Minicorso
Controllo Statistico di Processo

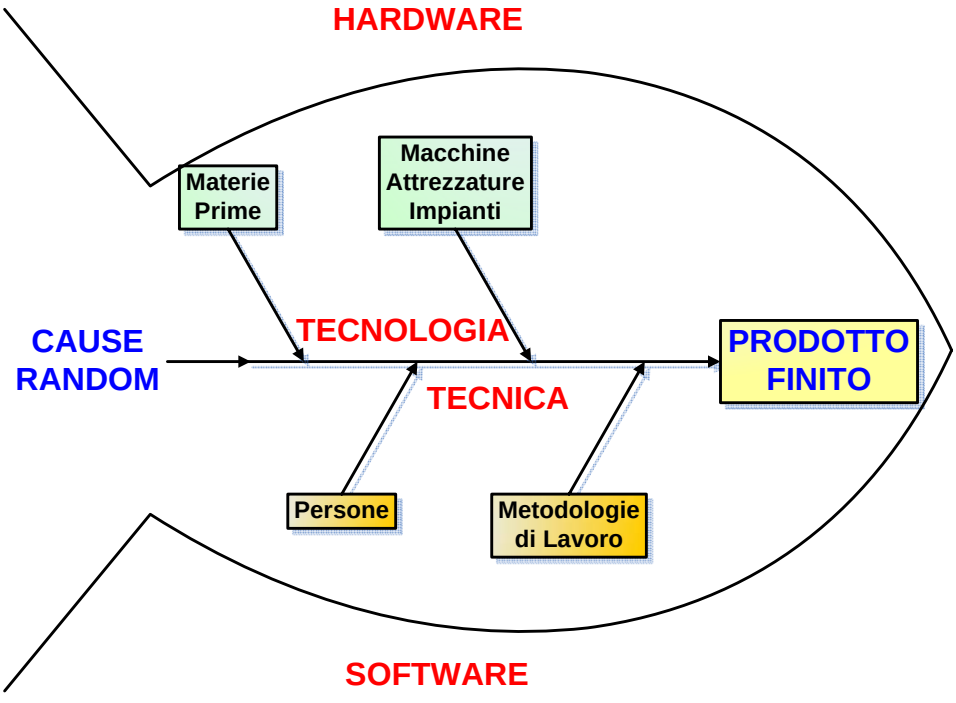
di Andrea Saviano

Parte 2

- Walter Andrew Shewhart, chi era costui, premessa
- Anche lei matematico, che combinazione!
- Probabilità... senza imprevisti
- Il 7 e ½ e altri giochi di "carte"
- Non poniamo limiti… solo alla divina Provvidenza
- Statistica, probabilità e… anomalie

Premessa

Walter Andrew Shewhart è conosciuto da tutti come il padre del **controllo statistico della qualità**, tuttavia pochi sanno che è stato anche il maestro di Edwards Deming (quello del **ciclo PDCA**). Il suo lavoro è stato essenzialmente correlato alle conoscenze nel campo della statistica, all'elaborazione della teoria delle cause comuni e a quella delle cause speciali.



Innanzitutto, bisogna individuare gli elementi casuali e causali che caratterizzano un processo. Si possono individuare tre tipologie di cause:

- **tecnologiche** (*hardware*), dovute alle caratteristiche delle materie prime impiegati o delle attrezzature, macchinari o impianti utilizzati;
- **tecniche** (*software*), dovute alle metodologie di lavoro applicate o alle persone che realizzano le attività;
- **casuali** (*random*), dovute alle normali oscillazioni casuali che esistono in ogni processo.

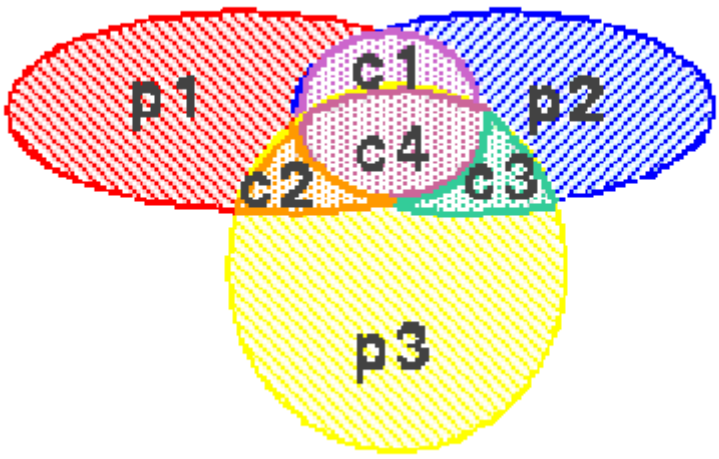
Il diagramma a spina di pesce utilizzato, volutamente divide gli ambiti tangibili, da quelli umani e intangibili, perché – come si vedrà più avanti – dalla parte bassa della "lisca" si originano le possibilità di realizzare dei **piani di controllo**, un'**analisi del modo e degli effetti con cui si presentano i difetti**, dei **piani di reazione**, mentre nella parte alta si generano sistemi di **controllo in accettazione** (materia prima) e **piani di manutenzione previsionali** (macchine, attrezzature e impianti). Come coda del pesce si pongono invece le normali oscillazioni casuali, perché la certezza non è altro che un'elevata probabilità o improbabilità che qualcosa accada.

Tangibile Vs intangibile

Spesso la questione della suddivisione degli elementi d'influenza genera una certa confusione, perché le metodiche vengono confuse con il loro campo d'applicazione, perdendo quella caratteristica d'intangibilità (non si può "toccare" il know-how) e di labilità soggettiva che invece gli elementi tangibili (quindi osservabili e misurabili) hanno. Una macchina è di per sé oggettiva, non ha percezione di chi sia il Cliente per cui sta lavorando, non cambia il suo modo di operare per questioni soggettive come l'avere il mutuo da pagare, non aver dormito la notte perché il bambino piangeva o aver subito un grave lutto (in questo caso l'ambito sono le persone).

Implicito Vs Esplicito

Le persone quando operano su macchine, impianti e attrezzature per attribuire valore aggiunto a materie prime e semilavorati utilizzano delle metodiche di lavoro che possono essere individuali, cioè patrimonio della persona (**professionalità**), o collettive, cioè patrimonio dell'azienda (**know-how**) oppure implicitamente diffuse. Consideriamo tre persone (**p1**, **p2**, **p3**), ognuna con il suo modo d'agire. Per esperienza o tradizione alcune metodiche sono tra loro condivise, seppur la cosa non sia esplicita (cioè non esistono procedure, istruzioni operative o istruzioni di lavoro scritte, non c'è stata una vera e propria formazione con addestramento e verifica).



Avendo la loro esperienza un'intersezione comune (**c4**) pur non esistendo esplicitamente un metodo esso, implicitamente esiste. L'area in comune è quindi metodo, mentre le aree non in comune rappresentano dal punto di vista causale la persona.

Testa o croce? Non c'è nulla di più certo del caso

Il caso, non essendo premeditato e non agendo in modo sistematico, ha una prerogativa rispetto a tutte le altre possibili cause: **risponde solo a modelli di tipo stocastico** alla cui base c'è la legge dei grandi numeri.
Tale legge si esplicita in due forme:

- forte, data una successione di n variabili casuali, all'aumentare di n, queste convergono su una medesima media μ e una medesima deviazione standard σ;
- debole, data una successione di n variabili casuali, all'aumentare di n, queste convergono in termini di probabilità alla medesima media μ.

Ciò garantisce che la media campionaria sia uno stimatore consistente della media di una popolazione; ciò equivale a dire che, grazie alla legge dei grandi numeri, la media che calcoliamo a partire da un numero sufficiente di campioni sia sufficientemente vicina alla media vera.
Questo non ci permette di prevedere l'esito del lancio di una monetina, né ci permette di prevedere le sequenze, ma ci assicura che, dopo un numero di lanci sufficientemente elevato, il numero di volte in cui sarà uscita una faccia piuttosto che l'altra sarà pressoché uguale.

Anche lei matematico, che combinazione!

Per poter parlare di statistica e di stocastica con un minimo di competenza è necessario avere almeno un'infarinatura di cosa siano:

- il calcolo combinatorio;
- i modelli di probabilità;
- gli indici di sintesi statistica.

Il calcolo combinatorio

Il calcolo combinatorio sviluppa degli strumenti matematici per valutare il numero di modi possibili in cui un certo numero di elementi, utilizzando precise regole, può essere disposto.
Questo, dal punto di vista pratico, permette di conoscere il numero di modi che 5 numeri, estratti da 90 (quindi tutti i numeri compresi tra 1 e 90) senza ripetizione, possono realizzare. Se si trattasse del lotto, il modo in cui questi numeri si ordinano non sarebbe rilevante perché la sequenza 21 34 52 13 65 sarebbe equivalente a quella ordinata 13 21 34 52 65, anche in questo caso il calcolo combinatorio ci viene in aiuto.

