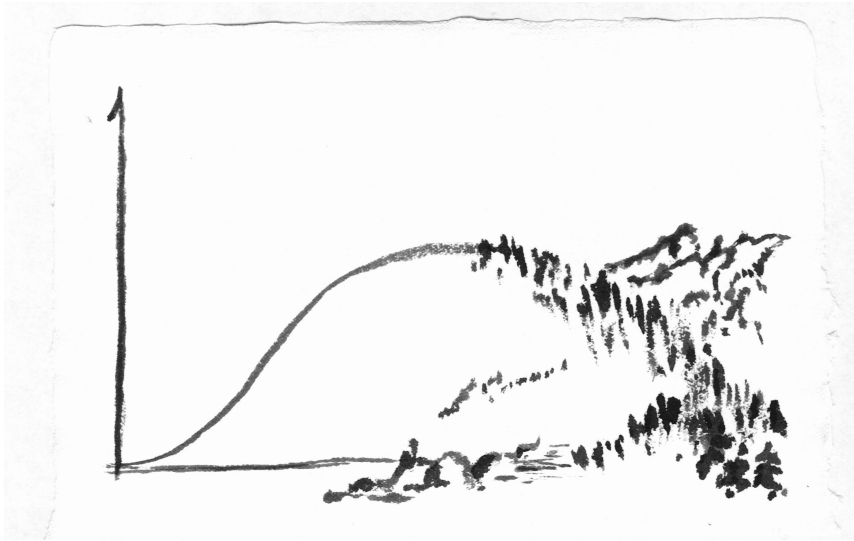
ORIGANOVA

Introduzione alla statistica con l'origami

Rel.1.6



Mario Cigada

ORIGANOVA

Indice

Cap.0 Introduzione^2 alla statistica con l'origami pag. 4

Cap.1 Un computer di carta pag. 5

Cap.2 Misurare la dispersione pag. 12

Cap.3 Misure di posizione, misure di dispersione e misure di associazione pag. 13

Cap.4 Le distribuzioni statistiche (infiniti masu) pag. 15

Cap.5 Altre distribuzioni pag. 18

Cap.6 Media campionaria e media di popolazione (Ma quanto succo di liquirizia ci hai messo?) pag. 19

Cap.7 Verifica di un test (Le caramelle mou extramorbide) pag. 21

Cap.8 ANOVA (ancora le caramelle mou) pag. 24

Cap.9 Un cenno sulla regressione pag. 29

Cap.10 Una storia vera pag. 31

Appendice per origamisti pag. 32

Appendice con le formule pag. 33

Bibliografia pag. 34

Capitolo 0
Introduzione^2 alla statistica con l'origami
ORIGANOVA

In queste poche pagine vorrei raccontarvi qualcosa sui numeri e sulla statistica, e mi piacerebbe farlo giocando insieme a voi con la carta.

Origami è la parola che definisce in giapponese l'attività di piegare la carta. ANOVA, invece, sta per ANalysis Of VAriance: in italiano analisi della varianza; uno strumento per l'analisi statistica molto importante e sofisticato. Da queste due parole è nato il buffo titolo che rivela l'idea, un po' strampalata, di spiegare alcuni importanti concetti della statistica come la media, la varianza o l'inferenza mentre giochiamo insieme con l'origami.

Non servono particolari requisiti teorici per seguire il testo; è invece indispensabile avere sotto mano qualche foglio A4 e qualche foglio quadrato di circa 10 cm di lato. I fogli di dimensione A4 sono quelli della comune carta per fotocopie (cm 21 x 29.7 circa; 80 g/mq); i foglietti quadrati si trovano nei negozi di giocattoli o nelle cartolerie come carta per origami, oppure vanno benissimo quei blocchi colorati per appunti, di forma approssimativamente cubica: basta controllare che ogni singolo foglietto sia esattamente quadrato; magari, se potete scegliere, prendete carta un po' più consistente di quella delle fotocopie. Ci servirà anche una riga o una squadra, una matita (e una gomma) forbici o taglierino.

Allora partiamo.

Capitolo 1
Un computer di carta

La statistica è uno strumento pratico, nato per manipolare numeri a fini pratici; ma possiamo anche usare la statistica per giocare, immaginando una situazione inventata, come in una favola.

C'era una volta un signore che fabbricava caramelle; dopo avere preparato le sue caramelle (di tutti i colori) le metteva in una macchina che preparava tanti sacchetti e li riempiva con le caramelle. La macchina confezionatrice era un po' scassata, piuttosto imprecisa, così a volte i sacchetti erano belli pieni a volte mezzi vuoti e i bambini si lamentavano. Allora il signore, per capire bene cosa stava succedendo alla sua macchina: prese tutti i sacchetti di caramelle che stavano in magazzino e li pesò uno per uno sulla bilancia. Il primo pesava 2 chili, il secondo 3 kg e così via, ecco tutti i pesi

2 3 3 5 2 3 3 2 2 3 2 3 1 2 3 3 4 3 4 2 4 3 1 5 1 3 1 2 2 2 4 3 2 2 4 3 5
3 2 1 4 3 2 3 2 3 1 4 5 1 1 3 3 1 2 2 1 4 3 2 2 2 2 2 3 4 2 2 2 1 2 2 3 2
2 3 4 1 2 3 3 4 2 2 2 1 3 3 1 4 1 2 1 2 1 2 2 4 2 2

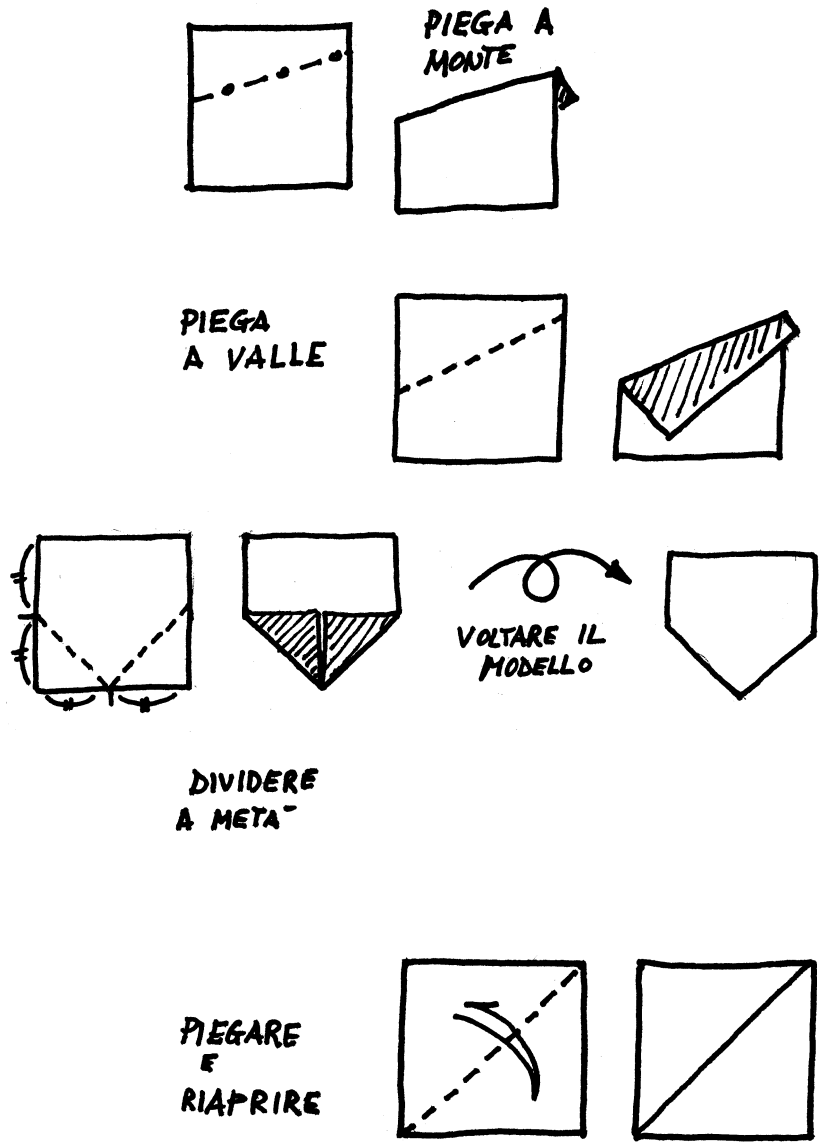
Come dite, sono troppi? Bé allora limitiamoci ai primi 5 sacchetti:

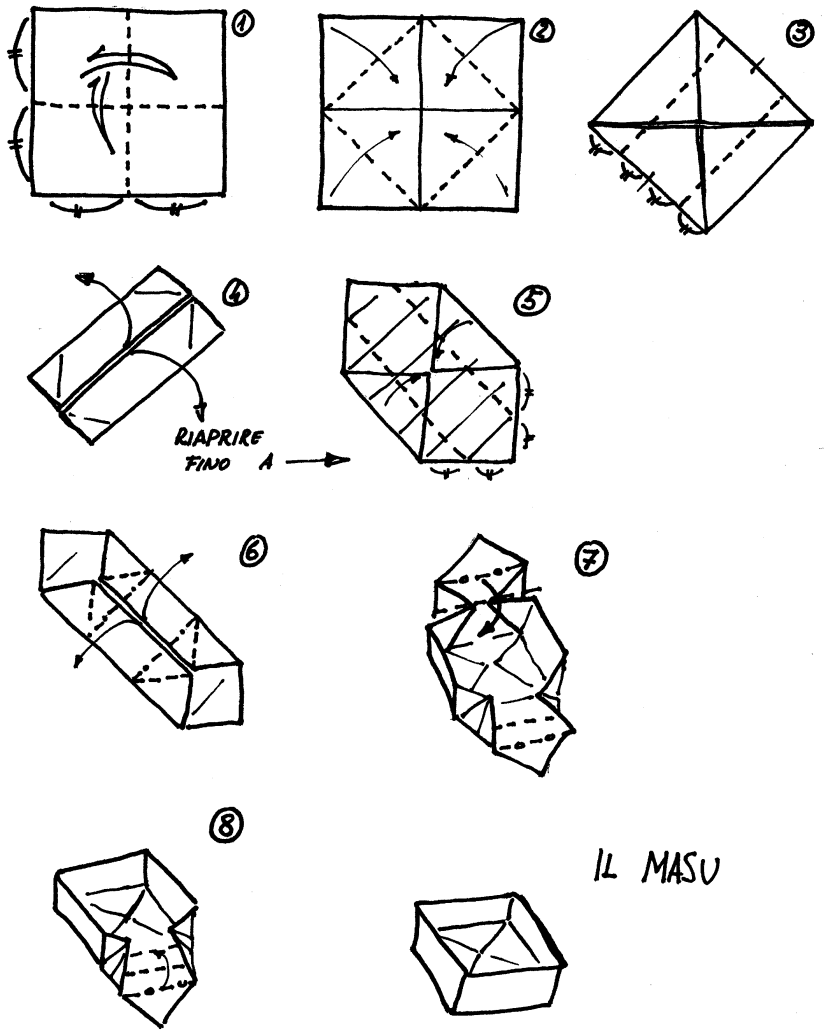
2 3 3 5 2

potremmo far finta che il signor Gervaso avesse un magazzino molto piccolo; vi avevo detto che il fabbricante di caramelle si chiamava Gervaso, vero? No ? Bé, ve lo dico adesso.

Allora, per rappresentare un kg ho deciso di usare una piega classica dell'origami tradizionale: il masu; rappresenta un contenitore che veniva usato proprio come unità di misura; dunque cominciamo a piegare un po' di masu, magari chiamate qualche amico per aiutarvi. Nelle prossime pagine trovate la spiegazione per piegare un masu; forse i disegni vi appariranno più chiari se tenete presente che in tutto il mondo per spiegare gli origami si utilizzano dei segni convenzionali; permettetemi di raccomandarvi una certa precisione nelle pieghe, come pure di premere bene la carta, ripassando le pieghe con il dorso dell'unghia.

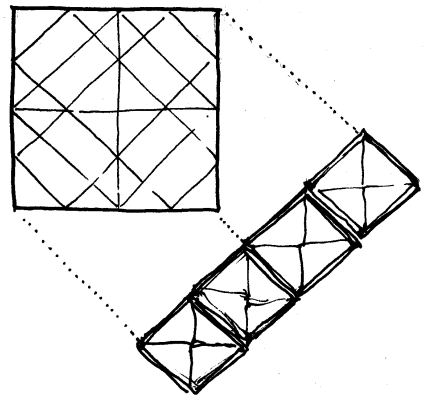
Ecco qui accanto i segni convenzionali più comuni





Il masu può essere usato come contenitore oppure, capovolto, lo si può usare per giocare alle costruzioni; è anche interessante notare che il lato di ciascun masu è uguale al lato del foglietto da cui siamo partiti moltiplicato per la radice quadrata di 2 e poi diviso per 4, se provate a riaprire un masu la cosa vi risulterà evidente osservando le pieghe, soprattutto se ricordate che $\sqrt{2}$ è la diagonale del quadrato. Quindi se avete usato fogli di 10 cm di lato viene

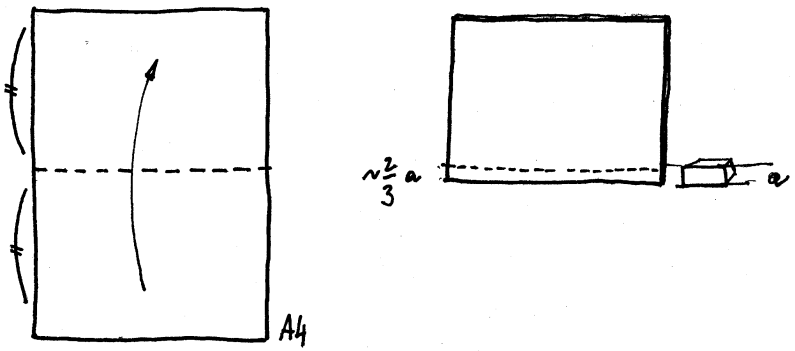
$10 \times \sqrt{2} \approx 14{,}1 \div 4 \approx 3{,}5$ cioè cm 3,5 circa.



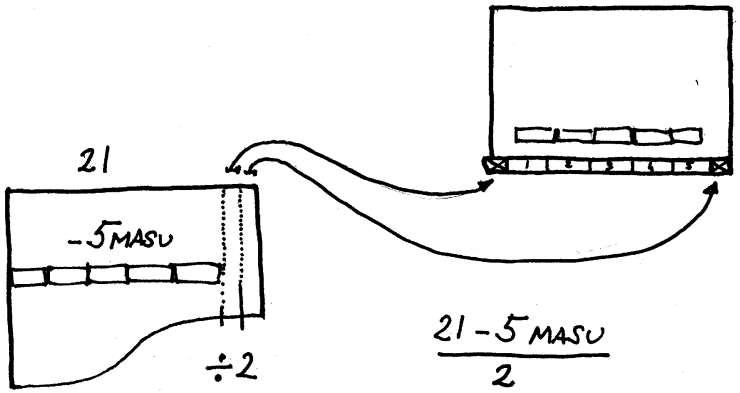
Già, ma quanti masu dobbiamo piegare? Vediamo un po'…
2 per rappresentare il primo sacchetto di caramelle
+ 3 per rappresentare il secondo sacchetto di caramelle
+ 3 per rappresentare il terzo sacchetto di caramelle
+ 5 per ...

