TEORIA DEGLI ERRORI

Quale è la definizione di errore? Durante questo corso approfondiremo meglio l'argomento è vero, tuttavia la sua accezione comune lo inquadra come "infrazione" nei confronti di una regola e/o di una consuetudine ma anche come azione inopportuna e svantaggiosa, in gergo "MISTAKE".
Noi però impareremo che l'errore è una proprietà intrinseca di una misura e lo tratteremo secondo la sua definizione scientifico-sperimentale, la quale associa il termine INCERTEZZA, e ne fa una insostituibile parte all'interno di una misurazione; difatti ripetendo più volte una di queste si incappa ne più ne meno, generalmente, in misure diverse.
Pertanto l'errore è informazione complementare alla misura.

L'INCERTEZZA

Trattiamo ora il nostro errore dal punto di vista sperimentale, ad esempio osserviamo il periodo di un pendolo attraverso varie misurazioni e ricaviamo il valore e la sua incertezza:

ES: Risultati: $\{2,3\,s;\ 2,4\,s;\ 2,5\,s;\ 2,4\,s\}$ e consideriamo la "Miglior Stima", ovvero il valore medio e il suo intervallo di incertezza:

$$x = \bar{x} \pm \delta x \xrightarrow{\text{PENDOLO ES.}} T = 2,4\,s \pm 0,1\,s$$

Osservando l'esempio sopra riportato possiamo vedere come l'intervallo di incertezza rappresenti: $T_{min} < T < T_{max}$ e come quest'ultima venga espressa tramite particolare notazione, incertezza generica di misura x: "δx".

Adesso che sappiamo a cosa serve e come viene scritta dobbiamo però capire come utilizzarla, ovvero conoscere e saper applicare la teoria sulle CIFRE SIGNIFICATIVE.
Per ogni misura difatti ha senso indicare solo le cifre "affidabili" e <u>non</u> andare oltre la precisione dello strumento, insomma:

- bisogna arrotondare sempre a 1 cifra significativa (tranne se = 1)
- l'ultima cifra significativa della misura è la stessa dell'incertezza

ES: "Velocità misurata di un aereo": $V = 875,43 \pm 30,43$ m/s

① Arrotondo l'incertezza $\Rightarrow$ $V = 875,43 \pm 30$ m/s
② Arrotondo la misura stimata $\Rightarrow$ $V = 880 \pm 30$ m/s

DISCREPANZA

Essa è la differenza tra la stima di due misure e può essere di due tipologie:

- SIGNIFICATIVA: quando i due intervalli di incertezze sulle relative misure non sono sovrapposti e ciò indica che le misurazioni effettuate sono inconsistenti
- NON SIGNIFICATIVA: quando i due intervalli di incertezze sulle relative misure si sovrappongono e ciò indica che le misurazioni effettuate sono consistenti e possono avere senso.

Riguardo invece alla propagazione dell'incertezza (differenza) dobbiamo porci una domanda: quanto vale l'incertezza se combiniamo diverse misure tra loro?

Consideriamo un peso iniziale e uno finale, ad esempio, pertanto l'incertezza della variabile differenza è data dalla somma delle incertezze delle singole misure.

MISURA

$$d = p - q = -0,7\ g$$

INCERTEZZA

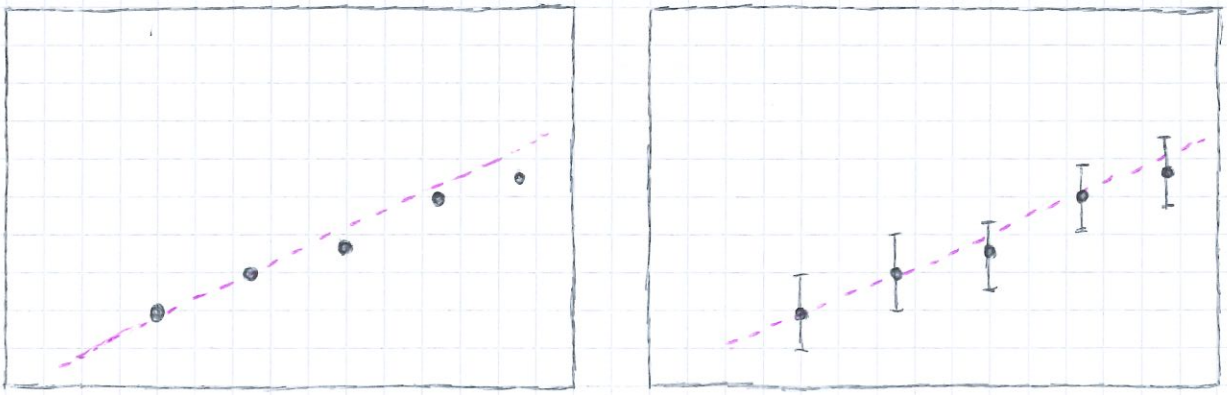
$$d_{min} = p_{min} - q_{max} = -1,6\ g$$
$$d_{max} = p_{max} - q_{min} = 0,2\ g$$

$$d = p - q \pm (\delta p + \delta q)$$

Tutto questo discorso a cosa può servire? Introduciamo così l'incertezza del LEGAME FUNZIONALE e per principio, le misurazioni in grafico non saranno <u>mai</u> sulla stessa retta. A cosa può essere dovuta la non perfetta linearità?

- Alla legge lineare errata
- Incertezza (errore) sperimentale

Tale errore può essere verificato aggiungendone le barre ai punti:

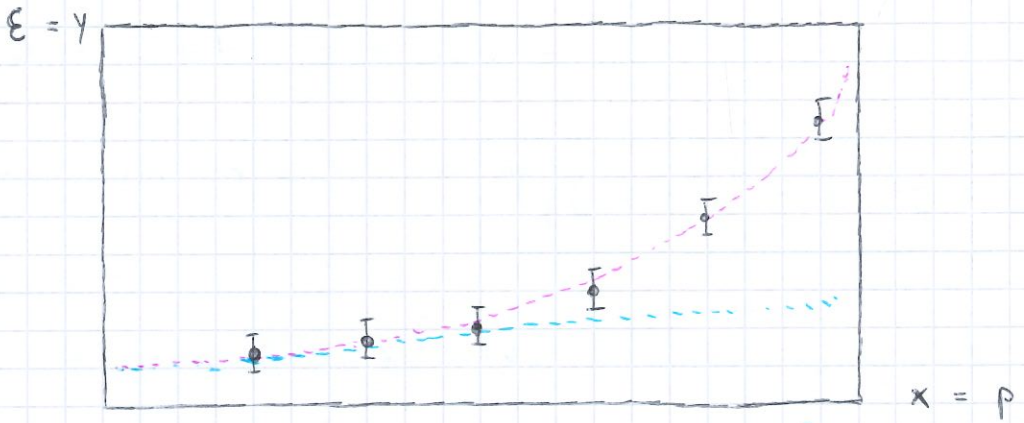


L'incertezza pertanto conferisce al problema un significato e pertanto l'ipotesi di proporzionalità è consistente coi risultati ottenuti
Nel caso più generale le barre di errore saranno anche orizzontali in modo che l'errore venga considerato su entrambe le variabili.

Se tuttavia, anche dopo l'aggiunta dell'incertezza non esiste una retta in grado di interpretare i dati posso fare solo una cosa:

CAMBIO LEGGE

Potrò ad esempio passare da una lineare ad una parabolica



Ad esempio il legame alternativo potrà essere scritto come legge parabolica e sarà COERENTE con l'andamento dei dati e delle incertezze

$$\varepsilon = C \cdot p^2 \longleftrightarrow y = k x^2$$

Idealmente tramite "cambio di variabile" si può rendere lineare ponendo $p' = p^2$ tale che $(x^2 = z)$

$$\varepsilon = C \cdot p' \longleftrightarrow y = k z$$

Ora che abbiamo affrontato le leggi che legano i dati possiamo occuparci del "peso" (= "significatività") dell'errore

INCERTEZZA ASSOLUTA vs INCERTEZZA RELATIVA

La significatività dell'errore dipende dalla variabile misurata pertanto parleremo di due cose distinte:

- INCERTEZZA ASSOLUTA: mero riferimento al valore tecnico
- INCERTEZZA RELATIVA: riferimento al "peso" del nostro errore sulla misurazione totale

$\Big\} \delta x$

Facciamo un esempio pratico: matita e uomo (m, u)

$$x_m = 14\ cm\ , \quad \delta x_m = 1\ cm \xrightarrow{\%} I.R. = \frac{\delta x_m}{|x_m|} = 7\%$$

$$x_u = 182\ cm\ , \quad \delta x_u = 1\ cm \xrightarrow{\%} I.R. = \frac{\delta x_u}{|x_u|} = 0{,}5\%$$

Alla luce di questi dati possiamo esprimere le nostre misurazioni come

$$x = \bar{x} \pm \delta x \begin{cases} x_m \pm \delta x_m = 14 \pm 7\%\ cm \\ x_u \pm \delta x_u = 182 \pm 0{,}5\%\ cm \end{cases}$$

In generale poi: $\boxed{\text{INCERTEZZA RELATIVA (I.R.)} = \frac{\delta x}{|x|} > 0}$ (adimensionale)

PROPAGAZIONE INCERTEZZA

Introduciamo ora le combinazioni di misure e osserviamo come calcolare anche le loro incertezze. Analizziamo il prodotto, conosciamo:

$$x = \bar{x} \pm \delta x = \bar{x}\left(1 \pm \frac{\delta x}{|\bar{x}|}\right)$$

$$y = \bar{y} \pm \delta x = \bar{y}\left(1 \pm \frac{\delta y}{|\bar{y}|}\right)$$

$$\rightarrow \begin{cases} p_{max} = \bar{x}\bar{y}\left(1 + \frac{\delta x}{|\bar{x}|}\right)\left(1 + \frac{\delta y}{|\bar{y}|}\right) \dots \\ p_{min} = \bar{x}\bar{y}\left(1 - \frac{\delta x}{|\bar{x}|}\right)\left(1 - \frac{\delta y}{|\bar{y}|}\right) \dots \end{cases}$$

Svolgendo i calcoli si ottiene che le I.R. si sommano e si ha:

$$\boxed{p = \bar{x} \cdot \bar{y}\left(1 \pm \left(\frac{\delta x}{|\bar{x}|} + \frac{\delta y}{|\bar{y}|}\right)\right)}$$

Se invece generalizziamo il concetto, ad esempio consideriamo un set di misure da sommare al quale ne va sottratto un'altro, quanto varrà l'incertezza?

ES: $$q = \sum_{i=1}^{m} x_i - \sum_{i=m+1}^{n} x_i$$ SOMMA / DIFFERENZA

L'incertezza totale si ottiene sommando tutte le incertezze (ASSOLUTE) e pertanto la propagazione è lineare

ES: $$\delta q = \sum_{i=1}^{m} \delta x_i + \sum_{i=m+1}^{n} \delta x_i$$

Se invece consideriamo un set di misure da moltiplicare e poi dividere per altre, la propagazione come avverrà?

ES: $$q = \frac{x_1 \cdot x_2 \cdot \ldots \cdot x_m}{x_{m+1} \cdot \ldots \cdot x_n}$$ PRODOTTO / QUOZIENTE

L'incertezza totale si otterrà sommando tutte le incertezze (RELATIVE)

ES: $$\frac{\delta q}{q} = \sum_{i=1}^{m} \frac{\delta x_i}{|x_i|} + \sum_{i=m+1}^{n} \frac{\delta x_i}{|x_i|}$$

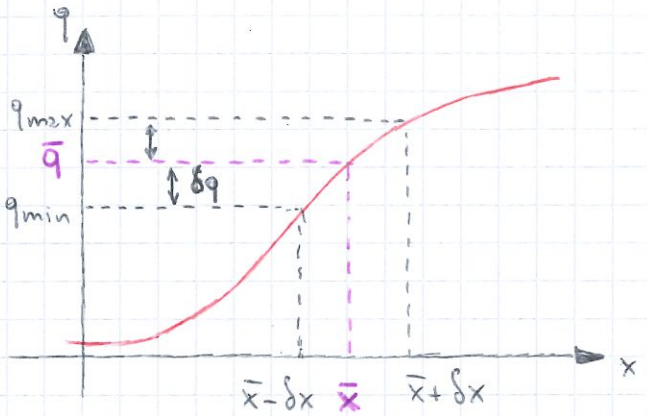
Consideriamo ora invece una misura moltiplicata per una costante (B)

ES: $q = B \cdot x \rightarrow \frac{\delta q}{|q|} = \frac{\delta B}{|B|} + \frac{\delta x}{|x|} = 0 + \frac{\delta x}{|x|} \rightarrow \delta q = \frac{\delta x}{|x|} |Bx| = B \cdot \delta x$

Infine possiamo osservare l'elevamento a potenza di grado "n"

ES: $q = x^n = x \cdot x \cdot x \ldots$ n volte $\rightarrow \frac{\delta q}{|q|} = \frac{\delta x}{|x|} + \frac{\delta x}{|x|} + \ldots = n \cdot \frac{\delta x}{|x|}$

A questo punto si può GENERALIZZARE la propagazione dell'incertezza a <u>qualsiasi</u> dipendenza funzionale?



Una qualsiasi dipendenza funzionale

$$q = f(x)$$

definita δx l'incertezza su x:

$$\delta q = q(\bar{x}+\delta x) - q(\bar{x}) \longrightarrow \frac{\delta q}{\delta x} = \frac{q(\bar{x}+\delta x) - q(\bar{x})}{\delta x} \approx \frac{dq}{dx}$$

se $\delta x \to 0$: <u>"limite del rapporto incrementale"</u>

e pertanto avremo:

$$\boxed{\delta q = \left|\frac{dq}{dx}\right| \delta x}$$

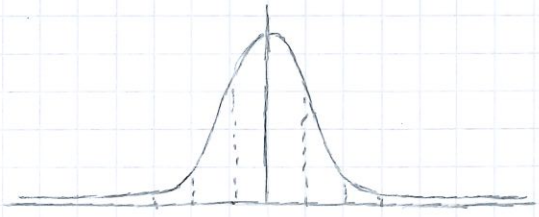
Se ci ritroviamo nel caso di funzioni a più variabili avremo, secondo i principi precedenti (ovvero a singola variabile):

$$q = f(x,y,z,w) \longrightarrow x,y,z,w \to \delta x, \delta y, \delta z, \delta w$$

$$\longrightarrow \delta q \approx \left|\frac{dq}{dx}\right| \delta x + \left|\frac{dq}{dy}\right| \delta y + \left|\frac{dq}{dz}\right| \delta z + \left|\frac{dq}{dw}\right| \delta w$$

Per non osservare i casi NON LIMITE e ridurre l'incertezza dobbiamo:

- avere variabili tra loro indipendenti
- errore nelle misurazioni distribuito in modo normale (o Gaussiano)



$$\longrightarrow \delta q \approx \sqrt{\left(\left|\frac{dq}{dx}\right| \delta x\right)^2 + \dots + \left(\left|\frac{dq}{dw}\right| \delta w\right)^2}$$

INCERTEZZA MINORE

DATI: ORGANIZZAZIONE E ANALISI

Per questo capitolo consideriamo innanzitutto un esempio da seguire, in tale caso osserviamo dei dati grezzi della lunghezza (in mm) di barre metalliche, circa 200 misurazioni.
Definiamo e trattiamo il procedimento di analisi di questi dati:

DEFINIZIONI

Osserviamo i concetti fondamentali della statistica industriale

- **DATO:** elemento statistico di base, osservazione, valore argomentale
- **POPOLAZIONE:** insieme completo, FINITO o INFINITO, dei valori di una variabile aleatoria
- **CAMPIONE:** insieme di osservazioni che rappresentano una porzione della popolazione, in base al quale si cerca di desumere le caratteristiche della popolazione
- **STATISTICA:** scienza che si occupa di "imparare i dati", 4 fasi:

Al fine di raggiungere i nostri obbiettivi prima **RACCOGLIAMO I DATI** e successivamente dovremo pre-processarli in modo tale che siano sistemati, pertanto passiamo all' **ORGANIZZAZIONE**. A tale scopo possiamo utilizzare diverse metodologie come l'ordinazione crescente o decrescente dei dati, per poi passare alla frequenza di occorrenza (f)

ES:

x	f
543	5
548	6
552	9
555	8
564	5

(*)

Il problema, a questo punto, è che la variabile può assumere una infinità di valori, per cui conviene valutare la frequenza di occorrenza di classi di dati (intervalli ad ampiezza costante). Queste sono contigue e non sovrapposte, tuttavia se sono

- Troppo AMPIE, forniscono poche informazioni sulla distribuzione dei dati
- Troppo RISTRETTE, rendono difficile la deduzione di conclusioni

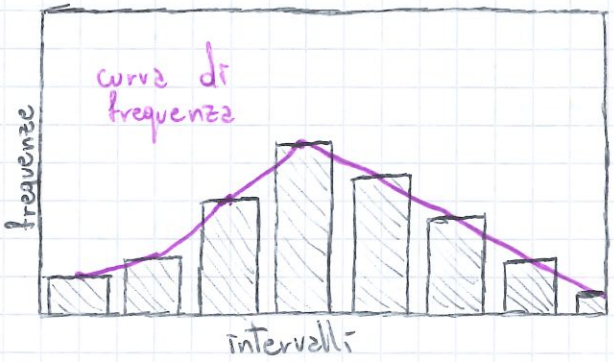
Vanno perciò stabiliti dei criteri per determinarne i confini:

- non sovrapposte, estremi inclusi → 516 - 519, 520 - 523 ....
- un solo estremo incluso → 516 - <520, 520 - <524 ...
- confini a valori non interi → 516,5 - 520,5, 520,5 - 524,5 ...
- "n" classi con ampiezza $\propto$ a $(x_{max} - x_{min})/n$ → 517 - 524,1, 524,1 - 531,2 ....

