Impianti di elaborazione

Anno Accademico
2025/26

Rocco Lo Russo
Agostino D'Amora

Contents

Introduzione

In questo documento verranno raccolti appunti presi a lezione del corso di *Impianti di elaborazione* tenuto nell'anno 2025-26 dai professori Cotroneo e Pietrantuono, con l'aggiunta di alcuni richiami e approfondimenti. Come riferimento principale sono stati consultati i seguenti testi:

- Montgomery, D. C., & Runger, G. C. (2010). *Engineering statistics* (5th ed.). Hoboken, NJ: John Wiley & Sons
- Jain, R. (1991). *The art of computer systems performance analysis: Techniques for experimental design, measurement, simulation, and modeling*. New York, NY: John Wiley & Sons
- Trivedi, K. S., & Bobbio, A. (2017). *Reliability and availability engineering: Modeling, analysis, and applications*. New York, NY: Cambridge University Press

Part I

Performance

Chapter 1

Valutazione delle Performance

La valutazione delle performance di un sistema(o system evaluation) è un argomento importante da dover trattare. Negli anni ci sono stati vari problemi ai sistemi che sono stati ideati, poichè o valutati in modo scorretto o progettati male. Lo scopo principale della system evaluation è quallo di misurare le prestazioni di un determinato sistema, in modo che sia anche possibile confrontare i parametri con quelli di valutazione di altri sistemi (nella maniera più oggettiva possibile).

1.1 System Evaluation

Per la system evaluation, quindi, è importante impostare e delineare un metodo formale di valutazione. Ciò ci permette di ridurre gli errori legati a particolari operazioni e di poter definire una serie di "passi" da seguire per effettuare una corretta valutazione delle performance. In linea formale, la system evaluation si divide un due principali categorie:

- **Performance Analysis**: Tale valutazione presuppone che il sistema non possa avere alcun tipo di fallimento (failure-free). Il che va a valutare solo le performance legate al suo funzionamento. (Bisogna stare attenti quando si effettua Performance Analysis di non andare a valutare in alcun modo i casi di fallimento)
- **Dependability Analysis**: Tale valutazione ci permette di valutare per quanto tempo il sistema sia in grado di funzionare e quindi anche il caso di problematiche ed errori. (tra le analisi di dependability rientra anche la **reiability**, che definisce un parametro per identificare il tempo medio in cui il sistema fallisce [System Mean Time To Failure]

*Un esempio pratico per capire i concetti di **Performance Analysis** e **Dependability Analysis** è quello di un auto di formula 1. Nel caso della Performance Analysis vado solo a valutare le specifiche performance (velocità massima, tenuta in curva, aerodinamica), senza tener conto in alcun modo di qualunque tipo di fallimento; mentre nel caso della Dependability Analysis si va a valutare quanto la macchina riesca a resistere in pista (durata delle gomme, tempo effettivo di funzionamento, casi di guasti imprevisti).*

1.1.1 Passi per la valutazione di un sistema

Come introdotto precedentemente, per la valutazione "corretta" (o di buona qualità) di un sistema, è ottimale definire una serie di passi da seguire, in modo da redere il criterio

Figure 1.1: Passi da effettuare per la CM e la CP

di valutazione il quanto più formale possibile. I passi per valutare un sistema sono i seguenti:

1. Definire cosa bisogna valutare
2. Ottenere informazioni sul sistema
3. Effettuare le misurazioni
4. Analisi dei risultati
5. Trovare e valutare il corretto feedback da dare sulle considerazioni iniziali

1.1.2 Performance Analysis

Nell'analisi delle performance, quindi, si va a presupporre che il sistema di base funzioni e che quindi bisogni valutare solo il "come" tale sistema funziona (failure-free). L'analisi delle performance può essere suddivisa formalmente in due principali parti:

- **Capacity Management**: Fa in modo che le odiene risorse disponibili diano le migliori performance possibili (quindi si basa solo sulla valutazione del caso presente senza aver fatto alcuna supposizione sul carico fututo)
- **Capacity Planning**: Si assicura che ci sia un allocazione delle risorse in base ai workload successivi (si va a fare delle previsioni sul possibile carico futuro)

Nella spiegazione dei termini precedenti si è fatto riferimento ad una parola, ovvero ***workload****. Il significato e la definizione vera e formale di workload sarà data nei capitoli successivi, per il momento possiamo definire workload come: richieste effettuate da utenti verso il sistema*

La Capacity Management e la Capacity Planning sono strutturate principalmente in 5 passi che portano alla corretta esecuzione della propria valutazione [1.1]

1.1.3 Errori comuni nella System Evaluation

Quando si fa systema evaluation, solitamente, si possono commettere degli errori, che possono portare poi ad avere uno scorretto parametro di confronto o giudizio del sistema. Gli errori che principalmente vengono fatti sono:

- **Nessun obiettivo**: senza obiettivi chiari non esiste un modello universale; gli obiettivi determinano tecniche, metriche e workload da usare, e non sono mai banali.

- **Obiettivi distorti**: porsi come obiettivo. Dimostrare che il nostro sistema è migliore di un altro porta a una valutazione di parte, in cui gli analisti fanno da giudici invece che da osservatori imparziali.
- **Approccio non sistematico**: condurre la valutazione senza un metodo strutturato (obiettivi → metriche → workload → esperimenti → analisi) porta a risultati incompleti o non riproducibili.
- **Analisi senza capire il problema**: raccogliere dati e produrre grafici senza aver compreso a fondo la natura del problema significa ottenere informazioni non utili alle decisioni.
- **Metriche di performance scorrette**: scegliere metriche che non riflettono gli aspetti importanti del sistema (es. guardare solo il throughput quando è cruciale la latenza) porta a conclusioni fuorvianti.
- **Workload non rappresentativo**: utilizzare un carico artificiale che non riproduce il comportamento reale degli utenti (picchi, mix di richieste, distribuzioni) rende i risultati poco affidabili.
- **Tecnica di valutazione sbagliata**: adottare un metodo inadeguato (analisi analitica, simulazione o misurazioni reali) rispetto agli obiettivi porta a risultati poco significativi o addirittura falsati.