Siete già stufi? Allora vi insegno un trucco: pieghiamo solo 5 masu, mettiamoli in fila e misuriamo quanto sono lunghi: viene un po' di più di
3.5 x 5 = 17.5
perché prima avevamo arrotondato per difetto e soprattutto perché le pieghe occupano un po' di spazio, ma non ha alcuna importanza; l'importante è che il valore della misura sia minore del lato corto di un foglio A4 che è di 21 cm (questo è il motivo per cui i foglietti quadrati devono avere il lato di circa 10 cm).

Adesso prendiamo un foglio A4 e pieghiamolo a metà così



e facciamo una piega che sia alta un po' meno di uno dei masu

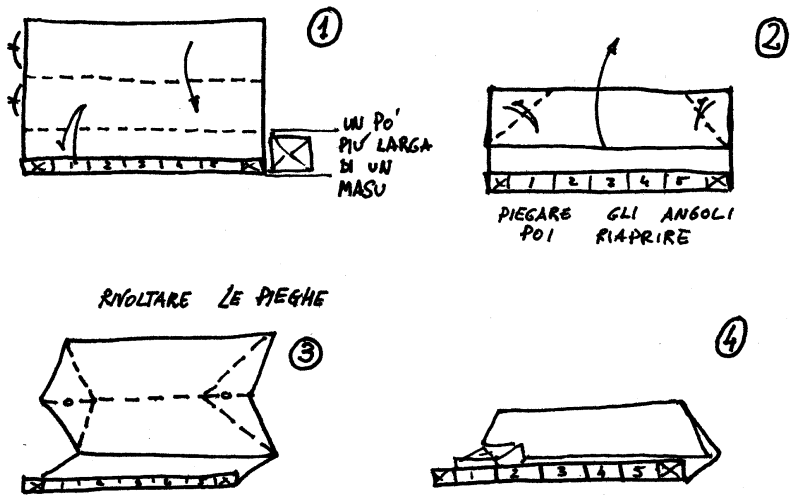


Adesso ho bisogno che qualcuno faccia un calcolo: il foglio A4, come abbiamo già detto, dovrebbe avere un lato di circa 21 centimetri, meno la lunghezza di 5 masu quanto fa? Ecco dividiamo questo numero per 2 e riportiamo il valore sul bordo appena piegato sia da una parte che dall'altra. Dividiamo poi la sezione centrale in 5 parti uguali e numeriamole, come se stessimo costruendo un righello. Ecco, ora possiamo stabilire che un masu posto nella posizione 1 valga 1 kg, mentre un masu posto nella posizione 2 valga 2 kg e così via. Non è una cosa così strana: anche nella comune aritmetica facciamo un uso posizionale (=della posizione) dei numeri; per esempio nel numero

371

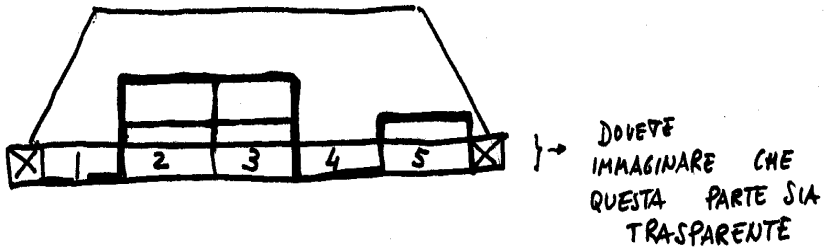
il 3 nella posizione delle centinaia vale trecento
il 7 nella posizione delle decine vale settanta
e una unità vale 1.

Allora torniamo al nostro gioco di costruzioni: per tener fermi i masu e per dare più solidità all'insieme è meglio fare anche queste pieghe

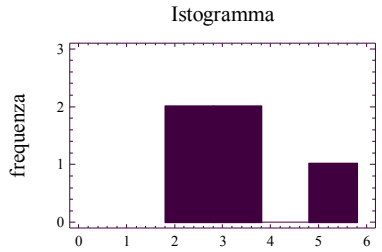


Ecco allora che con 5 masu possiamo rappresentare tutto il magazzino del sig Gervaso, basta disporli in questo modo:

- 2 masu sono i 2 sacchetti da 2 kg
- 2 masu per i 2 sacchetti da 3 kg
- 1 masu per il sacchetto da 5 kg.



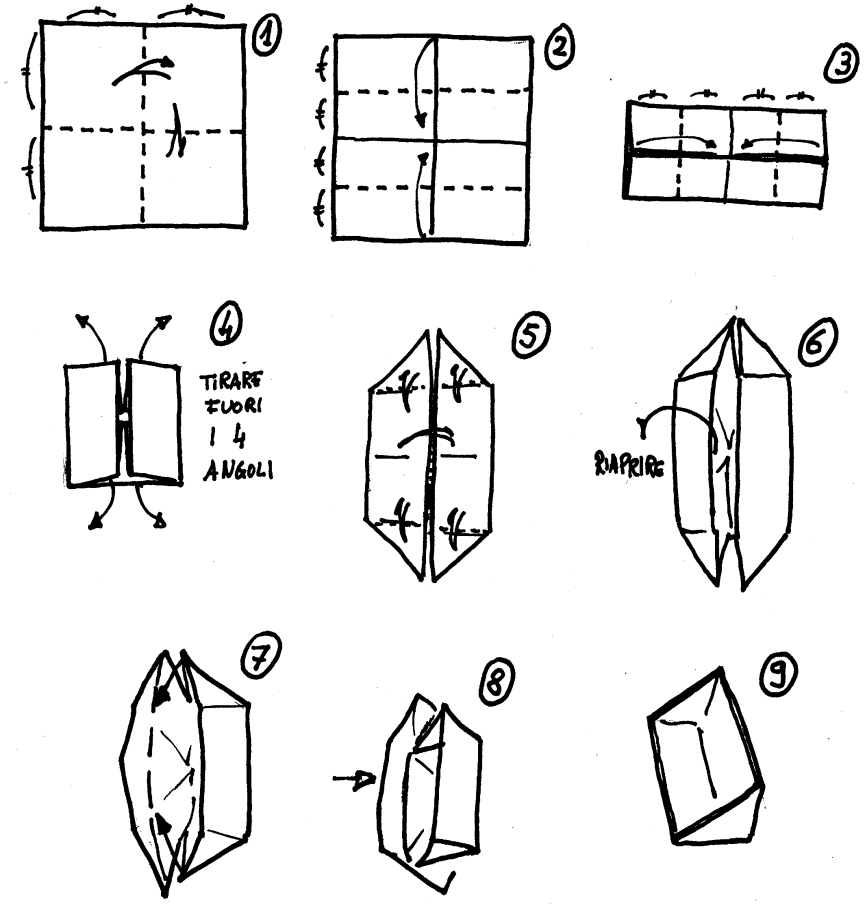
questo metodo di rappresentare i dati si chiama *istogramma*: usando i masu che avete costruito, potete divertirvi a rappresentare altri possibili insiemi di numeri.

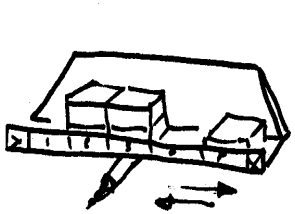


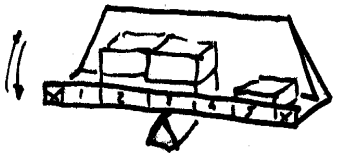
L'istogramma si può fare con i masu o lo si può disegnare sulla carta, ma in genere lo si fa disegnare ad un computer, per fare meno fatica. Quello che mi interessa raccontarvi è che l'istogramma ha una serie di caratteristiche proprio interessanti. Innanzitutto vi siete già accorti che fa risparmiare del lavoro: nel nostro piccolo esempio 5 masu ne rappresentano 15, ma in un problema più grosso un masu potrebbe anche rappresentare 100 sacchi di farina o mille cammelli o molti di più; in questo modo i dati vengono sintetizzati, vengono riassunti e si può vedere a colpo d'occhio come sono organizzati.
Ma adesso attenzione, provate a cercare il punto di equilibrio dell'istogramma che rappresenta il magazzino del sig Gervaso.

Si può fare in 2 modi: possiamo mettere sotto la costruzione una matita rotonda e farla rotolare a destra e a sinistra fino a trovare il punto di equilibrio

Oppure possiamo fare questa piega (che va bene anche come tetto di una casa-masu mentre si gioca alle costruzioni) ed usarla come fulcro.
Questa piega viene da un bellissimo libro che si chiama "Origami Omnibus", scritto da Kuniko Kasahara (vedi bibliografia [1]).







Allora, trovato il punto di equilibrio? Anche qui non è facile: dobbiamo accontentarci di un soluzione approssimata, ma va bene lo stesso: a me viene che il punto di equilibrio corrisponde al numero 3 della scala che abbiamo riportato sotto: ecco questo valore è *la media* dei pesi dei sacchetti di caramelle.
Magari qualcuno di voi sapeva già cos'è la media; probabilmente vi avevano insegnato a calcolarla sommando insieme i valori dei pesi dei sacchetti e dividendo per il numero dei sacchetti, così:

$$(2+3+3+5+2)\div 5=3$$

non è un caso, viene esattamente lo stesso numero perché la media è proprio il baricentro dell'istogramma.
Ripensate allo schema di pag 7, scriviamolo in modo un po' più ordinato così:

Peso (kg)	***Numero sacchetti***	***Peso x Numero***
1	0	0
2	2	4
3	2	6
4	0	0
5	1	5
tot	5	15

Ecco, in termini più generali la *media aritmetica* si calcola dividendo tra loro i 2 numeri dell'ultima riga; nel nostro caso : $15\div 5=3$
In altre parole si moltiplica il valore di ciascuna osservazione per la sua occorrenza, si sommano tutti questi prodotti e si divide per il numero delle osservazioni. Questo equivale a fare $(2+3+3+5+2)\div 5=3$

Ma attenzione a non confondersi; a volte il numero di volte cui è capitata una osservazione viene detta il suo *peso*, nel nostro esempio, caso ha voluto che le osservazioni rappresentino dei pesi (fisici) che moltiplichiamo quindi per dei pesi (matematici).

Tornando alla nostra piega: forse vi siete accorti che abbiamo costruito una macchina per calcolare le medie! Una specie di computer di carta che calcola le medie e che funziona senza pile! Basta mettere i masu, fare l'istogramma e trovare il punto di equilibrio: la media si legge sotto.

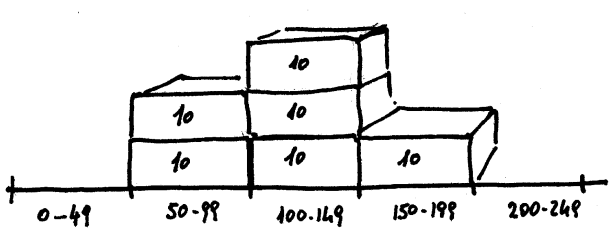
Come avete detto? Funziona solo con numeri che vanno da 0 a 5; bé tutti i computer hanno dei limiti di calcolo: il mio (che costa un sacco di soldi e che consuma corrente) non è capace di calcolare la differenza tra 10 alla 308 e 10 alla 308 meno 1 (provate sul vostro). Comunque per maneggiare numeri più grandi basta costruire dei masu più piccoli o impiegare un foglio più grande (o usare il foglio A4 piegato nell'altro senso).

Ma in effetti la nostra macchina per le medie sembra avere un'altra limitazione: lavora solo sui numeri interi. Questa è una osservazione interessante: è vero che in teoria basterebbe costruire dei masu più piccoli, ma pensate: se la bilancia del signor Gervaso pesasse i grammi oltre che i chili, questo ci costringerebbe a piegare dei masu grandi un millesimo di quelli che abbiamo fatto fin qui e vi garantisco che usare un foglio grande un decimo di millimetro per piegare un masu è piuttosto difficile.

Vi ricordo però che è un caso che nel nostro esempio 1 masu = 1 kg; nessuno ci vieta di rappresentare in istogramma questo insieme di numeri

138 113 134 195 87 70 75 195 91 116 145 126 174 149 131 83 53 138 173 163 104 129 121 51 144 50 72 76 194 137 112 136 96 146 142 131 135 132 113 132 69 102 76 137 167 83 60 103 118 120 52 69 149 56 52 161 83 158 153 136

In questo modo



E' più semplice capire come si fa se prima mettiamo i numeri in ordine crescente; non che questo sia indispensabile, è solo più comodo per me mostrarvelo,

50 51 52 52 53 56 60 69 69 70 72 75 76 76 83 83 83 87 91 96
102 103 104 112 113 113 116 118 120 121 126 129 131 131 132 132 134 135
136 136 137 137 138 138 142 144 145 146 149 149
153 158 161 163 167 173 174 194 195 195

ora se sottraiamo il più piccolo dal più grosso otteniamo *il range*

195-50=145

adesso dobbiamo decidere in quante classi dividere 145, a questo proposito esistono diverse regole empiriche, per esempio la tabellina seguente

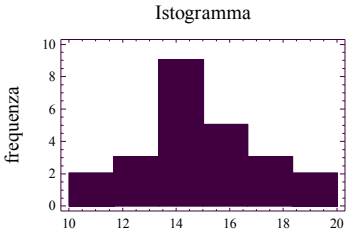
meno di 30 osservazioni l'istogramma serve a poco
meno di 100 osservazioni massimo 8 classi
da 101 a 250 massimo 10 classi
da 251 a 1000 massimo 12 classi

Nell'esempio io ho deciso di fare 4 classi (di cui una vuota), vi prego di notare che i limiti delle classi sono scelti in modo da non lasciare ambiguità nell'assegnazione delle osservazioni alle classi; inoltre ho deciso di mettere in ciascuna classe un masu ogni 10 osservazioni. D'accordo ho barato: il numero delle osservazioni per ciascuna classe è esattamente divisibile per 10 così non ho dovuto usare dei masu segati a metà; ma avrei sempre potuto stabilire che un masu vale 3 osservazioni o 13 o una.
Inoltre in appendice c'è un riferimento ad un modello di cubo, grande 2 masu, fatto con 2 fogli di carta (quindi pesa 2 masu); combinando cubi e masu si può rendere il sistema ancora più versatile.