In generale il metodo/criterio migliore è l'ultimo, dopodiché osserviamo la frequenza (f_i) di ogni classe e il suo valore intermedio (x_i^*)

ES:

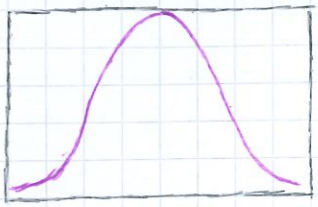
intervallo	x_i^*	f_i
517 - 524,1	520,55	4
524,1 - 531,2	527,65	7
531,2 - 538,3	534,75	17
538,3 - 545,4	541,85	31
⋮	⋮	⋮



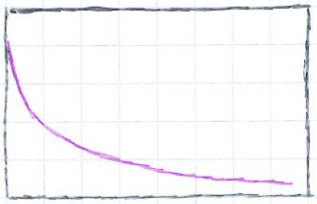
Quindi si ottiene un dato statisticamente significativo se è possibile individuare una distribuzione di dati caratteristica. La curva di frequenza la si ottiene congiungendo i valori intermedi di ogni classe tra di loro.

Durante il proseguio del corso incorreremo in una diversità di distribuzioni statisticamente rilevanti, come:

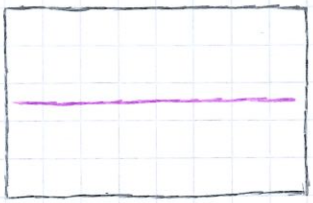
ES:



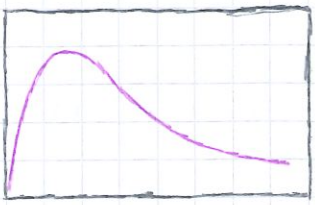
NORMALE



ESPONENZIALE



UNIFORME
(zeromodale)



ASIMMETRICA

Una informazione complementare a questa riguarda la curva di frequenza cumulata, ovvero la curva relativa alla frequenza di occorrenza cumulata di una classe, ottenuta dalla somma delle frequenze di occorrenza delle classi inferiori alla propria. Tale curva rappresenterà l'INTEGRALE della curva di frequenza e avrà come valore limite il numero totale di misurazioni (F_i)

Una volta finita l'organizzazione dei dati si passa all'ANALISI, ad esempio si andranno a cercare tutte le diverse misure di posizione centrale:

- MEDIA: considerando "n" dati (osservazioni) x_i, $i = 1, \dots, n$:

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

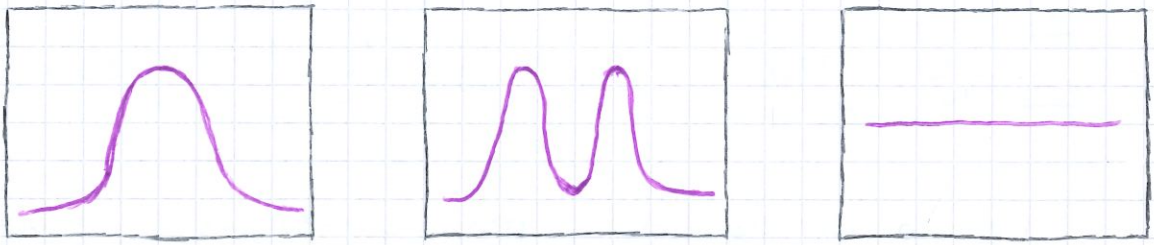
se si estende questa misura alle classi otterremo la seguente

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i^* f_i}{\sum_{i=1}^{N} f_i} = \frac{\sum_{i=1}^{N} x_i^* f_i}{n} = \sum_{i=1}^{N} x_i^* p_i$$

dove si considerano "n" dati in "N" classi x_i, $i = 1, \dots, n$ e p_i è la frequenza relativa di occorrenza (densità di probabilità)

- **MODA**: considerando "n" dati (osservazioni) x_i, $i = 1, \dots, n$, la moda è il valore argomentale con la massima frequenza di occorrenza. Ad esempio, osservando (*), noteremo che la moda è 552. Considerando "n" dati raggruppati in "N" classi, la moda è il valore argomentale intermedio appartenente alla classe con la massima frequenza di occorrenza: per esempio 541,85.

ES:



MONOMODALE BIMODALE ZEROMODALE

Siccome la moda può non essere un valore unico oppure non esistere proprio, le precedenti ne rappresentano le distribuzioni.

- **MEDIANA**: valore argomentale x_m che divide in due parti uguali l'istogramma ovvero $F_{m-1} < 50\%$ e $F_m > 50\%$ con x_{m-1} e x_m valori argomentali adiacenti. Vale sia per dati singoli che in classi.

ES:

intervallo	x_i^*	f_i	F_i	$F_i\%$
517 - 524,1	520,55	4	4	1,8
524,1 - 531,2	527,65	7	11	4,9
531,2 - 538,3	534,75	17	28	12,6
538,3 - 545,4	541,85	31	59	26,5
545,4 - 552,5	548,95	42	101	45,3
552,5 - 559,6	556,05	49	150	67,3
559,6 - 566,7	563,15	40	190	85,2
566,7 - 573,8	570,25	21	211	94,6
573,8 - 580,9	577,35	10	221	99,1
580,9 - 588	584,45	2	223	100.

(#)

dove la "$F_i\%$" è la frequenza relativa cumulata, ovvero F_i/n e la mediana si troverà dove quest'ultima arriva al 50%, peraltro 556,05

Vediamo ora cosa cambia tra queste misure di posizione centrale analizzandole insieme:

- MEDIA VS MEDIANA: facciamo l'esempio dei punteggi US Masters con determinati valori ($n = 10$), ovvero

" 277, 277, 278, 279, 280, 281, 282, 283, 284, 285 ", 382 (valore aberrante)

ES:

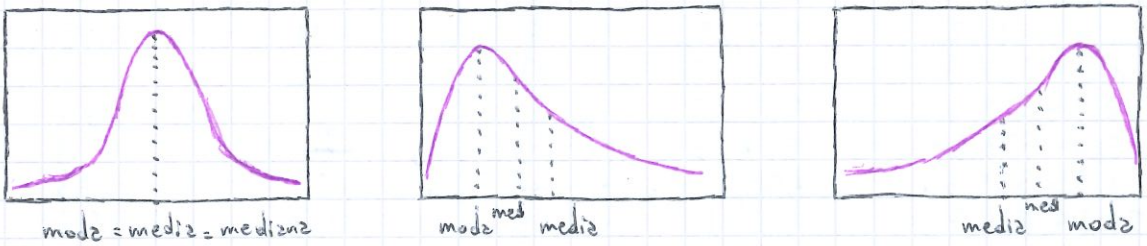
$$\bar{x} = \sum_{i=1}^{10} \frac{x_i}{10} = 280{,}4 \quad , \quad x_{med} = \frac{x_5 + x_6}{2} = 280{,}5$$

Inoltre se il numero di elementi è pari la mediana si calcola come media dei due elementi centrali. Ipotizziamo che nei dati precedenti, ora, venga messo un valore totalmente fuori logica togliendone uno: noteremo come la mediana è una stima ROBUSTA di posizione centrale

ES: $\bar{x} = \sum_{i=1}^{10} \frac{x_i}{10} = 290{,}5$, $x_{med} = \frac{x_5 + x_6}{2} = 280{,}5$

- MEDIA VS MEDIANA VS MODA: analizziamo grafici ipotetici cercando di capire il comportamento di queste nelle distribuzioni

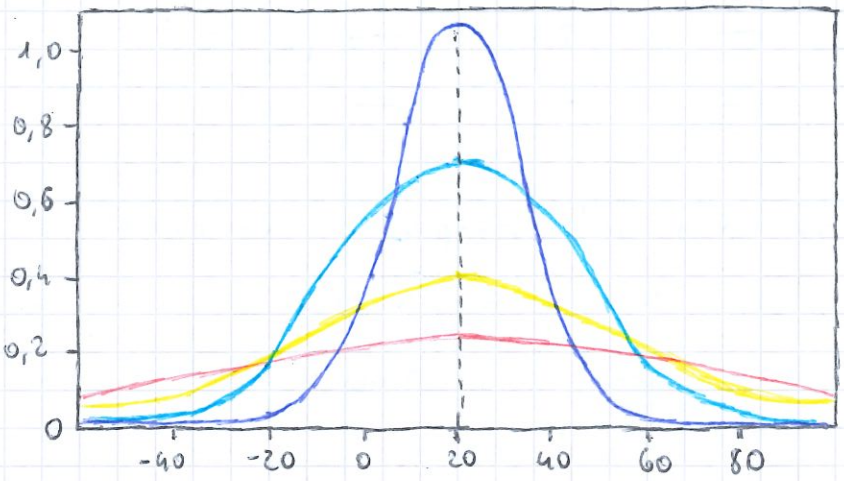
ES:



Ora che abbiamo visto le misure di posizione centrale trattiamo un ulteriore aspetto fondamentale e complementare

Affrontiamo quelle che vengono definite **misure di dispersione**, osserviamo subito un esempio e visualizziamolo:

ES:



media = moda = mediana

Ora capito il punto fondamentale di queste misure potremo elencarle, farne esempi e integrarle nel discorso statistico:

- **MEDIA DELLE DEVIAZIONI**: deviazione dei dati rispetto al valore centrale

ES:

$$\bar{d} = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{n} = \sum_{i=1}^{n} \frac{x_i - 280{,}4}{10} = 0$$

se, ad esempio, consideriamo nuovamente gli US Masters e proviamo a calcolare la deviazione media otteniamo lo zero poiché le deviazioni positive e negative si compensano, dimostriamolo

ES:

$$\bar{d} = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{n} = \sum_{i=1}^{n} \frac{x_i}{n} - \sum_{i=1}^{n} \frac{\bar{x}}{n} = \bar{x} - \frac{n\bar{x}}{n} = 0 \quad \text{(c.v.d)}$$

- **MEDIA DEI VALORI ASSOLUTI DELLE DEVIAZIONI**: deviazione dei valori assoluti dei dati rispetto al valore centrale

ES:

$$\bar{d} = \sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{n} = \sum_{i=1}^{n} \frac{|x_i - 280{,}4|}{10} = 2{,}4$$

come prima applichiamo l'esempio e notiamo la stima

- **MEDIA DELLE DEVIAZIONI QUADRATICHE**: media delle deviazioni del quadrato delle deviazioni dei dati rispetto al valore centrale

ES:
$$d^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{10} = 7,4$$

ricalcoliamo tenendo conto dell'esempio di partenza, otterremo un valore NON confrontabile con la media dei valori assoluti. Tuttavia questa misura di dispersione ha un problema di "gradi di libertà", difatti, poiché gli "n" dati ($n=10$) sono stati usati per calcolare la media "$\bar{x}$", l'ultima deviazione, conoscendo le "$n-1$" precedenti e "$\bar{x}$", è automaticamente determinata; ciò avviene perché l'ultima deviazione potrà assumere uno e un soltanto valore tale che la media non cambi e rimanga tale, perciò useremo altro

- **VARIANZA**: somma dei quadrati delle deviazioni dei dati rispetto al valore centrale. Sempre se consideriamo l'esempio:

ES:
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{9} = 8,3$$

se i dati sono raggruppati in classi avremo

ES:
$$s^2 = \sum_{i=1}^{n} \frac{(x_i^* - \bar{x})^2 \cdot f_i}{n-1} = \sum_{i=1}^{n} (x_i^* - \bar{x})^2 \cdot p_i$$

questa misura di dispersione si utilizza molto poiché la n-esima deviazione, essendo automaticamente determinata, risulta dipendente dalle altre conosciute e NON serve alla determinazione del risultando apparendo superflua. Essendo dipendente allora non fornirà informazioni utili, sarà in eccesso e non farà statistica.

- **DEVIAZIONE STANDARD (SCARTO QUADRATICO MEDIO)**: radice della somma dei quadrati delle deviazioni dei dati rispetto al valore centrale.

Consideriamo anche il solito esempio

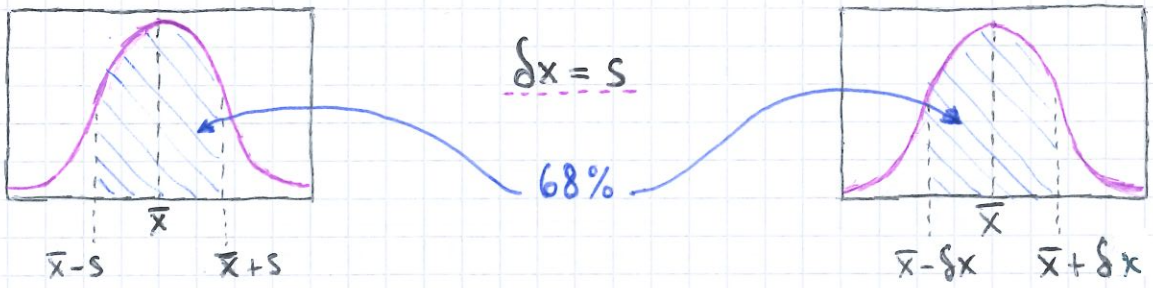
ES:
$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - 280{,}4)^2}{9}} = 2{,}9$$

Se i dati sono raggruppati in classi

ES:
$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i^* - \bar{x})^2 \cdot f_i}{n-1}} = \sqrt{\sum_{i=1}^{n} (x_i^* - \bar{x})^2 \cdot p_i}$$

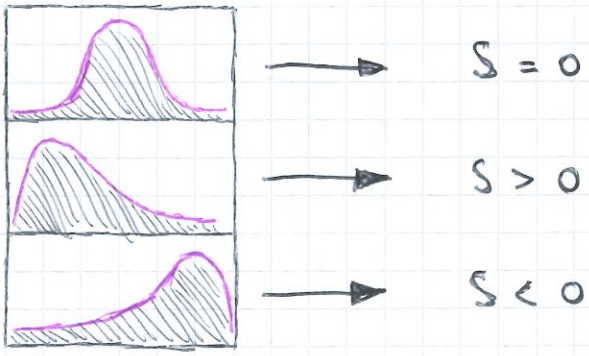
L'importanza di questa misura si vita al fatto che se le sorgenti di incertezze sono casuali e non sistematiche allora le misure sono distribuite in maniera **normale** (Gaussiana). In queste condizioni si dimostra che il 68% circa dei risultati ricade nell'intervallo $[\bar{x}-s, \bar{x}+s]$. Il vantaggio è appunto l'approssimazione dell'incertezza di "x" allo scarto quadratico medio e pertanto ci si "accontenta" di essere sufficientemente certi che la misura sia entro un certo limite del valore vero

ES:



Introduciamo infine l'**indice di skewness** (S)

ES:
$$S = 3 \cdot \frac{\bar{x} - x_{med}}{s}$$



PROBABILITÁ

In questa parte di corso analizzeremo da vicino il concetto di probabilità e le relazioni tra gli eventi. Iniziamo fornendo quelli che sono i concetti base su cui baseremo il nostro lavoro.