Dati tali errori di valutazione si comprende il motivo a cui è legato il bisogno di definire un path formale per la performance evaluation. Pertanto si va a definire una serie di passi sistematici che ci spiega come poter realizzare la syustem evaluation senza andare in contro alle problematiche descritte in precedenza. I passi da seguire sono i seguenti:

1. Definire gli obiettivi e descrivere il sistema
2. Elencare i servizi e i risultati attesi
3. Selezionare le metriche
4. Elencare i parametri
5. Selezionare i fattori da studiare
6. Scegliere la tecnica di valutazione
7. Selezionare il workload
8. Progettare gli esperimenti
9. Analizzare e interpretare i dati
10. Presentare i risultati
11. Ripetere il processo

Guardando tali passi si comprende che bisogna effettuare alcune scelte fondamentali. Le scelte che principalmente bisogna effettuare sono legate a: Tecniche di valutazione, Metriche di performance e Performance richieste

Tecniche di valutazione

Per valutare le performance di un sistema si possono utilizzare diverse tecniche che racchiudono una metodologia differente di approccio rispetto al sistema, che ci permette di poter valutare le prestazioni prescindendo dal sistema stesso (o in parte). Le tecniche principali di valutazione sono:

- **Modellazione Analitica**: Si va a ricostruire il sistema mediante un modello matematico. Tale tecnica permette di avere una soluzione in forma chiusa (utilizza formule matematiche senza dover simulare o replicare un sistema). Tali sistemi, però, fanno delle assunzioni sul sistema, che permettono la semplificazione e la modellazione matematica
- **Simulazione**: Tale tecnica cerca di combinare la modellazione analitica del sistema con il mondo reale, cercando di emulare quanto più è possibile il caso reale tramite particolati software. È una buona soluzione, poichè richiede un costo intermedio per essere effettuata; se non fosse per il tempo che bisogna dedicargli per la ricostruzione del sistema
- **Misura**: Si vanno a valutare le performance con misurazioni sul sistema reale. Tale approccio è il più costoso, sia in termini di carico che di costo, ma è il più efficiente poichè si è a contatto con il caso reale effettivo

Per selezionare la tecnica adatta al nostro caso bisogna fare una valutazione completa del sistema cercando quella che è la tecnica più adatta in base al nostro criterio di valutazione richiesto. Talvolta può essere anche possibile utilizzare più tecniche insieme. (Un esempio potrebbero essere i Twin Systems, dove vado ad effettuare modifiche prima in simulazione, e se noto miglioramento delle performance applico le stesse scelte anche al sistema reale andando ad analizzare ulteriormente le performance).

Selezione della metrica

Altro parametro importante da scegliere è la metrica da utilizzare. È importante capire il corretto modo di scegliere una metrica dato che è il concetto su cui si basa il confronto tra vari sistemi. Le metriche possono basarsi su diversi criteri, i principali possono essere classificati come:

- **Metriche per le performance**: Si vanno a valutare dei parametri di valutazione discreti (non probabilistici), dipendenti da: Tempo, Processing Rate, Consumo di risorse ecc.
- **Metriche per la Dependability**: Si va a valutare un sistema in base alla sua "efficenza", vista nel senso di probabilità di diversi eventi (guasti, eccezioni ecc.), esempi di tali metriche sono: Availability, Performability, Reliability ecc.
- **Metriche per i costi**: Si basano i criteri sui costi impiegati per l'implementazione di particolari sistemi

Uno dei criteri di selezione della metrica è quello presentato nell immagine [1.2], che ci permette di capire, in base a come reagisci il sistema, quale tipologia e quale classe di parametri andare a considerare. Fare attenzione all'immagine, essa suddivide le possibili

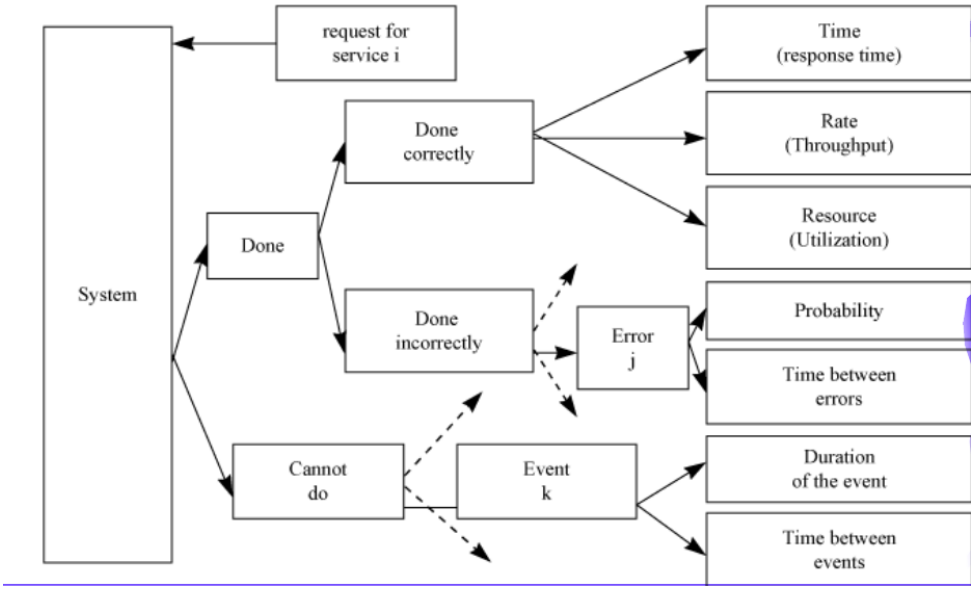


Figure 1.2: Criterio di selezione della metrica

metriche in tre macroaree, la prima è quella riguardanti le metriche deterministiche, ovvero quelle metriche che vengono valutate su casi effettivi e risultati non di natura probabilistica; a differenza delle altre 2 categorie che, avendo a che fare con errori ecc., ricadono nei casi di dover andare a valutare il sistema mediante dei valori di natura probabilistica.