Ecco come potremmo operare con numeri decimali, per esempio, i 30 numeri decimali

4,37348	19,2912	20,7221	12,1345	14,8025	22,2741
12,2369	15,5669	18,0976	17,748	14,5603	13,6388
8,89321	12,5463	15,3694	17,5275	13,7957	19,9823
14,256	15,1928	11,2705	20,9492	14,1695	23,5501
14,5677	15,3169	10,7054	14,4311	16,1116	13,4388

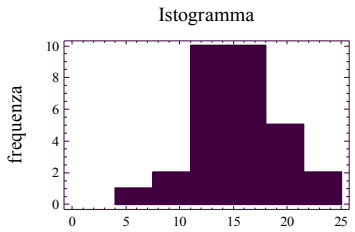
valore minimo = 4,37348 valore massimo = 23,5501 range = 19,1766
possono essere rappresentati in un istogramma così



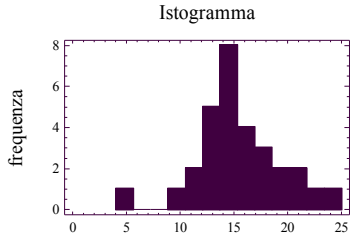
oibò dov'è finito il 4,37: il computer di sua iniziativa ha deciso di fare 6 classi a partire dal 10 ed ha anche chiuso l'ultima classe a 20
allora noi lo obblighiamo a partire da 4 e finire a 25
costruendo queste 6 classi

1	da 4,0 a 7,5	1 osservazione	1 masu
2	da 7,5 a 11,0	2 osservazioni	1 cubo
3	da 11,0 a 14,5	10 osservazioni	5 cubi
4	da 14,5 a 18,0	10 osservazioni	5 cubi
5	da 18,0 a 21,5	5 osservazioni	2 cubi + 1 masu
6	da 21,5 a 23,25	2 osservazioni	1 cubo

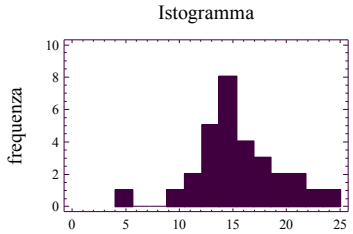
ecco il grafico



se poi cambiamo il numero delle classi da 6 a 13, l'istogramma diventa così:



attenzione, il computer ha cambiato la scala delle ordinate. Se rimettiamo la scala dell'istogramma precedente (da 0 a 10) questo istogramma diventa così:



Ora vi pregherei di prendervi qualche minuto per confrontare i grafici tra di loro; visto come sembrano diversi? Eppure rappresentano tutti lo stesso insieme di dati.
A questo punto spero di esser riuscito a farvi capire quanto sia importante definire con attenzione il numero delle classi di un istogramma e le dimensioni degli assi di un grafico questa è una regola generale molto importante, sia quando vi capita di fare un grafico, sia quando vi capita di guardare un grafico fatto da altri: attenzione alle scale.

Capitolo 2
Misurare la dispersione

Ma torniamo al nostro amico: allora il signor Gervaso ha scoperto che la sua macchina fa sacchetti che pesano mediamente 3 kg e qui mi sembra di sentire la famosa battuta: "La statistica dice solo bugie se un uomo mangia un pollo mentre un altro uomo resta digiuno, per la statistica mangiano ½ pollo a testa". Questa battuta, oltre ad essere vecchia, è proprio sbagliata : la media in effetti è ½ pollo a testa, ma la statistica non è fatta solo dalla media. E torniamo proprio all'esempio del magazzino del signor Gervaso; racconta la storia che dopo un po' quei sacchetti di caramelle andarono venduti, vennero dei bambini e se li comprarono: è vero che i sacchetti pesavano mediamente 3 kg, ma andatelo a raccontare ai due bambini cui capitarono i sacchetti da 2 kg, vi garantisco che rimasero piuttosto delusi, soprattutto quando il bambino cui era capitato il sacchetto da 5 chili cominciò prenderli in giro; era un bambino grasso che in seguito avrebbe avuto molti problemi col dentista; comunque Gervaso ci rimase male.
Allora è necessario inventare un modo per calcolare come sono dispersi i valori dei pesi dei sacchetti rispetto alla media. Ecco, potremmo calcolare quanto si discosta ciascun valore dalla media

$$33333-23352=100-21$$

cioè
3 – 2 = 1
3 – 3 = 0
3 – 3 = 0
3 – 5 = - 2
3 – 2 = 1

Adesso potremmo calcolare la media di questi scarti.
Ma acc! Viene zero!
Viene sempre zero qualunque insieme di numeri scegliate.

Se ci pensate un momento è ovvio: i numeri saranno un po' più grandi e un po' più piccoli della media, in modo esattamente bilanciato, vi ricordate: la media è il baricentro.
Ora una formula matematica che come risultato dà sempre zero serve poco; allora per avere qualcosa di più interessante potremmo elevare al quadrato le differenze: un quadrato non è mai negativo (se si eccettua il numero immaginario *i* che elevato al quadrato fa –1).

Ecco che viene: $1+0+0+4+1=6\div 4=1,5$

Hai sbagliato! Dirà qualcuno, hai diviso per 4, dovevi dividere per 5, i sacchetti erano 5.
Non è sbagliato, si divide per il numero delle osservazioni meno 1, e per complicarci la vita al risultato di questo calcolo (n-1) si da anche l'altisonante nome di *gradi di libertà*. Per sapere il perché dovete avere pazienza per qualche pagina, ve lo spiego dopo, ora credetemi sulla fiducia.
Il numero che abbiamo calcolato si chiama *varianza*, qualcuno lo chiama anche scarto quadratico medio, ma varianza è più semplice.
La sommatoria dei quadrati degli scarti dalla media si chiama anche *devianza*; quindi
varianza = devianza / gradi di libertà
(in appendice ho riportato tutte le formule scritte nella usuale notazione matematica).
Allora il signor Gervaso ha scoperto che la sua macchina confeziona sacchetti che pesano mediamente 3 kg con una varianza di 1.5 kg^2; kg al quadrato? Sì, avendo elevato al quadrato le differenze ora ci troviamo con dei chili al quadrato. Ma che cos'è un chilo al quadrato? Sono più buone le caramelle al quadrato? (Non cominciate a farmi confusione in testa: le caramelle al quadrato non sono necessariamente caramelle quadrate) Comunque sia, non si sa, semplicemente non si sa se le $caramelle^2$ siano più buone; però per semplificarci la vita potremmo estrarre la radice quadrata della varianza e trovare così una misura della dispersione con le stesse unità di misura di partenza; allora

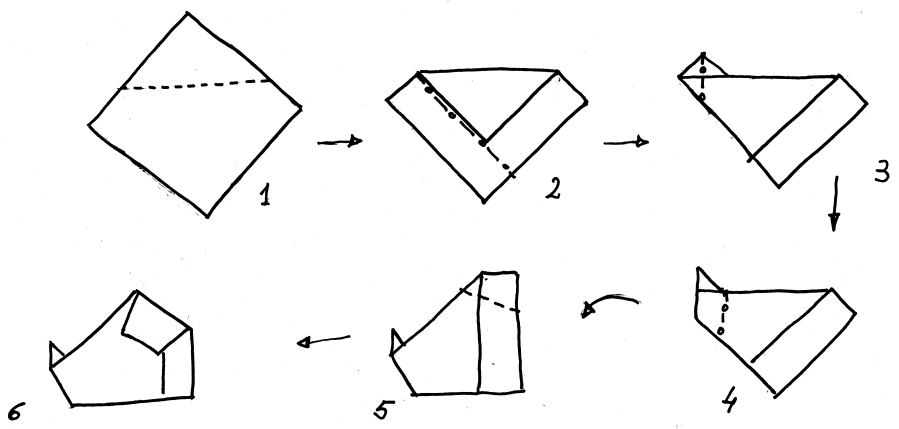
$\sqrt{1,5}\approx 1,2$ (si legge "la radice di 1,5 è circa uguale a 1,2")

quello che abbiamo calcolato si chiama *deviazione standard*

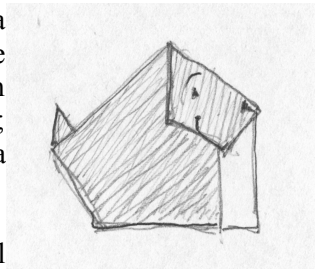
Dice Gervaso: "Va bene 3 chili di media, ma la dispersione è alta; la deviazione standard è di un chilo e 2 etti, quasi il 50% della media: la macchina è proprio scassata!"

Capitolo 3
Misure di posizione, di dispersione e di associazione

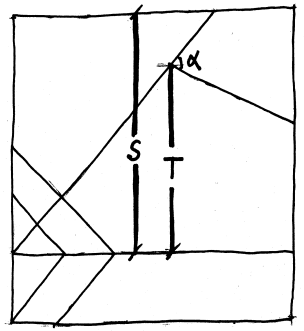
Ecco la media è una misura di *posizione* perché ci dice dove è posizionato l'istogramma; mentre la deviazione standard è una misura di *dispersione*. Esistono molte altre misure della posizione e della dispersione; per provare a conoscerne qualcun'altra torniamo a giocare con la carta. In bibliografia [5] c'è un libro di Nick Robinson da cui ho preso la piega di questo cagnolino.



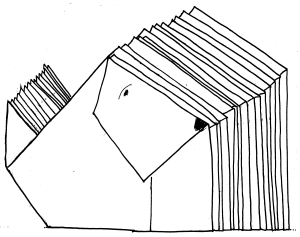
Ora vi prego di osservarla con attenzione: è una piega molto elegante nella sua essenzialità, viene meglio se usate carta per origami colorata su di un solo lato, iniziando con la faccia colorata sotto; volendo si possono aggiungere alcuni particolari a penna, così:



Questa piega ha una caratteristica particolare: il primo e l'ultimo passaggio non hanno dei riferimenti precisi, ma la decisione di dove fare le pieghe viene lasciata al senso estetico di chi le realizza. Questo non è raro in origami; l'arte non può essere vincolata da regole troppo rigide, anzi forse il succo dell'estetica sta proprio nel trovare il delicato equilibrio tra libertà e vincoli.
Ma lasciamo da parte la filosofia e torniamo al nostro cagnolino, provate a piegarne un po', diciamo una quindicina, provate a variare la prima e l'ultima piega, divertitevi ad osservare come cambia il risultato finale: è una piega semplice, la cosa non dovrebbe richiedervi troppo tempo.
Forse vi sarete già accorti che le variabili su cui giocare sono 3 (quattro se calcoliamo anche le dimensioni della carta) le ho riportate nella figura qui sotto identificandole con le lettere S, T ed α (si legge alfa).
(Secondo Nick Robinson se α supera i 90° il cagnolino diventa un mammouth).

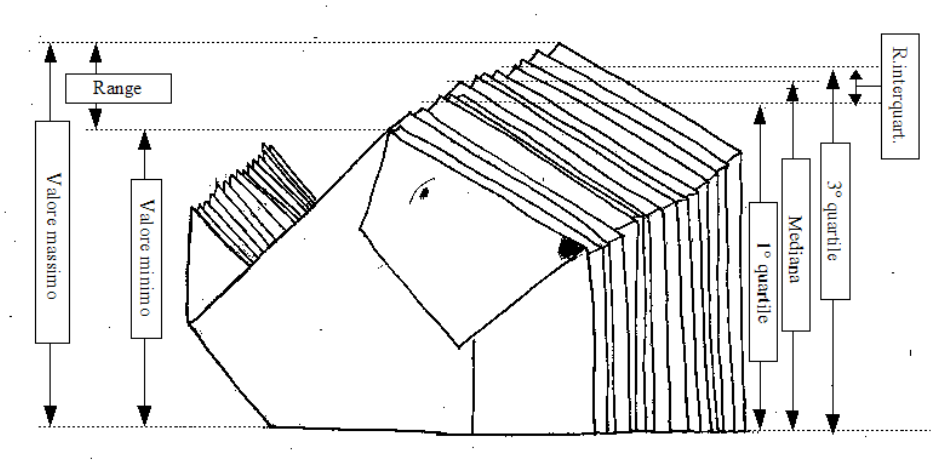


Riaprite un cagnolino ed osservate le pieghe: S dipende da dove avete fatto la piega al passo 1; mentre T ed α descrivono dove e come avete piegato la testa del cagnolino al passo 5.
E' affascinante pensare che 3 numeri possano descrivere in modo completo la "biometria" del nostro cagnolino, è come se stessimo studiando una nuova razza canina ed avessimo la fortuna di poterne modellizzare l'anatomia in modo completo con tre soli numeri. Questo suggerisce la possibilità di molti altri giochi ed esperimenti, ma non voglio divagare. Ora mi vorrei concentrare su una sola delle tre variabili: quella che nel disegno qui sopra viene identificata con la lettera T (come Taglia): l'altezza del cagnolino. Allora è come se avessimo "catturato" una quindicina di esemplari della nostra nuova razza (Canis Origamicus) ed ora volessimo studiarli, ovvero descriverli in base alla sola taglia.
Se mettete i vostri cagnolini in piedi, l'uno contro l'altro, non dovrebbe essere difficile ordinarli per altezza, come in figura



Adesso è semplice identificare il cagnolino di mezzo: il numero 8, quello per cui 7 cagnolini sono più piccoli di lui, 7 sono più grandi: ecco la taglia di questo cagnolino è la *mediana* del nostro campione di animali. Naturalmente qualcuno a questo punto protesterà: io ho fatto 14 animali e non c'è un cagnolino "di mezzo", è vero, dicendo "una quindicina" non pretendevo un numero esatto di esemplari, e infatti non c'è alcun problema: se il numero di cagnolini è pari, basta misurare i 2 cagnolini

"di mezzo" (nel caso di 14 esemplari il settimo e l'ottavo) sommare i valori e divider per 2: in altre parole la mediana del nostro campione è in questo caso la media dei due valori centrali.
La mediana è un bel modo per descrivere come è posizionato il nostro insieme di osservazioni senza stare a fare tanti conti; sì, avete indovinato, la mediana è un altro indice di posizione, come la media. Inoltre, se avessimo voluto calcolare la media delle taglie del nostro campione avremmo dovuto misurare tutti i cagnolini, mentre per avere la mediana di un campione, non importa quanto grande, basta fare una o 2 misurazioni.
Ricordate però che misurare solo la posizione di un campione ci espone al rischio di qualche svarione (se un uomo mangia un pollo, mentre un altro resta digiuno...)



Allora procediamo; la mediana divide un campione in 2 gruppi: la metà "dei piccoli" e la metà "dei grandi". Ma nulla ci vieta di prendere ciascuno dei 2 gruppi e ripetere l'operazione dividendo ciascuna metà in 2 quarti. La taglia del cagnolino che separa in quarti le due metà si chiama *quartile*; rispettivamente il *primo quartile* separa il quarto dei cagnolini "piccoli piccoli" dai 3 quarti dei più grandicelli, mentre il *terzo quartile* separa il quarto dei cagnolini più grandi dai 3 quarti più piccoli. E il *secondo quartile*? E' semplicemente un altro modo (poco usato) per chiamare la mediana.
Se poi calcoliamo la differenza tra il 3° ed il 1° quartile otteniamo il *range interquartile*, mentre il *range*, come avevamo già visto nel capitolo 1, è la differenza che passa tra il cagnolino più alto e quello più piccolo di tutti.

Range e range interquartile sono altri due indici di dispersione, come la deviazione standard.
In qualche caso, soprattutto quando i campioni sono molto grossi, si preferisce dividerli anziché in 4 parti in 100 parti; allora i valori che identificano queste parti prendono il nome di *percentili*. Stavo pensando che per esemplificarvi il concetto basterebbe piegare 2 o 300 cagnolini , metterli in ordine d'altezza e misurare la taglia di quelli che ….
Ma forse riuscite ad immaginare la cosa con la fantasia senza stare a piegare tutti quei cagnolini; basterà qualche esempio; il 3° percentile è la misura che separa il 3 per cento di osservazioni, più piccole di lei, dal 97 % di osservazioni, più grandi di lei; il 50° percentile è la mediana; il 90° percentile indica la misura superata solo dal 10% dei campioni, e così via.