DEFINIZIONI

Iniziamo innanzitutto dal titolo di questo argomento:

- **PROBABILITÁ**: valore numerico che esprime il grado di incertezza di un evento casuale ovvero "il numero, compreso tra 0 e 1, che esprime il grado di possibilità che l'evento si verifichi".
- **ESPERIMENTO CASUALE (PROVA)**: genera un risultato incerto
- **EVENTO**: risultato della prova
- **SPAZIO DEGLI EVENTI**: insieme di tutti i possibili eventi

Distinguiamo poi probabilità OGGETTIVA (sulla quale lavoreremo noi), ovvero basata su eventi pregressi e dati passati, dalla controparte SOGGETTIVA, ovvero fondata sulla mera esperienza ma senza base di dati.
La probabilità di un evento casuale riguarda una frequenza a lungo termine, se consideriamo eventi casuali equiprobabili e "N" eventi

ES:
$$Pr(E) = \frac{\text{Numero di eventi che danno origine ad } E}{\text{Numero Totale di eventi}}$$

Vale solo asintoticamente, ovvero per "N" sufficientemente grande

ES:
$$Pr(E) = \lim_{N \to \infty} \frac{n_E}{N} \quad \text{con} \quad \frac{n_E}{N}: \text{frequenza di occorrenza relativa di } E$$

La probabilità sarà: 0 (impossibile) $\leq Pr(E) \leq 1$ (certo)

EVENTI COMBINATI

Trattiamo ora le relazioni tra gli eventi probabilistici, iniziamo elencando le proprietà di questi:

- COMMUTATIVA: "$E_1 \cup E_2 = E_2 \cup E_1$"
- ASSOCIATIVA: "$(E_1 \cup E_2) \cup E_3 = E_1 \cup (E_2 \cup E_3)$"
- DISTRIBUTIVA: "$(E_1 \cup E_2) E_3 = E_1 E_3 \cup E_2 E_3$"

Affrontiamo ora quelli che sono degli **assiomi di base** della probabilità:

- $0 \leq Pr(E) \leq 1$
- $Pr(S) = 1$, con S = spazio degli eventi
- Considerando una serie di eventi esclusivi tra loro $Pr\left(\bigcup_{i=1}^{n}\right) = \sum_{i=1}^{n} Pr(E_i)$

Definiti questi tre ne discendono tutte le altre, le quali rispondono alla domanda: cosa succede se combiniamo più eventi? Avremo quindi:

- **Eventi complementari**: $Pr(\bar{E}) = 1 - Pr(E)$
- **Intersezione tra eventi**: prodotto logico, $Pr(E_1 \cap E_2) = Pr(\bar{E_1}) \cdot Pr(\bar{E_2})$
- **Unione $E_1 \cup E_2$**: somma logica, $Pr(E_1 \cup E_2) = Pr(\bar{E_1}) + Pr(\bar{E_2}) - Pr(E_1 \cap E_2)$
- **Eventi incompatibili**: eventi mutuamente esclusivi, $Pr(E_1 \cap E_2) = 0$

Facciamo ora un esempio e cerchiamo il numero totale di combinazioni tra i risultati (1. n° = m, 2. n° = n) di due esperimenti generici

ES: "m" possibili risultati
"n" possibili risultati
→ $(1,1) \dots (1,m)$ / $(2,1) \dots (2,m)$ / $(n,1) \dots (n,m)$ → "$m \times n$"

Questo discorso ci torna utile al fine di introdurre un concetto molto importante nell'ambito della probabilità

LE PERMUTAZIONI

Queste saranno utili al fine di ottenere probabilità su eventi combinati, anche in modi particolari

"a, b, c" $\xrightarrow[\text{permutazioni}]{\text{n° possibili}}$ abc, acb, bac, bca, cab, cba (n elementi)

ES:

$3 \cdot 2 \cdot 1 = 6$ permutazioni differenti date da prob. combinata (n_{tot})

Introduciamo così il **fattoriale** e definiamo le permutazioni come

ES:

$$n_{tot} = n!$$

Se ci spingiamo oltre con un ulteriore esempio, proviamo a calcolare le permutazioni, NON considerando l'ordine degli elementi, e generalizziamo ad una regola il concetto

"A, B, C, D, E" → elementi = "n = 5"

ES:

→ gruppi da 3 ($k=3$): $\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = \frac{5 \cdot 4 \cdot 3}{3!} \cdot \frac{2!}{2!} = \frac{5!}{3! \cdot 2!}$

Generalizzando possiamo osservare come la precedente rispecchi la formula risolutiva delle permutazioni di "n" elementi in tot. raggruppamenti di "k" elementi

ES:

$$\frac{n!}{k!\,(n-k)!} = \binom{n}{k}$$

Abbiamo appena definito il numero di possibili combinazioni di "n" oggetti presi per gruppi di "k", ovvero il **coefficiente binomiale**.

Torniamo ora al discorso di prima sulla probabilità

PROBABILITÀ CONDIZIONATA

Possiamo ora introdurre questo concetto il quale risponde alla domanda: quale è la probabilità che si verifichi E_2 dopo che E_1 sia avvenuto? La probabilità condizionata si indica con una sbarra e

ES: $$Pr(E_2|E_1) = \frac{Pr(E_2E_1)}{Pr(E_1)} \longrightarrow Pr(E_2E_1) = Pr(E_1) \cdot Pr(E_2|E_1)$$

Consideriamo l'esempio classico del dado

ES:
- E_1: dal lancio del primo dado si ottiene 3
- E_2: la somma dei due tiri è 8

$$\longrightarrow Pr(E_2|E_1) = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

EVENTI INDIPENDENTI

Si parla di eventi indipendenti quando l'accadimento di E_1 non influenza E_2. Questa proprietà è reciproca e biunivoca. Verifichiamo così un esempio pratico

ES: $$Pr(E_2|E_1) = Pr(E_2) \longleftrightarrow Pr(E_1|E_2) = Pr(E_1)$$

Troviamo la probabilità di estrarre una regina dopo una figura (Q, F) e di estrarre una regina dopo una carta di cuori (Q, H), avremo

ES:
- $Pr(Q|F) = \frac{4}{12} = \frac{1}{3}$ $\quad Pr(Q) = \frac{4}{52} = \frac{1}{13}$ $\quad$ dipendenti
- $Pr(Q|H) = \frac{1}{13}$ $\quad Pr(Q) = \frac{1}{13}$ $\quad$ indipendenti

EVENTI INCOMPATIBILI

Si noti come eventi incompatibili siano intrinsecamente dipendenti, dopo che il primo avviene il secondo avrà probabilità nulla

ES: $$Pr(E_1) \neq 0 \quad Pr(E_2 \cdot E_1) = 0 \quad Pr(E_2|E_1) = \frac{Pr(E_2 \cdot E_1)}{Pr(E_1)} = 0$$

L'esempio classico è con una carta di fiori e poi pescare l'asso di cuori, cosa non possibile e controsenso.

VARIABILI CASUALI

Andiamo ora a trattare quelle che chiameremo variabili casuali e analizziamole nelle loro due distinzioni: discrete e continue. Una volta capito cosa avremo davanti ne osserveremo le distribuzioni.

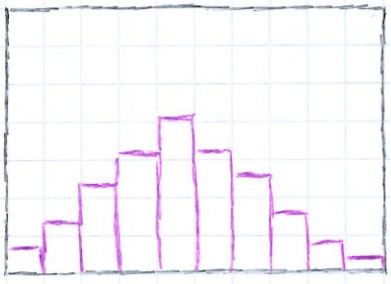
L'esempio Trattato parla di dadi e la variabile di interesse è la somma dei punteggi dopo due lanci di dado, la probabilità della singola combinazione sarà "$\frac{1}{36}$" ≡ "$\frac{1}{6} \cdot \frac{1}{6}$".

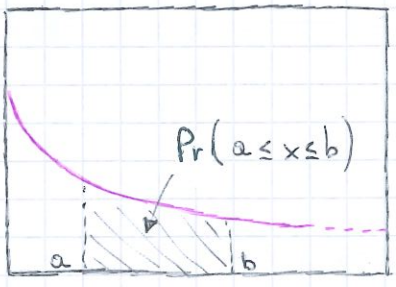
DEFINIZIONI

Consideriamo la base di questo capitolo:

- **VARIABILE CASUALE**: variabile ottenuta in seguito a un esperimento, che fornisce un risultato aleatorio, dall'esito sconosciuto
- **VARIABILE CASUALE DISCRETA**: variabile che può assumere un set limitato di valori, come i lanci di dadi/monete
- **VARIABILE CASUALE CONTINUA**: variabile che può assumere un set illimitato di valori, tipo le misurazioni di una barra metallica in centimetri

ES: (*)





Sia che si parli di variabili casuali discrete o continue ci sarà quella che viene chiamata **funzione cumulativa di distribuzione**. In entrambi i casi si parla di probabilità e saranno caratterizzate da

ES:

$$F(x_i) = Pr(x \leq x_i), \quad i = 0, \dots, n$$

VARIABILI CASUALI DISCRETE

La caratteristica di questo studio in particolare è che per ogni risultato di un dato evento aleatorio (E) esiste una probabilità finita ≥ 0. Ogni evento corrisponde una variabile casuale (x) che assume un valore assegnato (X) tale che

ES: $$E: x = X \longrightarrow Pr(E) = Pr(x = X)$$

dove "x" può assumere N valori X_i, $i = 1 \dots N$. Pertanto avremo che ogni valore che assumerà la nostra variabile avrà una relativa probabilità e che eventi corrispondenti a valori differenti di "X" sono INCOMPATIBILI

ES: $$Pr(x = X_i) = p_i(x), \quad Pr(x = X_i \cap x = X_j) = 0 \longrightarrow i \neq j$$

Otterremo così che la somma di tutto sarà l'intero spazio campionario con

ES: $$\sum_{i=1}^{N} Pr(x = X_i) = \sum_{i=1}^{N} p_i(x) = 1$$

Osserviamo l'esempio di prima

ES:

X	p(x)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36

X	p(x)
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

(*)

Ora che avremo una distribuzione, come potremo caratterizzarla? Ovvero come possiamo analizzarla e trarne conclusioni a riguardo?
Valutiamo così una misura di posizione e una di dispersione

VALORE ATTESO

Il valore atteso è la misura di posizione centrale di una distribuzione di probabilità. Questo si definisce come

ES:

$$E(x) = \sum_{i=1}^{N} X_i * Pr(x = X_i) = \mu$$

ed è la MEDIA PESATA dei possibili valori che "X" può assumere, in pratica lo si può vedere come «baricentro» della distribuzione. Non sempre, il valore atteso, avrà senso ma probabilmente stimerà soltanto la mera posizione centrale. Consideriamo un dado:

$$p(x_i) = \frac{1}{6} \quad , \quad i = 1 \dots 6$$

ES:

$$E(x) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3,5$$

Si osserva come $E(x)$ non sia necessariamente un valore che "X" può assumere. <u>NON</u> va inoltre fatta confusione tra questo operatore e la media ($E(x) \neq \bar{x}$) poiché:

- $\bar{x}$ è la media di un set di dati
- $E(x)$ è la media pesata di una distribuzione di probabilità

Inoltre il valore atteso può essere definito con un limite, ovvero

ES:

$$\lim_{N \to +\infty} \sum_{i=1}^{N} \frac{X_i}{n}$$

Abbiamo visto che cosa è, ora trattiamone le proprietà e i modi con i quali questo operatore si approccia ai calcoli. Ricordiamo inoltre che il valore atteso è un operatore LINEARE

Consideriamo le seguenti:

. Moltiplicazione per una costante: $E(Ax) = A * E(x)$

ES: $$E(Ax) = \sum_{i=1}^{N} A \cdot x_i \cdot p(x_i) = A \cdot E(x)$$

. Somma di una costante: $E(A+x) = A + E(x)$

ES: $$E(A+x) = \sum_{i=1}^{N} (A + x_i) \cdot p(x_i) = A \cdot \sum_{i=1}^{N} p(x_i) + \sum_{i=1}^{N} x_i \cdot p(x_i) = A + E(x)$$

. Somma e moltiplicazione di costanti:

ES: $$E(A + Bx) = A + B \cdot E(x)$$ ← linearità

. Somma di funzioni:

ES: $$E(f_1(x) + \dots + f_n(x)) = E(f_1(x)) + \dots + E(f_n(x))$$

VARIANZA

Analizziamo ora la varianza, ovvero una misura di dispersione di una distribuzione di probabilità. Si capisce particolarmente la funzionalità di questa dal momento che si hanno distribuzioni diverse ma con stesso valore atteso perchè se ne denotano le differenze

ES: $$V(x) = \sum_{i=1}^{N} \left[\left(X_i - E(x) \right)^2 \cdot p(X_i) \right] = E\left[\left(x - E(x) \right)^2 \right] = E(x^2) - \left(E(x) \right)^2$$

Il vantaggio di ciò è che una volta nota la distribuzione e il valore atteso, sarà possibile immediatamente trovare la varianza.

Vediamo ora il classico esempio del dado, consideriamo

$$E(x^2) = \sum_{i=1}^{N} x_i^2 \cdot p(x_i) = 1\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 9\cdot\frac{1}{6} + 16\cdot\frac{1}{6} + 25\cdot\frac{1}{6} + 36\cdot\frac{1}{6} = \frac{91}{6} \simeq 15,2$$

ES: $$(E(x))^2 = \left[\sum_{i=1}^{N} x_i \cdot p(x_i)\right]^2 = \left[1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6}\right]^2 = 3,5^2 \simeq 12,25$$

$$V(x) = E(x^2) - (E(x))^2 = 15,2 - 12,25 \simeq 2,95$$

Trattiamo ora le proprietà di questo operatore sapendo che la varianza NON è un operatore LINEARE. Consideriamo le seguenti:

. Moltiplicazione per una costante: $V(Ax) = A^2 * V(x)$

$$V(Ax) = E(A^2x^2) - [E(Ax)]^2 = A^2 \cdot E(x^2) - A^2[E(x)]^2$$

ES:
$$= A^2 \cdot (E(x^2) - [E(x)]^2) = A^2 \cdot V(x)$$

. Somma di una costante: $V(A+x) = V(x)$

$$V(A+x) = E((A+x)^2) - [E(A+x)]^2 =$$
$$= E(A^2 + 2Ax + x^2) - [A + E(x)]^2$$

ES:
$$= A^2 + 2AE(x) + E(x^2) - [A^2 + 2AE(x) + (E(x))^2]$$
$$= E(x^2) - [E(x)]^2 = V(x)$$

. Somma e moltiplicazione di costanti:

ES: $$V(A + Bx) = B^2 * V(x)$$

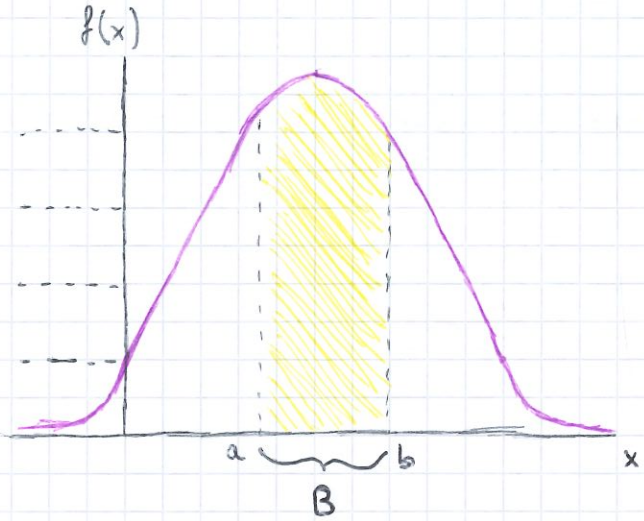
Possiamo ora ad altre variabili da considerare.