Le metriche presentate, quindi, sono le seguenti:

- **Response Time**: Tale parametro ci permette di capire ogni quanto di tempo un sistema produce un risultato valido. Per la valutazione di tale parametro, però, possono essere effettuate varie tipologie di valutazione:

  - **Response time**: Misurazione basata sul tempo tra l'inizio e la fine della richiesta
  - **Reaction time:** Fine della richiesta dall'inizio del suo processamento
  - **Turnaround time:** Inizio della richiesta fino alla fine della risposta

  Generalmente la Response Time cresce all'aumentare del carico sul sistema, per tale caso è stato definito quello che verrà chiamato **Strech Factor** (permette di capire quando non saranno più usabili un certo numero di risorse)

- **Processing Rate**: Il processing rate non rappresenta altro che il throughput associato al mio sistema, e che quindi calcola la quantità di lavoro svolto da un singolo componente per unitò di tempo.

Talvolta, utilizzare sia il thoroughput che il response time può risultare ridondante, pertanto si decide di utilizzare un unico parametro, dipendente da entrambi, chiamato **potenza**, che si può calcolare come $\frac{Throughput}{ResponseTime}$. Pertanto è giusto andare a ridefinire anche i diversi punti di evoluzione della potenza, che racciudono la nostra attenzione

- **Capacità nominale:** Rappresenta il throughput massimo raggiungibile in condizioni di carico di lavoro ideali. Tuttavia, a questo livello di throughput, il tempo di risposta è generalmente troppo elevato.
- **Capacità utilizzabile:** È il throughput massimo che si può ottenere, quindi è il punto massimo dopo il quale il sistema potrebbe andare in crash (quindi ad esempio ha tempi di risposta molto lunghi ma riesce a gestire molto più carico)
- **Capacità di "ginocchio":** È il punto operativo ottimale, considerato il miglior equilibrio tra un alto throughput e un basso tempo di risposta. Questo punto corrisponde al valore massimo della metrica Power.

Oltre la potenza un ulteriore parametro utile è la **fairness**, che ci permette di capire se un aprticolare sistema distribuisce bene il carico o meno. Ciò lo veniamo a scoprire mediante il calcolo del preciso valore di fairness che è normalizzato, e quindi compreso tra 0 ed 1.

A livello formale, si definisce:

- x_i, frazione del throughput associato ad x_i
- n, numero di utenti nel sistema

A questo punto comprendo se sto usando il throughput in maniera fair, calcolando la **fairness**, che mi dice che se ogni utente ha a disposizione una porzione eguale di throughput, allora sarà 1, mentre nel caso opposto sarà vicino allo 0. Per calcolare la fairness si utilizza la seguente formula:

$$fairness = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2}$$

Con tale formula si comprende che se tutti i throughput x_i sono uguali allora => $fairness = 1$.

Chapter 2

Workload

Un Workload, nella sua definizione di base, è: *Tutti i possibili input che un sistema riceve in un certo periodo di tempo*
Tale definizione, pertanto, prescinde dalla sola valutazione delle performance. Nel caso specifico i workload che sono utilizzati nella valutazione delle performance sono detti **test workload**, tali test workload possono essere generati e categorizzati secondo particolari tecniche di utilizzo e schematizzazione

2.1 Test Workload

I test Workload sono dei normali workload utilizzati per gli studi delle performance. In generale, i test workload possono essere classificati in due particolari categorie:

- **Workload Reali**: Workload che sono caratterizzati dall'osservazione di specifici casi reali
- **Workload Sintetici**: Workload che sono caratterizzati da pochi parametri ma che cercano di replicare quelli che potrebbero essere dei workload reali

Le due classificazioni differiscono fortemente sotto molti punti di vista. Il principale tipo di workload utilizzato è quello sintetico, dato che viene caratterizzato tramite pochi parametri e può essere replicato senza dover tenere memoria di un workload reale (che richiederebbe memorie molto ingenti per essere memorizzato).

2.1.1 Addition Istruction

L'**Addition Istruction** è una tecnica di workload sintetico per i computer systems. Essa è stata utilizzata in passato per la valutazione delle performance sui vari computer e comprendeva quello di valutare la velocità e l'efficienza dell'operazione di addizione (Operazione più usata all'interno dei programmi passati).

2.1.2 Instruction Mixes

L'**Istruction Mixes** è un evoluzione dell'Addition Istruction, poichè non va a considerare solo l'operazione di addizione intera, ma anche tutte le altre istruzioni possibili. Per avere uno schema più adatto alla sintetizzazione di un workload si hanno delle specifiche di

Istruction Mixes, che non sono altro che tabelle che listano le varie **classi di istruzioni** con la loro "percentuale" di utilizzo (frequenza). In questo modo sarà possibile, andare a selezionare a caso le istruzioni, rispettando la distribuzione descritta dall'Istruction Mixes.

Disadvantages

La complessità sempre crescente delle architetture e quindi delle classi di istruzioni, non viene riflessa all'interno delle tabelle di istruction mixes, pertento risulta difficile valutare completamente la totalità dell'architettura. Oltre alla complessità delle classi, incidono anche i tempi di esecuzione, che è altamente variabile e dipendente da diversi parametri quali:

- **Modalità di indirizzamento**: Come le modalità di indirizzamento diretto o indiretto di un architettura
- **Cache Hit**: Probabilità di trovare il dato su cui si vuole lavorare in Cache
- **Pipeline Efficiency**: Se la pipeline mantiene per molto tempo l'esecuzione di un comando per ciclo di clock (ad esempio gestendo bene la questione dei salti e della branch prediction)
- **Interferenza dei dispositivi esterni durante i cicli di accesso processore-memoria**: Ad esempio concorrenza dei bus mentre si accede alla memoria per il prelievo di un dato

Oltre a problematiche di tipo puramente architetturale e strutturale, il tempo di esecuzione può variare anche in base alla forma e alle operazioni che devono essere fatte sui dati, come operazioni del tipo:

- La frequenza con cui compare lo zero come parametro
- Quante volte compare lo zero in operazioni di moltiplicazione
- il numero medio di spostamenti richiesti in un'operazione in virgola mobile
- numero di volte in cui un ramo condizionale viene eseguito

Oltretutto, le combinazioni di istruzioni delle istruction mixes non riflettono le funzionalità di indirizzamento virtuale della memoria

Considerazioni

Nonostante le varie problematiche che porta con se, l'istruction mixes, ci permette comunque di poter avere un singolo parametro di confronto tra sistemi diversi. Il parametro è un numero che esprime l'inverso del tempo di esecuzione e può essere espresso come:

- **MIPS (Millioni di Istruzioni Per Secondo)**
- **MFLOPS (Millioni di FlOaring Point istructionS)**

Però un'altro problema legato a questo valore è che stima le prestazioni solo del **processore** e quindi **non dell'intero sistema**. Il divario tra le performance reali e non dei sistemi che vengono valutati con tali modelli è fatto, quindi, solo dalla differenza dei programmi che vengono eseguiti, e quindi non è una statistica affidabile per una tipologia generale di applicazioni.