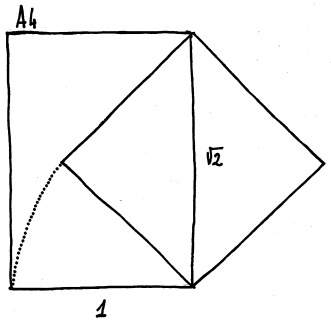
Abbiamo visto quindi alcuni indici di posizione e di dispersione; esistono anche gli indici di *associazione.* Infatti a volte è utile avere qualcosa che ci indichi quanto due misure siano rappresentate l'una dall'altra, mi spiego meglio con un esempio; nel fabbricare il torrone, la pasta di zucchero e mandorle viene trafilata tra 2 cilindri, poi la striscia che ne esce viene tagliata a pezzi lunghi un tot. E' probabile che sia più comodo controllare il peso dei pezzi di torrone piuttosto che la loro lunghezza, perché il metro si appiccica al torrone e diventa noioso fare le misure (il metro non si lecca, per favore!).
E' ragionevole pensare che, se la sezione rimane costante, ci sia un legame tra lunghezza e peso del torrone, anche se il legame non può essere esatto; per esempio dipende da quante mandorle son capitate in quel singolo pezzo. Allora con la statistica possiamo valutare quanto siano associate le misure dei pesi con le misure delle lunghezze di un campione di torroni. Per esempio possiamo calcolare quale percentuale della deviazione standard della lunghezza possa essere spiegata dalla deviazione standard del peso. Ecco questa è una misura di associazione: in genere la si indica con R^2. Un'altra misura di associazione è il coefficiente di correlazione, detto anche "r"; un numero che vale 0 quando le due variabili non sono correlate proprio per niente; vale 1 quando data una misura possiamo ricavare esattamente l'altra e tanto più una cresce, tanto cresce l'altra; mentre vale -1 quando se una cresce l'altra cala, ancora con un legame matematicamente esatto.

Se approfondirete questa branca della statistica, magari vorrete provare a vedere se c'è una qualche associazione tra i valori di T, di S e di α dei cagnolini che avete piegato, ma questo lo lasciamo per un'altra storia.

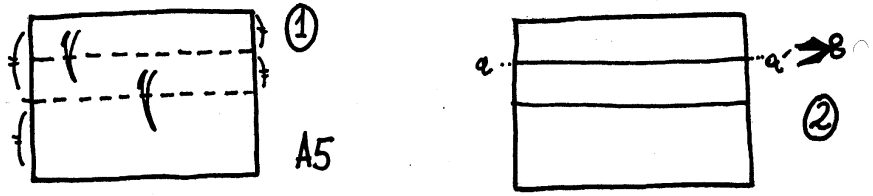
Capitolo 4
Le distribuzioni statistiche
(infiniti masu)

Sicuramente ricorderete che per semplificarci la vita avevamo immaginato che il magazzino del signor Gervaso contenesse solo 5 sacchetti di caramelle, ma a me piace lavorare anche con numerosità molto grandi; allora immaginiamo di poter avere il peso in grammi di tutti i sacchetti di caramelle che la macchina di Gervaso ha prodotto e di tutti quelli che farà in futuro, anzi, mi voglio rovinare: tutti gli infiniti sacchetti prodotti e da produrre, pesati con una precisione assoluta; poi facciamo l'istogramma. E' impossibile da fare, direte voi, ci vogliono infiniti masu e per fare infiniti masu ci vuole un tempo infinito e una pazienza infinita; allora, prima che io esaurisca la vostra, bisogna che mi inventi qualche trucco. Benissimo, torniamo a giocare con la carta: prendiamo un foglio A4 e tagliamolo a metà, otterremo 2 fogli A5.

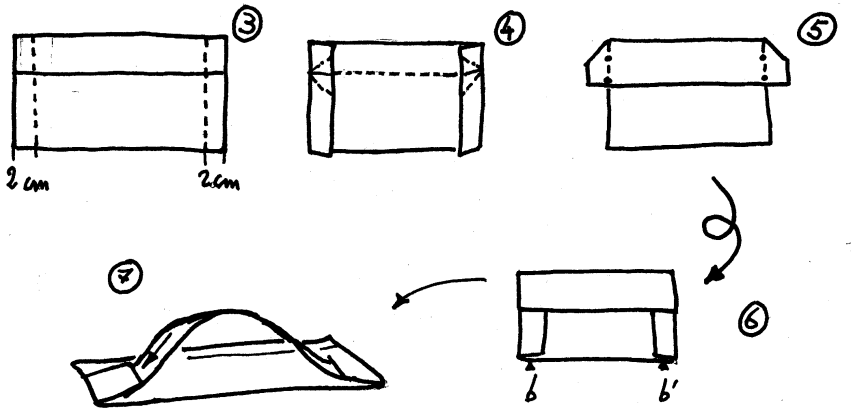
Una caratteristica interessante dei fogli che utilizziamo in Europa per fare le fotocopie (formato UNI) è che il lato lungo è lungo come il lato corto moltiplicato per √2 (lo so che è brutto detto così ma fa niente). Che sarebbe come dire che il lato corto misura come il lato del quadrato di cui il lato lungo è la diagonale (sembra uno scioglilingua eh?). Ma la cosa interessante è che, dividendo un foglio come abbiamo fatto noi le proporzioni restano esattamente le stesse. Allora ciascuno dei 2 fogli A5 ha le stesse proporzioni dell'originale foglio A4: sono rettangoli simili.



Prendiamo dunque un foglio A5 e pieghiamolo così

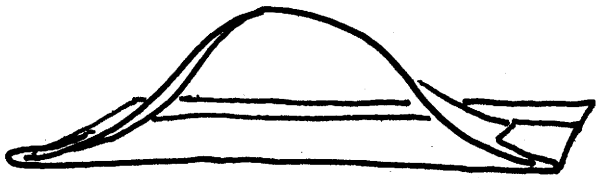


Poi tagliamo lungo la linea a-a' tenendo da parte la striscia di carta e proseguiamo la piegatura così.



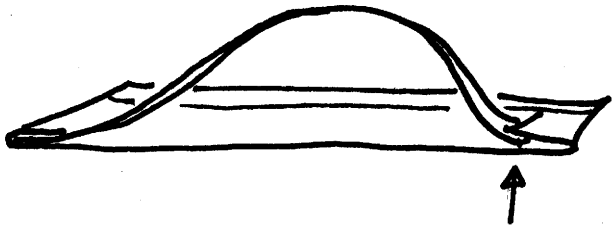
Infilando infine la striscia che avevamo messo da parte nelle due tasche b e b'. Poi bisogna girare il modello in modo da poterlo guardare di fianco; così.

Adesso osservate il profilo della striscia di carta; disegna una curva particolare che è molto importante in statistica e che definisce proprio la forma che avrebbe un istogramma come quello che volevamo realizzare: con infiniti masu infinitamente piccoli; mica male eh? Zic, zac, 2 tagli, 3 pieghe invece di infiniti masu piegati con carta infinitamente piccola.



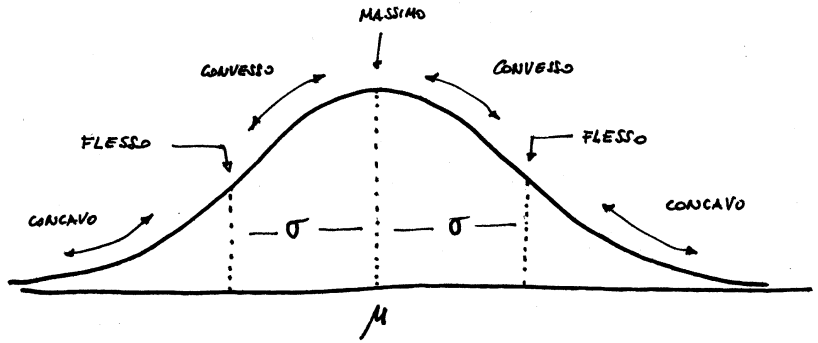
Un bel po' di lavoro risparmiato, mi sembra quasi di poter percepire la vostra immensa gratitudine.
In effetti c'è qualche precisazione da fare; la curva che abbiamo realizzato descrive una distribuzione: come si distribuirebbero i masu (infiniti) se la macchina scassata di Gervaso sbagliasse a fare i sacchetti in modo *normale*. Cosa vuol dire sbagliare in modo normale? Diciamo che sono "normali" gli sbagli che avvengono in modo assolutamente casuale. Questo indipendentemente da dove avvengano: potrebbero essere "sbagli" della macchina confezionatrice oppure errori nello strumento di misura, ma sempre in modo casuale, non c'è nulla che alteri le misure in modo sistematico.

Per esempio la distribuzione è simmetrica, il che vuol dire che Gervaso è in buona fede, infatti se provate a spostare un pochino l'estremo di destra della striscia, così



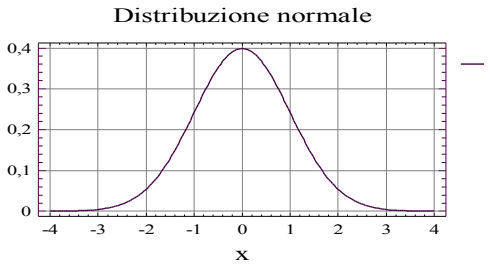
Otteniamo una distribuzione diversa, asimmetrica, come se Gervaso, ogni tanto, accorgendosi che un sacchetto è troppo pieno lo togliesse dal magazzino, ma togliesse solo i sacchetti troppo pieni, mica anche quelli troppo vuoti (chiamalo fesso). Ora rimettete a posto i 2 lembi in modo da ritornare ad avere una distribuzione simmetrica; osservate con attenzione, per favore, le estremità della striscia; a seconda di come avete piegato il modello possono toccare o meno la superficie del tavolo. Attenzione: la vera distribuzione normale ha una differenza importante con il modello che abbiamo realizzato: la striscia arriva a toccare il tavolo, ma solo ad una distanza infinita, del resto la striscia ha una lunghezza infinita (come il tavolo) ma sono sicuro che questo potete immaginarlo con la fantasia senza abbattere infiniti alberi per avere a disposizione infinita cellulosa per costruire una striscia infinita di carta. A proposito, quando avrete finto di giocare con gli origami, per cortesia ricordatevi di gettare la carta nei contenitori per la raccolta differenziata.

Ma ritorniamo alla distribuzione normale, che viene chiamata anche *gaussiana* in onore del famoso matematico Johann Carl Friedrich Gauss (1777- 1855), se la osservate bene potete notare che la curva sale prima con una concavità verso l'alto, poi la curvatura cambia e diviene convessa, raggiunge un massimo, poi scende convessa e poi di nuovo concava, ecco: il punto più alto della curva corrisponde alla media (provate a trovare il baricentro della gaussiana che avete costruito); mentre i punti in cui la curva da concava diviene convessa distano dalla media esattamente il valore di una deviazione standard.



Bello eh! La prima volta che me l'hanno raccontato mi sono divertito un sacco, sapete io sono uno che si diverte con poco.
Nella formula della gaussiana (in appendice) compaiono i simboli μ e σ (si leggono mi e sigma), dove μ è la media e σ è la deviazione standard; questi vengono chiamati *parametri* della gaussiana, perché date una media ed una

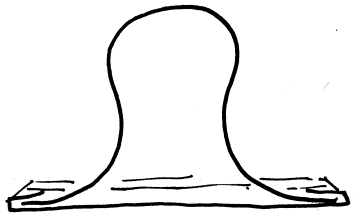
deviazione standard si ottiene una ed una sola gaussiana. Quando $\mu=0$ $\sigma=1$ la gaussiana si chiama gaussiana standardizzata.



Adesso provate a costruire un'altra gaussiana come quella che abbiamo appena montato, ma, prima di inserire la striscia nelle 2 taschine accorciate la striscia di 2 centimetri. Ecco dovreste ottenere una cosa con questa forma.

Così appare una gaussiana con una deviazione standard più grande. Ora confrontando le 2 gaussiane e facendole scivolare sul tavolo potete simulare quello che accade quando cambia la media: la gaussiana si sposta a destra o a sinistra (si sposta il baricentro); o quando cambia la deviazione standard: la gaussiana si "allarga" o di "restringe"; in effetti non è proprio che si allarghi, ricordate che le estremità raggiungono il tavolo all'infinito, quindi le gaussiane sono tutte larghe infinito, quindi sono tutte larghe uguali; diciamo che se σ aumenta diventano un po' più spampanate (che ci crediate o no sullo Zingarelli "spampanato" c'è).

Se provate a pasticciare un po' con strisce di carta lunghe e corte, vi renderete presto conto che il sistema di allungare la striscia per simulare la riduzione della deviazione standard funziona solo fino ad un certo punto, dopo di che la curva assume una forma come questa,



che non è una gaussiana; del resto stiamo solo facendo un modello di una funzione matematica piuttosto complicata, e il modello, come tutti i modelli ha delle limitazioni.

Ad ogni buon conto ci sono alcune proprietà della gaussiana che sono proprio interessanti; per esempio la media, visto che divide esattamente in 2 la distribuzione, è uguale alla mediana; inoltre nell'intervallo che va dal valore della media meno la deviazione standard al valore della media più la deviazione standard sono comprese circa il 68% delle osservazioni e, più in generale si può calcolare che:

μ	$\pm 1\ \sigma$	= 68,26 %
μ	$\pm 2\ \sigma$	= 95,45 %
μ	$\pm 3\ \sigma$	= 99,86 %

Quindi si possono calcolare i percentili in base alla deviazione standard.

Ma torniamo al signor Gervaso che, viste le scarse prestazioni della sua macchina confezionatrice, decise di effettuare delle drastiche operazioni di manutenzione straordinaria. Per cui, dopo aver convinto Adalgisa, la gallina ovaiola a scegliere per la cova un luogo diverso dai contrappesi della bilancia; dopo aver tolto dai leverismi per la chiusura del sacchetto gli addobbi natalizi che da anni vi stazionavano e dopo aver sistemato altri piccoli dettagli; raccolse un nuovo campione di sacchetti di caramelle ed ecco i pesi in grammi

2995 3010 3007 2999 2998 2994 3006 3003 2998 2992
3002 3004 3005 2997 3002 3003 3006 3002 3009 3008
3000 3001 2995 2990 3011

media= 3001
deviazione standard=5,7

media – 3 deviazioni standard =2984
media + 3 deviazioni standard =3018

quindi, concluse Gervaso, mi aspetto che circa il 99,8% dei miei sacchetti di caramelle pesi tra i 2 chili e 984 grammi e i 3 chili e 18 grammi; adesso penso che i bambini non dovrebbero più lamentarsi.

Capitolo 5
Altre distribuzioni

Ora dovrebbe essere chiaro quanto sia comodo poter utilizzare la distribuzione gaussiana, quando si può, come modello di eventi casuali. In effetti non abbiamo una garanzia assoluta che gli errori nel confezionamento delle caramelle abbiano esattamente una distribuzione gaussiana, più avanti vedremo come accertarcene.
Comunque avere una distribuzione di riferimento è una cosa così comoda che gli statistici hanno cercato molte altre distribuzioni, adatte a descrivere eventi differenti. Per esempio, abbiamo detto che la gaussiana descrive una misura continua cui è applicato un errore casuale; invece la distribuzione *binomiale* può essere usata per descrivere eventi con 2 sole possibili alternative; come quella volta che Gervaso decise di sistemare la produzione delle caramelle col buco: che probabilità c'è di avere caramelle col buco e che probabilità c'è che le caramelle riescano senza buco.
La distribuzione *poissoniana* in genere si dice che è adatta per descrivere eventi rari, per esempio fu usata dal colonnello von Bortkiewicz (1868-1931) alla fine del 1800 per descrivere i morti da calcio di cavallo, per ciascun anno, per ciascun corpo d'armata dell'esercito prussiano.
La distribuzione *uniforme* descrive eventi che hanno tutti la medesima probabilità; mentre la distribuzione di *Weibull* viene spesso impiegata per descrivere l'andamento dei guasti.