VARIABILE CASUALE CONTINUA

Sapendo che, per definizione, una variabile casuale continua può assumere un set ∞ di valori e fissato un evento $E(x)$ tale che

ES: $$E: \quad x = X$$

possiamo dire che la probabilità di "x" (misurazione) di assumere un certo "X" (valore) è **NULLA**, a causa degli ∞ valori. Dobbiamo quindi vedere la probabilità di avvenimento di un evento come fosse contenuta in un intervallo $[a, b]$ e descrivere la nostra variabile casuale continua come funzione non negativa $f(x)$, definita $\forall x \in (-\infty, +\infty)$, chiamata funzione di densità di probabilità



$$p(a \leq x \leq b) = \int_a^b f(x)\,dx$$

$$\updownarrow$$

$$p(X \in B) = \int_B f(x)\,dx$$

ES:

Questa distribuzione $f(x)$ di una variabile casuale continua "x" è la curva che delimita la probabilità che un certo evento cada in un intervallo assegnato. Scriviamo così VALORE ATTESO e VARIANZA per i quali valgono le stesse proprietà delle variabili discrete:

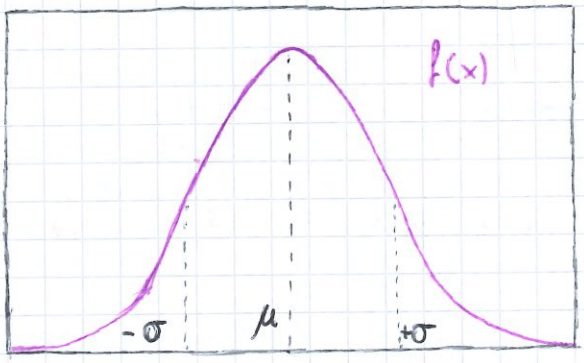
- $E(x) = \int_{-\infty}^{+\infty} x\, f(x)\,dx$
- $V(x) = \int_{-\infty}^{+\infty} (x - E(x))^2\, f(x)\,dx$

DISTRIBUZIONE NORMALE

Considereremo in questo capitolo la distribuzione più famosa, ovvero quella normale (o Gaussiana). Essa è tale per tre motivi

- le misurazioni sperimentali danno spesso, come risultati, valori distribuiti in modo normale
- spesso distribuzioni più complesse possono essere approssimate a quella normale
- consente di fare **inferenza** sui valori attesi di popolazioni sulla base di medie campionarie

ES:



Avremo che la nostra curva di distribuzione di probabilità $f(x)$ viene rappresentata dalla seguente formula matematica

ES:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Se si volessero considerare valore atteso e varianza allora avremo, con "$E(x) = \mu$" e "$v(x) = \sigma^2$", $x \sim N(\mu, \sigma^2)$

ES:

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

dove l'asse di simmetria è il "μ" mentre i flessi in "$\mu \pm \sigma$"

Il punto di massimo, ovvero con "$x=\mu$", dipenderà così solamente dalla varianza σ. Questo accade perché tutte le distribuzioni normali hanno il vincolo che l'integrale indefinito corrisponda all'unità:

$$f(\mu) = \frac{1}{\sqrt{2\pi}\,\sigma} = \frac{0{,}399}{\sigma} \quad ; \quad \text{"pt. massimo} \propto \frac{1}{\sigma}\text{"}$$

ES:

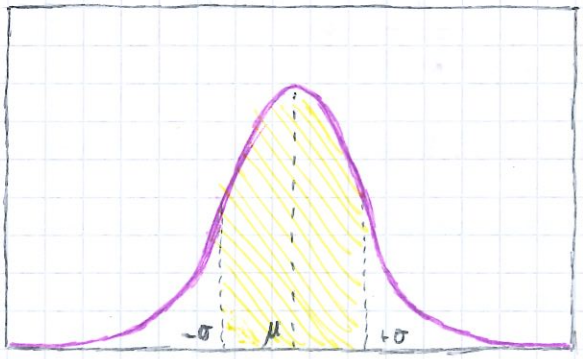
$$\text{Distribuzione normale: } \int_{-\infty}^{+\infty} f(x)\,dx = 1$$

Introduciamo ora una variabile ausiliaria "$Z = \frac{x-\mu}{\sigma} = -\frac{\mu}{\sigma} + \frac{1}{\sigma}x$" e troviamone valore atteso e varianza:

- $E(z) = E\left[-\frac{\mu}{\sigma}, \frac{1}{\sigma}x\right] = -\frac{\mu}{\sigma} + \frac{\mu}{\sigma} = \boxed{0}$
- $V(z) = \frac{1}{\sigma^2} \rightarrow V(Z) = V\left[-\frac{\mu}{\sigma}, \frac{1}{\sigma}x\right] = \frac{\sigma^2}{\sigma^2} = \boxed{1}$

Pertanto introduciamo la distribuzione normale standard: $Z \sim N(0,1)$ e sostanzialmente, tramite cambio di variabile, tutto può esservi riportato. Il vantaggio di ciò è che l'integrale non è più parametrico ma solo dipendente da "z"

ES:



- σ piccolo → Alta precisione
- σ grande → Bassa precisione

Come sappiamo il 68% dell'area sottesa dalla curva si trova in $[\mu-\sigma, \mu+\sigma]$ quindi se si assume la deviazione standard (σ) come incertezza si ha fiducia di essere 68/100 volte nei paraggi del risultato esatto. La probabilità, infine, che la misura cada entro "$k\sigma$" → 100% al k molto rapidamente.

ES

$$Pr(\mu-\sigma \le x \le \mu+\sigma) = \int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi}\,\sigma} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx \approx 0{,}68$$

Per trovare le varie probabilità facciamo riferimento alla tabella integrale, con valori tabulati solo per $z > 0$; i rimanenti verranno trovati per SIMMETRIA. Potremo avere diverse richieste del problema:

- CASO 1: $P_r(z < z_0)$, $z_0 > 0$, usiamo una tabella random

Es: $P_r(z < 0,32) = F(0,32) = 0,6255$

- CASO 2: $P_r(z > z_0)$, $z_0 > 0$

" $P_r(z > z_0) = 1 - F(z_0)$ ovvero il complemento a 1 "

Es: $P_r(z > 0,18) = 1 - F(0,18) = 1 - 0,5714 = 0,4286$

- CASO 3: $P_r(z < -z_0)$, $z_0 > 0$, usiamo la simmetria

" $P_r(z < -z_0) = P_r(z > z_0) = 1 - F(z_0)$ "

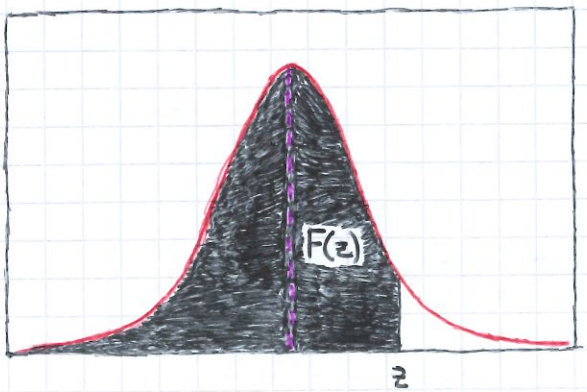
$P_r(z < -0,24) = P_r(z > 0,24) = 1 - F(0,24) = 1 - 0,5948 = 0,4052$

Consideriamo sempre la primitiva della distribuzione di probabilità:

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \cdot e^{\left(-\frac{z^2}{2}\right)} \, dz$$

Tenendo conto poi che la distribuzione normale standard è univoca, potremo così tabulare la completezza delle probabilità in modo tale da averla sott'occhio durante gli esercizi

Es:



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8863	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633

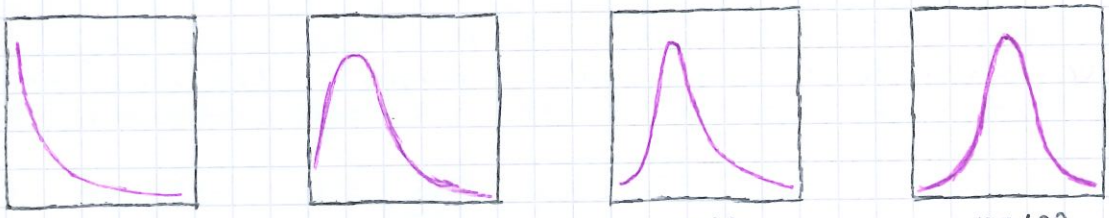
Andiamo ora a trattare un punto fondamentale dello studio statistico, ovvero un teorema riguardante una distribuzione

TEOREMA DEL LIMITE CENTRALE (TLC)

Questo teorema è il fulcro di questo corso e si può dire che sia il ponte tra probabilità e statistica insieme alla distribuzione normale.
Tale teorema ci dice che, considerata una generica popolazione con valore atteso "μ" e varianza "σ^2" da cui si estraggono n osservazioni indipendenti $(x_1, \dots x_n)$, per $n \to +\infty$ la variabile casuale $\bar{x}$ (media) tende ad essere distribuita in modo **asintoticamente normale**, indipendentemente dalla popolazione di origine

ES: $X \sim$ casualmente $(\mu, \sigma^2) \xrightarrow{\text{se } n \to +\infty} \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

A livello empirico già un buon numero di osservazioni è $n = 30$ per il quale la distribuzione è ragionevolmente più simile a una normale e si dice **regola del pollice**. Al variare di n osserviamo



$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \longrightarrow X_{TOT} = \sum_{i=1}^{n} x_i = n\bar{x} \sim N\left(n\mu, n\sigma^2\right)$$

Osserviamo ora quali sono i passaggi coinvolti nella determinazione del valore atteso e della varianza di una distribuzione con "$\bar{x}$":

- $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^{n} x_i / n$
- $E(\bar{x}) = E\left(\sum_{i=1}^{n} \frac{x_i}{n}\right) = \frac{\sum_{i=1}^{n} E(x_i)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \frac{n\mu}{n} = \mu$
- $V(\bar{x}) = V\left(\sum_{i=1}^{n} \frac{x_i}{n}\right) = \frac{1}{n^2} \sum_{i=1}^{n} V(x_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

INFERENZA

Fare INFERENZA significa trarre conclusioni su una popolazione più ampia o su un fenomeno generale basandosi sui dati raccolti da un campione rappresentativo di quella popolazione

STIMA E STIMATORE

« La media del campione, ovvero la media delle misure, è la miglior stima del valore vero o del valore atteso della popolazione ». Questo pensiero è ragionevole e intuitivo, ma è possibile giustificarlo ?
Diamo così due definizioni di base

- **STIMA**: valutazione di un parametro caratteristico della popolazione a partire da una serie di misure
- **STIMATORE**: variabile campionaria utilizzata per stimare un determinato parametro della popolazione
- **STIMA PUNTUALE**: valore assunto dallo stimatore in corrispondenza di un determinato campione

Lo stimatore è una variabile CASUALE poiché dipende dal campione selezionato e, pertanto, la stima puntuale varierà a seconda di quest'ultimo. Infine lo stimatore sarà soggetto a incertezza, analizziamone le **proprietà**:

1. **CORRETTEZZA**: il valore atteso dello stimatore corrisponde al parametro della popolazione da stimare (es: E(stimatore) = E(popolazione))

. la media è uno stimatore corretto del valore atteso della popolazione?

ES:
$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} \longrightarrow E(\bar{x}) = \sum_{i=1}^{n} \frac{E(x_i)}{n} = \frac{n\mu}{n} = \mu$$

SI lo è perché $E(x) = E(\bar{x}) = \mu$.

. La varianza campionaria è uno stimatore corretto della varianza della popolazione?

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

ES:

$$\longrightarrow = \frac{1}{n-1}\left(\sum_{i=1}^{n} (\mu^2+\sigma^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \ldots = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

Sì lo è perché $E(s^2) = V(x) = \sigma^2$. Tuttavia questo ci indica che la "media delle deviazioni quadratiche" <u>non</u> lo è poiché al denominatore avrebbe "n" e non "n-1" come la varianza.

② **EFFICIENZA**: la stima deve avere la minima varianza possibile.

Senza le dimostrazioni scriviamo

. $E(\bar{x}) = E(x_{mediana}) = \mu \longrightarrow$ entrambi CORRETTI

ES:

. $V(\bar{x}) = \frac{\sigma^2}{n}$, $V(x_{mediana}) = \pi \frac{\sigma^2}{n}$

Osserviamo come la media campionaria sia più efficiente come stima del valore atteso della popolazione rispetto alla mediana. Equivalenza solo asintotica

③ **CONSISTENZA**: precisione e affidabilità della stima devono crescere al crescere della numerosità del campione

ES:

$$\lim_{n \to +\infty} Pr\{|T_n(X) - \theta| > \varepsilon\} = 0 \qquad \forall \varepsilon > 0$$

Dato un campione di "n" dati X, al crescere di "n", la probabilità che la differenza, in valore assoluto, tra stimatore e parametro, risulti maggiore di un valore ε piccolo, tende a 0.

MASSIMA VEROSIMIGLIANZA

Dati n valori campionari indipendenti $x_1, \dots, x_n$ le migliori stime di parametri della popolazione sono quei valori per cui $x_1, \dots, x_n$ assumono la massima probabilità. L'obbiettivo di questo metodo è trovare quei valori dei parametri del modello che massimizzano la probabilità di osservare i dati effettivamente osservati. Troviamo la funzione di verosimiglianza

ES1 $$L = Pr(x_1, \dots x_n) = Pr(x_1) \cdot \dots \cdot Pr(x_n) = p(x_1) \cdot \dots p(x_n)$$

dove $p(x_i)$ è la funzione densità di probabilità: probabilità che una misura cada tra x_i e $x_i \pm dx_i$. Ora consideriamo la seguente e ricaviamo il valore atteso e la varianza: (con $x_1 \dots x_n$)

ES2 $$L = \left(\frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right)} \right) * \dots * \left(\frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)} \right)$$

Conoscendo quindi $x_1, \dots, x_n$ calcoliamo i valori di "μ" e "σ^2" per i quali $x_1, \dots, x_n$ apparterranno alla distribuzione dalla quale derivano. Usando il criterio della massima verosimiglianza e dopo diversi calcoli otteniamo:

ES3 $$\begin{cases} \frac{\partial L}{\partial \mu} = 0 \\ \frac{\partial L}{\partial \sigma} = 0 \end{cases} \xrightarrow[\text{parziali uguali a 0}]{\text{ponendo le derivate}}$$ "Sistema di 2 equazioni in μ, σ^2"

Ricaviamo quindi i risultati

- $\frac{\partial L}{\partial \mu} = 0 \Rightarrow \dots \sum_{i=1}^{n} \frac{x_i}{n} = \boxed{\bar{x} = \mu}$
- $\frac{\partial L}{\partial \sigma} = 0 \Rightarrow \dots \boxed{\sigma^2 = \left(\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} \right)}$

Quindi la miglior stima per "μ" sarà la media mentre per "σ^2" sarà la media delle deviazioni quadratiche delle n misure

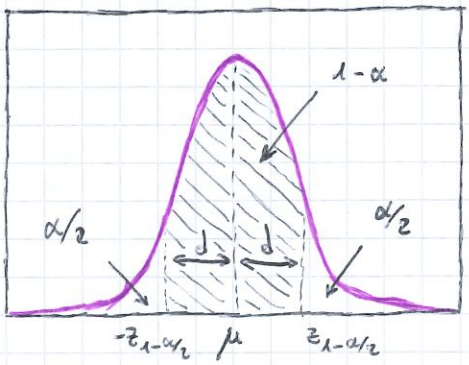
Intervalli Fiduciari

Ora che abbiamo trovato che $\bar{x}, s^2$ sono stimatori corretti, efficienti e consistenti dei parametri μ, σ^2 della popolazione dobbiamo chiederci: ma quale è la probabilità che corrispondano ai rispettivi «valori veri»?
La risposta è 0 poiché le stime puntuali hanno intrinsecamente probabilità di accadimento nulle ⟷ "distribuzioni continue"

Es: $$Pr(\bar{x} = \mu) = 0 \qquad Pr(s^2 = \sigma^2) = 0$$

È più ragionevole parlare di intervalli di valori, al cui interno siamo sufficientemente sicuri che sia presente il «valore vero» di μ e σ^2.
Questi verranno chiamati intervalli fiduciari (o di confidenza) e verranno costruiti nell'intorno della stima puntuale in modo tale da contenere, con una certa affidabilità, il valore vero. Ad ogni intervallo fiduciario è associato un grado di affidabilità: con quanta probabilità $(1-\alpha)\%$ il valore vero è contenuto all'interno dell'intervallo

Es:



Dal teorema del limite centrale, $\bar{x}$ è distribuito normalmente con media "μ" e varianza "σ^2/n"

Es: $$z = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})} \sim N(0,1)$$

Da cui:

Es: $$Pr\left(-z_{1-\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

ES:
$$Pr\left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

ovvero $\mu = \bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ con affidabilità $(1-\alpha)\%$.