2.1.3 Kernels

I **Kernels** sono un'evoluzione dell'istruction mixes, poichè non vanno a considerare più le istruzioni nella loro singolarità, ma vanno a considerare dei gruppi di istruzioni (delle funzioni). Le funzioni, in particolare, sono dette **kernel** e sono implementare solo per il consumo della CPU, quindi nessuna prevede o fa uso dei dispositivi di I/O (almeno nelle loro versioni iniziali, dato che oggi tale classe di kernel è chiamata processing kernel). Il nome **kernel** viene dal fatto che si vuole identificare una serie di passaggi chiave che poi sono utilizzati nelle più comuni applicazioni. Ad esempio si possono utilizzare tutti i passaggi che servono per il calcolo dell'inverso di una matrice o di tutte le operazioni che vengono richieste da un algoritmo di sorting. Difatti le tecniche più utilizzare ad oggi che rispettano un modello kernel sono:

- **Sieve**: Algoritmo per trovare tutti i numeri primi fino ad N (Crivello di Eratostene)
- **Puzzle**: Algoritmi per la risoluzione del gioco del 15, le N-Regine o il Sudoku
- **Tree Searching**: Operazioni che possono essere effettuate all'interno di un'albero
- **Ackermann's Function**: Funzione matematica e molto ricorsiva che permette di valutare la reazione e la gestione di tali chiamate (funzione di Ackermann)
- **Matrix Invertion**: Va a valutare il comportamento del sistema rispetto alle operazioni che bisogna effettuare per ottenere l'inverso di una matrice NxN (anche tramite diversi metodi di calcolo)
- **Sorting**: Agglomerato di algoritmi di ordinamento differenti

Però, molti dei problemi che ritroviamo all'interno dell'istruction mixes si ripercuotono anche sull'utilizzo dei kernels, quali tutti i problemi dipendenti dall'applicazione e dalla forma dei dati e non instrinsecamente dall'architettura (dove vi è sempre la mancanza però della gestione dei dispositivi di I/O)

2.1.4 Application Benchmarks

Gli **Application Benchmarks** sono dei workload che vengono costruiti in base all'applicazione che si sta andando a testare. Quindi si vanno a verificare i casi d'uso di un'applicazione in base all'impiego che ne devo fare. Nella letteratura, in realtà, benchmark viene utilizzato come sinonimo di workload, pertanto molte volte le tecniche come quella dei kernel (test di funzioni e non di singole istruzioni), vengono visti come benchmark. Il processo che vuole valutare le performance in base ad un determinato benchmark viene detto **benchmarking**. L'application benchmark, quindi, fa riferimento e cerca di replicare un workload reale in base alla tipologia di applicazione che voglio andare a valutare, ad

esempio, se voglio valutare un servizio bancario è inutile che vada a testare delle funzioni di high performance sul processore (dato che non vengono mai fatte), ma vada a valutare la qualità di utilizzo e di controllo del database.

In generale, per confrontare due sistemi, posso utilizzare i benchmark, oltretutto, una cosa importante di caratterizzazione dei benchmark, sono le proprietà che essi devono mantenere, ovvero:

- **Representativeness**(Rappresentatività): Si garantisce che il benchmark sia rappresentativo di un workload reale che si vuole andare a valutare
- **Portability**: Il benchmark deve poter essere eseguito su piattaforme diverse, e quindi non dipende dalla macchina e dall'hardware di un sistema specifico
- **Repeatability**: Eseguendo più volte lo stesso benchmark nelle stesse condizioni, i risultati devono essere coerenti. Questo garantisce l'affidabilità e la robustezza delle misure
- **Scalability**: Il benchmark deve poter funzionare su sistemi di dimensioni diverse (ad esempio da un singolo nodo a un cluster) e adattarsi a diversi livelli di carico, senza perdere significato
- **Non-intrusiveness**: L'esecuzione del benchmark non deve alterare significativamente il comportamento del sistema misurato. Deve misurare senza influenzare in modo rilevante le prestazioni stesse
- **Easy-to-use**: Deve essere semplice da configurare, avviare ed eseguire, in modo che chiunque possa utilizzarlo senza particolari complessità tecniche
- **Easy-to-understand**: I risultati prodotti devono essere chiari e facilmente interpretabili, anche da chi non è un esperto tecnico

La cosa importante quando si sceglie un benchmark e trovare uno specifico **agreement**, e quindi un accordo su quale tipologia di applicazione andare a testare. In generale un benchmark nasce per poter comparare diverse tipologie di strutture, di componenti e di architetture, rispettando però lo specifico agreement, che oltre a dare un ordine a quello che si vuole testare, permette di comparare le diverse architatture per lo specifico compito che andranno a svolgere (sempre in linea con l'agreement).

2.1.5 Esempi di benchmark

Sieve

Algoritmo che utilizza il criterio di Eratostene per la determinazione dei numeri primi da 0 ad N, con N dato in ingresso all'algoritmo. Il suo funzionamento principale è quello di partire da tutti i numeri interi da 1 ad N, e poi eliminare tutti i multipli dei valori (in ordine), da 1 a $\sqrt{N}$. I valori che però vengono considerati sono solo quelli non eliminati.
Per esempio:
N = 20, $\sqrt{20} \approx 5$

Passo 0: $[\underline{1}, \underline{2}, \underline{3}, \underline{4}, \underline{5}, \underline{6}, \underline{7}, \underline{8}, \underline{9}, \underline{10}, \underline{11}, \underline{12}, \underline{13}, \underline{14}, \underline{15}, \underline{16}, \underline{17}, \underline{18}, \underline{19}, \underline{20}]$

