{IQR} = Q_3 - Q_1 = 87 - 52 = 35$

b



c $Q_1 - 1.5(\text{IQR}) = 52 - 1.5(35) = 52 - 52.5 = -0.5$ — *Outliers are more than 1.5 × IQR below Q_1 or above Q_3.*

$\therefore$ 18 is not an outlier.

Frequenza cumulativa: è la somma delle frequenze precedenti a quella del dato in esame.

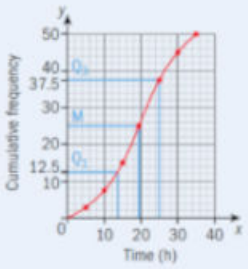
L'esempio sottostante (importante) è riferito alla durata di un insieme di batterie.

50 batteries were tested to see how long they lasted. The results (in hours) are shown in the table. Draw a cumulative frequency diagram and find the median and interquartile range.

Time (h)	f
$0 \le h < 5$	3
$5 \le h < 10$	5
$10 \le h < 15$	8
$15 \le h < 20$	10
$20 \le h < 25$	12
$25 \le h < 30$	7
$30 \le h < 35$	5

Answer

Time (h)	f	Cumulative frequency
$0 \le h < 5$	3	3
$5 \le h < 10$	5	8
$10 \le h < 15$	8	16
$15 \le h < 20$	10	26
$20 \le h < 25$	12	38
$25 \le h < 30$	7	45
$30 \le h < 35$	5	50



Median = 19 hours

IQR = (25 – 13) hours = 12 hours

Add a 'cumulative frequency' column to the table. Work out the cumulative frequency by adding up as you go.

f	Cumulative frequency
3	3
5	3 + 5 = 8
8	3 + 5 + 8 = 16
10	3 + 5 + 8 + 10 = 26
12	3 + 5 + 8 + 10 + 12 = 38
7	3 + 5 + 8 + 10 + 12 + 7 = 45
5	3 + 5 + 8 + 10 + 12 + 7 + 5 = 50

3 batteries lasted less than 5 hours

8 batteries lasted less than 10 hours

38 batteries lasted less than 25 hours

Plot the upper limit of the time intervals against the cumulative frequencies. The first two points are (5, 3) and (10, 8)

$n = 50$

Median $= \frac{50}{2} = 25$*th data value*

Draw a line across from 25 on the vertical axis to the graph then down to the time axis.

Read off Q_3 *and* Q_1 *from the graph in the same way.*

$Q_3 = 25$, $Q_1 = 13$

$IQR = Q_3 - Q_1$

For large data sets, the median is the $\frac{n}{2}$ th value.

Values of the median and quartiles from a GDC may be different from values read from a cumulative frequency graph.

Varianza e deviazione standard (SD): Rispetto a IQR costituiscono una misura **più significativa** della dispersione di un insieme di dati rispetto alla **media μ**. Si indicano rispettivamente con σ^2 e **σ**.

La **varianza** è la **media dei quadrati** delle differenze tra ciascun dato e la media dei dati**. La deviazione standard non è altro che la radice quadrata della varianza. In questo modo ha la stessa unità di misura del dato.**

L'elevazione al quadrato ha **due vantaggi**: dà più peso alle differenze maggiori e rende tutti i termini positivi, evitando la cancellazione rispetto alla media.

→ The formulae for the variance and standard deviation are:

$$\sigma^2 = \text{Population variance} = \frac{\sum_{i=1}^{n}(x-\mu)^2}{n}$$

$$\sigma = \text{Population standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x-\mu)^2}{n}}$$

Thirty farmers were asked how many farm workers they hire during a typical harvest season. Their responses were:

4, 5, 6, 5, 3, 2, 8, 0, 4, 6, 7, 8, 4, 5, 7, 9, 8, 6, 7, 5, 5, 4, 2, 1, 9, 3, 3, 4, 6, 4

Calculate the mean and standard deviation for this data.

Answer

Solution – 'by hand'

Workers (x)	Frequency (f)	(fx)	($x-\mu$)	$(x-\mu)^2$	$f(x-\mu)^2$
0	1	0	−5	25	25
1	1	1	−4	16	16
2	2	4	−3	9	18
3	3	9	−2	4	12
4	6	24	−1	1	6
5	5	25	0	0	0
6	4	24	1	1	4
7	3	21	2	4	12
8	3	24	3	9	27
9	2	18	4	16	32
	30	150			152

To calculate the mean:

$\mu = \frac{150}{30} = 5$

To calculate the standard deviation:

$\sigma = \sqrt{\frac{152}{30}} = 2.25$

The IB syllabus covers 'Calculation of standard deviation/variance using only technology'. This is how you would calculate the standard deviation for a discrete variable 'by hand'.

1. *Calculate the mean.*
2. *Subtract the mean from each observation.*
3. *Square each of the results from step 2.*
4. *Add these squared results together.*
5. *Divide this total by the number of observations. This gives the variance, σ^2.*
6. *Take the positive square root to get the standard deviation, σ.*

$\mu = \frac{\sum fx}{n}$

$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$

Bassi/alti valori di SD indicano che la maggior parte dei dati si discosta **poco/molto** dalla media.