Facciamo un esempio e consideriamo $n = 40$, $\bar{x} = 1{,}7$ cm, $\sigma = 0{,}5$ cm: dovremo calcolare l'intervallo fiduciario per μ con $(1-\alpha) = 95\%$

1. Calcolo $\frac{\alpha}{2} = 0{,}025$ con $\alpha = 1 - 0{,}95 = 0{,}05$
2. Valuto $z_{1-\frac{\alpha}{2}} = z_{1-0{,}025} = z_{0{,}975} \xrightarrow{\text{(tabelle)}} z_{0{,}975} = 1{,}96$
3. Calcolo $z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} = z_{0{,}975} * \frac{\sigma}{\sqrt{n}} = 1{,}96 * \frac{0{,}5}{\sqrt{40}} = 0{,}15$

Quindi $\mu = \bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 1{,}7 \pm 0{,}15$ con 95% di affidabilità

PRECISIONE VS AFFIDABILITÁ

Vediamo ora le relazioni tra precisione, affidabilità e parametri di studio come ad esempio "d":

- valori grandi di "d" maggiore sarà l'affidabilità
- valori piccoli di "d" maggiore sarà la precisione

$$d = z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

Osserviamo poi la formula inversa

ES:
$$n = \left(\frac{z_{1-\frac{\alpha}{2}} * \sigma}{d}\right)^2$$

per aumentare entrambe posso solo lavorare sulla NUMEROSITÁ del campione con

- "n" direttamente proporzionale a $z^2_{1-\frac{\alpha}{2}}$ (affidabilità)
- "n" inversamente proporzionale a d^2 (precisione)

TEST STATISTICI

I test statistici sono lo scopo di tutto ciò che abbiamo studiato prima e pertanto, ora, andremo ad analizzarli. Permettono di dedurre conclusioni generali a partire da informazioni campionarie utilizzando i "tool" studiati nei capitoli precedenti. Al fine di spiegare meglio l'argomento ci aiuteremo con due esempi:

- **Esempio 1:** un'impresa edile ha acquistato una fornitura di cavi d'acciaio, con un carico di rottura dichiarato a 500 bar. Per verificare questa ipotesi, l'impresa seleziona un campione casuale di 10 cavi, e ne misura i carichi di rottura

- **Esempio 2:** viene sviluppato un motore automobilistico per il quale è richiesta una benzina con un numero di ottano almeno pari a 98. Per verificare che la benzina selezionata ha le caratteristiche necessarie, si prelevano 10 campioni e ne si valuta la resistenza alla detonazione

Al fine di trarre conclusioni si imbandisce un **test statistico** che, però, per essere fatto, necessita di **IPOTESI** ("H" da "Hypotesis") da verificare. Chiariamo subito che "accettare" un ipotesi **non** per forza significa che questa sia vera poiché il Tutto dipende dal campione CASUALE sul quale si effettua il test

Es: "ACCETTAZIONE" $\neq$ "VERITÀ"

Formuliamo ora le ipotesi del problema (2) e consideriamo come ipotesi **nulla** (H_0) che il numero di ottano sia 98 mentre, di contro, consideriamo come ipotesi **alternativa** (H_1) l'opposto. I dati raccolti ci aiuteranno a determinare se accettare o rifiutare la prima, e quindi la seconda

Es: $H_0 \Rightarrow \mu = 98$ $\quad$ $H_1 \Rightarrow \mu \neq 98$

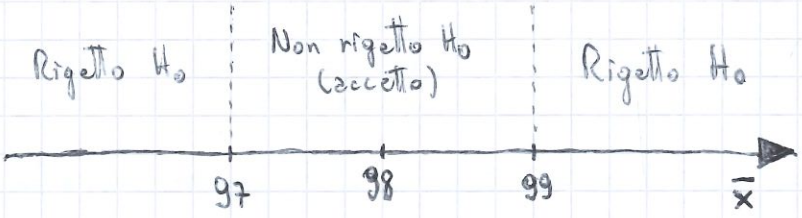
Perché il test statistico sia consistente richiede che le ipotesi siano

. Esaustive: H_0 e H_1 coprono l'intero spazio dei risultati

. Mutuamente esclusive: H_0 e H_1 non possono verificarsi contemporaneamente

A questo punto sulla base del campione di "n" misure è possibile fare inferenze sui parametri della popolazione: calcolo $\bar{x}$ e stimo μ. Così facendo otterremo il seguente, dove vi sarà un intervallo arbitrario di comodità grazie al quale trarre conclusioni sensate

Es:



A questo punto si prendono decisioni, Tuttavia però queste possono anche essere errate. Avremo due tipologie di errori possibili:

. Errore del I Tipo: rigetto H_0 ma H_0 è vera

. Errore del II Tipo: non rigetto H_0 ma H_0 è falsa

Si tratta di errori dovuti a variabili casuali, per cui hanno una probabilità associata

. $\alpha = Pr$ (errore del I Tipo)

. $\beta = Pr$ (errore del II Tipo)

Facciamo un esempio e supponiamo che l'errore σ nella misura del numero di ottani sia noto e pari 1,7. Supponiamo inoltre di avere $n = 10$ misure e il Teorema del limite centrale applicabile, avremo quindi che:

Es: $$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \longrightarrow \bar{x} \sim N(98, 0{,}53)$$

A questo punto soffermiamoci sul primo errore e definiamo un livello di significatività α: rigetto H_0 solo se dai campioni è disposizione è molto improbabile che H_0 sia vera. Rigetto H_0 se la stima di "μ", ovvero $\bar{x}$, è molto diversa dal valore de verifresve

Es: $$\alpha = Pr(\bar{x} < 97 \text{ se } \mu = 98) + Pr(\bar{x} > 99 \text{ se } \mu = 98)$$

Continuando con lo stesso esempio passiamo alla distribuzione normale STANDARD

$$z_1 = \frac{97-98}{0,53} = -1,9 \qquad z_2 = \frac{99-98}{0,53} = 1,9$$

Es: $Pr(z < -1,9) = 0,029$ $\qquad Pr(z > 1,9) = 0,029$

$$\longrightarrow \alpha = 0,029 + 0,029 = 0,058 \longrightarrow 5,8\,\%$$

Si ottiene che il 5,8 % di tutti i campioni casuali porterebbe a rigettare l'ipotesi H_0 ($\mu = 98$) quando il valore vero del numero di ottano è 98. Inoltre è possibile ridurre α:

. Ampliando la regione di accettazione (es: 96,7 e 99,3)

$z_1 = \ldots = -2,45$, $z_2 = \ldots = 2,45$

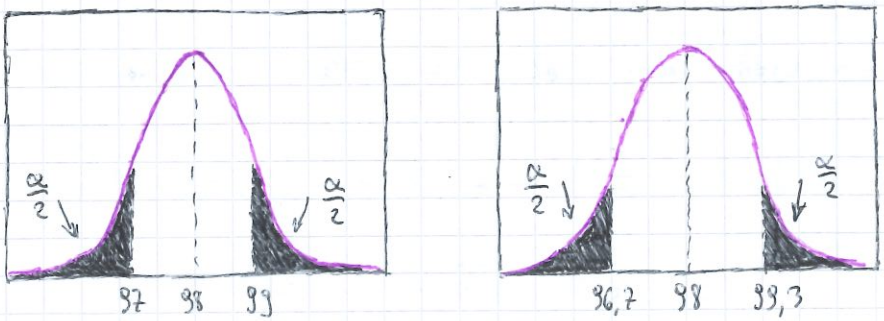
Es: $\longrightarrow \alpha = Pr(z < -2,45) + Pr(z > 2,45) = 0,007 + 0,007 = 1,4\,\%$

. Aumentando la numerosità del campione (es: da $n = 10$ a $n = 16$)

$\frac{\sigma}{\sqrt{n}} = 0,43 \longrightarrow z_1 = \ldots = -2,3$, $z_2 = \ldots = 2,3$

$\longrightarrow \alpha = Pr(z < -2,3) + Pr(z > 2,3) = 0,008 + 0,008 = 1,6\,\%$

Es:



E per quanto riguarda l'errore del II tipo? In questo caso la probabilità di "β" non esisterà in quanto, non rigettare H_0 quando è falsa può andare bene, ma per trovare la probabilità "β" abbiamo bisogno di formulare una specifica ipotesi alternativa H_1: con ciò non si può lavorare in quanto si avranno INFINITE POSSIBILITÀ.
Viene così verificata solo l'ipotesi nulla H_0, non potendo analizzare tutte le ipotesi alternative.
Di contro a questa, quella precedente, ovvero rigettare l'ipotesi H_0, è quindi una conclusione forte!

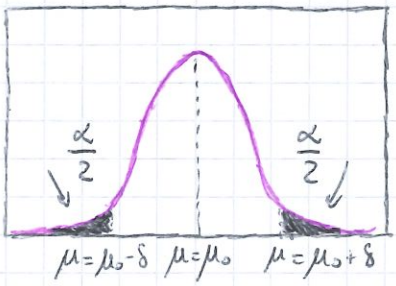
Il test statistico viene formulato in modo da cautelarsi contro un errore del primo tipo.

TIPI DI TEST STATISTICI

Distinguiamo ora due tipologie di questi

- Test a due code:

$$H_0: \mu = \mu_0 \quad , \quad H_1: \mu \neq \mu_0$$

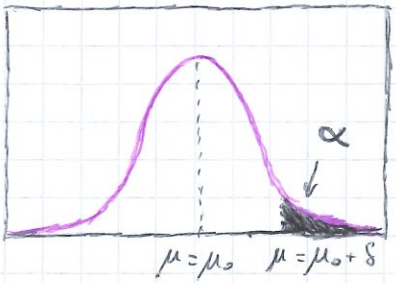


- Test a una coda:

$$H_0: \mu \leq \mu_0 \quad , \quad H_1: \mu > \mu_0$$

oppure

$$H_0: \mu \geq \mu_0 \quad , \quad H_1: \mu < \mu_0$$



Per il test a una coda come basis a fissare l'area sottesa migliore: dovrà fare in modo di averne una tale da cautelarmi da errori del I tipo. ovvero rigettare H_0 quando è vera. Ad esempio scartare una benzina quando potrebbe essere appropriata per il motore considerato.

Impostiamo quello che è il nostro test statistico attraverso 5 fasi:

1. Formulazione delle ipotesi H_0 e H_1
2. Individuazione della statistica coinvolta
3. Definizione livello di significatività di accettazione α

} Fase progettuale

4. Raccolta dati campionari
5. Decisione

} Fase esecutiva

Facciamo un esempio di test statistico a due code: consideriamo l'esempio due di partenza e aggiungiamo che una raffineria vuole verificare se una certa benzina prodotta è in grado di soddisfare questa specifica con un livello di significatività $\alpha = 0,05$. Viene fatta un'analisi attraverso $n = 25$ misure casuali, ottenendo un $\bar{x} = 99$. È possibile quindi mettere in commercio questa benzina?

1. Formulazione delle ipotesi H_0 e H_1

Es: $H_0 : \mu = 98$ $\quad H_1 : \mu \neq 98$

2. Individuazione della statistica coinvolta

Es: T.L.C.: $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ e $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

3. Definizione livello di significatività di rigetto α

Es: $\alpha = 0,05 \rightarrow \frac{\alpha}{2} = 0,025$ con $\alpha = Pr(z < -z_{\alpha/2}) + Pr(z > z_{\alpha/2}) = 0,05$

$\rightarrow z_{\alpha/2} = z_{0,025} = 1,96$ (da tabelle)

4. Raccolta dati campionari

Es: $\bar{x} = 99$ $\quad z = \frac{99 - 98}{(1,7/\sqrt{25})} = 2,94$

5 Decisione

ES: $z = 2,84 \longrightarrow z > z_{\alpha/2}$

Quindi l'ipotesi nulla H_0 viene rigettata con livello di significatività del 5%

Se invece facciamo lo stesso esempio di test statistico ma ad una coda, con la richiesta che il numero di ottano debba essere di ALMENO pari a 98, avremo

1 Formulazione delle ipotesi

ES: $H_0: \mu \geq 98 \quad H_1: \mu < 98$

2 Individuazione della statistica coinvolta

ES: T.L.C.: $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ e $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

3 Definizione livello di significatività di rigetto α

ES: $\alpha = 0,05$ con $\alpha = Pr(z < -z_\alpha)$ ovvero coda a sinistra

$\longrightarrow -z_{0,05} = 1,64$ (da Tabelle)

4 Raccolta dati campionari

ES: $\bar{x} = 99 \qquad z = \frac{99 - 98}{(1,7/\sqrt{25})} = 2,84$

5 Decisione

ES: $z = 2,84 \longrightarrow z > -z_{0,05}$

L'ipotesi nulla non viene rigettata con livello di significatività del 5% ma se $\bar{x} = 97$ allora $z = \ldots = -2,84$ e $z < -z_\alpha$, dove $-z_\alpha = -z_{0,05}$ e pertanto l'ipotesi nulla H_0 verrebbe rigettata con livello di significatività del 5%

Si parla inoltre di test statistico a **una coda in alto** nel caso, ad esempio, che si voglia il numero di attesa inferiore a 98 e pertanto le ipotesi saranno

ES: $H_0: \mu \leq 98 \qquad H_1: \mu > 98$

e il calcolo sarà simmetrico all'esempio subito precedente

P VALUE

Spesso, nei test statistici il livello di significatività α non è impostato in anticipo, ma si ragiona a ritroso: si guarda ai dati per individuare il valore minimo di α che porterebbe a rigettare l'ipotesi H_0. Si indica p-value, o valore P e si valuta a seconda del tipo di test:

- Test a due code

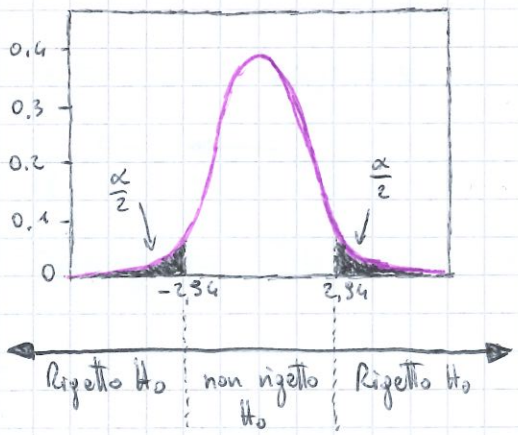
  $H_0: \mu = \mu_0$
  $H_1: \mu \neq \mu_0$

  $$P = 2\left[1 - Pr\left(z < z_{\alpha/2}\right)\right]$$

- Test a una coda in alto

  $H_0: \mu \leq \mu_0$
  $H_1: \mu > \mu_0$

  $$P = 1 - Pr\left(z < z_{\alpha}\right)$$



- Test a una coda in basso

  $H_0: \mu \geq \mu_0$
  $H_1: \mu < \mu_0$

  $$P = Pr\left(z < -z_{\alpha}\right)$$

Nell'esempio di prima con $\bar{x} = 99$ avremo

ES:

$$z = \frac{99 - 98}{(1{,}7/\sqrt{25})} = 2{,}94 \xrightarrow{\text{(da tabelle)}} Pr\left(z < 2{,}94\right) = 0{,}9984$$

$$\rightarrow P = 2\left[1 - Pr\left(z < 2{,}94\right)\right] = 2\left[1 - 0{,}9984\right] = 0{,}0032$$

H_0 sarebbe rigettata per ogni livello di significatività $\alpha \geq 0{,}0032$: per esempio l'ipotesi nulla sarebbe rigettata per $\alpha = 0{,}01$ ma non per $\alpha = 0{,}001$

TEST STATISTICI E DISTRIBUZIONI

La struttura in 5 fasi dei test statistici è applicabile a diverse tipologie di problemi nel campo statistico, con obiettivo comune di verificare una determinata ipotesi.