Passo 1: $[1, \underline{2}, \underline{3}, \underline{4}, \underline{5}, \underline{6}, \underline{7}, \underline{8}, \underline{9}, \underline{10}, \underline{11}, \underline{12}, \underline{13}, \underline{14}, \underline{15}, \underline{16}, \underline{17}, \underline{18}, \underline{19}, \underline{20}]$

Passo 2 (eliminazione multipli di 2): $[1, \underline{2}, \underline{3}, 4, \underline{5}, 6, \underline{7}, 8, \underline{9}, 10, \underline{11}, 12, \underline{13}, 14, \underline{15}, 16, \underline{17}, 18, \underline{19}, 20]$

Passo 3 (eliminazione multipli di 3): $[1, \underline{2}, \underline{3}, 4, \underline{5}, 6, \underline{7}, 8, 9, 10, \underline{11}, 12, \underline{13}, 14, 15, 16, \underline{17}, 18, \underline{19}, 20]$

Risultato finale: $[1, \underline{2}, \underline{3}, 4, \underline{5}, 6, \underline{7}, 8, 9, , 10, \underline{11}, 12, \underline{13}, 14, 15, 16, \underline{17}, 18, \underline{19}, 20]$

Algoritmo di Ackermann

Tale algoritmo è adatto per lo studio delle chiamate ricorsive, dato che costituisce la funzione ricorsiva più pura. In generale con tale tipologia di struttura si può andare a valutare bene:

- **Tempo di esecuzione medio per le call**
- **Numero di istruzioni eseguite per call**
- **Stack space per le call**

SPEC Benchmark Suite

Corporazione non profit che ha stilato una serie di bechmark (circa una decina), che possono essere utilizzati per valutazioni di varia natura. Essi sono una base per la valutazione più generale dei sistemi a prescindere dalla loro struttura architetturale

Chapter 3

Tecniche di caratterizzazione dei workload

I workload reali sono la migliore opzione per andare a valutare le performance specifiche di un dato sistema, il problema è che mantenere tale workload richiede un ingente quantità di memoria. Quindi è nata la necessità di trovare e ricercare delle tecniche che permettessero di poter caratterizzare un workload reale e ridurre la quantità di dati da memorizzare per poterne avere una statistica quanto meno affidabile

3.1 Terminologia

Gli argomenti che saranno affrontati durante tale capitolo richiedono una conoscenza della specifica terminologia utilizzata. In generale i concetti fondamentali da conoscere inerenti al dispositivo ed al componente da testare sono:

- **DUT**(Device Under Test): Sistema che viene sottoposto ad uno specifico test (es. CPU o un processo di transazione)
- **CUT**(Component Under Test): Componente del sistema di cui si vogliono conoscere le performance (es.ALU o Unità Disco)
- **Metrica**: Metrica che si vuole andare a valutare per il CUT (es. MIPS o T/s)

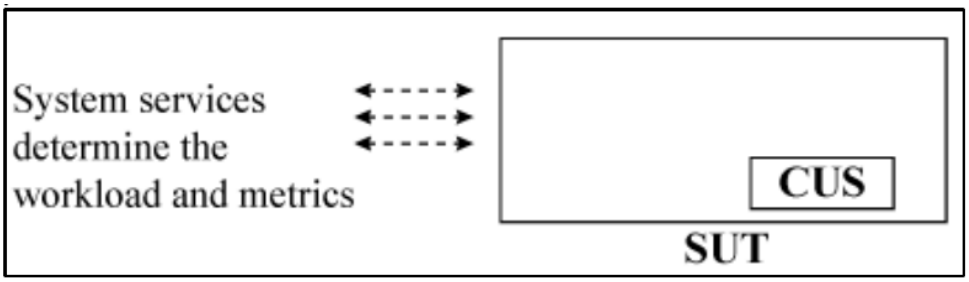


Figure 3.1: Struttura del sistema di test

Oltre la terminologia intrinseca al sistema di test bisogna definire anche la terminologia inerente ad altre entità interagenti. Quindi si definiscono i seguenti termini:

- **User**: entità che esegue le richieste di servizio

- **Workload components**(Qualitative identifiers): Componenti basilari che mi permettono di capire a livello qualitativo cosa fa un workload, quindi la natura e la struttura delle attività che vengono svolte dalle varie componenti (es. transazione in un database, una query in un motore di ricerca o Un processo o thread in un sistema operativo)

- **Workload parameters**(Quantitative Identifiers): Sono parametri quantitativi associati al workload e che quindi descrivono come le componenti si comportano. Servono principalmente per avere una misura numerica delle caratteristiche del workload (es. Arrival rate [Quante richieste al secondo], il service time [quanto tempo serve per completare un compito], Resource Usage [quante risorse vengono consumate], I/O operations [quante lettura/scritture su disco avvengono])

3.2 Workload Characterization Tecniques

Info: La **Workload Characterization** è un processo che permette di definire un workload di test di dimensione ridotta, ma che conservi tutte le caratteristiche e le proprietà (sia statiche che dinamiche), del workload reale. Ciò ne permette la replicazione e l'utilizzo in ambito di performance analysis.

C'è però da capire come sia possibile estrarre il workload sintetico dal workload reale, pertanto sono state utilizzare e definite diverse tecniche negli anni. L'obbiettivo principale di tali tecniche è quello di trovare dei parametri ridotti con cui cercare di poter descrivere il workload reale a meno di una certa quantità di informazione persa (tale quantità saraà valutata in vari modi, solitamente si utilizzerà la varianza o la devianza).

3.2.1 Averaging

La tecnica dell'**Averaging** è molto semplice, cerca di ridurre il workload in una singola istanza i cui parametri vengono valutati con la media dei parametri presenti nel workload reale. Quindi quello che si va a fare è:
Siano: $x_1, x_2, \ldots, x_n$ i valori assunti nel workload dal parametro x, allora si può approssimare tale parametro tramite la media aritmetica definita come:

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

Talvolta però non è proprio l'ideale utilizzare la media, ciò dipende fortemente dalla tipologia di dati che si ha. Solitamente si potrebbe pensare di utilizzare altre tecniche come: la mediana (permette di prendere un valore appartenente al workload più centrale), oppure il 50-percentile, la media geometrica ecc.