Permutazioni semplici o con ripetizioni

Si parla di **permutazioni**, quando si desidera conoscere il numero di modi in cui è possibile ordinare un certo numero di oggetti. Ad esempio: in quanti modi è possibile ordinare le lettere **A**, **B** e **C**?
Consideriamo i casi che si possono realizzare tenendo costante il primo elemento:

ABC, **ACB**
BAC, **BCA**
CAB, **CBA**

Si hanno in tutto 6 possibilità. Più in generale, le possibilità di ordinare n oggetti diversi sono:

$$P_n = n!$$

Dove $n!$ si legge: "n fattoriale" ed è così definito:

$$n! = 1 \cdot 2 \cdot \ldots \cdot n$$

Nel caso in oggetto: 3! = 1·2·3= 6 (c.v.d.).

*Dato un insieme di **n** elementi distinti,*
si dicono permutazioni di tali elementi
tutti i possibili diversi ordinamenti di tali elementi.

Quando invece gli ***n*** oggetti sono composti da n_a elementi uguali di tipo ***a***, n_b elementi uguali di tipo ***b*** e così via, tali che:

$$n = n_a + n_b + \ldots$$

Si parla di **permutazioni con ripetizioni** (senza nessun riferimento al fatto che servano delle lezioni supplementari per comprenderle a fondo) e si ha:

$$P_n^{n_a, n_b, \ldots} = \frac{n!}{n_a! \cdot n_b! \cdot \ldots}$$

Di particolare interesse la suddivisione in soli due gruppi, perché molti eventi vengono giudicati in maniera binaria 1/0 (sì/no, vero/falso, presente/assente, etc.) in cui:

$$P_n^{x, n-n} = \frac{n!}{x! \cdot (n-x)!} \Leftrightarrow \begin{cases} n_0 = x \\ n_1 = n - x \end{cases}$$

Disposizioni semplici

Si parla di **disposizioni**, quando si desidera conoscere il numero di modi in cui è possibile ordinare un certo numero di oggetti che provengono da un insieme più ampio. Ad esempio: in quanti modi è possibile ordinare 3 numeri utilizzando le cifre da 0 a 9?
Consideriamo i casi che si possono realizzare tenendo costante come primo elemento lo 0:

012, 013, 014, 015, 016, 017, 018, 019
021, 023, 024, 025, 026, 027, 028, 029
…
…
…
…
091, 092, 093, 094, 095, 096, 097, 098

Sono 8×9= 72 modi. Essendo 10 le cifre utilizzabili, si hanno 720 modi differenti. Più in generale, le possibilità di ordinare ***r*** oggetti diversi provenienti da un insieme di ***n*** oggetti differenti sono:

$$D_{n,r} = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+1)$$

*Dato un insieme di **n** elementi distinti*
*si dicono disposizioni semplici di classe **r** (**r ≤ n**)*
*tutti i diversi ordinamenti di **r** elementi scelti tra gli **n** dati.*

Combinazioni semplici

Si parla di **combinazioni**, quando si desidera conoscere il numero di modi in cui è possibile associare un certo numero di oggetti che provengono da un insieme più ampio, ma senza ordinarli e senza riutilizzarli. Ad esempio: in quanti modi è possibile associare 5 numeri che vanno da 1 a 90, estraendoli da un insieme che ne contiene 90?
Consideriamo i casi che si possono realizzare tenendo costante come primo elemento l'1:

1 2 3 4 5, 1 2 3 4 6, 1 2 3 4 7, 1 2 3 4 8… 1 2 3 4 90 = 86 casi
1 3 4 5 6, 1 3 4 5 7, 1 3 4 5 8… 1 3 4 5 90 = 85 casi
…
…
1 85 86 87 90, 1 85 86 88 90, 1 85 86 89 90 = 3 casi
1 86 87 88 90, 1 86 87 89 90 = 2 casi
1 87 88 89 90 = 1 caso

2 3 4 5 6, 2 3 4 5 7, 2 3 4 5 8… 2 3 4 5 90 = 85 casi
…
…
86 87 88 89 90 = 1 caso

Si nota che i casi scemano, mano a mano che si prosegue con l'algoritmo da usare come primo numero 1 ad usare invece 89, per un totale di *43 949 268* combinazioni possibili, insomma indovinare la combinazione uscente è un po' come… vincere al lotto! Dalle disposizioni è possibile estrapolare il numero di casi a prescindere dall'ordine, mentre dalle permutazioni è possibile conoscere i modi "disordinati" in cui si possono ordinare. Più in generale, le possibilità di scegliere ***r*** oggetti diversi provenienti da un insieme di ***n*** oggetti differenti sono:

$$C_{n,r} = \frac{D_{n,r}}{P_r} = \frac{\dfrac{n!}{(n-r)!}}{r!} = \binom{n}{r}$$

*Dato un insieme di **n** elementi distinti*
*si dicono combinazioni semplici di classe **r***
*i diversi sottoinsiemi formati con **r** elementi scelti tra gli **n** dati.*

Sempre più difficile! Con ripetizioni

Innanzitutto il concetto "con ripetizioni" non è connesso al fatto che occorrano delle ripetizioni di matematica per capire quanto esposto, ma nel fatto che è possibile rimettere in gioco l'elemento ad ogni "estrazione".
Se riconsideriamo le permutazioni anche **AAB** o **AAA** sarebbe, ad esempio, un caso possibile. Ad esempio: in quanti modi è possibile anagrammare il nome **ANNA**?

AANN, **ANAN**, **ANNA**
NNAA, **NANA**, **NAAN**

Si ottengono 6 casi ammissibili. Posto $\boldsymbol{n}_A$=2 e $\boldsymbol{n}_N$=2 si ha $\boldsymbol{n}=\boldsymbol{n}_A+\boldsymbol{n}_N$=4. In generale, le possibilità di ordinare ***n*** oggetti diversi – tra i quali ***m*** possono essere presenti $\boldsymbol{n}_1$, $\boldsymbol{n}_2$… $\boldsymbol{n}_m$ volte – sono:

$$P_{n,n_1\ldots n_m} = \frac{n!}{n_1!\cdot\ldots\cdot n_m!}$$

Consideriamo, invece, le disposizioni possibili di ***r*** componenti provenienti da un insieme di ***n*** elementi, potendoli riutilizzare. Il tipico esempio sono i numeri da 0 a 9 per realizzare numeri a 3 cifre. Quanti numeri si ottengono?
Nulla di più semplice, la risposta è: 1000, tutti i numeri a partire da 0 fino ad arrivare a 999!

$$DR_{n,r} = n^r$$

Consideriamo, infine, le combinazioni possibili di ***r*** componenti provenienti da un insieme di ***n*** elementi, potendoli riutilizzare. In quanti modi è possibile associare 5 numeri che vanno da 1 a 90, estraendoli da un insieme che ne contiene 90 e rimettendo ogni volta dentro il numero estratto?
Si tratta quindi di valutare le disposizioni semplici ottenibili attraverso 5 numeri estratti da un insieme di 90, dividendo il numero ottenuto delle volte in cui cambia l'ordine, ma i numeri estratti sono sempre gli stessi:

$$\frac{D_{90,5}}{P_5} = \frac{\frac{90!}{(90-5)!}}{5!}$$

Da cui si ottiene *43 949 268* casi. In generale, le possibilità di estrarre ***r*** oggetti da un insieme di **n**, reinserendoli ad ogni estrazione, sono:

$$CR_{n,r} = \frac{D_{n,r}}{P_r} = \frac{n!}{(n-r)!\cdot r!}$$

Probabilità... senza imprevisti

Al fine di poter formulare dei giudizi su eventi futuri in base alle conoscenze a disposizione, si può ricorrere al calcolo delle probabilità che un dato evento ha di realizzarsi rispetto alla casistica possibile.
Un esempio banale è quello del dado a 6 facce. Consideriamo il lancio di due dadi (per convenzione uno rosso e uno blu). Gli esiti possibili, intesi come somma dei punteggi, sono:

- **2**, 1+1, 1 combinazione;
- **3**, 1+2 e 2+1, 2 combinazioni;
- **4**, 1+3, 2+2, 3+1, 3 combinazioni
- **5**, 1+4, 2+3, 3+2, 4+1, 4 combinazioni;
- **6**, 1+5, 2+4, 3+3, 4+2, 5+1, 5 combinazioni;
- **7**, 1+6, 2+5, 3+4, 4+3, 5+2, 6+1, 6 combinazioni;
- **8**, 2+6, 3+5, 4+4, 5+3, 6+2, 5 combinazioni;
- **9**, 3+6, 4+5, 5+4, 6+3, 4 combinazioni;
- **10**, 4+6, 5+5, 6+4, 3 combinazioni;
- **11**, 5+6 e 6+5, 2 combinazioni;
- **12**, 6+6, 1 combinazione;

per un totale di $DR_{6,2}=6^2=36$ combinazioni possibili.
Se si costruisce una tabella (approssimata) delle probabilità si ha:

2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
2.8%	5.6%	8.3%	11.1%	13.9%	16.7%	13.9%	11.1%	8.3%	5.6%	2.8%

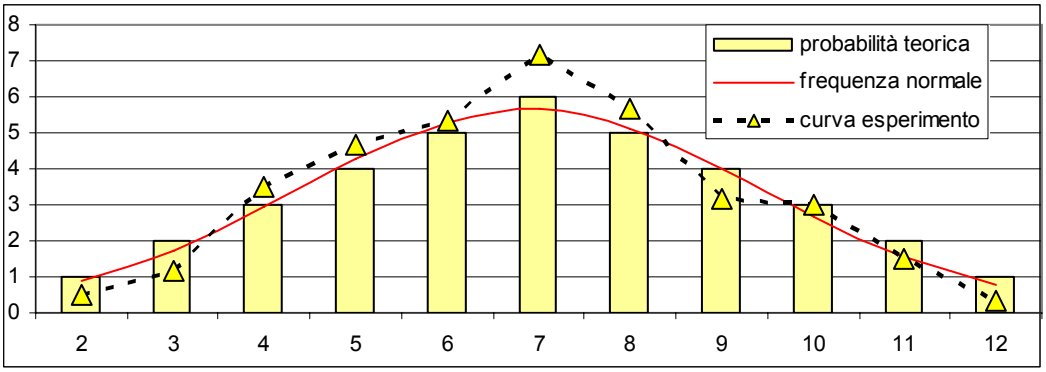
A questo punto si passa all'**esperimento**, che consiste in 6 cicli di 36 lanci, annotando le frequenze rapportate ai 36 lanci. Infine si calcolano le medie aritmetiche e campo di variabilità. Io ho ottenuto:

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	11	4	6	6	10	11	11	6	8	6	10	8	11	8	7	11	8	7	7	7	4	5	8	8	8	11	7	4	6	4	6	5	4	5	4	7
2	9	7	9	7	11	4	8	10	10	7	5	7	6	7	8	5	6	12	8	9	7	5	11	9	9	7	10	4	10	7	12	6	10	6	5	5
3	8	3	8	5	6	8	5	4	5	7	7	6	7	7	9	8	5	7	9	3	4	8	5	8	9	10	3	4	6	5	8	8	8	10	9	6
4	7	5	8	5	5	4	2	4	9	10	10	6	4	8	7	8	5	7	10	7	9	7	6	7	7	5	8	9	5	6	7	8	6	7	10	7
5	7	6	4	7	7	7	7	3	8	6	8	8	4	10	9	6	9	10	8	6	10	5	6	4	5	5	9	5	8	7	6	3	6	4	4	7
6	6	6	5	7	7	6	2	7	2	3	8	5	8	4	4	10	8	3	11	9	5	7	9	7	9	7	9	5	7	8	10	6	6	6	8	6

Una volta effettuato l'esperimento è consigliabile porrei i valori in una tabella riassuntiva in modo da poter facilmente analizzare gli esiti dell'esperimento:

	02	03	04	05	06	07	08	09	10	11	12
1	0	0	6	3	6	6	7	0	2	6	0
2	0	0	2	5	4	8	3	5	5	2	2
3	0	3	3	6	4	5	9	4	2	0	0
4	1	0	3	6	4	10	5	3	4	0	0
5	0	2	5	4	7	7	5	3	3	0	0
6	2	2	2	4	7	7	5	4	2	1	0
X	0,5	1,2	3,5	4,7	5,3	7,2	5,7	3,2	3,0	1,5	0,3
R	2	3	4	3	3	5	6	5	3	6	2

Si possono riportare questi dati in forma grafica in modo da verificare se l'andamento dell'esperimento ha seguito l'andamento previsto, in particolare è possibile in base al valore medio e alla dispersione dei dati tracciare la curva di probabilità di un evento caratterizzato dai medesimi parametri e che avesse una distribuzione di tipo normale.



Qualche nozione e alcuni assiomi

Supponiamo di condurre un esperimento ***N*** volte il cui esito sia ***n*** volte l'evento ***a***. Si definisce come probabilità che si verifichi l'evento ***a*** la funzione così definita:

$$\Pr(a) = \lim_{N\to\infty} \frac{n}{N}$$

ATTENZIONE: non è detto che questo limite esista ed è necessario che esistano le condizioni di ripetitività dell'evento ***a***!
È invece evidente che:

- se ***a*** è un evento, la probabilità che si verifichi ***a*** è 0 ≤**Pr**(***a***)≤ 1;
- se ***a*** è un evento, la probabilità che non si verifichi ***a*** è data da **Pr**(¬***a***)=1-**Pr**(***a***);

Posto che **S** è l'insieme di tutti gli eventi possibili, consideriamo quindi i seguenti assiomi:

- se **Pr**(***a***)=1, allora ***a*** è un evento certo, ne deriva che **Pr**(**S**)=1;
- se **Pr**(***a***)=0, allora ***a*** è un evento impossibile, ne deriva che **Pr**(**Ø**)=0;
- se ***a*** e ***b*** sono due eventi incompatibili, allora **Pr**(***a***∪***b***)=0 e **Pr**(***a***∩***b***)=**Pr**(***a***)+**Pr**(***b***);

- se ***a*** e ***b*** sono due eventi, la probabilità che si verifichi ***a*** in presenza dell'evento ***b*** è data da **Pr**(***a***|***b***) così definito:

$$\Pr(a \mid b) = \frac{\Pr(a \cap b)}{\Pr(a)}$$

- se ***a*** e ***b*** sono due eventi compatibili, allora **Pr**(***a***∪***b***)≠0 e **Pr**(***a***∩***b***)=**Pr**(***a***)+**Pr**(***b***)-**Pr**(***a***∪***b***);

Da suddetti assiomi derivano alcuni teoremi fondamentali, quali:

- il **teorema della probabilità totale**:

$$\Pr(a \cup b) = \Pr(a) + \Pr(b) - \Pr(a \cap b)$$

- il **teorema della probabilità composta**:

$$\Pr(a \cap b) = \Pr(a) \cdot \Pr(b \mid a) = \Pr(b) \cdot \Pr(a \mid b)$$

- il **teorema di Bayes**:

$$\Pr(a) = \Pr\left[(a \cap b) \cup (a \cap \neg b)\right]$$

nonché concetti chiave come

- la **probabilità condizionata**: se ***a*** è un evento dipendente dall'evento ***b***, allora **Pr**(***a***) dipende dal fatto che l'evento ***b*** si verifichi o meno, cioè **Pr**(***a***)=**Pr**(***a***|***b***).
- la **probabilità totale**: se **S**={$\boldsymbol{a}_1$, $\boldsymbol{a}_2$, … , $\boldsymbol{a}_n$}, dove gli eventi $\boldsymbol{a}_i$ sono mutuamente esclusivi ed esaustivi, allora

$$\Pr\left(\bigcup_{i=1}^{n} a_i\right) = 1$$

- l'**indipendenza stocastica**: se ***a*** e ***b*** sono due eventi stocasticamente indipendenti, allora **Pr**(***a***∪***b***)=**Pr**(***a***)·**Pr**(***b***) ovvero

$$\Pr(a \cap b) = \Pr(a) \cdot \Pr(b) \Leftrightarrow \begin{cases} \Pr(a \mid b) = \Pr(a) \\ \Pr(b \mid a) = \Pr(b) \end{cases}$$

Evitiamo brutte figure

Supponiamo d'avere un mazzo di carte e di voler determinare la probabilità che pescando una carta ci capiti una carta di picche o una figura.
Poiché ci sono 52 carte in un mazzo di carte francesi. 13 carte sono di picche per cui **Pr**(picche) = 13/52, mentre le figure sono 3 per ognuno dei 4 semi, cioè 3·4 = 12 per cui **Pr**(figure) = 12/52.
Ora che una carta possa essere una figura e che il seme di questa sia di picche non è un evento incompatibile perché ci sono ben 3 carte nel mazzo che soddisfano questa condizione, **Pr**(picche ∪ figura) = 3/52.
Ne deriva che la **probabilità totale** risulterà essere data da: **Pr**(picche ∩ figura) = **Pr**(picche) + **Pr**(figura) - **Pr**(picche ∪ figura) = 22/52~**35,5%**.
Vediamo un caso di **probabilità condizionata** e di **probabilità composta**: in una popolazione di 100 ragazze avvenenti, un ragazzo sa attraverso un questionario anonimo che 40 di loro accetterebbero di buon grado un suo invito ad uscire. A questo punto lo studente decide di proporre l'invito a quattro di loro in modo d'avere – a suo avviso – molte più probabilità.
Misuriamo la probabilità che il ragazzo vada in bianco (evento sgradito), che solo una delle quattro ragazze accetti l'invito (evento desiderato) e che più di una accetti contemporaneamente (evento non desiderato).