Tante storie sono state scritte su queste ed altre distribuzioni, ma non fanno parte di questo libro.

Capitolo 6
Media campionaria e media di popolazione
(quanto succo di liquirizia ci ha messo?)

Una volta Gervaso partì per un viaggio, doveva partecipare ad uno stage sui canditi organizzato dal suo amico Barbadigesso. Aveva lasciato il laboratorio alle amorevoli cure del suo migliore collaboratore: Tonio. In quel momento era in corso la lavorazione di una partita di super giusoni: more di liquirizia secondo una ricetta segreta di Gervaso. Solo che, nel trambusto della partenza, Gervaso si era dimenticato di lasciare precise disposizioni su come procedere con la lavorazione; in particolare Tonio non riusciva a ricostruire quanto succo di liquirizia il maestro avesse già messo nel pentolone in cottura.
In realtà Gervaso aveva messo esattamente 500 millilitri di succo di liquirizia nel pentolone da 50 litri di sciroppo in preparazione quindi 500 millilitri diviso 50 litri (vale a dire 50000 ml), fa esattamente 0,01 cioè una concentrazione dell 1%, ma questo Tonio non lo sapeva e non voleva disturbare Gervaso per chiederglielo con un piccione viaggiatore (i telefoni cellulari allora non si usavano ancora).
Allora Tonio decise di prelevare un piccolo campione dal pentolone e di analizzarlo per determinare la concentrazione esatta di succo di liquirizia; sul campione effettuò 5 analisi ottenendo questi risultati

0,01 0,015 0,02 0,008 0,022

con una media di 0,015
ma come, non doveva fare 0,01? Già noi sappiamo che la media è esattamente 0,01, ma questo Tonio non lo sa ed è possibile che, vuoi per una miscelatura imperfetta degli ingredienti, vuoi per qualche imprecisione negli strumenti di misura, la media su un piccolo campione di misure non faccia esattamente 0,01. La vita è piena di tante assurdità che, sfacciatamente, non hanno neppure bisogno di parere verosimili, perché sono vere (e così abbiamo citato anche Pirandello [6]). Così Tonio pensa che nel pentolone siano stati messi 750 ml di succo di liquirizia mentre noi sappiamo che non è vero. "Lo dicevo io che la statistica imbroglia" mi sembra quasi di sentire una vocina; ma anche questa volta non è vero: semplicemente dobbiamo fare molta attenzione a non confondere la media calcolata sulla base di un campione con la vera media del pentolone. La media ricavata da un campione e calcolata in un modo qualsiasi: a mano, con un computer o col nostro computer di carta, non fa differenza; resta una media calcolata su di un campione, per questo si chiama *media campionaria* e in genere la si indica con una piccola linea sopra il nome della variabile per es. la media di x è $\bar{x}$
La vera media del pentolone, invece viene di solito chiamata *media di popolazione o media vera* e in genere non si riesce mai a conoscerla esattamente, la si indica con la lettera greca μ ed è uno dei parametri della gaussiana. Già perché è logico pensare che se anche analizzassimo tutto il pentolone, con infiniti campioni non otterremmo sempre lo stesso valore, ma avremmo delle misure distribuite in modo normale, secondo una gaussiana, quindi con una media ed una deviazione standard; questo perché il movimento delle molecole di succo di liquirizia in un pentolone è intrinsecamente variabile e può essere descritto solo con metodi statistici, in gergo si dice che è un fenomeno *stocastico*.

Vi assicuro che, quando ho iniziato a studiare la statistica, da autodidatta, ci ho messo un sacco di tempo a capire perché in una parte del libro per la media si usava il simbolo $\bar{x}$, mentre altrove si usava il simbolo μ.
Ecco, ora la cosa dovrebbe esservi evidente: $\bar{x}$ è semplicemente il risultato di un calcolo, mentre μ è qualcosa che non conosciamo e che vorremmo proprio poter stimare, perché in questo modo potremmo usarlo come parametro di una gaussiana e quindi come modello di tutto l'universo dei dati che stiamo analizzando.
In genere noi siamo nella stessa condizione di Tonio, non possiamo conoscere la media di popolazione (non possiamo analizzare tutto il pentolone); possiamo solo calcolare una media su di un campione. Ma, direte voi, un rapporto tra le due cose ci sarà bene! Se non altro perché si è deciso di chiamarle con lo stesso nome! E infatti uno degli scopi della statistica è proprio quello di aiutarci a stimare il valore della media vera; in gergo l'operazione di stimare dei parametri si chiama *inferenza* e per questo questa branca della statistica viene chiamata statistica inferenziale.

E come si fa? Semplicissimo, prima di tutto si calcola la deviazione standard sul campione (la chiameremo s) e si divide s per la radice quadrata della numerosità del campione. Tonio aveva preso 5 campioni dal pentolone allora la deviazione standard calcolata sui 5 campioni è circa 0.0060 , diviso la radice quadrata di 5 fa circa 0,0027, questo valore si chiama *errore standard*

Allora Tonio non sa dove sta la media vera, ma la statistica gli dice che la probabilità di trovarla è distribuita (anche lei!) come una gaussiana con media uguale alla media campionaria e deviazione standard uguale all'errore

standard. Quindi (secondo la tabella a pag. 17) c'è il 95% di probabilità che la media vera stia tra 0,015 più o meno 2 volte l'errore standard quindi tra 0,0204 e 0,0096 e in effetti, la media vera (che noi conosciamo) è compresa tra questi 2 valori. In conclusione Tonio sa che col 95% di probabilità Gervaso ha messo nel pentolone tra i 1020 ed i 480 ml di succo di liquirizia. Come dite, un po' vago? Non è colpa della statistica: o si riduce la deviazione standard o si aumenta la numerosità del campione. E' ovvio che se il campione diventa molto grosso la stima migliora, se n diventa uguale a infinito l'errore standard diventa zero e la media campionaria è uguale alla media vera. Così come è ovvio che se si mescola meglio il pentolone o si impiegano per l'analisi metodi più precisi s diventa più piccolo e la stima migliora. Ma spesso i ricercatori hanno bisogno della statistica proprio perché i fenomeni che stanno studiando sono intrinsecamente incerti e s non si può ridurre in alcun modo.

Riassumendo c'è un universo, che non possiamo conoscere (se lo potessimo conoscere non avrebbe senso usare la statistica inferenziale) da questo noi estraiamo un campione di n osservazioni, su questo campione facciamo dei calcoli (per esempio la media e la deviazione standard) poi stimiamo qual è l'errore che potremmo commettere decidendo che le misure che abbiamo calcolato siano i parametri della distribuzione che descrive l'universo da cui siamo partiti. Vi prego di notare che questo ragionamento generale vale per qualsiasi distribuzione, l'importante è decidere qual'è la distribuzione giusta da usare.

Un'altra cosa interessante è che c'è un teorema: il Teorema centrale del limite che dimostra che qualsiasi sia la distribuzione del nostro universo di partenza (bé σ^2 non deve essere infinito), se estraiamo tante volte n campioni ed ogni volta calcoliamo una media campionaria, tutte queste medie tenderanno comunque a distribuirsi secondo una gaussiana e abbiamo visto quanto sia comodo usare la gaussiana come modello.

Adesso, per cortesia, andate a riprendere l'istogramma che avevamo fatto con i masu a pag. 15 e la gaussiana di pag. 8 io vi avevo detto che quella gaussiana era come un istogramma fatto con un numero infinito di masu infinitamente piccoli, ma ora possiamo essere più precisi: in effetti la nostra gaussiana di carta ha un σ che vale circa 1 masu, mentre la s calcolata a pag. 12 era di 1,2 questo perché avevamo usato carta A4, foglietti di circa 10 cm e perché avevamo piegato la carta a 2 cm … tutti valori stabiliti ad hoc per far tornare (all'incirca) i conti.

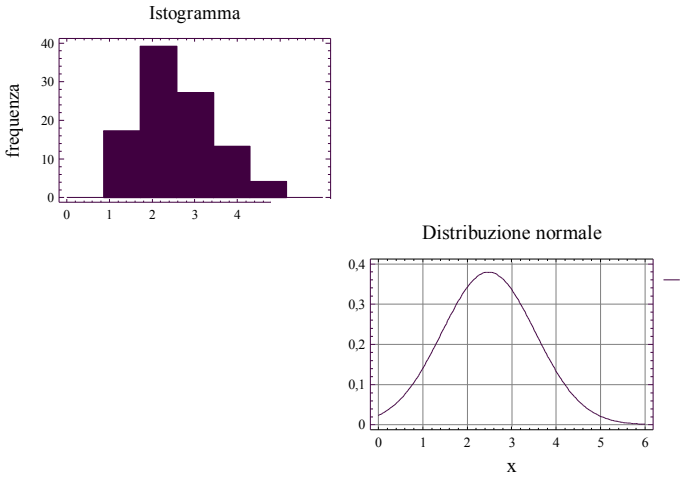
Inoltre adesso possiamo provare a sovrapporre la gaussiana all'istogramma, sapendo che la nostra stima μ=3 è affetta da un errore standard di 1,2/ $\sqrt{5}$ = 0,54
quindi la media vera potrebbe essere tra 1,92 e 4,08 kg (col 95% di confidenza circa).

Se poi ricordate il magazzino dei sacchetti di caramelle di Gervaso, quello di pag. 5 , quello troppo grosso …
ora tutti quei numeri non ci fanno più paura

la media è = 2,48
la deviazione standard è = 1,05
e l'errore standard è= 0,10

quindi la media vera dovrebbe essere tra 2,28 e 2,68 kg (col 95% di probabilità circa).

ecco l'istogramma relativo ed una gaussiana con μ=2,48 σ=1,05



Capitolo 7
La verifica di un test
(le caramelle mou ultramorbide)

Dovete sapere che una delle specialità assolute di Gervaso erano delle caramelle mou che, oltre ad essere buonissime, erano veramente morbidissime e succosissime, ma le cose non erano sempre andate così. Una volta, quando Berta non aveva ancora imparato a filare, Gervaso produceva ancora caramelle mou di quelle normali, che si attaccano ai denti, finché un giorno un suo giovane collaboratore (Tonio, sempre lui) trovò per caso un ingrediente che, unito all'impasto, produceva caramelle mou morbidissime e scioglievolissime. Sono certo che i più curiosi tra di voi vorranno sapere di che cosa si trattava; purtroppo è passato tanto di quel tempo che se n'è persa la memoria. Comunque Tonio preparò un po' di queste nuove caramelle e le fece assaggiare a Gervaso che le apprezzò molto ma, Tonio – disse – io ho sempre fatto le caramelle con la mia ricetta tradizionale, prima di cambiare voglio essere sicuro; come facciamo ad essere certi che queste nuove caramelle siano più morbide proprio per merito del tuo succo.
Tu hai usato uno stampo diverso, hai un modo tutto tuo di regolare il fuoco, poi c'è la temperatura di raffreddamento, e poi lo sai, ogni caramella ha la sua scioglievolezza.
Allora facciamo così: prepariamo due lotti di caramelle uno con la vecchia ricetta, uno con la nuova ricetta, cercando di usare lo stesso fuoco, lo stesso stampo e lo stesso modo di farle raffreddare, poi misuriamo esattamente la scioglievolezza e facciamo i confronti.
Va bene – rispose Tonio.
Volete sapere come si fa a misurare la scioglievolezza di una caramella mou? Facile, si fa il test della della sputazza di drago: si mette la caramella in un bicchiere pieno di sputazza di drago e si cronometra quanto tempo ci mette a sciogliersi completamente.
Ecco i valori di scioglievolezza delle 10 caramelle fatte con la ricetta di Tonio, li chiameremo A, come Tonio, che in realtà si chiamava Antonio
A= 72 82 65 83 50 61 83 68 52 75 media = 69,1
E di 10 caramelle fatte con la ricetta tradizionale, che chiameremo G, come Gervaso
G= 89 71 76 81 75 79 60 62 70 61 media = 72,4
Ma come si fa a dire quali sono le più succose: la media dei tempi di scioglimento in sputazza di drago di A è inferiore a G medio, ma c'è un paio di valori in G inferiori alla media di A: è un caso? E allora? Dovremmo fare un'altra prova? E la sputazza che serve la procurate voi? Attenzione poi che non si può scaricare la sputazza usata dove e come si vuole: è un rifiuto speciale, altamente inquinante.

In realtà le cose sono molto più semplici: basta applicare un test, in statistica si parla di test di ipotesi perché, in effetti, si fa un'ipotesi e si verifica quell'ipotesi; o sarebbe meglio dire si cerca di falsificare quell'ipotesi. Infatti l'ipotesi è sempre una ipotesi di non differenza, nel nostro caso l'ipotesi è che G sia uguale ad A; e la si chiama *ipotesi nulla*, per gli amici H0. Quindi H0 : A = G
Ma se A è uguale a G la differenza tra le loro medie dovrebbe essere uguale a zero, anche se dobbiamo ricordarci che le due medie sono solo delle stime e quindi dovremo tener conto di questo calcolando l'errore standard della differenza tra le medie.

Allora facciamo un calcolo: la differenza tra le medie diviso l'errore standard di questa differenza, so benissimo che non abbiamo ancora imparato a calcolare l'errore standard di una differenza tra medie, ma non vi facevo così appassionati alle formule; la formula, come le altre, è in appendice.

La cosa veramente interessante è che il numero che vien fuori segue anche lui una distribuzione nota, cioè è distribuito secondo una funzione matematica nota. La distribuzione è stata descritta per la prima volta da William S. Gosset (1876- 1937) nel 1908 mentre lavorava per la Alec Guinness & Co. sì proprio quella della birra, vi avevo detto che la statistica è uno strumento pratico, serve anche a fare la birra.
Il dottor Gosset pubblicò risultati dei suoi studi con lo pseudonimo di "studente" in inglese Student, per cui la distribuzione da allora viene chiamata t di Student. Questa distribuzione ci permette di calcolare che probabilità c'è che un certo valore di t (di Student) ci sia capitato per caso; ma siccome t è una differenza tra le medie (diviso un errore), dire che è un caso che c'è una differenza è come dire che non c'è differenza, che sarebbe un po'come dire che probabilità c'è che H0 sia vera.
Allora, se questa probabilità è sufficientemente bassa possiamo concludere che H0 probabilmente è falsa, quindi A è diverso da G.

Lo so che vi sembra di aver perso il filo del ragionamento, proviamo a ricapitolare con uno schema.

In pratica qualcuno si è già preso la briga di

1. inventare una formula che misura la differenza tra 2 campioni
2. dimostrare che il risultato segue una distribuzione
3. calcolare i valori di questa distribuzione
4. sistemarli in una tabella, ordinati per probabilità che H0 sia vera, cioè che sia vero che non ci son differenze tra le medie

Allora a noi non resta che

a) definire una H0 (qualcosa è uguale a qualcos'altro)
b) calcolare la statistica test (t, nel nostro esempio)
c) cercare nella tabella se per quel valore (di t) la probabilità che H0 sia vera è alta o bassa
d) se la probabilità è bassa allora si respinge H0 (le due cose sono diverse)
e) se la probabilità è alta allora si dice che non si può respingere H0 (probabilmente le 2 cose non sono diverse)

Come tutte le scienze la statistica ha il suo gergo, e come tutti i linguaggi il gergo della statistica ha la sua ragione di esistere; in effetti dire che si respinge H0 (punto "d") è un po' come dire che H0 è falsa, ma è più corretto dire che è probabilmente falsa. Mentre dire che se la probabilità è alta "non si può respingere H0" (punto "e") sembra un inutile bizantinismo, ma in effetti non è così: perché vi ho detto una imprecisione: la probabilità che si trova in tabella non è la probabilità che H0 sia vera, ma la probabilità di sbagliare dicendo che è falsa, e questa non è la stessa cosa.

ecco allora il valore del t- di Student calcolato usando G ed A
t = -0,675284
che corrisponde ad una probabilità (P) di 0,508078, vale a dire circa il 50%
quindi c'è il 50% di probabilità di sbagliare dicendo che G non è uguale ad A (respingere H0), per cui non ci conviene respingere l'ipotesi nulla: l'ingrediente di Tonio non modifica la scioglievolezza in modo significativo.