Specifying Dispersion

Caratterizzare un workload reale mediante un workload sintetico prevede di avere, intrinsecamente, degli errori. Ciò accade principalmente per la limitatezza che ho nei parametri che voglio andare a considerare (non posso avere memoria di tutte le istanze del workload reale, ma solo di alcune di esse). Si potrebbe pensare di andare a stimare l'errore mediante la somma di tutte le deviazioni, ovvero:

$$errore\_totale = \sum_{i=1}^{n}(x_i - \overline{x})$$

Tale rappresentazione, però, non è proprio utile, il problema principale risiede nel segno che possono avere le deviazioni (immaginiamo il caso di avere 5 e 15, la media è 10, ma l'errore, se calcolato con la formula sopra è 0 [(5-10) + (15-10)]), il che porta a rendere tale ragionamento errato. Un'altra soluzione potrebbe essere quella di andare a considerare la deianza, ovvero la somma degli errori quadratici

$$errore\_totale = \sum_{i=1}^{n}(x_1 - \overline{x})^2$$

La devianza, quindi, risolve il problema del segno delle deviazioni, ma dipende fortemente dal numero di dati. Data quindi tale dipendenza dal numero di dati della devianza, si preferisce utilizzare la **varianza campionaria**, che va a dividere la devianza per n - 1. Si va a considerare n - 1 per via dei gradi di libertà, ovvero, il numero di deviazioni linearmente indipendenti [warning successivo]

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

Warning:

Tale dimostrazione non è stata fatta in aula, per quanto sia semplice non è richiesta ai fini dell'esame ma solo per questione di conoscenza personale.

Tale formula ci fa capire perchè dividiamo per n-1 nella varianza capionaria e perchè tale divisione è giustificata come il numero di **gradi di libertà**

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

$$= \sum_{i=1}^{n}x_i - n\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = \sum_{x=1}^{n}x_i - \sum_{x=1}^{n}x_i = 0$$

Dopo tale dimostrazione si comprende che le devianze linearmente indipendenti sono n-1, dato che una sarà esprimibile come somma delle altre

Oltre al concetto di varianza, solitamente, si preferisce parlare di deviazione standard, dato che è espressa nell'unità di misura della grandezza che si sta andando a

valutare. La deviazione standard si calcola come radice quadrata della varianza campionaria, ovvero:

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

Oltre a tale valore, per rendersi conto dell'incertezza rispetto ai dati effettivi (essere in grado di confrontare un sistema piccolo con un sistema grande cercando di evitare un confronto rispetto alle grandezze di misura), è utile considerare il **coefficente di variazione**, che viene definito come:

$$COV = \frac{s}{\overline{x}}$$

Un esempio di utilizzo di tale valore è il seguente: Immaginiamo di avere due sistemi e di averne valutato il tempo di risposta e la deviazione standard associata ad ogni sistema. Si avrà il seguente scenario:

- **Sistema 1**: response time = 10 ms, dev. standard = 2 ms
- **Sistema 2**: response time = 200 ms, dev. standard = 15 ms

Se mi chiedessi quale dei due sistemi risulta più **stabile**, allora intuitivamente andrei a confrontare le dev. standard e valuterei quella minore come più stabile. Ma questo, però, non viene messo a confronto con gli andamenti medi (non guardo la larghezza della campana rispetto alla sua altezza [gaussiana]). Pertanto se calcolo i coefficenti di variazione, avrò che per il sistema 1: COV = 0,2 = 20%; mentre per il sistema 2: COV = 0,075 = 7,5%. Il che mi dimostra che il sistema 2 è più stabile rispetto al sistema 1 (completamente il contrario rispetto alla decisione iniziale).

3.2.2 Single Parameter Histogram

Il **Single Parameter Histogram** si occupa di costruire un istogramma delle occorrenze che vada a caratterizzare il **singolo parametro** considerato ed analizzato. La costruzione di un istogramma viene effettuata andando a valutare la frequenza di occorrenza di un dato valore in base ad una sua distribuzione discreta.

Più formalmente quello che si va a costruire è una funzione $f(x)$ che mi dice quante volte il parametro x assume un certo valore. Tale funzione viene costruita andando a suddividere l'intervallo di valori che il parametro può assumere in **buckets** (o bins), ovvero sottointervalli. Quindi si va a contare quante volte il parametro assume un valore compreso in un certo bucket e si va a riportare tale conteggio sull'asse delle ordinate, mentre sull'asse delle ascisse si riporta il bucket considerato.

Vi sono però dei problemi nell'utilizzare tale tecnica, utilizzare un istogramma per ogni parametro vuol dire utilizzare grandi quantità di memoria, a livello numerico: Consideriamo $n$bucket per ogni valore (intervalli di cui si deve tenere traccia), m il numero di parametri per ogni componente, ed k il numero di componenti, allora la memoria richiesta per memorizzare tale istogramma sarà:

$$Memoria = nmk$$

Ciò risulta troppo dettagliato per la reppresentazione del workload, oltretutto, tale metodo non tiene conto delle correlazioni tra i vari parametri, dato che ogni istogramma

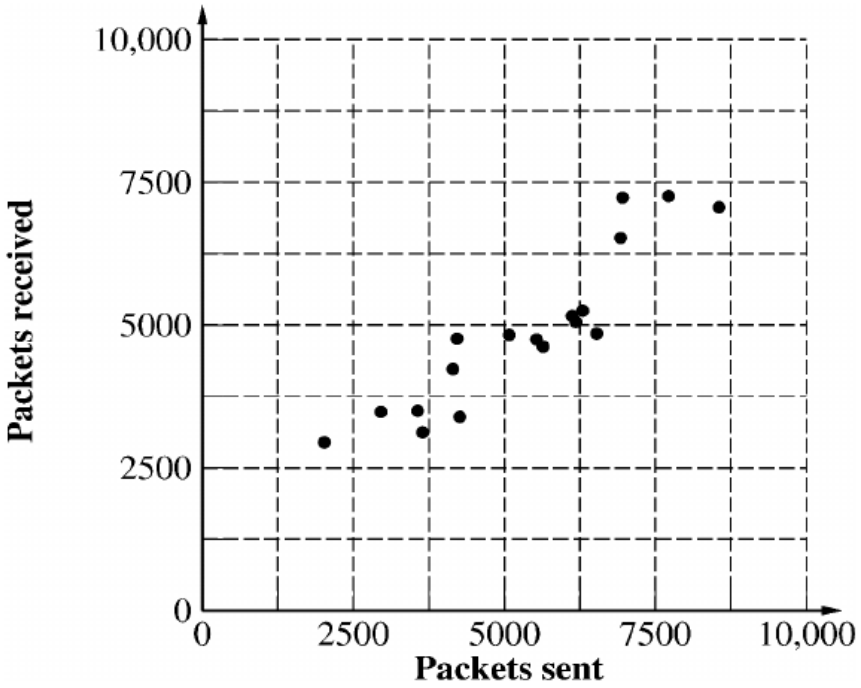
viene costruito in maniera indipendente dagli altri, ciò porta quindi a portare all'interno della descrizione del workload anche parametri che potrebbero essere deducibili da altri (ridondanza di informazioni). Questo pregiuduca anche la possibilità di selezionare i parametri da considerare mediante la varianza, dato che dati che non sono indipendenti porterebbero la stessa quantità di informazione.