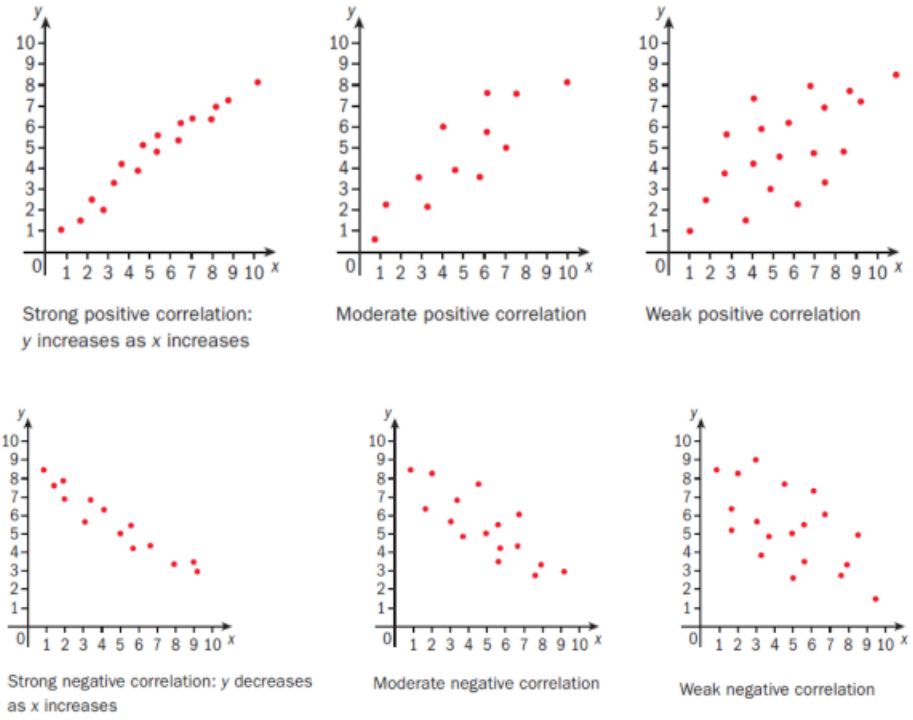
Indicatori di Correlazione

Un **grafico di dispersione** o **scatter plot** è un tipo di grafico in cui **due variabili** di un set di dati, ad esempio X e Y, sono riportate su uno spazio cartesiano.

I dati sono visualizzati tramite una collezione di punti ciascuno con una posizione sull'asse orizzontale determinato da una variabile (X) e sull'asse verticale determinato dall'altra (Y).

La **correlazione è una misura del grado di associazione** tra due variabili: è l'unico caso che vediamo qui di **analisi bivariata**.

In figura esempi di **correlazione lineare**: **positiva** e **negativa**, rispettivamente *forte*, *moderata* e *debole*.



Non bisogna confondere la **correlazione** con la **causalità**. Ecco un esempio: Osserviamo che nei giorni di pioggia molte persone portano l'ombrello: c'è un'alta correlazione tra le due variabili. Potremmo aprire un blog di informazione alternativa e dire alla gente di aprire gli occhi, perché c'è un complotto che funziona così: gli ombrelli causano la pioggia e per colpa degli ombrelli che causano la pioggia tanta gente rimane a casa a farsi indottrinare dalla TV di regime. (ovviamente la causalità qui è quella che la pioggia causa il portare con sé l'ombrello). Ci sono anche studi correlazionali che dimostrano una correlazione fra numero di nati in una città ucraina e il numero di cicogne in volo!

Coefficiente di correlazione: è il più importante indicatore di correlazione ed è l'unico che consideriamo qui ed è riportato con la lettera "**r**".

Il coefficiente di correlazione di due insiemi di dati X e Y si trova con la **formula di Pearson**. Esso varia tra -1 e +1.

→ The formula for finding Pearson's correlation coefficient is

$$r = \frac{S_{xy}}{S_x S_y}$$

where

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n},\ S_x = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \text{ and}$$

$$S_y = \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}.$$

→ A quick way to interpret the *r*-value is:

r-value	Correlation
$0 < \lvert r \rvert \leq 0.25$	Very weak
$0.25 < \lvert r \rvert \leq 0.5$	Weak
$0.5 < \lvert r \rvert \leq 0.75$	Moderate
$0.75 < \lvert r \rvert \leq 1$	Strong

Esempio

Sara vuole determinare la correlazione tra numero di cucchiai di concime vegetale (x) che utilizza e numero extra di orchidee (y) che si ottiene da una delle 4 piante A, B, C, D. Così usa la formula di Pearson per trovare il coefficiente di correlazione e interpretare la relazione tra concime e orchidee.

Plant	x	y	xy	x^2	y^2
A	1	2	2	1	4
B	2	3	6	4	9
C	3	8	24	9	64
D	4	7	28	16	49
Total	10	20	60	30	126

Answer

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$
$$= 60 - \frac{10 \times 20}{4} = 10$$

$$S_x = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}$$
$$= \sqrt{30 - \frac{10^2}{4}} = \sqrt{5}$$

$$S_y = \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}$$
$$= \sqrt{126 - \frac{20^2}{4}} = \sqrt{26}$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{10}{\sqrt{5}\sqrt{26}} \approx 0.877$$

Un **coefficiente di correlazione positivo** indica in questo esempio che aumentando il **numero di cucchiai di concime** aumenta anche il numero di **orchidee supplementari**. Il valore **r** di **0,877** indica **forte correlazione**, ossia relazione abbastanza stretta tra le due variabili.

Conclusione della prima parte

Abbiamo esaminato rapidamente alcuni concetti basilari della statistica molti dei quali si ritroveranno nella Teoria della probabilità, che occuperà la Seconda parte. In essa dovrebbe emergere il legame tra le due discipline, un legame non sempre chiaro, al punto che a volte vengono uniformate o sopvrapposte. Spero quindi che oltre al legame emerga anche la dovuta distinzione. Come si sarà notato, per motivi di semplicità non ho parlato degli indicatori di concentrazione e di forma.