In effetti ho tralasciato di precisare una cosa. Io ho fatto calcolare al mio computer sia il valore di t che quello della P riportati qui sopra; magari qualcuno di voi vuol cimentarsi nel provare a fare i conti a mano: bell'esercizio. Allora la formula per calcolare t sta in appendice (si chiama t di student per dati non appaiati, dopo vi spiego perché), ma come si calcola la P?

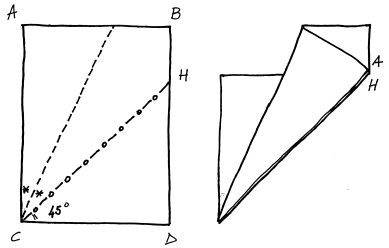
Calcolare il valore esatto della probabilità è piuttosto complicato, allora si impiegano delle tavole, come già accennato. La tavola del t di student riporta i valori critici di t per alcune probabilità, in genere almeno 0,05 e 0,01. Basta confrontare il nostro valore con quelli tabulati per capire se corrisponde ad una probabilità inferiore all'1% ; compresa tra l'1 ed il 5% o superiore al 5%.
Per comprendere bene la cosa è necessario che vi procuriate una tavola del t di Student per osservarla; ne trovate una in appendice a qualsiasi libro serio di statistica; non correte all'appendice di questo libro, non la troverete: ho detto libro serio. Se non ne avete uno scaricate da internet il manuale di Lamberto Soliani [8]. Attenzione! Per ciascun valore della probabilità ci sono tanti valori di t che sono ordinati secondo i gradi di libertà.
Gradi di libertà, questo nome non mi è nuovo... Infatti l'avevamo già incontrato nella formula della varianza (cap.2), vi avevo promesso che avrei spiegato perché si divide per n-1 anziché per n, per quello dovete pazientare fino al capitolo sull'ANOVA; ora invece cerchiamo di capire cosa c'entrano i gradi di libertà nella tavola del t di Student.
Se ci pensate un attimo è una cosa intuitivamente semplice: non può essere la stessa cosa confrontare tra loro 2 campioni di 10 osservazioni ciascuno o 2 campioni di 1000 osservazioni ciascuno. Ecco allora che i gradi di libertà ci aiutano a "tarare" il test sulla misura del nostro campione.

Quello fatto sin qui è un discorso generale che vale per moltissimi test statistici, dove ciò che cambia è solo la statistica test e la sua distribuzione di riferimento. Facciamo qualche esempio.

Nel nostro caso stiamo confrontando la scioglievolezza di 2 campioni di super giusoni di liquirizia, le misure del primo campione sono indipendenti dalle misure del secondo campione allora il test t- di Student va bene.

Ma una volta Gervaso si trovò nei guai con i folletti che, golosissimi di giusoni di liquirizia rischiavano di farsi venire la pressione alta. Infatti qualcuno sostiene che mangiando troppa liquirizia la pressione arteriosa si alzi; allora Gervaso chiamò un suo amico dottore che misurò la pressione dei folletti prima e dopo il turno di lavoro ai super giusoni; si sa che i folletti quando lavorano ai super giusoni, assaggia qua, assaggia là, un chilo di giusoni non glie li leva nessuno: questa sembrava una buona "prova da carico".
I questo caso non possiamo trascurare il fatto che le pressioni arteriose misurate si riferiscono sempre agli stessi folletti, cioè la x-esima misura

prima del carico di liquirizia corrisponde alla x-esima misura dopo la liquirizia: sono tutte e 2 misure del folletto numero x. In questo caso non si può usare il t di Student nella forma che abbiamo visto, ma si deve usare un'altra formula che qualcuno chiama *t di Student per dati appaiati*; la distribuzione di riferimento resta la stessa, ma cambia il modo di calcolare la t ed i gradi di libertà.
Vi assicuro che non è un bizantinismo: facciamo un esempio.
Prendiamo 2 variabili
M= 5 23 18 9 12 25 19 14
N= 7 24 19 11 15 25 20 15
Se le vogliamo confrontare tra di loro con il t di student dobbiamo prima sapere da dove provengono i dati per decidere se usare il t di student per dati appaiati o il t di student per dati non appaiati. Allora vediamo; abbiamo già visto che in un foglio A4 i lati sono in proporzione $1 \text{div} \sqrt{2}$ questo può essere verificato con la piega qui sopra. (NB la piega a valle divide l'angolo ACH esattamente a metà).



Immaginiamo di voler confrontare tra loro l'accuratezza di 2 fabbricanti di carta, allora M e N sono le misure, in micron, della distanza tra A e H presi su 10 fogli di 10 risme differenti dei fornitori Manuelo e Nando. Applicando il t per dati non appaiati t=-0,419 P=0,68 gdl=16 quindi non si può respingere H0.

Se invece immaginassimo di avere 10 fornitori per la carta e di avere loro proposto un prezzo per la carta legato alla loro precisione nel taglio, potremmo cercare di verificare se l'incentivo economico ha avuto effetto confrontando la precisione su un foglio per ciascun fornitore prima (M) e dopo (N) il cambiamento di contratto,in questo caso bisogna usare il t per dati appaiati: t=-4,24 P=0,003 gdl=8 per cui respingiamo H0.

Visto?
Una diversa provenienza dei dati porta (in questo caso) a conclusioni opposte. Questo è molto importante: nessun computer può sapere da dove avete pescato i vostri dati e come avete effettuato le misure: questo dovete assolutamente deciderlo voi.

Altre volte capita di analizzare dati che non sono misure, ma conteggi: si riferiscono a variabili qualitative: vi ricordate le caramelle col buco? Una caramella il buco o ce l'ha o non ce l'ha. Non si può misurare ½ buco, o 2 buchi virgola 7. Anche in questo caso non si può usare il t di Student ma bisogna usare altre statistiche test, come per esempio il *chi- quadrato* (per gli amici χ^2).

Un'applicazione interessante del test del chi quadrato è quella di verificare se è ragionevole ipotizzare che un certo campione di osservazioni provengano da una popolazione distribuita in modo normale.

Se i dati non risultano distribuiti in modo normale, allora non è possibile utilizzare molti test statistici; si deve far ricorso ad una nuova famiglia di test: i *test non parametrici*.

La cosa importante è che tutti questi test si comportano alla stessa maniera: si definisce un H0, si cerca di falsificarla, si calcola una statistica test e si va a vedere la sua distribuzione su una tavola o con un computer. capito il meccanismo una volta va bene sempre; solo bisogna fare attenzione a scegliere il test giusto.

Spesso, parlando della probabilità ho scritto probabilità piccola, probabilità alta, si, ma quanto? Di solito si usa il 5% (in qualche caso l%), vale a dire 0,05 (o 0,01) se P è minore di questi valori ci si sente autorizzati a respingere H0.
Ma attenzione ci resta sempre un 5 % di probabilità di prendere lucciole per lanterne, vale a dire di considerare diversi 2 campioni che invece non lo sono: ecco questo si chiama *errore di primo tipo* detto anche *errore alfa* (α). Ovviamente esiste anche un altro tipo di errore: quando H0 è falsa ma noi non la respingiamo, questo si chiama *errore di II° tipo* o *errore beta* (β).

La realtà	La mia decisione	
	respingo H0	non respingo H0
in realtàH0 è vera	errore α	OK
in realtà H0 è falsa	OK	errore β

Capitolo8
ANOVA
(ancora le caramelle mou ultramorbide)

Un giorno Bortolo, l'altro assistente di Gervaso, forse perché un po' geloso, insisteva che l'ingrediente ideale per migliorare la scioglievolezza delle mou fosse la salsapariglia e così, tanto disse e tanto fece, che Gervaso preparò altre 10 caramelle con la ricetta di Bortolo, ecco i valori
B= 79 52 80 68 61 68 74 71 76 73
Ora abbiamo un problema: con cosa dobbiamo confrontare B:

- con G ?
- Con A?
- E' lo stesso, tanto abbiamo già "dimostrato" che non c'è differenza tra i 2?

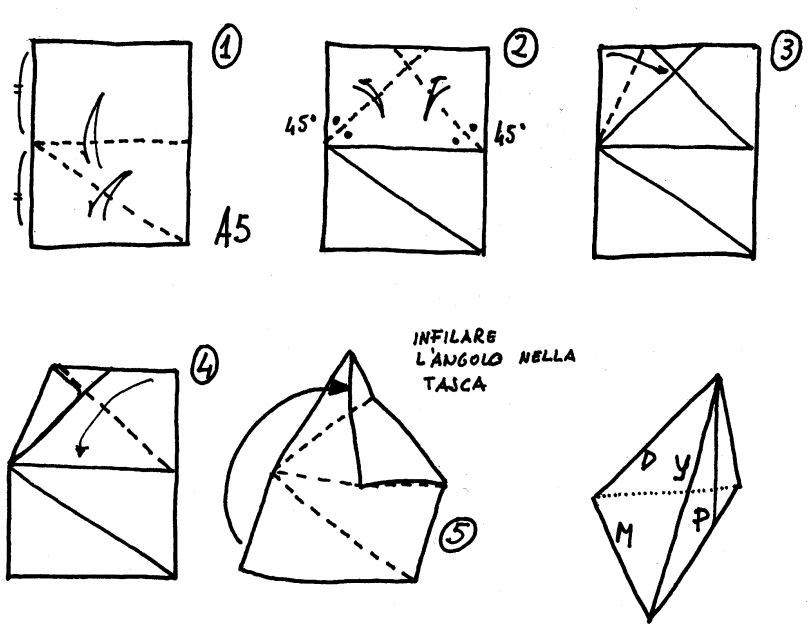
Ma in effetti non abbiamo dimostrato che non c'è differenza; abbiamo stabilito che non conviene respingere l'ipotesi nulla che non ci sia differenza tra loro. Per fortuna esiste un bellissimo strumento della statistica che sembra proprio fatto per toglierci di impaccio, infatti può aiutarci a rispondere a domande di questo genere (anche quando i problemi sono più seri o più complessi) è *l'analisi della varianza*; ANOVA per gli amici.

Prima di continuare il nostro discorso sull'ANOVA, però devo presentarvi un nuovo concetto: il concetto di *vettore*.
E' semplice, si prendono un po' di numeri messi in fila insieme: ecco un vettore. Quindi B è un vettore come anche G ed A. In genere i vettori si scrivono in grassetto ecco allora scriviamo **B** è un vettore, come anche **G** ed **A**; così è più corretto e il pignolo che c'è in me è più contento.
Può darsi che qualcuno di voi abbia già sentito parlare dei vettori, studiando le forze, in fisica; allora forse vi avevano spiegato che i vettori sono delle specie di frecce con una lunghezza, un orientamento ed un verso. Non c'è contraddizione tra le due definizioni: se disegnate una di queste frecce su un sistema di assi cartesiani, mettendo l'origine della freccia nel punto 0 0; allora la punta del vettore–freccia si troverà in un certo punto x y; per esempio x=13 y=78. Questi 2 numeri insieme formano un piccolo vettore, il vettore
13 78
Se poi la freccia si trovasse in uno spazio tridimensionale il suo vettore avrebbe 3 elementi, sarebbe fatto di 3 numeri (x,y e z); ma allora un vettore di 5 elementi è come una freccia in uno spazio a 5 dimensioni; come dite? Non esistono spazi a 5 dimensioni? Bè nulla ci vieta di immaginarli a 5, 7 e 256 dimensioni, immaginare non costa niente.

Allora prendiamo i 3 vettori **G, A** e **B** ed "attacchiamoli insieme" per fare un vettore più lungo che chiameremo **Y** = 89 71 76 81 75 79 60 62 70 61 72 82 65 83 50 61 83 68 52 75 79 52 80 68 61 68 74 71 76 73

Adesso basterebbe costruire con la carta uno spazio a 30 dimensioni e metterci il nostro vettore. Vi confesso che non sono capace di piegare un origami a 30 dimensioni; ma non è una cosa così grave: non avete mai visto una carta geografica? Cosa c'entra una carta geografica: bé una carta geografica è un esempio di una proiezione in uno spazio a 2 dimensioni di qualcosa di tridimensionale. Ecco allora noi costruiremo una proiezione tridimensionale (3d) di un qualcosa a 30 dimensioni (**Y**).

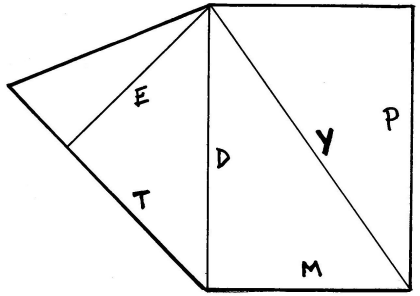
Partiamo da un foglio A5 costruiamo questa piega



Ecco, osserviamo un momento la figura: è una piramide triangolare e tutte le sue facce sono triangoli rettangoli.
Gli spigoli della piramide seguono le proporzioni 1; √2 ; √3

6 piramidi come questa formano un cubo; o meglio, con 3 piramidi come questa + altre 3 speculari si può fare un cubo.