Ad esempio si hanno una serie di osservazioni e si vuole verificare se queste sono descrivibili attraverso una certa distribuzione. Per fare ciò si confrontano le frequenze di appartenenza a un certo intervallo osservate (O_K) con le equivalenti attese se le misure appartenessero a quella distribuzione (E_K). È intuibile che le deviazioni "$O_K - E_K$" devono essere piccole affinché le osservazioni seguano la distribuzione assunta.

Si usa allo scopo la distribuzione χ^2 (chi-quadro), definita come

ES:

$$\chi^2 = \sum_{n=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Introduciamo ora il concetto di gradi di libertà, ovvero il numero di dati osservati (k), o classi in cui si suddivide l'intervallo totale, meno il numero di parametri calcolati (c) dai dati usati nel calcolo

ES:

$$glc = k - c$$

Facciamo un esempio pratico: abbiamo un forno e vengono eseguite 40 misure di temperatura (°C), vediamone i valori

ES:

731	772	771	681	722	688	653	757	733	742
739	780	709	676	760	748	672	687	766	645
678	748	689	810	805	778	764	753	709	675
698	770	754	830	725	710	738	638	787	712

Il test statistico che andremo a fare risponderà a: "è ragionevole considerare queste misure come normalmente distribuite?" (con media e varianza stimate dal campione).

ES: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 730{,}1\ °C$; $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = 2190{,}2\ °C^2$; $s = \sqrt{s^2} = 46{,}8\ °C$

Dal T.L.C. se questi dati sono distribuiti normalmente allora

ES: $$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad (*)$$

Pertanto la domanda sarà questa (*). A questo punto svolgiamo il test statistico passando per le 5 fasi risolutive:

1. Formulazione delle ipotesi H_0 e H_1

ES: H_0: distribuzione normale H_1: distribuzione non normale

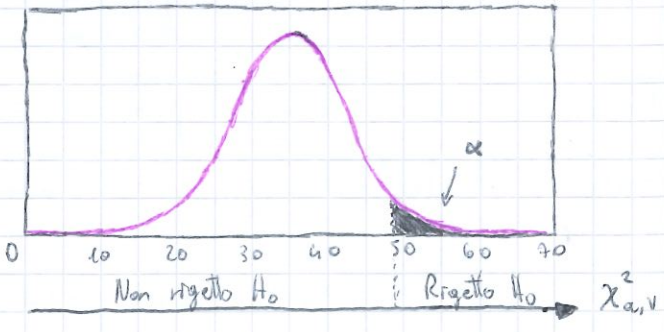
2. Individuazione della statistica coinvolta

ES: "se $\bar{x} \sim N \Rightarrow \exists\ \chi^2$" con $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

3. Definizione livello significatività di rigetto α (arbitraria)

$$\alpha = Pr(\chi^2 > \chi^2_{\alpha,\nu}) = 0{,}05 = 5\%$$

ES: "Se χ^2 è piccolo (differenze $O_i - E_i$ piccole) non rigettiamo l'ipotesi H_0: consideriamo che la distribuzione dei dati sia NORMALE"



4. Raccolta dati campionari

"La misura x è una variabile continua: $Pr(x = X) = 0$, per cui bisogna ragionare su intervalli di dati, valutando per ciascuno di essi $Pr(a < x < b)$"

"k intervalli, valore arbitrario, purché maggiore del numero di vincoli, ovvero i gradi di libertà (con gdl > 0). Arbitrari sono anche i relativi VALORI LIMITE:

- si conta il numero di osservazioni di ciascun intervallo (O_i)
- li si confronta con il numero atteso di osservazioni (E_i): prodotto del numero totale di osservazioni (n) per la probabilità della distribuzione NORMALE di cadere in quell'intervallo"

"È opportuno scegliere i valori degli intervalli in maniera opportuna, ovvero ad esempio in base alle probabilità note:

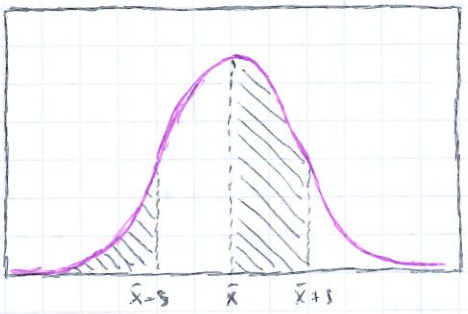
ES:

$x < \bar{x} - s$ ~ 16%

$\bar{x} - s < x < \bar{x}$ ~ 34%

$\bar{x} < x < \bar{x} + s$ ~ 34%

$x > \bar{x} + s$ ~ 16%



"

"

i	Intervallo	Osservati O_i	Pr (normale)	Attesi (E_i)
1	$x < \bar{x} - s$ ($T < 683,3$)	8	0,16	6,4 ($40 \times 0,16$)
2	$\bar{x} - s < x < \bar{x}$ ($683,3 \leq T < 730,1$)	10	0,34	13,6 ($40 \times 0,34$)
3	$\bar{x} < x < \bar{x} + s$ ($730,1 \leq T < 776,8$)	16	0,34	13,6 ($40 \times 0,34$)
4	$x > \bar{x} + s$ ($T \geq 776,8$)	6	0,16	6,4 ($40 \times 0,16$)

ES:

Fatta la tabella, calcoliamo il "Chi-quadro": ($k = 4$)

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(8-6,4)^2}{6,4} +$$

$$\frac{(10-13,6)^2}{13,6} + \frac{(16-13,6)^2}{13,6} + \frac{(6-6,4)^2}{6,4} = 1,8$$

"

5 Decisione

"Stabiliamo i gradi di libertà: $gdl = k - c$ dove

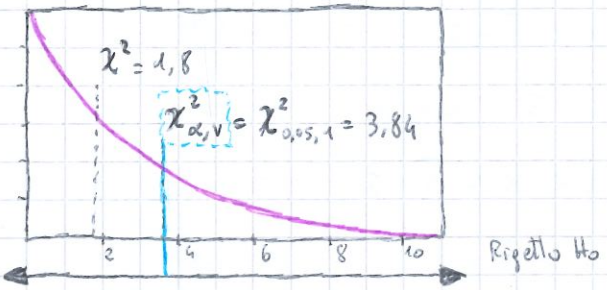
- $k = 4$ (intervalli)
- $c = 3$ perché abbiamo trovato tre vincoli "di studio statistico" ovvero media ($\bar{x}$), varianza → dev. standard ($s^2 \to s$) e numero osservazioni (n)

Quindi $gdl = k - c = 4 - 3 = 1$ e ora (tramite tabella) troviamo il $\chi^2_{\alpha,\nu}$ critico che ci fa da spartiacque nella nostra decisione. Una volta trovato vediamo se è maggiore, minore del χ^2 trovato e calcolato con i dati di partenza del problema:

ES:

$$\chi^2_{\alpha,\nu} = \chi^2_{0,05,1} = 3,84$$

$$\chi^2 = 1,8 < \chi^2_{0,05,1}$$



A questo punto deduciamo: non c'è forte evidenza che l'ipotesi H_0 sia da rigettare e pertanto possiamo assumere che la Temperatura sia NORMALMENTE DISTRIBUITA"

Se invece di $k = 4$ intervalli avessi preso $k' = 6$ intervalli, per le probabilità (dato che non è nota), bisognava ricondursi alla distribuzione NORMALE STANDARD al fine di trovare la probabilità classe per classe.
In qualsiasi caso, anche se variassero i gradi di libertà e di conseguenza il $\chi^2_{\alpha,\nu}$ critico, riscontreremo lo stesso risultato del test statistico precedente, ovvero con $k = 4$ intervalli.

df	$\chi^2_{0,995}$	$\chi^2_{0,990}$	$\chi^2_{0,975}$	$\chi^2_{0,950}$	$\chi^2_{0,900}$	$\chi^2_{0,100}$	$\chi^2_{0,050}$	$\chi^2_{0,025}$	$\chi^2_{0,010}$	$\chi^2_{0,005}$
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589

CONFRONTO TRA POPOLAZIONI

In questo capitolo analizzeremo popolazioni diverse, insieme. Confrontiamo generalmente le variabili di due popolazioni P_1 e P_2:

- P_1: $E(P_1) = \mu_1$, $V(P_1) = \sigma_1^2$, numerosità campione $= n_1$
- P_2: $E(P_2) = \mu_2$, $V(P_2) = \sigma_2^2$, numerosità campione $= n_2$

A questo punto, per n_1, n_2 sufficientemente elevati, grazie al **Teorema del limite centrale**, avremo e potremo assumere

ES: $\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ e $\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$

Da ciò ne deriviamo che

ES: $D \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$, $D = \bar{x}_1 - \bar{X}_2$

Generalmente le varianze delle popolazioni <u>non</u> sono note, per cui la varianza della differenza va stimata in base alle VARIANZE CAMPIONARIE. Il metodo empirico diffuso è il seguente

- $n_1, n_2 > 30$: $s_D^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$
- n_1 e/o $n_2 < 30$: $s_D^2 = \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

A questo punto dovremo rispondere alla domanda iniziale, esistono due <u>modi</u> per farlo e li vedremo entrambi

INTERVALLI FIDUCIARI

La teoria per questo metodo ci dice che per trovare gli intervalli dovremo rapportarci alla distribuzione **normale standard** con:

ES: $$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_D}$$

Gli intervalli fiduciari con probabilità saranno

$$Pr\left[(\bar{x}_1 - \bar{x}_2) - d \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + d\right] = 1 - \alpha$$

ES:

$$Pr\left[-z_{1-\frac{\alpha}{2}} \leq Z \leq +z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

Sapendo rapportarci alla distribuzione NORMALE STANDARD otterremo

ES: $$Pr\left[(\bar{x}_1 - x_2) - z_{1-\frac{\alpha}{2}}\sigma_D \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_{1-\frac{\alpha}{2}}\sigma_D\right] = 1 - \alpha$$

e pertanto l'intervallo fiduciario con affidabilità "$1-\alpha$" sarà:

ES: $$\mu_1 - \mu_2 = (\bar{x}_1 - x_2) \pm z_{1-\frac{\alpha}{2}}\sigma_D$$

Facciamo un esempio pratico: misurazione del calore di neutralizzazione in [kJ/mol] di NaOH con HCl. Abbiamo raccolto campioni con due metodi diversi, osserviamone i dati riscontrati:

- Metodo 1: $n_1 = 65$, $\bar{x}_1 = 57,3$ kJ/mol, $\sigma_1^2 = 2,7\ (kJ/mol)^2$
- Metodo 2: $n_2 = 32$, $\bar{x}_2 = 57,0$ kJ/mol, $\sigma_2^2 = 1,1\ (kJ/mol)^2$

Calcolare l'intervallo fiduciario per il valor vero della differenza tra i due metodi con affidabilità del 90%, quindi

$$1 - \alpha = 90\% \longrightarrow \alpha/2 = 0,05 \longrightarrow z_{1-\alpha/2} = 1,645$$

$$D = (\bar{x}_1 - \bar{x}_2) \sim N(\mu_1 - \mu_2, \sigma_D^2) \quad \text{con} \quad \sigma_D = \sqrt{\frac{2,7}{65} + \frac{1,1}{32}} = 0,3$$

ES:

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm z_{1-\frac{\alpha}{2}}\sigma_D = (57,3 - 57) \pm 1,645 \cdot 0,3 = 0,3 \pm 0,5$$

$$-0,2 \leq \mu_1 - \mu_2 \leq 0,8$$

L'intervallo fiduciario contiene il valore 0: non ho evidenze per dire che le due stime sono diverse.

TEST STATISTICI PER CONFRONTI

Se al posto degli intervalli fiduciari si vuole utilizzare un altro metodo, si potrà usare il test statistico. Il risultato dello studio non cambierà.

Facciamo un esempio pratico: un'azienda vuole verificare se una nuova verifica automatica controllo qualità sia più veloce di quella manuale. A tale scopo organizza due serie di prove eseguendo:

- 50 verifiche automatiche (n_1)
- 100 verifiche manuali (n_2)

Da precedenti dati la varianza (σ) è la stessa, con 14 s.

1. Formulazione delle ipotesi H_0 e H_1

es: $H_0: \mu_1 \geq \mu_2 \longrightarrow \mu_1 - \mu_2 \geq 0$, $H_1: \mu_1 < \mu_2 \longrightarrow \mu_1 - \mu_2 < 0$

2. Individuazione della statistica coinvolta

es: $D = (\bar{x}_1 - \bar{x}_2) \sim N(\mu, \sigma_D^2)$ con $\sigma_D^2 = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ e $z = \frac{D}{\sigma_D}$

3. Definizione livello di significatività di rigetto α

es: $\alpha = Pr(z > z_\alpha) = 0{,}01 = Pr(z < -z_\alpha) \Rightarrow z_\alpha = -2{,}33$

4. Raccolta dati campionari

es: $\bar{x}_1 = 45{,}6$ s, $\bar{x}_2 = 59{,}6$ s $\longrightarrow z = \frac{D}{\sigma_D} = \dots = -5{,}6$

5. Decisione

es: $$z < -z_\alpha$$

L'ipotesi nulla va rigettata: la probabilità che il risultato di maggior velocità della verifica automatica sia dovuto al caso è minore dell'1%.

Vediamo ora un caso particolare, ovvero il test statistico su campioni accoppiati: ogni coppia di dati viene raccolta in condizioni OMOGENEE, che però possono variare da una coppia all'altra. In questo modo è possibile verificare l'effetto di un certo trattamento senza effetti distorsivi di altro genere.

Facciamo un esempio pratico e vediamone lo studio per mezzo dei due metodi precedenti: si hanno due forni e si cerca di capire quale dei due riscalda nel minor tempo possibile. Per rendere omogeneo il test si useranno gli stessi materiali, divisi in due parti equivalenti, contemporaneamente nei due forni. Dai dati otteniamo:

- Forno 1: $\bar{x}_1 = 211,3$, $s_1 = 26,1$
- Forno 2: $\bar{x}_2 = 202,9$, $s_2 = 23,7$
- $D = (x_1 - x_2)$: $\bar{d} = 8,4$, $s_d = 9,3$ (*)

Analizziamo il problema prima con gli intervalli fiduciari: quale è l'intervallo della differenza tra i due forni con affidabilità 95%?

$$\bar{d} \sim N\left(\mu_d, \frac{s_d^2}{n}\right) \quad \text{grazie a T.L.C.}$$

ES: $$\mu_d = \bar{d} \pm z_{1-\alpha/2} \frac{s_d}{\sqrt{n}} \quad \text{e} \quad z_{1-\alpha/2} = z_{0,975} = 1,96$$

$$\mu_d = 8,4 \pm 1,96 \cdot \frac{9,3}{\sqrt{7}} = 8,4 \pm 1,96 \cdot 3,5 = 8,4 \pm 6,9$$

Poiché nell'intervallo non è presente lo 0 e data la differenza

ES: $$x_1 - x_2 > 0 \longrightarrow x_1 > x_2$$

è ragionevole affermare che il primo forno sia meno efficiente e scaldi più lentamente del secondo forno.