Affrontiamo per primo l'ultimo dei tre casi, perché in realtà è il più semplice. Infatti, basta che due ragazze accettino perché si realizzi la situazione di fare una brutta figura (a quel punto, qualsiasi sia la risposta delle altre due, la frittata è fatta).
È assodato che la probabilità **Pr**(***e***, ***t***), espressa come esito ***e*** in conseguenza dei tentativi ***t***, nel caso di un esito favorevole con un solo tentativo è **Pr**(1; 1) = **40%**.
Si utilizza la seguente formula:

$$\Pr(x) = \frac{(n-k)!\cdot r!}{n!\cdot(r-k)!}$$

dove:

- ***n*** è il numero di casi possibili, nel caso specifico 100;
- ***r*** è il numero degli eventi a:
  - per l'evento ***a***=sì, è 40,
  - per l'evento ¬***a***=no, è 60;
- ***k*** è il numero di tentativi effettuati.

Per primo caso affrontiamo il primo caso, cioè quello in cui tutte le ragazze daranno buca. **Pr**($\neg 1_{sì}$)=60/100, **Pr**($\neg 2_{sì}|\neg 1_{sì}$)=59/99 e così via, per cui **Pr**(0; 4) = **12,4%**.
Per analizzare gli altri due casi utilizziamo la seguente formula:

$$P(x) = \frac{\binom{p}{x}\cdot\binom{q}{k-x}}{\binom{n}{k}}$$

dove:

- ***n*** è il numero di casi possibili, nel caso specifico 100;
- ***p*** è il numero degli eventi favorevoli:
- ***q*** è il numero degli eventi sfavorevoli:
- ***k*** è il numero di tentativi effettuati.
- ***x*** è il numero di tentativi favorevoli tra i tentativi effettuati.

Ricerchiamo le probabilità legate al fatto che solo una delle quattro ragazze dia una risposta affermativa secondo lo schema:

sì	*no*	*no*	*no*
no	*sì*	*no*	*no*
no	*no*	*sì*	*no*
no	*no*	*no*	*sì*

Ovviamente si potrebbe utilizzare l'analisi delle singole probabilità per i quattro eventi ma la formula proposta semplifica di molto le cose. **Pr**(1; 4) = **34,9%**.
Infine, affrontiamo il caso 2, 3, o tutte le risposte positive: **Pr**(2∧3∧4; 4) = 35,2%+15,1%+2,3% = **52,7%**.
Fermo restando il fatto che lo studente farebbe meglio a non invitare quattro ragazze contemporaneamente, ma attendere di volta in volta l'esito della proposta d'invito, si vede come la probabilità che si realizzi la combinazione desiderata è inferiore a quella del tentativo secco, ciò che in realtà aumenta è la possibilità di fare una brutta figura ottenendo più di una risposta positiva contemporaneamente.
In base a queste valutazioni è interessante valutare la cosiddetta **speranza matematica** di ottenere un 7 mediante il lancio di 2 dadi.
ATTENZIONE: in questo caso l'esito di un lancio non influenza in alcun modo l'esito di quello successivo.
La formula per il calcolo della probabilità in questo caso è:

$$\Pr(x) = 1 - \left(1 - \frac{r}{n}\right)^k$$

dove:

- ***n*** è il numero di casi (combinazioni) possibili, nel caso specifico 36;
- ***r*** è il numero dei casi favorevoli, nel caso specifico 6;
- ***k*** è il numero di tentativi effettuati.

É interessante notare innanzitutto che la **speranza matematica** di ottenere un 7 effettuando 36 lanci è solo del 99,9%, cioè esiste una probabilità dello 0,1% che in 36 lanci non compaia mai il 7!
L'esperienza comune legata al concetto erroneo di "*numero ritardatario*" porterebbe invece a utilizzare la formula precedente che fornirebbe la certezza di un 7 dopo soli 22 lanci.
Riprendiamo la tabella della prova precedente e consideriamo i valori in colonna anziché lungo le righe (il riempimento delle celle è stato comunque un evento aleatorio). Applicando la formula corretta, circa 26 colonne su 36 dovrebbero aver almeno un 7. Applicando il criterio che i lanci siano influenzati dagli esiti precedenti (concetto del numero ritardatario), circa 34 colonne dovrebbero presentare almeno un 7. Nel mio esperimento ho ottenuto 27 colonne (solo una in più della previsione).

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	11	4	6	6	10	11	11	6	8	6	10	8	11	8	7	11	8	7	7	7	4	5	8	8	8	11	7	4	6	4	6	5	4	5	4	7
2	9	7	9	7	11	4	8	10	10	7	5	7	6	7	8	5	6	12	8	9	7	5	11	9	9	7	10	4	10	7	12	6	10	6	5	5
3	8	3	8	5	6	8	5	4	5	7	7	6	7	7	9	8	5	7	9	3	4	8	5	8	9	10	3	4	6	5	8	8	8	10	9	6
4	7	5	8	5	5	4	2	4	9	10	10	6	4	8	7	8	5	7	10	7	9	7	6	7	7	5	8	9	5	6	7	8	6	7	10	7
5	7	6	4	7	7	7	7	3	8	6	8	8	4	10	9	6	9	10	8	6	10	5	6	4	5	5	9	5	8	7	6	3	6	4	4	7
6	6	6	5	7	7	6	2	7	2	3	8	5	8	4	4	10	8	3	11	9	5	7	9	7	9	7	9	5	7	8	10	6	6	6	8	6

Conseguenze del teorema di Bayes

Fin qui s'è compreso che la teoria delle probabilità è una scienza matematica che s'occupa di eventi aleatori, cioè di situazioni che si presentano ogni volta diversamente, se ripetute più volte, come l'esito del lancio di uno o più dadi.
In ogni esperimento condotto per comprendere il fenomeno si riconoscono delle condizioni essenziali, necessarie per il compiersi dell'esperimento, e delle condizioni secondarie, che variano da un esperimento all'altro e che sono la causa delle variazioni aleatorie dei risultati.
Si supponga quindi l'esistenza di un evento ***e*** (effetto) che non può mai verificarsi se non in presenza di altri ***n*** eventi c_i (cause). Questo è il tipico caso in cui i fattori aleatori secondari – intimamente legati gli uni agli altri – giocano un ruolo importante.
Grazie alle relazioni fin qui esaminate, sviluppiamo la formula di Bayes:

$$\Pr(a) = \Pr\left[(a \cap b) \cup (a \cap \neg b)\right] = \Pr(a \cap b) + \Pr(a \cap \neg b) = \Pr(b) \cdot \Pr(a \mid b) + \Pr(\neg b) \cdot \Pr(a \mid \neg b)$$

Applichiamo tutto ciò ad un esempio pratico: si supponga che il 30% della popolazione sia predisposto ad una malattia, si sappia da indagini statistiche che tra gli individui predisposti se ne ammalano annualmente il 50%, mentre per l'altro gruppo la percentuale risulta essere del 20%. Qual è la probabilità che un individuo qualsiasi s'ammali?
In questo caso:

- evento ***a***: un individuo s'ammala;
- evento ***b***: un individuo è predisposto.

$$\Pr(a) = \Pr(b) \cdot \Pr(a \mid b) + \Pr(\neg b) \cdot \Pr(a \mid \neg b) = 30\% \cdot 50\% + (100\% - 30\%) \cdot 20\% = 29\%$$

Si è così ottenuta la probabilità complessa che un individuo qualsiasi s'ammali.
In termini più generici del tipo cause/effetto:

$$\Pr(e) = \sum_{i=1}^{n} \Pr(c_i) \cdot \Pr(e \mid c_i)$$

Si dice che nella formula di Bayes si combinino tre tipi di probabilità:

- la probabilità **a priori**, che assegna una probabilità di verificarsi alla causa c_i a prescindere del verificarsi dell'effetto e;
- la probabilità **a posteriori**, che identifica la probabilità che la causa c_i abbia influito sull'effetto e;
- la probabilità **probativa**, che rappresenta l'effettiva probabilità di verificarsi dell'effetto e in presenza della causa c_i.

Le incongruenze apparenti e il gioco d'azzardo

Affiniamo l'analisi delle probabilità. Ci sono tre carte:

- la prima ha i due lati neri;
- la seconda ha un lato nero e l'altro bianco;
- la terza ha entrambi i lati bianchi.