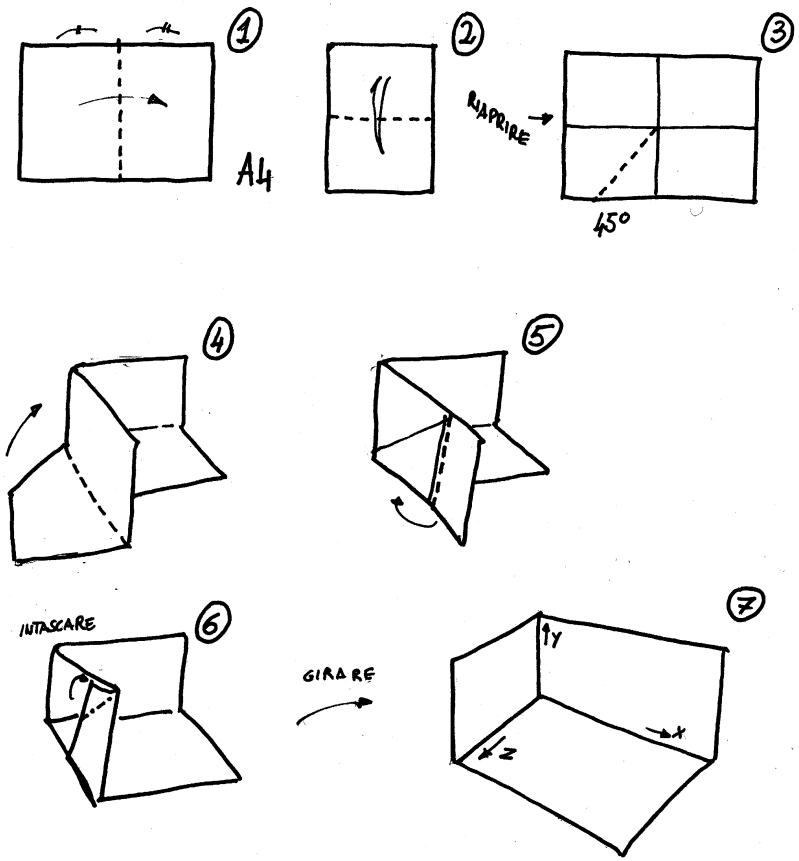
Per comodità è meglio identificare gli spigoli della piramide con delle lettere, riaprite il modello fino al punto 5, poi giratelo in modo da avere verso l'alto le pieghe a monte e contrassegnate le pieghe, che poi sarebbero gli spigoli della piramide, seguendo questo schema:



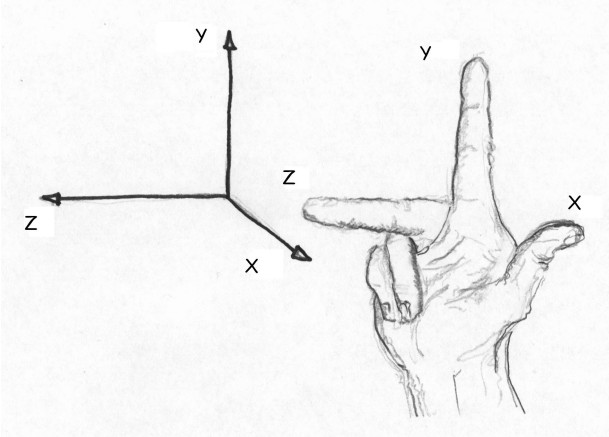
Adesso rimontate la piramide e osservate per il momento solo il triangolo formato dagli spigoli Y, D e M. Immaginiamo che lo spigolo che abbiamo contrassegnato con la lettera "Y" rappresenti il nostro vettore **Y**, per essere più precisi dovremmo dire una proiezione 3d del vettore **Y**, ma sicuramente avevate capito lo stesso.

Ora immaginate **Y** nello spazio, provate a muoverlo nello spazio cartesiano: il vertice compreso tra **Y** e **M** lo mettiamo nell'origine, mentre l'altra estremità di **Y** (dove c'è l'angolo con **D**) può muoversi dove vuole.

Potete costruire lo spazio cartesiano con un foglio A4 e magari scrivere i nomi dei 3 assi delle 3 dimensioni x,y, e z.



Oppure potete usare 3 dita della vostra mano, messe come in figura



Attenzione, i vettori possono anche assumere valori negativi, per cui se usate il modello di carta dello spazio cartesiano dovete immaginare che la piramide possa anche penetrare attraverso i piani xy yz e xz costruiti con la carta.

Dicevamo che l'estremità di **Y** può assumere qualsiasi posizione nello spazio; mi spiego meglio, immaginate che Gervaso stia dettando a Tonio i valori sperimentali uno per volta, mentre quest'ultimo sistema il vettore nello spazio a 30 dimensioni… già perché nel mondo della fantasia esiste uno spazio 30d come qualcosa di tangibile, non chiedetemi se è fatto di carta, di compensato o di marzapane, io non lo so. Comunque sia è evidente che fino a quando Gervaso non ha dettato a Tonio tutti e 30 i valori, questi non sa dove mettere la freccia –vettore: ogni nuova osservazione specifica dove collocarlo rispetto ad un certo asse (una certa dimensione) e solo quando tutto **Y** è stato esaminato, è possibile collocare il vettore con precisione nello spazio. Si può dire che **Y** ha la libertà di trovarsi in un qualunque punto dello spazio, quindi in uno spazio a n dimensioni ha n gradi di libertà.

Adesso immaginiamo che lo spigolo **M** rappresenti la media di **Y**; ma la media è un singolo numero, e come si rappresenta in uno spazio 30d? Semplice, così:
70,6 70,6 70,6 70,6 70,6 70,6 70,6 70,6…….trenta volte

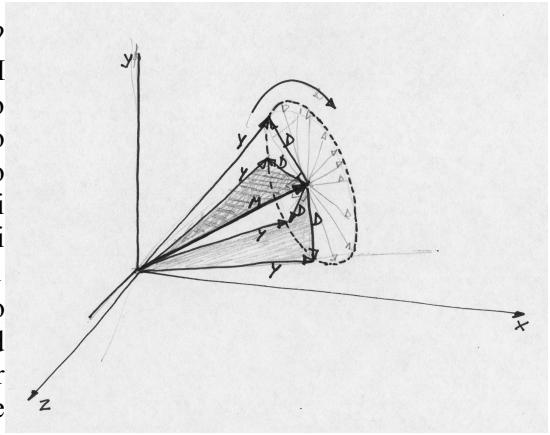
30 volte lo stesso numero. E dove sta un vettore con qualsiasi numero di dimensioni, fatto tutto da numeri uguali? Sta per forza su di una retta che passa per l'origine e che è equidistante da tutti gli assi: se gli assi sono 2, siamo su di un piano (2d) e **M** sta sulla bisettrice dell'angolo tra l'asse x e l'asse y; se siamo in 3d, **M** sta sulla diagonale di un cubo che ha un vertice nell'origine e così via, in 30d **M** sta sulla diagonale di un ipercubo a 30d. Quindi, qualunque sia il numero n delle dimensioni, **M** può muoversi solo lungo una retta che passa per l'origine: in una sola dimensione, infatti, appena Gervaso ha detto a Tonio uno dei valori di **M**, ecco che Tonio sa dove mettere il vettore: tanto sa che gli altri 29 valori sono tutti uguali. Per cui **M** ha sempre e comunque un solo grado di libertà.
Sono sicuro che a questo punto morite dalla voglia di sapere che cos'è **D**; un attimo di pazienza, prima devo dirvi ancora un paio di cose sui vettori.
I vettori hanno alcune particolarità che riguardano il modo di fare le operazioni matematiche:

il valore di un vettore si ottiene sommando il quadrato di tutti i suoi elementi (se pensate ad un vettore nel piano 2d ed al teorema di Pitagora, la cosa vi apparirà evidente)

la somma e la sottrazione tra due vettori si fanno sommando e sottraendo gli elementi corrispondenti dei 2 vettori (l'avevamo già fatto a pagina 12 senza sapere che stavamo facendo operazioni con i vettori: già eravamo più in gamba di quello che credevamo di essere). Oppure, se i vettori sono ortogonali, il vettore somma si può ricavare semplicemente disegnando il vettore che unisce le due estremità dei vettori di partenza (le punte delle 2 frecce) .

Per cui
Y = M + D quindi **D = Y - M** cioè **D** è la differenza tra le singole osservazioni e la media.
Il valore di **D** (la sommatoria degli scarti dalla media elevati al quadrato) è una nostra vecchia amica: la devianza.

Ora **M** sta sulla bisettrice dell'angolo tra gli assi cartesiani, quindi teniamolo fermo lì come in figura;



allora **D** come si può muovere? Può solo ruotare intorno ad **M** come intorno ad un asse. Può muoversi solo in un piano perpendicolare a **M**, cioè in uno spazio 2d, cioè con un grado di libertà in meno di quelli di partenza, in generale con n-1 gradi di libertà. Vi avevo promesso che ci sarei arrivato ed eccoci qua, ecco perché per calcolare la varianza si deve dividere per n-1, il motivo è semplice : per calcolare la media (che ci serve per calcolare gli scarti dalla media) abbiamo già impiegato un grado di libertà ed ora ne abbiamo solo n-1 per calcolare la devianza (e quindi la varianza e la deviazione standard). Infatti se Gervaso detta a Tonio i valori di **D**, appena è arrivato al penultimo valore, ecco che Tonio, abile matematico, lo ferma e gli dice:
«scommettimo che indovino l'ultimo valore?»
«Facile: so che la sommatoria di tutti i valori di D (se non li elevo al quadrato) fa zero (ricordate a pag.12), quindi basta sommare tutti i valori che mi hai detto e vedere quanto manca a zero»
In generale se io conosco la media di un campione e conosco n-1 valori posso ricavare matematicamente l'n-esimo, quindi questo n-esimo valore non è libero di assumere tutti i valori che vuole: la media si è "mangiata" il suo grado di libertà; questo capita perché, non conoscendo la media vera, siamo costretti ad usare una sua stima, la media campionaria, anche per stimare la varianza.

Ma torniamo alla nostra piramide; adesso prendiamo in considerazione il triangolo **Y**, **P** ed **E**.
P sta per **P**revisione: è il vettore con i valori più probabili dei 3 vettori **G**, **A** e **B**; la media è la stima migliore (come abbiamo detto), quindi è anche il valore più probabile, allora **P** è fatto con le medie di **G**, **A** e **B**; eccolo qui:

72,4 72,4 72,4 …[10 volte] 69,1 69,1 69,1 …[10 volte] 70,2 70,2 70,2…[10 volte]

Y meno **P** ci dà un vettore con lo scarto dalla previsione; ci indica come variano le misure, per effetto del caso, all'interno dei 3 gruppi (**E**ntro gruppi); a volte viene anche chiamato errore: **E**; perché indica l'errore nelle nostre stime.

Mentre il triangolo che sta di sotto, quello tra **P**,**M** e **T** ci dice che sottraendo la **M**edia dalla **P**revisione otteniamo **T,** cioè è il vettore con i contributi di ciascuna nuova ricetta alla scioglievolezza delle caramelle; in genere lo si chiama effetto del **T**rattamento o variazione **T**ra i gruppi.

Ecco allora che la nostra piramide ci mostra come possiamo scomporre la devianza totale (e quindi la varianza totale) in una deviazione dovuta al trattamento **T** ed in una deviazione dovuta all'errore, alla componente casuale: **E**.
Basta guardare il triangolo **DTE**.

L'idea geniale di Sir Ronald A. Fisher (1890 1962) è stata quella di calcolare la distribuzione della statistica test che si ottiene dividendo il valore di **T** per il valore di **E**. Infatti se il contributo della nuova ricetta è grande quanto la componente casuale è logico ipotizzare che la nuova ricetta non aggiunga niente alla scioglievolezza delle caramelle, mentre se **T** è molto più grosso di **E** possiamo aspettarci di avere trovato qualcosa di interessante.
Questo test prende il nome di analisi della varianza o ANalysis Of VAriance; quasi tutti identificano la distribuzione come F di Fisher o al più come F di Snedecor- Fisher dato che Snedecor propose dei miglioramenti all'originale metodo di Fisher.

cosi' appare una classica tabella ANOVA stampata dal computer

Analisi della Varianza

	SSq	Gdl	Varianza	F	P
Tra gruppi	56,4667	2	28,2333	0,27	0,7641
Entro gruppi	2804,9	27	103,885		
Totale (Corr.)	2861,37	29			

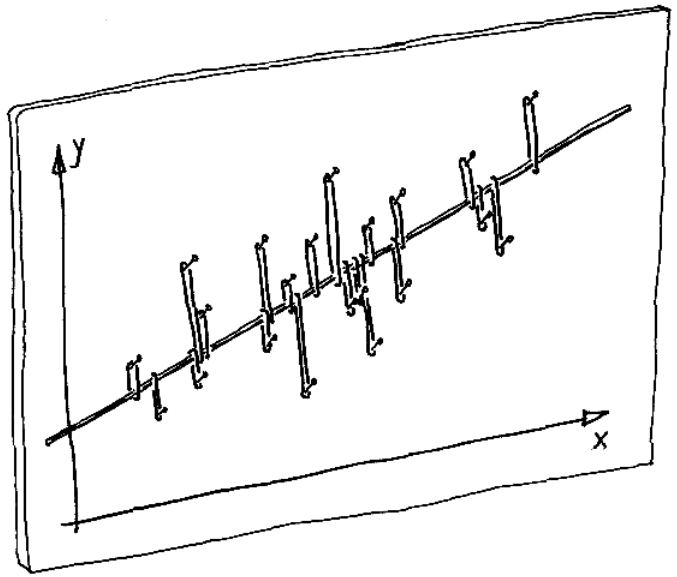
SSq è la sommatoria degli scarti quadratici (la devianza)
Gdl sono i gradi di libertà
la Varianza è la Varianza (SSq / Gdl)
F è la F di Fisher (rapporto tra le 2 varianze)
P la probabilità di sbagliare dicendo che **A**, **G** e **B** sono diversi.

Nel nostro esempio circa il 76 %
Per cui non c'è una differenza significativa tra le 3 ricette.

Probabilmente avrete già notato che come si sommano i vettori, così si sommano i gradi di libertà, per cui

vettore	gradi di libertà
Y	n (numero delle osservazioni)
M	1
D	n-1
P	k (numero dei trattamenti)
T	k-1
E	(n-1)-(k-1)

E le caramelle mou extramorbide?
Ah già, dimenticavo, quelle Gervaso le scoprì per caso.



Capitolo 9
Un cenno sulla Regressione

Se la somma tra 2 vettori è una operazione non dissimile dalla somma tra 2 numeri, il prodotto o il rapporto tra 2 vettori richiedono operazioni matematiche un po' più complesse, ma che danno risultati molto interessanti.
Ora, prima di proseguire, devo presentarvi un nuovo personaggio: *la matrice*. Se un vettore è un insieme di numeri messo in fila, una matrice è una serie di vettori "messi in fila", cioè una tabella di numeri organizzati per righe e per colonne (e magari pagine, ipercubi eccetera: una matrice può anche avere più di 2 dimensioni).

Adesso immaginate una matrice un po' particolare: fatta tutta di zeri e di uno ma disposti così

1 0 0
1 0 0
1 0 0
1 0 0
1 0 0
1 0 0
1 0 0
1 0 0
1 0 0
1 0 0
0 1 0
0 1 0
0 1 0
0 1 0
0 1 0
0 1 0
0 1 0
0 1 0
0 1 0
0 1 0
0 0 1
0 0 1
0 0 1
0 0 1
0 0 1
0 0 1
0 0 1
0 0 1
0 0 1
0 0 1

la chiameremo **X**

Nella prima colonna gli uno stanno nelle prime 10 righe, nella seconda colonna gli uno stanno nelle righe dalla 11 alla 20 e nella terza colonna gli uno occupano le ultime 10 posizioni

Ora se immaginiamo di mettere il vettore **Y** dell'esempio dell'ANOVA (Cap.9) a fianco di questa matrice risulta chiaro che gli uno della prima colonna stanno in corrispondenza delle misure fatte con la ricetta di Tonio (il vettore **A**) gli uno della seconda colonna corrispondono al vettore **G** e quelli della terza al vettore **B**.

X	**Y**	
1 0 0	89	
1 0 0	71	
1 0 0	76	
1 0 0	81	
1 0 0	75	
1 0 0	79	
1 0 0	60	
1 0 0	62	
1 0 0	70	
1 0 0	61	**A**
0 1 0	72	
0 1 0	82	
0 1 0	65	
0 1 0	83	
0 1 0	50	
0 1 0	61	
0 1 0	83	
0 1 0	68	
0 1 0	52	
0 1 0	75	**G**
0 0 1	79	
0 0 1	52	
0 0 1	80	
0 0 1	68	
0 0 1	61	
0 0 1	68	
0 0 1	74	
0 0 1	71	
0 0 1	76	
0 0 1	73	**B**

Adesso se prendiamo un computer e gli facciamo dividere il vettore **Y** per la matrice **X** otteniamo 3 numeri: uno per ciascuna colonna di **X:** vi avevo detto che non era una divisione qualsiasi.