Cosa succede se utilizzo la "DIFFERENZA DELLE DISTRIBUZIONI" anziché la DISTRIBUZIONE DIFFERENZA? Per fare ciò non dovremo tenere conto di (*) ma soltanto fare i calcoli partendo dalle altre due.

$$s_D = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 13,3 \text{ min}$$

ES:

$$\mu_1-\mu_2 = (\bar{x}_1-\bar{x}_2) \pm z_{1-\frac{\alpha}{2}} \cdot s_D = 8,4 \pm 1,96 \cdot 13,3 = 8,4 \pm 26,07$$

Stesso valore centrale, ma l'accoppiamento riduce la variabilità interna, dato che considera le coppie di prove come eseguite in condizioni omogenee!

Ora facciamo la stessa cosa ma con il test statistico e separiamo:

1. Formulazione ipotesi H_0 e H_1

ES: $H_0: \mu_1-\mu_2 \leq 0$, $H_1: \mu_1-\mu_2 > 0$

2. Identificazione della statistica coinvolta

ES: $d \sim N(\mu_d, \sigma_d^2)$ | "$D=(\bar{x}_1-\bar{x}_2) \sim N(\mu_d, \sigma_d^2)$, $z = \frac{D}{s_D}$"

3. Definizione livello significatività di rigetto α

ES: $\alpha = Pr(z > z_\alpha) = 0,05$, $z_\alpha = 1,65$ test a una coda in alto

4. Raccolta dati campionari

ES: $z = \frac{8,4-0}{(9,3/\sqrt{7})} = 2,4$ | "$z = \frac{211,3-202,9}{13,3} = 0,63$"

5. Decisione

ES: $z > z_\alpha$ | "$z < z_\alpha$"

"H_0 viene rigettata" | "H_0 non viene rigettata"

Non considerare le condizioni OMOGENEE per ogni prova aumenta la variabilità e modifica la decisione finale!

REGRESSIONE LINEARE SEMPLICE

Considerando n variabili casuali, per calcolare la probabilità dell'evento $\bar{E}$: « le variabili casuali $(x_1, \dots, x_n)$ risultano tutte minori di una serie di valori assegnati $(X_1, \dots X_n)$

ES: $\bar{E} = x_1 \leq X_1$ AND ... AND $x_n \leq X_n$

È possibile inoltre associare una probabilità a ciascun evento solo se le variabili sono indipendenti tra loro. In tal caso:

ES: $Pr(\bar{E}) = \dots = Pr(x_1, \dots, x_n) = p(x_1) \cdot p(x_2) \cdot \dots \cdot p(x_n)$

È possibile così quantificare il grado di correlazione, nonché la dipendenza, tra due variabili casuali mediante la covarianza:

$$\sigma_{xy} = E(xy) - E(x)\,E(y)$$

ES:

$$\sigma_{xy} = E\left[(x-\mu_x)(y-\mu_y)\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-\mu_x)(y-\mu_y)\, p(x,y)\, dx\, dy$$

Quindi avremo che:

- se le variabili sono INDIPENDENTI, $E(xy) = E(x)\,E(y) \longrightarrow \sigma_{xy} = 0$
- se le variabili sono DIPENDENTI è possibile che sia $E(xy) \neq E(x)\,E(y)$ ma <u>non necessariamente</u>

Vediamo ora le proprietà dell'operatore covarianza:

- $Cov(xy) = Cov(yx)$
- $Cov(aX + bY) = a * Cov(XY)$
- $Cov(X+Y)Z = Cov(XZ) + Cov(YZ)$

dove a, b costanti $\in \mathbb{R}$ e X, Y, Z variabili

La covarianza è parte integrante del coefficiente di correlazione lineare

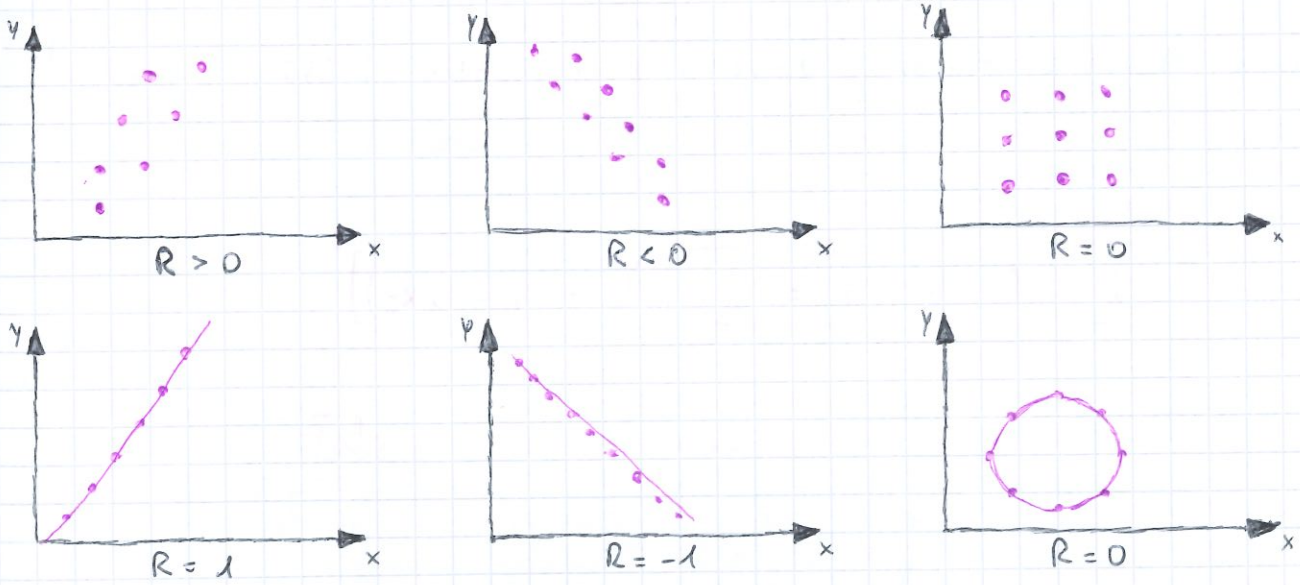
ES: $$R = \frac{S_{xy}}{S_x S_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$ adimensionale $-1 \leq R \leq 1$

Ovvero:

ES: $$R = \frac{E(xy) - E(x)\,E(y)}{\sqrt{\left[E(x^2) - (E(x))^2\right]\left[E(y^2) - (E(y))^2\right]}}$$

Accade che due variabili perfettamente correlate (es.: $y = 2x$) avranno un coefficiente di correlazione lineare unitario ($R = 1$). Tuttavia non è sempre così, per certe funzioni si parla di anticorrelazione ($R = -1$) come per "$y = -x$". Per questo motivo spesso si parla di "R^2" e $0 \leq R^2 \leq 1$ dove la perfetta correlazione lineare avrà "$R^2 = 1$"

(*)



Considerando un'indagine di laboratorio: all'aumentare della pressione aumenterà la conversione (%) di fase di un gas. Sembra intuibile un legame lineare tra queste variabili, tuttavia non avremo mai un "$R = 1$" poiché le misurazioni mostreranno deviazioni dovute a incertezze sperimentali. Quindi come RAZIONALIZZIAMO il legame e l'andamento?

MODELLO DI CORRELAZIONE LINEARE

Osservando i dati, disposti in modo analogo a (*), sembra logico attendersi che la conversione sia linearmente legata alla pressione, in generale

ES:
$$Y = b_0 + b_1 X$$

Se pertanto si assume ciò, quale è la retta che MEGLIO rappresenta questo legame? Che valori hanno "b_0 e b_1"?
Prendiamo in considerazione la seguente ipotesi: << le misure y sono soggette a incertezze tutte tra loro uguali (σ_y), mentre le misure di x sono più precise e si può RAGIONEVOLMENTE TRASCURARE l'incertezza di queste misure >>.
Riguardo a pressione (p) e conversione (c) scriviamo dei valori

ES: TOTALE: $\sum_p = 369$, $\sum_c = 103{,}8$, $\bar{x}_p = 36{,}9$, $\bar{x}_c = 10{,}38$, $s_p = 5{,}00$, $s_c = 2{,}60$

Dal momento che esistono infinite possibili correlazioni, al fine di scegliere la retta migliore, dovremo **minimizzare** l'SSE (ovvero il "Summed Squared Error") quantificando la BONTÀ di una retta:

ES:
$$\sum_{i=1}^{n} \left[Y_i - y(x_i) \right]^2 = \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 x_i) \right]^2 \quad \text{(residuo)}$$

Questo si chiamerà pertanto **metodo dei minimi quadrati**, ma come si può fare a minimizzare questo indice? Ebbene con le derivate parziali sulle nostre incognite, nello specifico b_0 e b_1, poste uguali a **zero**

ES:
$$\begin{cases} \frac{\partial}{\partial b_0} SSE = 0 \\ \frac{\partial}{\partial b_1} SSE = 0 \end{cases} \Longrightarrow \begin{cases} b_1 = \frac{n \sum x_i y_i - \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} \\ b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} \end{cases}$$

Questo che abbiamo trovato è quindi un minimo o un massimo? Dopo diversi calcoli troviamo comunque che si tratta di un MINIMO

Ora che abbiamo le formule calcoliamo i coefficienti dell'esempio proposto

$$b_1 = 0{,}44 \ [\%/\text{bar}] \ ; \ b_0 = -6{,}0 \ [\%]$$

ES:

$$\longrightarrow \hat{Y}_i = 0{,}44 X_i - 6{,}00$$

Abbiamo trovato così la retta migliore. A questo punto, se questo è un buon modello e descrive bene il processo, possiamo dire la singola misura Y_i sia distribuita **normalmente** Tale che

ES:

$$Y_i \sim N(b_0 + b_1 x_i, \sigma_y^2)$$

Come vediamo i valori Y_i con $i = 1, 2, \dots, n$ presentano una loro incertezza σ_y, generalmente non nota, la quale si propagherà alle costanti e pertanto b_0 e b_1 saranno stime dei valori veri di β_0 e β_1. Ora trattiamo quella che l'**incertezza dei parametri** e mostriamo

ES:

$$b_1 \sim N\left(\beta_1, \frac{\sigma_y^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right) \ ; \ b_0 \sim N\left(\beta_0, \frac{\sigma_y^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}\right) \quad (\not\exists)$$

Anche se saltando i passaggi abbiamo trovato valore atteso e varianza dei nostri parametri. Ora invece stimiamo la variabilità della popolazione a partire dalla deviazione media dei dati del campione poiché i residui chiudono a zero (somma deviazioni positive e negative). Useremo "n-2" gradi di libertà poiché, se "n = 2", si avrebbero solamente due punti per far passare una retta e di conseguenza non si potrebbe studiare la variabilità essendo univoca (minimo "n=3")

ES:

$$S_y = \sqrt{\frac{\sum [Y_i - (b_0 + b_1 X_i)]^2}{n-2}} = \frac{SSE}{n-2}$$

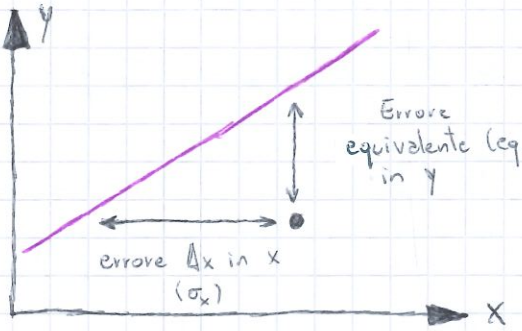
Non è possibile valutare quindi l'errore con **meno di 3 dati**: per 2 punti passa esattamente una retta e quindi il metodo dei minimi quadrati non può dare informazioni sullo scarto

Tornando all'ipotetico esercizio considerato troviamo quindi i valori:

ES: $$SSE = \sum \left[Y_i - \hat{Y}_i \right]^2 = 7,44 \quad ; \quad S_y = 0,96$$

Tutto questo ragionamento è stato fatto considerando "$\sigma_x \to 0$" mentre "$\sigma_y > 0$" calcolandoci i parametri e la loro varianza. La domanda è: «cosa succede se ho **incertezza su y** e **incertezza su x**?». Rispondiamo che possiamo sfruttare la propagazione dell'errore mediante derivate

ES: $y = b_0 + b_1 x$ → deriviamo → $\frac{dy}{dx} = b_1$



Applicata la derivazione troviamo mediante PROPAGAZIONE DELL'ERRORE, una delle prime lezioni del corso, l'errore equivalente (eq.)

ES: $$\sigma_{y,eq.} = \frac{dy}{dx}\sigma_x = b_1 \sigma_x$$

Se nessuno dei due errori (σ_x, σ_y) prevale sull'altro dovremo allora combinarli e, pertanto, se avremo misure indipendenti

ES: $$\sigma_{y,eq.\,tot} = \sqrt{\sigma_y^2 + \sigma_{y,eq.}^2} = \sqrt{\sigma_y^2 + (b_1 \sigma_x)^2}$$

Questo discorso vale SOLAMENTE quando (σ_x, σ_y) sono UNICHE e COSTANTI durante tutto lo studio del campione di partenza. Se ciò non vale esiste un metodo ancora più generale ("pesato") con l'introduzione di

ES:
$$\begin{cases} b_1 = \dfrac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2} \\ b_0 = \dfrac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2} \end{cases}$$

; $w_i = \frac{1}{\sigma_i^2}$ peso di ciascuna i-esima misura ma i valori di σ_i devono essere noti

A questo punto, noto un set di dati e stimati tutti i valori dei parametri, è possibile stimare gli intervalli fiduciari di b_0 e b_1. Sappiamo da (#) che i nostri parametri saranno

ES: $$b_0 \sim N(\beta_0, \sigma_{b_0}^2) \quad ; \quad b_1 \sim N(\beta_1, \sigma_{b_1}^2)$$

e di conseguenza, fissata una certa affidabilità "$1-\alpha$" otterremo

ES: $$\beta_0 = b_0 \pm t_{1-\alpha/2} \cdot \sigma_{b_0} \quad \text{e} \quad \beta_1 = b_1 \pm t_{1-\alpha/2} \cdot \sigma_{b_1}$$

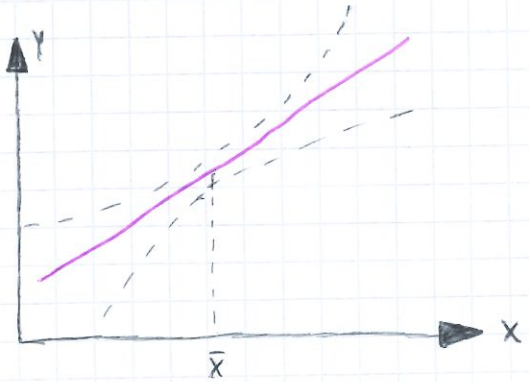
dove $(\sigma_{b_0}, \sigma_{b_1})$ non sono note, ma stimate da s_y sulla base di $n-2$ gradi di libertà.

A questo punto data la previsione della retta andrezziamo l'errore: abbiamo una retta costruita a partire da una serie di dati sperimentali quindi, se prendo un punto casuale x_p, quale è l'incertezza sulla previsione y_p derivante dal modello?