Viene posta una carta a caso sul tavolo. La scommessa che propongo è:

scommetto che anche l'altro lato è del medesimo colore

Vi conviene accettare?
Provate ad ottenere empiricamente la risposta.
SOLUZIONE: la scommessa non è equilibrata, ma è a mio favore.
Indico con B il bianco e con N il nero.
Le carte sono:

carta 1: B1-B2
carta 2: B3-N1
carta 3: N2-N3

Supponiamo sia uscito un lato bianco, si potrebbe trattare di uno dei tre casi: lato bianco, quindi potrebbe appartenere alla carta 1 oppure alla carta 2.
Dunque i casi possibili per il lato visibile sono 3:

B1, B2, B3.

I casi favorevoli all'evento che anche l'altro lato sia bianco sono 2:

B2, B1

Dunque la probabilità che anche l'altro sia bianco è circa del 66,7% e non del 50%!

* * *

Vi faccio un'altra proposta, ed è la seguente:

indovinate tra 90 numeri (da 1 a 90) i 5 che estrarrò

pago:

- per 2 numeri indovinati 250 volte la posta;
- per 3 numeri indovinati 4.250 volte la posta,
- per 4 numeri indovinati 80.000 volte la posta;
- per 5 numeri indovinati 1.000.000 volte la posta.

Vi conviene accettare?
Provate ad ottenere empiricamente la risposta.
SOLUZIONE: segnalo solo che si hanno:

- 4.005 ambi;
- 117.480 terni;
- 2.555.190 quaterne;
- 43.949.268 cinquine;

quindi, se vi inducessi a questo gioco sarei o un gran farabutto o, in alternativa, lo Stato italiano.

* * *

Ultimo quesito, si gioca a poker. Ci sono 4 giocatori e un mazzo che contiene 32 carte con otto valori differenti (7, 8, 9, 10, J, Q, K, A) e quattro semi (♠ ♣ ♥ ♦). Le combinazioni possibili sono:

- **coppia**, si realizza avendo 2 carte dello stesso valore;
- **doppia coppia**, si realizza avendo 2 coppie;
- **tris**, si realizza avendo 3 carte dello stesso valore;
- **scala semplice**, si realizza avendo 5 carte di seme diverso, ma in ordine;
- **full** si realizza avendo un tris e una coppia;
- **colore**, si realizza avendo 5 carte dello stesso seme, non in ordine;
- **poker**, si realizza avendo 4 carte dello stesso valore;
- **scala reale** si realizza avendo 5 carte dello stesso seme, in ordine.

Vi faccio la seguente proposta:

a prescindere dalle vostre carte, vi pago alla pari se "servito" non ho almeno una coppia

Vi conviene accettare?
SOLUZIONE: calcoliamo, la probabilità di realizzare, le suddette combinazioni, usando la definizione classica di probabilità come rapporto tra i casi favorevoli e i casi possibili.

7♠	8♠	9♠	10♠	J♠	Q♠	K♠	A♠
7♣	8♣	9♣	10♣	J♣	Q♣	K♣	A♣
7♥	8♥	9♥	10♥	J♥	Q♥	K♥	A♥
7♦	8♦	9♦	10♦	J♦	Q♦	K♦	A♦

Ora, il numero dei casi possibili, sono le combinazioni delle 32 carte a gruppi di 5:

$$C_{32,5} = \frac{D_{32,5}}{P_{32}} = \frac{32!}{(32-5)!\cdot 5!} = \binom{32}{5} = 201\,376$$

La **scala reale**: dato un seme, ci sono solo 5·casi di scala reale, dalla minima alla massima:

A♠	7♠	8♠	9♠	10♠
7♠	8♠	9♠	10♠	J♠
8♠	9♠	10♠	J♠	Q♠
9♠	10♠	J♠	Q♠	K♠
10♠	J♠	Q♠	K♠	A♠

Essendoci 4 semi, si ha un totale di 5·4=20 casi.

$$\Pr(\text{scala reale}) \approx 0.010\%$$

Il **poker**: ci sono solo 8 casi di avere 4 carte uguali (si faccia riferimento alle colonne della precedente tabella) che vanno moltiplicati per i 28 casi in cui si può presentare la quinta carta, per un totale di 224 casi.

$$\Pr(\text{poker}) \approx 0.111\%$$

Il **colore**: si hanno 32 casi di avere la prima carta, quindi restando nel medesimo seme si hanno 7 carte, poi 6, poi 5 ed infine 4. Ora, poiché l'ordine non conta si ha 32·7·6·5·4=224 casi a cui si devono togliere le scale reali che, come si è visto, sono 20, per un totale "ridotto" di 204 casi.

$$\Pr(\text{colore}) \approx 0.101\%$$

La **scala**: si hanno 20 casi con il medesimo seme che vanno moltiplicati per tutte le combinazioni ottenibili con i 4 semi (4^4, deriva dal tenere fissa la prima e mutevole le altre 4 carte), a questo totale (5 120) vanno tolte le 20 scale reali, per un totale di 5 100 combinazioni.

$$\Pr(\text{scala}) \approx 2.533\%$$

Il **tris**: la prima carta può essere scelta in 32 modi, a questo punto si hanno per la seconda carta solo 3 modi per averne 2 del medesimo valore e poi solo 2 modi per ottenerne 3 del medesimo valore, per cui (32·3·2)/3!=32 combinazioni.
A questo punto si considerano le carte rimanenti, occorre escludere l'eventualità di una quarta carta uguale (altrimenti si avrebbe un poker), restano 32 carte del mazzo da cui escludere le 3 per il tris, più la quarta che farebbe il poker, ottenendo 32-3-1=**28** possibili scelte.
Per la quinta carta non solo questa non può essere uguale alla prima per la questione del poker, ma non può nemmeno valere quanto la quarta (altrimenti si avrebbe un full), si hanno così le 32 carte del mazzo da cui escludere le 3 per il tris, più la quarta che farebbe il poker, meno la carta già pescata ed escludendo le altre 3 uguali che altrimenti darebbero un full, ottenendo 32-3-1-1-3=24.
Ne derivano (28·24)/2!=**336** combinazioni.
Mettendo insieme le due cose si ha 32·336=10 752 combinazioni.

$$\Pr(\text{tris}) \approx 5.339\%$$

La **coppia**: la prima carta può essere scelta in 32 modi, a questo punto si hanno per la prima carta solo 3 modi per averne 2 del medesimo valore, per cui (32·3)/2!=**48** combinazioni.
A questo punto si considerano le carte rimanenti, occorre escludere l'eventualità di una terza carta uguale (altrimenti si avrebbe un tris), restano allora 32 carte del mazzo, da cui escludere le 2 per la coppia, più le altre 2 che farebbero il tris, ottenendo 32-2-2=28 possibili scelte.
Per la quarta carta si hanno le 32 carte del mazzo da cui escludere le 2 per la coppia e le altre 2 rimanenti, meno la terza e le 3 che con questa farebbero una doppia coppia, ottenendo 32-(2+2)-(1+3)=24 possibili scelte.
Per la quinta carta non solo questa non può essere uguale alla coppia, ma nemmeno alle altre due (altrimenti si avrebbe di nuovo una doppia coppia), si hanno così le 32 carte del mazzo da cui escludere le 2 per la coppia, più le altre 2 che farebbe il tris e le due rimanenti, la terza e le tre a lei uguale, la quarta e le 3 ad essa uguali, ottenendo 32-(2+2)-(1+3)-(1+3)=20 possibili scelte.
Ne derivano (28·24·20)/3!=**2240** combinazioni.
Mettendo insieme le due cose si ha 48·2240=107 520 combinazioni.

$$\Pr(\text{coppia}) \approx 53.393\%$$

La **doppia coppia**: la prima carta può essere scelta in 32 modi, a questo punto si hanno per la seconda carta solo 3 modi per averne 2 del medesimo valore, per cui (32·3)/2!=**48** combinazioni.
A questo punto si considerano le carte rimanenti, occorre escludere l'eventualità di una terza carta uguale (altrimenti si avrebbe un tris), restano allora 32 carte del mazzo, da cui escludere le 2 per la coppia, più le altre 2 che farebbero il tris, ottenendo 32-2-2=28 possibili scelte.
Per la quarta carta si hanno le 3 carte del mazzo che forniscono la seconda coppia, cosicché per la seconda coppia si ha (28·3)/2!=**42** possibili scelte.
Poiché non è rilevante l'ordine tra le due coppie si ha (48·42)/2!=**1 008**.
Per la quinta carta non solo questa non può essere uguale alla prima coppia, ma nemmeno alla seconda, si hanno così le 32 carte del mazzo da cui escludere le 4 per la prima coppia, più le altre 4 che della seconda, 32-(2+2)-(2+2)=**24** possibili scelte.
Mettendo insieme le due cose si ha 1008·24=24 192 combinazioni.

$$\Pr(\text{doppia coppia}) \approx 12.013\%$$

Il **full**: la prima carta può essere scelta in 32 modi, a questo punto si hanno per la seconda carta solo 3 modi per averne 2 del medesimo valore, per cui (32·3)/2!=**48** combinazioni per la coppia.