72.4 69.1 70.2

Li riconoscete? sono le medie di **A**, **G** e **B** rispettivamente.

Non tutti i software sono capaci di dividere tra loro due matrici (o un vettore e una matrice), per cui è possibile che su qualche libro troviate un procedimento leggermente differente: prima si calcola l'inversa della matrice **X**, che si scrive $\mathbf{X}^{-1}$; poi si moltiplica questa matrice per **Y**.
Per calcolare una matrice inversa è necessario prima calcolare una cosa che si chiama *determinante* della matrice, se avete studiato l'algebra matriciale sicuramente vi avranno insegnato a calcolare un determinante e, altrettanto sicuramente vi sarete domandati a che cosa servivano tutti quei calcoli; ecco questa è un'applicazione pratica dell'algebra matriciale; se non avete mai studiato l'algebra matriciale, non importa: come vi ho detto io in genere faccio fare tutti i calcoli al computer.

Ad ogni buon conto fin qui nulla di nuovo, abbiamo solo scoperto un altro modo per calcolare la media; ma adesso incominciamo ad occuparci di qualcosa di differente.
Ricordate l'esempio del torrone di pag 14? Se costruiamo un vettore **L** con una serie di lunghezze di torroni , poi costruiamo una matrice **P** con una colonna tutta di 1 ed una seconda colonna con i pesi degli stessi torroni, possiamo facilmente cercare di valutare la relazione tra peso e lunghezza. Infatti se dividiamo il vettore **L** per la matrice **P** otteniamo 2 numeri che corrispondono rispettivamente ai parametri a e b della nota funzione della retta:

$$\{1\}\quad y = a + bx$$

a è l'intercetta tra la retta e l'asse delle y
b è la pendenza della retta.

In questo modo implicitamente abbiamo ipotizzato che il peso e la lunghezza siano in un rapporto lineare (il che e' ragionevole solo se lo spessore del torrone resta sempre lo stesso). Immaginate di procurarvi un piano cartesiano di legno; piantare un chiodo all'intersezione di ogni coppia di valori x-y e appenderci un elastico; poi bisogna procurarsi una bella bacchetta dritta dritta ed infilarla attraverso tutti gli elastici, ora basta lasciare andare e... il gioco è fatto! La bacchetta descrive la funzione della retta $\{1\}$.

Volendo descrivere la cosa in modo più formale diciamo che di tratta di render minima la sommatoria dei quadrati delle differenze tra ciascuna y_i ed il rispettivo valore $\hat{y}_i$ previsto dalla formula $\{1\}$. Naturalmente non è indispensabile conoscere l'algebra matriciale per applicare questo metodo, basta applicare una formula che si trova su qualsiasi libro di statistica (vedi per es. [8] in bibliografia).
Questo metodo si chiama appunto *metodo dei minimi quadrati*; allora y (la lunghezza del torrone) viene chiamata *variabile dipendente*, perché speriamo proprio che dipenda da x (il peso, che chiamiamo *variabile indipendente*), così non sporcheremo più il metro di pasta di torrone (vi ho già detto di non leccare il metro - per favore!) ma ci basterà pesare i torroni e poi ricavare matematicamente quanto sono lunghi; anche se la lunghezza ricavata matematicamente sarà necessariamente affetta da un errore.

E in effetti è più preciso scrivere così:

$$y = a + bx + \varepsilon$$

dove , nel nostro esempio
y è la lunghezza del torrone
x è il peso del torrone
a è l'intercetta con l'asse y, cioè quanto dovrebbe essere lungo un torrone che pesa zero grammi (ci aspettiamo che sia un numero molto vicino a zero)
b è il coefficiente angolare (nel nostro caso e' il peso di un torrone lungo 1)
ε è l'errore statistico (si legge epsilon).

Se ci pensate un attimo, è chiaro che non c'è nulla nei calcoli fatti fin qui che ci dica quanto possiamo fidarci della a e della b che abbiamo appena calcolato; in altre parole potremmo aver messo in L e in P dei numeri scelti a casaccio, senza alcun legame tra di loro e il computer, diligentemente, ci restituirebbe una a ed una b senza alcun senso; ecco perché, oltre ad una specifica ed attenta conoscenza dei propri dati, a volte è utile calcolare una misura dell'associazione tra la variabile dipendente e quella indipendente. A questo scopo si usano gli indici di associazione, come il coefficiente di correlazione r o R^2.Spero ve li ricordiate dal cap.3

Naturalmente non pretendo di avervi spiegato che cosa è la regressione, ma, forse a che cosa serve, e soprattutto spero di avervi lasciato il desiderio di approfondire lo studio di questo argomento veramente affascinante.

10. E per finire..
Una storia vera

"Qualche" tempo fa il responsabile di stabilimento di una grossa azienda italiana mi raccontò che, parecchi anni prima aveva deciso di provare ad ottimizzare una certa fase della produzione utilizzando la Stepwise Regression: una particolare applicazione della regressione che qualcuno traduce in italiano come regressione passo- passo. Con la Stepwise Regression è possibile di scegliere tra tante variabili indipendenti quelle che siano più importanti per prevedere il comportamento di una variabile dipendente.

Ora dovete sapere che "parecchi anni prima" di "qualche tempo fa", pur non essendo proprio il tempo dei dinosauri, è pur sempre un epoca in cui gli strumenti di calcolo non erano comodi come adesso. I computer erano oggetti del peso di qualche tonnellata, costosissimi, che dovevano stare in locali climatizzati e che si programmavano attraverso pacchi di schede di cartoncino perforate. Per poter utilizzare un computer bisognava avere tutte le necessarie autorizzazioni, disporre del necessario "tempo macchina" e soprattutto non era facile trovare software semplici da usare per problemi di puro calcolo. Così il mio interlocutore aveva deciso che sarebbe stato più pratico fare i calcoli "a mano"; vale a dire impiegando le macchinette calcolatrici (quelle le usavamo già, come anche le lavatrici il motore a scoppio e le biciclette). Allora organizzò due squadre che avrebbero dovuto lavorare in parallelo sul problema in questione.
Dopo due settimane di lavoro a fare calcoli le due squadre arrivarono entrambe al risultato: ma i risultati erano diversi!
Presi dallo scoramento decisero che per mettere a punto l'impianto andava benissimo il metodo che avevano usato fino a quel momento e che la Stepwise Regression poteva restare nei libri di statistica.

Oggi invece è abbastanza facile trovare un programma per calcolare una Stepwise Regression, farlo girare su di un qualsiasi PC ed ottenere il risultato in meno di un secondo; ma questa enorme disponibilità di calcolo rischia qualche volta di non lasciarci il tempo per capire quello che il computer stia facendo. Ecco, con queste pagine non intendevo certo convincervi che piegare la carta sia il modo più pratico per risolvere dei problemi di statistica; ma desideravo solo aiutarvi a far conoscenza con alcuni concetti della statistica. Nel frattempo la mia speranza era quella di cercare di farvi provare almeno un po' del divertimento che ho provato io piegando la carta e facendo i disegni.

Se siete riusciti ad arrivare fin qui devo proprio ringraziarvi per la pazienza e l'attenzione che avete voluto dedicarmi. Colgo quindi l'occasione per ringraziare anche Guido Pacchetti, Piergiorgio Duca, Giorgio e Chiara Cigada, Carlo Alberto Spinicci, Mauro Sette e Remo Cacciafesta che mi hanno dato incoraggiamento e consigli preziosi; ma soprattutto mia moglie Flavia e le mie figlie Irene ed Anna che hanno sopportato per mesi la casa piena di pezzi di carta, che sono state costrette ad ascoltare per innumerevoli volte idee, frasi, pensieri e... hanno resistito. Ovviamente il solo responsabile di eventuali errori o imprecisioni sono io e mi scuso in anticipo per ogni inesattezza.

Appendice
Per origamisti

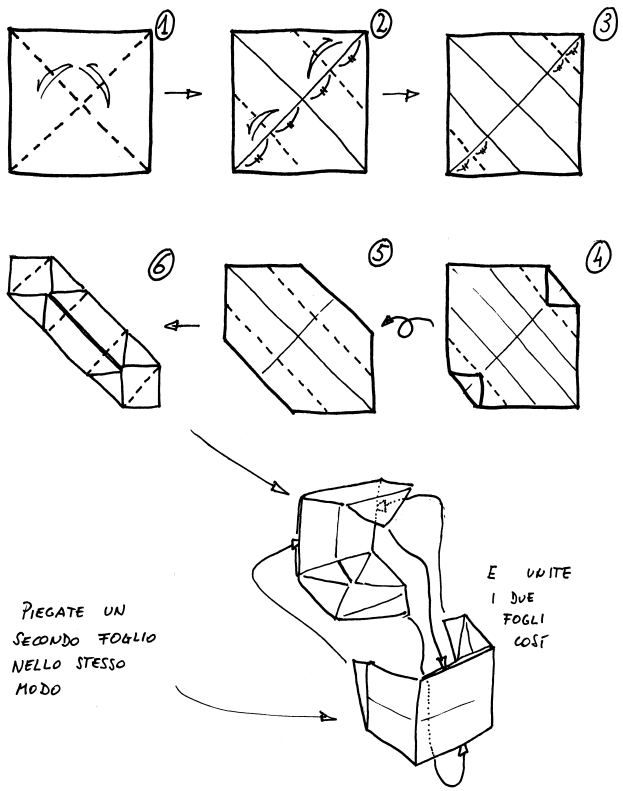
In questo libretto mi sono limitato a modelli di carta molto semplici, immaginando che il lettore fosse completamente digiuno di origami. In questo modo è possibile che gli origamisti più esigenti siano rimasti un po' delusi da alcuni modelli forse troppo spartani. Allora ho pensato di dare qualche riferimento per pieghe solo un pochino più complesse, ma utilizzabili allo stesso modo di quelle presentate nel libro.

Il compianto Thoki Yenn nel libretto citato in bibliografia [3], ha presentato una piega per fare il tetraedro sghembo di pag 24 con un foglio 2×1 .
In questo modo, ci fa notare Thoki, si ricava un sesto di cubo da mezzo quadrato!
Il suo tetraedro è più bello del mio perché ha tutte le facce (anche quella di sotto).
Thoki è stato un personaggio straordinario, dopo la sua morte la British Origami Society ha deciso di ospitare tra le sue pagine web il sito di Thoki, che rischiava di venire smantellato
http://www.britishorigami.org.uk/thok/origami.html

Nel libro di Kasahara, già citato [1], è presente la spiegazione di diversi modelli di cubo con il lato uguale a $\sqrt{2} \div 4$ (come il masu) , realizzabili con 2 fogli di carta. In questo modo si possono fare gli istogrammi combinando i cubi e i masu. A fianco uno schema.

Se vi piace l'origami e non conoscete il CDO, contattateli subito; hanno carte bellissime ed un sacco di libri, anche stranieri, a buon prezzo.

Centro Diffusione Origami
casella postale 42
21040 Caronno Varesino (VA)
www.origami-cdo.it/



Formule

$$\overline{x}=\left(\sum_{i=0}^{n} x_i\right)\div n$$

Media: la sommatoria da i a n , degli n elementi di x , divisa per n

$$R=x_{Max}-x_{Min}$$

Range: valore massimo meno valore minimo

$$D=\sum_{i=1}^{n}\left(\overline{x}-x_i\right)^2$$

Devianza: sommatoria da i ad n dei quadrati delle differenze tra gli n valori di x e la media di x

$$S^2=D\div(n-1)$$

ovvero

$$S^2=\left(\sum_{i=1}^{n}\left(\overline{x}-x_i\right)^2\right)\div(n-1)$$

Varianza: Devianza diviso gradi di libertà

$$S=\sqrt{S^2}$$

ovvero

$$S=\sqrt{\left(\sum_{i=1}^{n}\left(\overline{x}-x_i\right)^2\right)\div(n-1)}$$

Deviazione standard: Radice quadrata della Varianza

$$ES=S\div\sqrt{n}$$

Errore standard: deviazione standard diviso radice quadrata di n

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\left[\frac{(x-\mu)^2}{(2\sigma^2)}\right]}$$

Gaussiana

$$t=\frac{(\overline{x}_1-\overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

$$gdl=n_1+n_2-2$$

t di Student per dati non appaiati: la differenza tra le 2 medie diviso l'errore standard della differenza; cioè la radice quadrata della somma tra le due varianze diviso n. NB è possibile che $n_1\neq n_2$

$$t=\frac{\sum(x_1-x_2)\div n}{\sqrt{s^2\div n}}$$

$$gdl=n-1$$

t di Student per dati appaiati: la media delle differenze tra i 2 campioni diviso l'errore standard di queste differenze. NB $n_1=n_2=n$

Bibliografia (ragionata?)

[1] Origami Omnibus di Kuniko Kasahara è pubblicato da Japan Publications a Tokyo, in Italia potete trovarlo se contattate il Centro Diffusione Origami (www.origami-cdo.it/) . E' un libro bellissimo, in lingua inglese con molti spunti per la riflessione oltre a tanti modelli divertenti.

[2] Per chi vuole iniziare l'origami consiglio caldamente, sempre di Kuniko Kasahara, Origami Facile, ed. il Castello, 1978 Milano.

[3] 13 Thoki Yenn Orikata e' un libretto di poche pagine, ma veramente stimolante, pubblicato dalla British Origami Society (www.britishorigami.org.uk/) nell'aprile dell' 1985. E' stato ristampato nel 1987 in formato A4. Forse ne trovate ancora qualche copia al Centro Diffusione Origami. Oppure potete acquistarlo dal sito della BOS.

[4] Parlando di origami e geometria, un bel libro è quello di Tomoko Fuse: Origami Modulare, il Castello, 1988 Milano.

[5] Super quick origami animals di Nick Robinson, Sterling Publisher Co. Inc (New York 2002) raccoglie molte pieghe di animali, geniali nella loro semplicità ed essenzialità.

[6] Luigi Pirandello, Sei personaggi in cerca d'autore in "Maschere Nude"editori vari, per es. Garzanti.

[7] Sul libro di Box, Hunter e Hunter che si intitola Statistics for Experimenters ed John Wiley & Sons 1978 New York; potete trovare una trattazione più formale del modello geometrico dell'ANOVA.

[8] Per studiare seriamente la statistica esistono molti libri, volendo sceglierne uno che tratti in modo molto più serio di me gli argomenti con cui ho giocato in questo libretto, allora vi consiglio il libro di Lamberto.Soliani che potete trovare gratis in rete all'URL http://www.dsa.unipr.it/soliani/soliani.html
è un libro molto bello e completo.

[9] L'obiettivo di questo libriccino è quello di rendervi più familiari alcuni concetti della statistica; altro è imparare ad interpretare i risultati di una analisi statistica. A questo proposito esiste un bel libro (introduttivo all'argomento) scritto da G.Gigerenzer: Quando i numeri ingannano. Imparare a vivere con l'incertezza. Raffaello Cortina Editore, 2002 Milano.

[10] L'idea di realizzare una regressione lineare con una tavola di legno, chiodi ed elastici non è mia, ma viene da un articolo pubblicato qualche anno fa sulla rivista "Le Scienze" N. 204 agosto 1985 a pag 112; nella rubrica (Ri)creazioni al calcolatore a cura di A.K.Dewdney, l'articolo è intitolato "Congegni analogici che risolvono problemi di varia natura e sollevano un sacco di domande".