ES: $$\hat{y}_p = b_0 + b_1 x_p \quad , \quad \sigma_{y_p} = \sigma_y \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x_i^2 - (1/n)(\sum x_i)^2}}$$

Da quest'ultima formula si riezva che se "$x_p = \bar{x}$" allora "$\sigma_{y_p} = \sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}}$" e pertanto la variabilità minore si trova proprio nel valore medio di y seguendo le stesse formule del Teorema del limite Centrale. Inoltre si nota che l'errore dipende da x_p: quanto più la distanza aumenta dal valor medio $\bar{x}$ tanto più l'errore è maggiore → i limiti fiduciari diventano più ampi

ES:



Per poter fare statistica e attuare un'analisi regressionale bisogna stare attenti all'**estrapolabilità** delle conclusioni: i risultati di questo lavoro valgono **SOLO** all'interno del campo investigato e la linearità può non essere più tale al di fuori dell'intervallo considerato

Infine per capire se il modello è buono si analizzano i residui

ES: $$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

più i valori di "e_i" sono vicini allo zero e meglio è (positivi e negativi). Se questi sono <u>distribuiti normalmente</u> è possibile valutare i **limiti fiduciari** essendo il loro valore atteso nullo (nell'ipotesi del modello corretto)

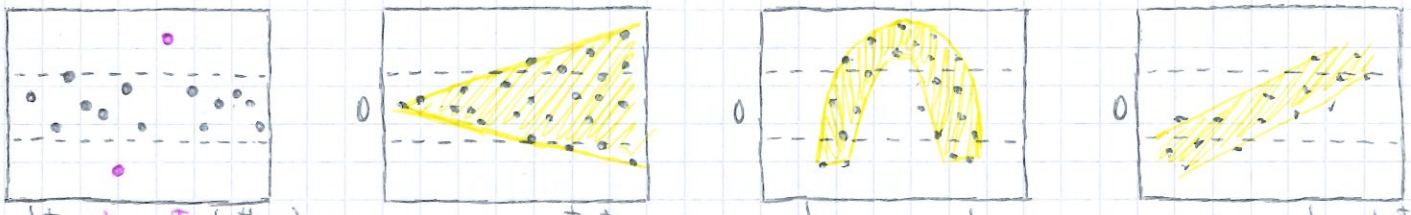
$$e_i \sim N(0, \sigma_y^2) \longrightarrow e_i = \pm z_{1-\alpha/2}\, \sigma_y$$

ES:

se $\alpha = 5\%$ allora $z_{0,975} = 1,96$ e $e_i = \pm 1,96 * 0,96 = \pm 1,9$

L'analisi dei residui permette di individuare comportamenti anomali di dati singoli o dell'intera serie

ES:



dati **aberranti** (outliers) — errore non costante — relazione non lineare — errori non indipendenti

Riconsiderando il **coefficiente di correlazione lineare** definiamo

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

devizione totale (SSTO) — devizione <u>non</u> spiegata (SSE) — devizione spiegata (SSR)

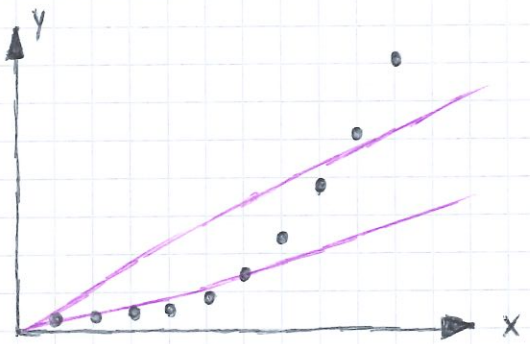
ES:

$$R^2 = \frac{\text{Devizione spiegata}}{\text{Devizione Totale}} = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO}$$

RELAZIONI NON LINEARI LINEARIZZABILI

Se ad esempio consideriamo il numero di cani inquinati (y) col passare degli anni (x) dopo il rilascio di una sostanza tossica

ES:



Per questi dati una relazione lineare non è corretta: il legame è di tipo esponenziale $y = b_0 b_1^x$. Se applichiamo il logaritmo otteniamo che

ES: $$\ln(y) = \ln(b_0 b_1^x) = \ln(b_0) + \ln(b_1^x) = \ln(b_0) + x \ln(b_1)$$

pertanto la nuova relazione, questa volta lineare sarà

$$\ln(y) = \ln(b_0) + x \ln(b_1)$$

ES:

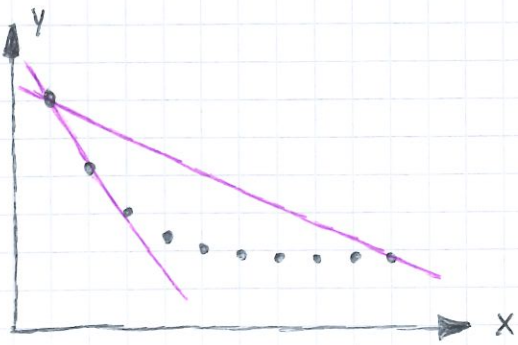
$$y_{new} = b_{0,new} + x\, b_{1,new}$$

Per ritrovare i parametri iniziali si farà l'inversa con l'esponenziale

ES: $$b_0 = e^{b_{0,new}} \quad \text{oppure} \quad b_1 = e^{b_{1,new}}$$

Se invece, ad esempio di nuovo, consideriamo il valore della densità dell'aria (Y) misurata a differenti temperature in gradi Kelvin (x)

ES:



Per questi dati una relazione lineare non è corretta: il legame è di tipo inverso $y = b_0 + \frac{b_1}{x}$. Si andrà a valutare allora il reciproco di x

ES:
$$y = b_0 + b_1 \left(\frac{1}{x}\right)$$

pertanto la nuova relazione, anche stavolta lineare sarà

ES:
$$y = b_0 + b_1 x_{new}$$

Mostriamo ora, qua sotto, esempi di modelli non lineari e le loro trasformazioni in modelli lineari ($\ldots_{new} = \ldots'$)

Modello NON lineare	Trasformazioni		Modello trasformato $y' = b_0 + b_1 x'$	
$y = ab^x$	$y' = \ln(y)$	$x' = x$	$b_0 = \ln(a)$	$b_1 = \ln(b)$
$y = ax^b$	$y' = \ln(y)$	$x' = \ln(x)$	$b_0 = \ln(a)$	$b_1 = b$
$y = a + \frac{b}{x}$	$y' = y$	$x' = \frac{1}{x}$	$b_0 = a$	$b_1 = b$
$\frac{1}{y} = a + bx$	$y' = \frac{1}{y}$	$x' = x$	$b_0 = a$	$b_1 = b$
$\frac{1}{y} = a + \frac{b}{x}$	$y' = \frac{1}{y}$	$x' = \frac{1}{x}$	$b_0 = a$	$b_1 = b$
$\frac{1}{y} = a + b\sqrt{x}$	$y' = y$	$x' = \sqrt{x}$	$b_0 = a$	$b_1 = b$

REGRESSIONE LINEARE MULTIPLA

Per introdurre questo capitolo partiamo dello stesso esempio precedente: effettuiamo una nuova serie di misure sul legame pressione - conversione e si calcolano le seguenti

$$b_1 = \dots = 0,39 \quad ; \quad b_0 = \dots = -4,22 \quad ; \quad S_y = \dots = 0,8$$

ES: $e_i \sim N(0, \sigma_y^2)$: $\alpha = 5\%$, $z_{1-\alpha/2} = 1,96$, $e_i = \dots = \pm 1,6$

$$R^2 = \frac{SSTO - SSE}{SSTO} = 0,88$$

e si osserverà una buona correlazione lineare. Ora invece trattiamo un altro set di misure/dati per lo stesso problema ma con l'unica differenza che per ogni misura si riscontrano tracce di ferro [mg] differenti. Siccome il ferro può agire da catalizzatore per la conversione, avremo rese migliori e pertanto, oltre alla variabile pressione (x_1), aggiungiamolo come variabile ferro (x_2): dato che poi entrambe condizionano la y in maniera ~ lineare si può proporre una legge lineare a più variabili come:

ES: $$y = b_0 + b_1 x_1 + b_2 x_2$$

Per calcolare "b_0, b_1, b_2" si usa il solito metodo dei minimi quadrati sostanto che si tratterà di un sistema lineare 3×3 più complesso del solito

ES:
$$\begin{cases} n b_0 + b_1 \sum x_{1,i} + b_2 \sum x_{2,i} = \sum y_i \\ b_0 \sum x_{1,i} + b_1 \sum x_{1,i}^2 + b_2 \sum x_{1,i} x_{2,i} = \sum y_i x_{1,i} \\ b_0 \sum x_{2,i} + b_1 \sum x_{1,i} x_{2,i} + b_2 \sum x_{2,i}^2 = \sum y_i x_{2,i} \end{cases}$$

Risolvendo i conti dell'esempio di problema:

ES: $$\hat{y} = -1,48 + 0,29 x_1 + 0,11 x_2$$

Allo stesso modo della regressione lineare semplice, in questa verrà utilizzato il metodo per l'analisi dei residui. Tuttavia cambia qualcosa per la variabilità della popolazione (σ_y) stimata dalla deviazione media dei dati del campione (s_y)

ES:
$$s_y = \sqrt{\frac{\sum [y_i - (b_0 + b_1 x_{1,i} + b_2 x_{2,i})]^2}{n-3}} = \sqrt{\frac{SS_{RES}}{n-3}}$$

da ciò si vede come servano almeno 4 punti per calcolare "s_y". A questo punto generalizziamo: cosa succede se le variabili sono "m-1" e non 2? Ricordiamo quindi i parametri ("m") e set di "n" dati allora otterremo la seguente

ES:
$$\hat{y}_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \dots + b_{m-1} x_{i,m-1}$$

dovremo pertanto passare ad una forma vettoriale, matriciale

ES:
$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{m-1} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,m-1} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,m-1} \\ 1 & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,m-1} \end{bmatrix}$$

moltiplicando entrambi i membri per la TRASPOSTA di X:

ES:
$$\hat{Y} = BX \longrightarrow X^T \hat{Y} = (X^T X) B \longrightarrow B = (X^T X)^{-1} X^T \hat{Y}$$

calcoliamo la deviazione media dei dati del campione

ES:
$$s_y = \sqrt{\frac{\sum [y_i - (b_0 + b_1 x_{1,i} + \dots + b_{m-1} x_{m-1,i})]^2}{n-m}} = \sqrt{\frac{SS_{RES}}{n-m}}$$

dove "n-m" saranno i gradi di libertà (dati utilizzati m volte per calcolare b_0, b_1, b_{m-1}). Per modelli NON lineari in "x" ma lineari nei parametri "$b_0, b_1, b_2 \dots$" si può effettuare un cambio di variabile e procedere analogamente, tenendo però conto nei calcoli di ciò: pertanto vi sarà immediatamente una estensione a ordine superiore

F - TEST

Trattiamo ora l'ultima tipologia di test statistico del corso, nonché l'ultimo argomento pre esame. Senza andare troppo nel dettaglio teorico, la distribuzione F, risulta dal rapporto di due varianze campionarie. Pertanto consideriamo due set di dati

ES: $A: (n_A, \bar{x}_A, s_A^2)$ $\quad B: (n_B, \bar{x}_B, s_B^2)$

Questo test di Fischer può risultare utile ad esempio nel caso di dover stabilire con un certo grado di significatività una caratteristica di un set di dati rispetto a un altro e si applica all'interno dei test statistici per verificare la diversa variabilità tra campioni

ES: $$F_{n_A-1, n_B-1} = \frac{s_A^2}{s_B^2} \qquad F_{v_1, v_2, \alpha} = \frac{1}{F_{v_2, v_1, \alpha}}$$

Facciamo un esempio: un'azienda vuole verificare se la variabilità del tempo necessario all'imballaggio del prodotto è inferiore attraverso la procedura A rispetto a quella B, avendo fatto una campagna di 10 prove per la prima e 15 prove per la seconda, che hanno prodotto i seguenti risultati

ES: • $s_A^2 = 29,3\ (s^2)$ • $s_B^2 = 10,4\ (s^2)$

Come si può giustificare il fatto che la variabilità di A sia maggiore? Imbastiamo un test statistico a una coda e procediamo

1. Formulazione ipotesi H_0 e H_1

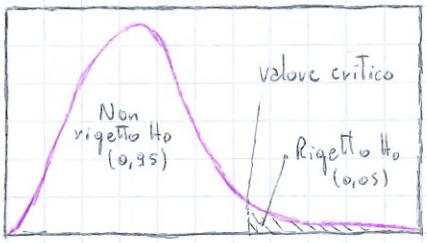
ES: $H_0: s_A^2 \leq s_B^2$ $\quad H_1: s_A^2 > s_B^2$

2. Individuazione della statistica coinvolta

ES: $$F_{n_A-1, n_B-1} = \frac{s_A^2}{s_B^2}$$

③ Definizione livello di significatività di rigetto α

ES: $\alpha = 0{,}05$, $n_A = 10$, $n_B = 15$ $\Rightarrow F_{9,14,0.05} = 2{,}65$ (tabelle)



④ Raccolta dati campionari

ES: $\frac{s_A^2}{s_B^2} = \frac{29{,}3}{10{,}4} = 2{,}82$

⑤ Decisione

ES: $\frac{s_A^2}{s_B^2} > F_{9,14,0.05}$ H_0 viene rigettata

La probabilità che l'aver ottenuto una variabilità della procedura A maggiore di quella B sia dovuta al caso è inferiore al 5%

La distribuzione F trova immediato utilizzo nella verifica dei modelli di regressione: affinché sia corretto, il modello adottato non deve aggiungere ulteriore variabilità nei dati rispetto a quella sperimentale.
Nota "s_Y" da altre "N_s" informazioni sperimentali e "m = n° param. modello"

ES: $\frac{[SSE/(N-m)]}{s_Y^2} \sim F_{N-m,\,N_s}$

Impostiamo il test statistico

ES: $H_0 : \frac{SSE}{(N-m)} \leq s_Y^2$ $\qquad H_1 : \frac{SSE}{(N-m)} > s_Y^2$

Se il modello è <u>CORRETTO</u> non introduce ulteriore variabilità nei dati oltre a quella sperimentale

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

ES:

$$\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{n} p_i(\bar{y}_i-\hat{y}_i)^2 + \sum_{i=1}^{n} p_i(\hat{y}_i-\bar{y})^2$$

Il tutto è, rispettivamente

$$SST = SSEE + SSLF + SSR$$

ES:

$$SSE = \underline{SSEE} + \underline{SSLF} \quad ; \quad [N-m] = [N-n] + [n-m]$$

dove uno è l'errore sperimentale e l'altro l'errore del modello. Il modello risulta corretto se il contributo di "SSLF" è trascurabile rispetto a quello dell' "SSEE". Inoltre

ES: $$R^2 = \frac{\text{deviazione spiegata}}{\text{deviazione totale}} = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO}$$

Facciamo un esempio: abbiamo dei dati riguardanti processi di produzione dove vengono correlati tempo di produzione e resa, avremo

ES: $N = 10$, $n = 6$, $\bar{x} = 21,5$, $\bar{y} = 82,1$

Avremo anche come dato il modello lineare "$y = 86,4 - 0,20x$", procediamo

ES:

	SST	$=$ $SSEE$	$+$ $SSLF$	$+$ SSR
gdl	$N-1 = 9$	$N-n = 4$	$n-m = 4$	$m-1 = 1$

1. Formulazione delle ipotesi

ES: $$H_0 : \frac{[SSLF/(n-m)]}{[SSEE/(N-n)]} \leq 1 \qquad H_1 : \frac{[SSLF/(n-m)]}{[SSEE/(N-n)]} > 1$$

2. Individuazione della statistica coinvolta

ES: $$\frac{[SSLF/(n-m)]}{[SSEE/(N-n)]} \sim F_{n-m,\, N-n}$$

3. Definizione livello di significatività di rigetto α

ES: $\alpha = 0,05$ $\qquad F_{4,4,\,0.05} = 6,4$

④ Raccolta dati

ES: $SSEE = 27$, $SSLF = 659$: $\frac{[SSLF/(n-m)]}{[SSEE/(N-n)]} = 24$

⑤ Decisione

ES: $\frac{[SSLF/(n-m)]}{[SSEE/(N-n)]} > F_{4,4,0.05}$ H_0 viene rigettata

Pertanto la probabilità che sia il caso a determinare l'aumento di incertezza della risposta è inferiore al 5%

Se invece implementiamo il modello quadratico "$y = 35,8 + 5,3x - 0,13x^2$" e pertanto i parametri "$m = 3$", seguendo lo stesso test giungeremo alla conclusione che esso è corretto e H_0 <u>non</u> viene rigettata

Si può applicare lo stesso test per capire se è meglio utilizzare un modello a diverso numero "m" di parametri (vedi slide)
In generale non è detto che più parametri equivalgano a un miglior modello ma anzi, se si arriva anche al 5°, 6° ordine si potrebbero avere problemi di estrapolabilità