A questo punto si considerano le carte rimanenti, occorre escludere l'eventualità di una terza carta uguale (altrimenti si avrebbe il tris che si vuole realizzare con le prossime 3 carte), restano allora 32 carte del mazzo, da cui escludere le 2 per la coppia, più le altre 2 che farebbero il tris, ottenendo 32-2-2=28 possibili scelte.
Per la quarta carta si hanno le 3 carte del mazzo che forniscono la seconda coppia, quindi altre 2 scelte per la quinta carta in modo da ottenere il tris, ovvero (28·3·2)/3!=**28**.
Mettendo insieme le due cose e tenendo conto che l'ordine della coppia e del tris non importa, si ottengono 48·28=1 344 combinazioni.

$$\Pr(\text{full}) \approx 0.667\%$$

Riassumendo si ha:

Coppia	53.393%
Doppia coppia	12.013%
Tris	5.339%
Scala	2.533%
Full	0.663%
Colore	0.101%
Poker	0.111%
Scala reale	0.010%
TOTALE	74.167%

È evidente che le probabilità che in una mano non si abbia almeno una coppia sono molto basse, circa 1 su 4.

* * *

Insomma, scommettere è sempre un gioco un po' azzardato se non si conoscono le effettive possibilità di vincere (evento favorevole) o perdere (evento sfavorevole).

Statistica

La statistica è la scienza che ha come fine lo studio quantitativo e qualitativo di un insieme collettivo di informazioni. Studia i modi (descritti attraverso formule matematiche) in cui una realtà fenomenica – limitatamente ai fenomeni collettivi – può essere sintetizzata e quindi compresa oppure come – tramite un numero limitato di esemplari – possa essere stimata.
Oltre a questi due impieghi, tipici della statistica descrittiva e di quella inferenziale, s'affianca anche un terzo utilizzo, fortemente stocastico, che è la **statistica esplorativa**, tramite la quale si tenta di verificare le ipotesi su un fenomeno analizzato definendo i criteri di accettabilità o inammissibilità di un insieme di dati più ampio.
Molte sono le critiche che i "*non esperti*" muovono alle informazioni che derivano dalla statistica, ma questo è dovuto essenzialmente alla manipolazione "*esperta*" dei dati a disposizione, secondo quella che io chiamo la **legge dell'avvocato**:

solo chi conosce alla perfezione le regole,
è in grado di infrangerle
senza che gli altri se n'accorgano.

In statistica descrittiva, dove oggettivamente si osserva l'intera popolazione e se ne sintetizzano le informazioni, gli elementi di sintesi non hanno alcun significato se poi non si usano degli strumenti quali, ad esempio, la stratificazione (cioè la massa di dati omogenei che sovrapposti formano quelli della popolazione) o la correlazione (in che modo le informazioni si legano tra loro). Tanto per dare un esempio, l'inflazione sull'intera popolazione ha veramente un misero significato, può essere più utile una **stratificazione** per fasce di reddito, perché è possibile così effettuare una **correlazione** tra fascia di reddito e paniere d'acquisti di riferimento.
In statistica inferenziale scegliere grossolanamente il campione di riferimento e non verificare l'ipotesi di probabilità se corrisponde o meno a quella norma o di Gauss (in gergo: **pianificare degli esperimenti**) può portare a conclusioni fuorvianti.
Utilizzare la stocastica per valicare un'ipotesi può poi portare ad uno dei due errori caratteristici della statistica esplorativa:

- invalidare un'ipotesi valida;
- validare un'ipotesi falsa.

Per questi e molti altri motivi, diventa fondamentale conoscere le basi della statistica ma anche le regole che bisogna rispettare affinché l'esame e l'analisi dei dati porti poi ad un'informazione affidabile.

Mean: la media aritmetica semplice

La media aritmetica semplice è la media per antonomasia. Viene adoperata per riassumere con un solo numero un insieme di dati su un fenomeno misurabile. Definito un insieme di ***n*** valori $\{x_1, x_2 \ldots x_n\}$ la formula per ottenere la media aritmetica semplice è:

$$\overline{x} = \frac{x_1 + x_2 + \ldots x_n}{n}$$

Dove il trattino sopra a ***x*** serve ad indicare il concetto di media.
In maniera sintetica si può anche scrivere:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In questa formula s'introduce il simbolo di sommatoria, con il quale si indica che si deve effettuare la somma di tutti i generici valori x_i al variare dell'indice ***i*** da 1 a ***n***.
Quest'indice è fondamentale per gli aspetti geometrici della statistica ovvero quelli legati alla rappresentazione grafica dell'insieme dei dati mediante indicatori di posizioni e la determinazione di alcune proprietà geometriche di tali indicatori.

Min: il valore minimo

Si tratta dal valore minore tra gli $\boldsymbol{n}$ che compongono l'insieme dei dati $\{\boldsymbol{x}_i\}$ e rappresenta il limite inferiore dell'insieme di variabilità.

Max: il valore massimo

Si tratta dal valore massimo tra gli $\boldsymbol{n}$ che compongono l'insieme dei dati $\{\boldsymbol{x}_i\}$ e rappresenta il limite superiore dell'insieme di variabilità.

Range: l'intervallo di variazione

Si tratta dell'intervallo Δ di variabilità degli $\boldsymbol{n}$ che compongono l'insieme dei dati $\{\boldsymbol{x}_i\}$ ottenibile come differenza tra il valore massimo e quello minimo.

$$\Delta = \max\{x_i\} - \min\{x_i\}$$

Interval: la classe

Al fine di raggruppare in termini di frequenza un insieme dei dati $\{\boldsymbol{x}_i\}$ si ricorre alla formazione di **classi**, raggruppando i dati in sottoinsiemi $\mathbf{Y}_j$ il cui rappresentante è il valore $\boldsymbol{y}_j$ intermedio tra il massimo e il minimo del sottoinsieme.

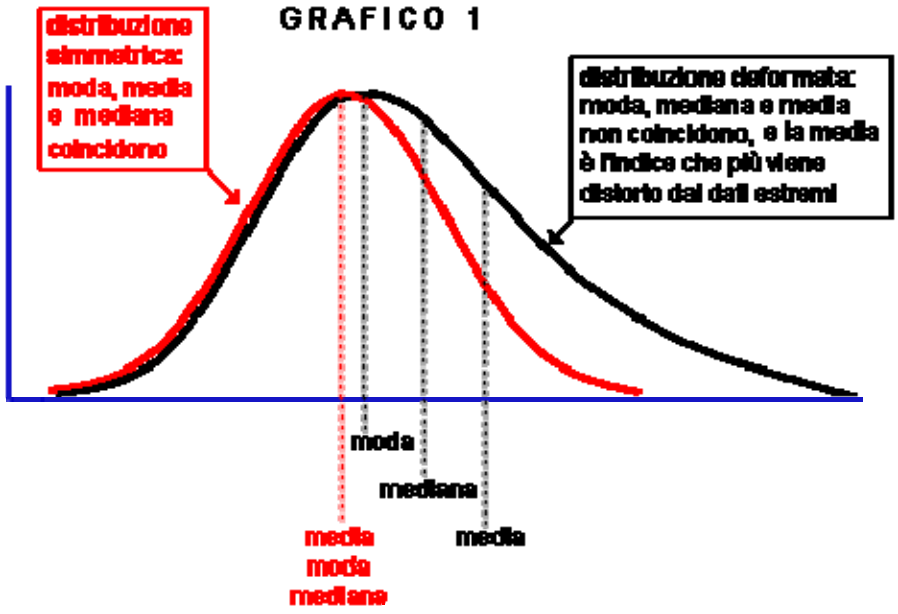
$$Y_j = \{x_i \mid y_j - a < x_i \le y_j + a\}$$

Ogni classe è caratterizzata da:

- valore del rappresentante, $\boldsymbol{y}_j$;
- numerosità in termini di frequenza, $f(\boldsymbol{y}_j)$;
- ampiezza, $2 \cdot \boldsymbol{a}$.

Mode: la moda

Per **moda** (o norma) di un insieme di dati $\{\boldsymbol{x}_i\}$ s'intende il rappresentante $\boldsymbol{y}_j$ della classe che è caratterizzata dalla massima frequenza, in altre parole si tratta del valore che compare più frequentemente.



La moda può non esistere o, se esiste, può non essere unica. Una distribuzione che abbia una sola moda viene detta: **unimodale**.