3.2.3 Multiparameter Histogram

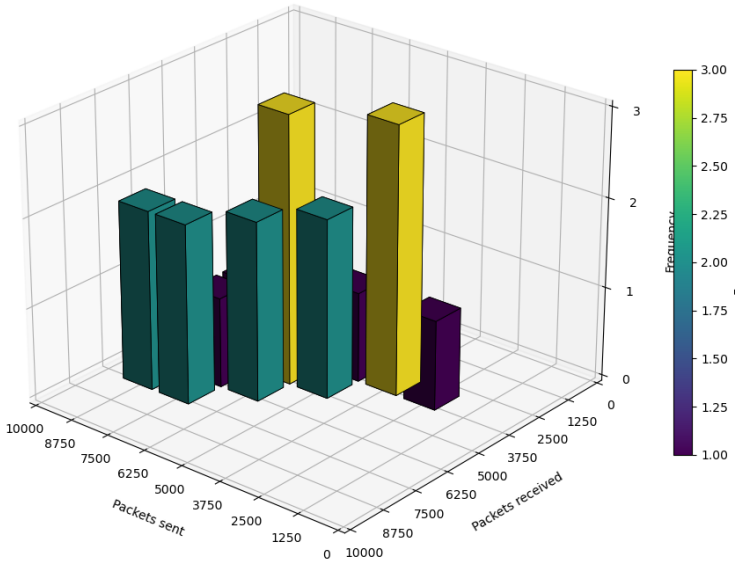
Il **Multiparameter Histogram** è una tecnica che cerca di risolvere i problemi del single parameter histogram, andando a considerare le correlazioni tra i vari parametri. Quello che si va a fare è costruire un istogramma che consideri non swolo le frequenze di un singolo parametro, ma le frequenze di n-parametri, correlate tra di loro. Per comprendere la complessità nella costruzione di un istogramma a n-variabili, vediamo un esempio con due variabili, che sarebbe quello mostrato in figura [**??**]. Per leggere tale grafico dobbiamo considerare 3 assi, che nel nostro caso sono rappresentati come:

- **Asse X**: rappresenta il range di valori del primo parametro
- **Asse Y**: rappresenta il range di valori del secondo parametro
- **Asse Z**: Viene rappresentato mediante la griglia quadrettata, ed il valore considerato è il numero di punti che sono presenti per un dato intreccio di valori.

Ciò ci aiuta a capire come si possono trovare dei pattern tra i vari parametri, data la specifica distribuzione dei parametri (in questo caso le due variabili crescono l'una rispetto all'altra). La problematica principale risiede nella quantità di parametri che bisognerebbe incatenare per trovare delle correlazioni tra le variabili, ed oltretutto, per workload molto grandi risulta complicato andare a memorizzare tale quantità di dati anche andando a considerare un filtro con la varianza.



(a) Multiparameter Histogram on 2D space



(b) Multiparameter Histogram on 3D space

Figure 3.2: Esempi di istogrammi multiparametrici

3.2.4 Principal Component Analysis (PCA)

Un modo utilizzato per ridurre il numero di parametri con cui andare a rappresentare le istanze del workload è la **Principal Component Analysis (PCA)**. Tale tecnica ha come compito quello di trasformare l'insieme di istanze in altre istanze, le nuove istanze vengono definite sulla base delle componenti principali (che differiscono dai parametri reali), poichè nel nuovo spazio, tali parametri sono tutti linearmente indipendenti, e non ci sono correlazioni. Il funzionamento matematico della PCA è basato sull'effettuazione di una media pesata per ogni parametro. Precisamente:
Dati i parametri $x_1, x_2, \ldots, x_N$, si vuole trovare un nuovo insieme di parametri $y_1, y_2, \ldots, y_N$, voglio però che l'insieme di parametri y_i sia linearmente indipendente, e che la varianza di ogni parametro y_i sia massimizzata. Di base per vedere come andare a costruire la matrice di trasformazione della PCA, si dovrebbe calcolare la matrice di covarianza, una volta calcolata si valutano gli autovettori di tale matrice, che costruiranno poi la matrice di trasformazione finale. Gli autovettori, quindi, rappresentano le direzioni principali e sono disposti in maniera che la prima componente principale sia quella con la varianza maggiore. Difatto si sta andando a fare una media pesata dei parametri per costruire il valore della nuova componente principale. Precisamente:

$$y_j = \sum_{i=1}^{N} w_{ij} x_i$$

Tale formula va letta in questo modo:

- y_j: rappresenta la j-esima componente principale
- x_i: rappresenta il valore del parametro i della componente da trasformare
- w_{ij}: rappresenta il peso associato al parametro i per la j-esima istanza

Per effettuare la PCA bisogna seguire i seguenti passi:

1. Andare a calcolare la matrice di covarianza dei dati (prima si potrebbero effettuare anche operazioni di normalizzazione)
2. Andare a calcolare gli autovettori e gli autovalori della matrice di covarianza
3. Costruire la matrice di trasformazione mediante gli autovettori ordinati secondo gli autovalori in maniera decrescente (gli autovalori portano con loro la quantità di varianza, quindi ordinando gli autovettori si avrà uno spazio in cui il primo parametro copre la maggior varianza [utile per la selezione di un minor numero di parametri])