Quartiles: i quartili

Un **quartile** è quel valore x_q per il quale la somma di tutte le frequenze (o l'integrale della funzione di densità) è uguale al valore q (compreso tra zero e uno). Quando q assume valori pari a

- $0{,}25 = {}^1/_4 \rightarrow Q_1$;
- $0{,}5 = {}^2/_4 \rightarrow Q_2$;
- $0{,}75 = {}^3/_4 \rightarrow Q_3$;

si parla di quartile.
Allora:

$$\begin{aligned} Q_0 &= \min\{x_i\} \\ Q_4 &= \max\{x_i\} \end{aligned} \Rightarrow \Delta = Q_4 - Q_0$$

Similmente q assume valori in decimi pari, ad esempio, a:

- $0{,}05 = 5\% \rightarrow P_5$;
- $0{,}10 = 10\% \rightarrow P_{10}$;
- $0{,}90 = 90\% \rightarrow P_{90}$;
- $0{,}95 = 95\% \rightarrow P_{95}$;

si parla di percentili.

Median: la mediana

In statistica, dato un insieme dei dati $\{x_i\}$ di un carattere quantitativo oppure qualitativo ordinabile (ovvero le cui modalità possano essere ordinate in base a qualche criterio), si definisce la **mediana** il valore corrispondente a Q_2 (secondo quartile).

InterQuartile Range (IQR): scarto interquartile

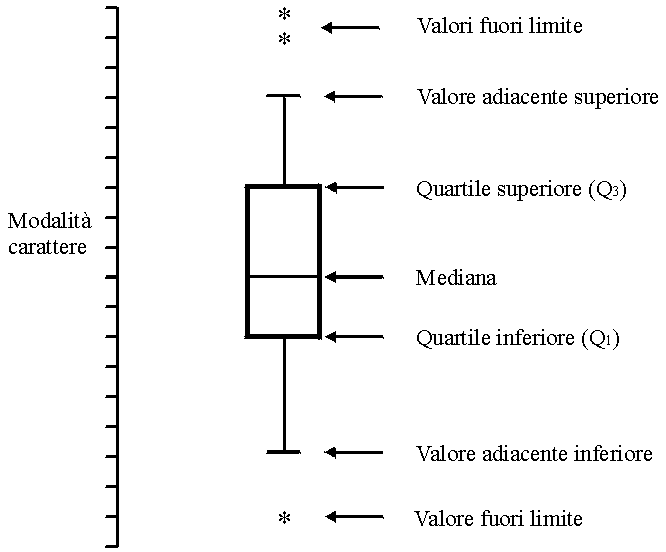
Lo **scarto interquartile** (IQR) è la differenza tra il primo e il terzo quartile:

$$IQR = Q_3 - Q_1$$

Tale parametro viene utilizzato come un indice di dispersione dei dati.

Box-plot: diagramma a scatola

I quartili vengono spesso utilizzate per sintetizzare gli elementi di disperazione di un insieme di dati $\{x_i\}$ tramite costruzione grafica detta **box-plot**:



Il box-plot è una rappresentazione grafica che serve per descrivere in modo compatto e grafico la distribuzione di una funzione. È il disegno su un piano cartesiano di un rettangolo, i cui estremi sono il primo e terzo quartile (Q_1 e Q_3), è tagliato da una linea all'altezza della mediana (Q_2). Il minimo della distribuzione viene indicato con (Q_0), mentre il massimo con (Q_4). Abitualmente vengono aggiunte due righe (detti anche **baffi**) corrispondenti ai valori distanti 1,5·IQR dal primo e dal terzo quartile.
A volte vengono anche rappresentati nel grafico i valori che fuoriescono dall'intervallo delimitato dalle due righe come **punti isolati**.

Variance (Var): la varianza

La **varianza reale** esprime il quadrato della distanza media dei dati di un insieme $\{\boldsymbol{x}_i\}$ dal valor medio.

$$Var = \sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}$$

La **varianza campionaria** è invece uno stimatore della varianza reale ed è dato dalla formula:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$

Standard deviation (SD): la deviazione standard

La **deviazione standard** (o scarto quadratico medio) è un indice di dispersione derivato direttamente dalla varianza che ha per unità di misura la stessa unità di misura dei valori osservati.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$$

La **deviazione standard campionaria** è data invece dalla formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Coefficient of variation (CV): indice di dispersione

Il **coefficiente di variazione** è un indice di dispersione che permette di confrontare misure riferite a unità di misura differenti, in quanto si tratta di un numero adimensionale.

$$CV = \frac{\sigma}{\left|\bar{x}\right|}$$

Skewness: l'asimmetria

L'**asimmetria** di un insieme dei dati $\{\boldsymbol{x}_i\}$ fornisce il grado di scostamento della curva di frequenza associata rispetto alla simmetria. Un sistema rapido per verificare la presenza di un'asimmetria in una distribuzione di dati unimodale è quella di fare la differenza tra la moda e la media (**primo coefficiente di asimmetria di Pearson**).

$$\frac{\bar{x} - \text{moda}}{\sigma}$$

Per evitare di usare la moda (che non sempre esiste o è unica) si può utilizzare la mediana (**secondo coefficiente di asimmetria di Pearson**).

$$3 \cdot \frac{\overline{x} - \text{mediana}}{\sigma}$$

L'asimmetria in termini di **skewness** è fornita dalla formula:

$$sk = \frac{n}{(n-1)\cdot(n-2)} \cdot \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sigma^3}$$

oppure, per $n > 300$, utilizzando la più semplice **formula di Pearson**:

$$sk = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n \cdot \sigma^3}$$

Posto che ***sk*** è una condizione necessaria, ma non sufficiente per la simmetria, si ha:

- ***sk***=0, nel caso di perfetta simmetria;
- ***sk***<0, per l'asimmetria a destra;
- ***sk***>0, per l'asimmetria a sinistra.

Kurtosis: la curtosi

La **curtosi** di un insieme dei dati $\{x_i\}$ fornisce il grado di altezza raggiunto dalla curva di frequenza associata. Un sistema rapido per verificare la curtosi di una distribuzione è basata sui quartili e sui percentili:

$$3 \cdot \frac{Q_3 - Q_1}{P_{90} - P_{10}}$$

La curtosi in termini di **kurtosis** è fornita dalla formula:

$$ku = \frac{n\cdot(n+a)}{(n-1)\cdot(n-2)\cdot(n-3)} \cdot \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{\sigma^4} - \frac{3\cdot(n-1)^2}{(n-2)\cdot(n-3)}$$

oppure, per $n > 300$, utilizzando la più semplice **formula di Fisher**:

$$ku = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n \cdot \sigma^4} - 3$$

Se:

- ***ku***>0 la curva si definisce *leptocurtica*, cioè più alta di una normale;
- ***ku***<0 la curva si definisce *platicurtica*, cioè più bassa di una normale;
- ***ku***=0 la curva si definisce *normocurtica*, cioè simile ad una normale.

Il 7 e ½ e altri giochi di "carte"

Pianificare l'indagine di un processo attraverso l'impiego di:

- statistica inferenziale;
- stocastica;
- euristiche;

non è una cosa banale, cioè non si tratta di accumulare dati per elaborarli e sintetizzarli a testa bassa tramite indici o grafici. Se non c'è un lavoro di analisi di forma e di distribuzione e la costruzione di tutta una serie

di esperimenti di verifica, **dalla semplice elaborazione e sintesi di dati di fatto veri si può giungere a conclusioni errate**.
Vengono quindi in aiuto **sette strumenti statistici** indispensabili per chiunque voglia seriamente fare della statistica e del controllo di processo:

- i diagrammi a: torta, barre, radar, bolle;
- i fogli di riscontro e i diagrammi di concentrazione dei difetti;
- il raggruppamento in classi e i grafici di Pareto;
- i diagrammi causa-effetto;
- l'analisi per stratificazione;
- i grafici a dispersione e l'analisi di correlazione;
- l'analisi dei processi tramite diagrammi di flusso;

a cui s'affiancano tutte le tecniche di problem-solving. In caso contrario si finisce per concepire il controllo come una mera conta dei vivi e dei morti e non un potente strumento per:

- comprendere i meccanismi che condizionano un processo per scoprire le giuste contromisure;
- controllare la regolarità di una distribuzione di dati con un piano di campionamento molto leggero;
- giudicare una popolazione in base all'esame di un limitato campione rappresentativo;

ovvero, in poche parole, per migliorare la qualità, diminuendo al contempo i costi.
Ne deriva che, nello svolgersi di un processo, i dati non affluiscono come un unico insieme, ma gradualmente, mano a mano che si generano. Il loro andamento subisce una variabilità dovuta a:

- **fenomeni naturali o accidentali** che si manifestano come l'effetto cumulato di un gran numero di piccole cause inevitabili ed incontrollabili;
- **fenomeni sistematici o intenzionali** che indicano la presenza di una distorsione nel processo che può essere dovuta solo a elementi tecnici o tecnologici.

Anticipando il fatto che la linea centrale individua un valor medio e che a σ è un indice di dispersione dei dati, nel controllo statistico di processo sono note le:

10 REGOLE DI SENSIBILITÀ

- 1 o più punti cadono al di fuori dei limiti di controllo
- 2 punti consecutivi su 3 cadono oltre i limiti di sorveglianza posizionati a 2σ, ma rimangono dentro i limiti a 3σ,
- 4 punti su 5 consecutivi cadono oltre la distanza di 1σ, dalla linea centrale,
- 8 punti consecutivi cadono dalla stessa parte della linea centrale,
- 6 punti consecutivi sono in ordine crescente o decrescente,
- 15 punti consecutivi sono nella zona a 3σ, (sia sopra che sotto la linea centrale),
- 14 punti consecutivi si alternano a zig-zag,
- 8 punti consecutivi si alternano intorno alla linea centrale, ma nessuno è nella zona 3σ,
- si manifesta un comportamento non casuale dei dati,
- 1 o più punti si posizionano vicino ai limiti di sorveglianza e di controllo,

esse, ad esempio, segnalano l'insorgere di fenomeni sistematici o intenzionali.
L'obiettivo è allora quello di individuare la presenza nel processo di eventuali variabilità sistematiche, poiché si tratta di cause rimovibili. La variabilità naturale, invece, è impossibile da eliminare, ma d'altro canto non influenza particolarmente i processi, introduce solo il concetto di limite di tolleranza.
Se all'interno di un processo di produzione è presente solo una variabilità naturale, il processo si dice **sotto controllo**, mentre se si rileva una variabilità sistematica il processo è detto **fuori controllo**.

Tipologia delle carte di controllo

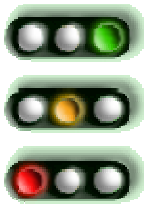
La tipologia delle carte di controllo è molto variegata e si può proporre una generale distinzione tra:

- **carte tradizionali** (di tipo Shewhart), definite come carte senza memoria ed insensibili a piccoli cambiamenti e che a loro volta possono essere suddivise in:
    - carte di controllo per variabili,
    - carte di controllo per attributi;
- **carte alternative** (a memoria illimitata uniforme o non uniforme) adatte a evidenziare piccoli cambiamenti.

A prescindere dalla tipologia adottata, **le carte di controllo non devono essere intese come una sintesi grafica in output dei dati, ma un foglio di raccolta dati in input, sul quale deve essere direttamente osservabile la presenza di un'anomalia statistica**.
Sono un attrezzo operativo da utilizzare in reparto e non uno strumento grafico da utilizzare in riunione per esporre la riuscita o meno di un processo.
Se un qualsiasi primate è in grado di:

- azionare leve,
- premere pulsanti,
- spostare oggetti

in risposta ad un determinato stimolo luminoso, allora qualsiasi operatore[1] è in grado di recepire una sequenza di colori del tipo:



verde **!** il processo è sotto controllo

giallo **…** il processo è a rischio fuori controllo

rosso **?** il processo è fuori controllo

Non poniamo limiti… solo alla divina Provvidenza

Nel momento in cui si decide di monitorare un processo dal punto statistico si hanno una serie di linee di riferimento che si possono tracciare sulla carta di controllo, in tal modo l'operatore è in grado di comprendere al volo come si stia comportando il processo di cui è il responsabile. In questo modo non si basa più su un unico controllo finale (sperando nell'intervento benevolo della divina Provvidenza), ma l'intero processo, lungo le attività che lo caratterizzano viene monitorato con continuità.

Standard limits: limiti di tolleranza interni

Qualsiasi processo ha come obiettivo implicito il conseguimento di un valore nominale che corrisponde al valore atteso in output dal processo. Tale attesa tiene conto anche delle possibili oscillazioni, per cui, più in generale, si hanno:

- un valore atteso massimo;
- un valore atteso minimo;
- un valore atteso nominale.

Questi tre valori definiscono quello che dovrebbe essere l'output standard del processo. Ogni elemento in uscita inserirà un valore sequenziale che, secondo le attese, andrà a collocarsi all'interno della fascia descritta dai limiti di massimo e minimo. Se il processo è soggetto solo ad una variabilità naturale, il succedersi dei punti tracciati sulla carta di controllo dovrà rispettare delle precise regole di probabilità, in caso contrario si dovrà assodare l'esistenza di una qualche causa che sta alterando questa distribuzione.

[1] Non ci sono posizioni razziste in questa frase, infatti, non s'asserisce che gli stupidi siano gli operatori, ma coloro che non reputano gli operatori sufficientemente intelligenti da poter capire concetti elementari.

Questi limiti sono solitamente attribuiti o in base all'esperienza o dall'analisi di uno storico, comunque vengano originati sono parametri tipicamente interni, cioè definiti da chi "possiede" il processo.

Statement limits: limiti di specifica esterni

Qualsiasi cliente si attende che il proprio fornitore soddisfi delle specifiche (caratteristiche speciali) in termini di:

- minore (o uguale) di;
- maggiore (o uguale) di.

Al cliente non importa quanto la scheda tecnica sia rispondente a tali specifiche, al cliente basta solo che tutte le specifiche siano soddisfatte.
Questo porta all'esistere di più fornitori che sono in grado di attenersi alle specifiche richieste ognuno con un livello di sicurezza e un'affidabilità differente in base a dove si pongono i limiti di tolleranza rispetto ai limiti di specifica imposti.
Si possono avere così situazioni del tipo:

- **una parte non trascurabile** dell'output del processo non è conforme alle specifiche, sicché è molto probabile che il cliente si accorgerà che c'è del non conforme;
- **una parte trascurabile** dell'output del processo non è conforme alle specifiche, sicché è molto improbabile che il cliente si accorgerà che c'è del non conforme;
- **tutto** l'output del processo è conforme alle specifiche;

tutto dipende da dove si collocano questi limiti rispetto a quelli standard.

Control limits: limiti di tolleranza naturali

Qualsiasi processo, ogni volta che viene replicato, non si ripropone uguale a se stesso, ne deriva che i propri limiti di oscillazione sono notevolmente inferiori alle tolleranze attribuite al suo output, cosicché le oscillazioni che di volta in volta non rendono ripetibile il processo non vanno ad inficiare il risultato finale.
Ne deriva che ogni processo è caratterizzato da:

- un **valore centrale** corrispondente al valor medio;
- un **limite superiore** spostato rispetto alla media di +3 deviazioni standard;
- un **limite inferiore** spostato rispetto alla media di -3 deviazioni standard.

I punti di volta annotati non necessariamente cadranno tutti all'interno della fascia descritta dai due limiti di controllo naturali, alcuni (**anomalie**) potranno uscire costituendo dei "fuori controllo".

Warning limits: limiti d'attenzione

Chiunque debba controllare un processo preferirebbe avere un margine di sicurezza tra il sapere che tutto va bene/male. Si tratta della medesima logica che concede ad un autista la possibilità di rallentare e fermarsi ad un semaforo rosso, perché ha visto passare il medesimo dal verde al giallo.
Come non esiste per il giallo semaforico una durata minima[2] che molti giudicano insufficiente, alla stessa maniera il tracciamento dei limiti d'attenzione è in realtà arbitraria e dipende da:

- quanto il processo sia solitamente sotto controllo;
- quanto lunghi siano i tempi di intervento;
- quanto prolungati siano i fenomeni d'inerzia del processo.

Statistica, probabilità e… anomalie

Qualsiasi appassionato di cucina, quindi non necessariamente un cuoco professionista, se vuol essere sicuro che ciò che sta preparando sia gradevole al palato e certo che non sia necessario aggiungere un pizzico di qualcosa per regolare il sapore, sa che è più utile verificare la cosa durante la preparazione, piuttosto che a cose fatte.
Il buon senso induce a ritenere che sia necessario assaggiare un vino, prima d'aggiungerlo ad una pietanza, piuttosto che scoprire che sapeva d'aceto, una volta che ha servito il tutto ai commensali.
Nello svolgimento di un processo questi "assaggi" consistono in una serie di verifiche relativamente al rispetto di semplici **precetti probabilistici** che, nel

[2] Tre secondi.

loro insieme, danno origine alle cosiddette **regole di sensibilità delle carte di controllo**.

Drift/trend: le derive

I valori migrano progressivamente in una direzione manifestando un innalzamento o abbassamento progressivo.

Dispersion: la variabilità

I valori oscillano sopra e sotto il valor medio posizionandosi sempre nei pressi dei limiti esterni di variabilità.

Shift: gli scostamenti

I valori migrano improvvisamente in una direzione manifestando un innalzamento o abbassamento improvviso, dando luogo a due differenti livelli.

Run: le serie

I valori si susseguono permanendo al di sotto o al di sopra del valor medio in maniera prolungata o denotando un aumento o diminuzione costante.

Cycle: i cicli e le autocorrelazioni

I valori tendono ad oscillare secondo sequenze ricorrenti.