In maniera più compatta, quindi, effettuata la PCA, si avrà che:

- I nuovi parametri y sono calcolabili come combinazioni lineari dei parametri x
- I parametri y sono linearmente indipendenti, dato che il prodotto interno tra due parametri y è 0
- Il nuovo set di parametri y è ordinato in maniera tale che la varianza del primo parametro è maggiore della varianza del secondo e così via

Z-Score Normalization

La **z-score normalization** è una tecnica che permette di andare a normalizzare i dati, in maniera tale che ogni parametro abbia media 0 e deviazione standard 1. Tale tecnica è utile per poter effettuare la PCA, dato che si vuole evitare che parametri con range di valori molto diversi tra di loro possano influenzare in maniera sproporzionata il risultato finale. La formula per effettuare tale normalizzazione è la seguente:

$$x'_s = \frac{x_s - \overline{x}_s}{s_{x_s}}$$

Tale operazione permette di poter confrontare i parametri con un distribuzione normale, il che, quindi, andrà a rappresentare i dati come distanza da 0 e rappresentato secondo la deviazione standard s. Ciò ci permette anche di poter confrontare i dati tra di loro, senza andarea a considerare il range di valori che possono assumere.

3.2.5 Clustering

Mediante l'utilizzo della PCA si è andata ad effettuare la riduzione della quantità di parametri rappresentativi di un istanza (o componente) del workload. Per ridurre ancora di più la quantità di dati di rappresentazione del workload, si può andare ad effettuare una riduzione della quantità di istanze stesse. Per effettuare tale riduzione si fa utilizzo del **Clustering**, che è una tecnica di apprendimento non supervisionato che permette di andare a raggruppare le istanze in base alla loro similarità. Ciò richiede anche che la rappresentazione delle istanze sia adeguata per trovare degli specifici agglomerati di dati. (Per comprendere meglio, guardare la figura [3.3], se i dati non fossero ben divisi non potrei creare i cluster per bene dato che avrei molta confusione). Una volta raggruppate le istanze, si può andare a rappresentare ogni cluster mediante un'unica istanza rappresentativa (solitamente la media dei parametri delle istanze che compongono il cluster). In questo modo si va a ridurre la quantità di istanze da memorizzare, andando a perdere però una certa quantità di informazione (che può essere valutata mediante la varianza o la devianza).

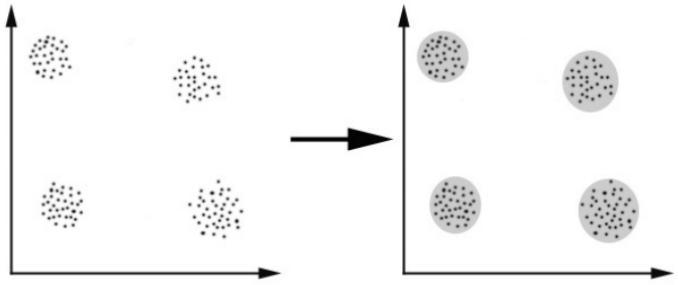


Figure 3.3: Esempio di clustering

Per effettuare il clustering, generalmente, si eseguono i seguenti passaggi:

1. Prendere un insieme di istanze appartenenti al workload (o ad una sua trasformazione mediante PCA)
2. Scegliere i parametri del workload da considerare per effettuare la clusterizzazione (omettere i parametri che portano una quantità di varianza bassa)

3. Selezionare una metrica di distanza (utile per la valutazione della similarità tra le istanze)
4. Trattazione degli outliers (dati che sono molto distanti dagli altri, e che potrebbero falsare il risultato finale)
5. Data scaling (utile per evitare che parametri con range di valori molto diversi tra di loro possano influenzare in maniera sproporzionata il risultato finale)
6. Effettuare il clustering (utilizzando uno degli algoritmi di clustering esistenti)
7. Interpretazione dei risultati (andare a valutare la qualità del clustering effettuato)
8. Cambiare i parametri e/o il numero di cluster e ripetere i passi dal 3 al 7
9. Selezionare un componente rappresentativo per ogni cluster

Warning: Negli appunti si fa utilizzo del termine **istanza**, che per il seguente ambito è sinonimo di **componente**. Tale termine viene utilizzato per indicare un'entità del workload che viene caratterizzata da un insieme di parametri.

Campionamento delle componenti

Il **Campionamento delle istanze** è una tecnica che va a selezionare un sottoinsieme di componenti del workload prima di proseguire nell'algoritmo di clustering. Ciò è dovuto al fatto che gli algoritmi di clustering sono computazionalmente costosi, e pertanto si cerca di ridurre la quantità di dati da considerare. La tecnica di selezione delle componenti da andare a considerare può essere di varia natura. Precisamente si studiano (per il seguente corso), le seguenti tecniche:

- **Random Sampling**: Si va a selezionare un sottoinsieme di componenti in maniera casuale
- **Resource Consuption Based Sampling**: Si va a selezionare un sottoinsieme di componenti in base alla quantità di risorse che esse consumano (si vanno a selezionare le componenti che consumano più risorse)

Per capire se ho effettuato un buon campionamento, una volta effettuato il clustering vado a vedere, sul workload intero, se posso associare le componenti mancanti ai cluster trovati. Se il numero di componenti non assignabili è alto allora è indice di un cattivo campionamento.

Selezione dei parametri

Vado a selezionare un sottoinsieme di parametri rtappresentativi delle componenti. Tali parametri devono essere scelti in maniera tale che siano in grado di rappresentare la maggior quantità di varianza possibile o rispetto all'impatto sulle performance. Per fare ciò si può utilizzare la PCA, andando a selezionare i parametri che portano più varianza. Facendo in questo modo introduco una quantità di varianza persa a discapito, però, della riduzione della dimensionalità del problema, e di conseguenza, della riduzione del tempo speso per effettuare la clusterizzazione.

Metrica di distanza

Una metrica di distanza mi permette di calcolare la ditanza tra varie componenti in spazi n-dimensionali, dove n è il numero di parametri considerati. Le metriche più utilizzate sono:

1. **Distanza Euclidea**: La distanza euclidea è la distanza più intuitiva, ed è definita come:
$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{n}(X_{ik} - X_{jk})^2}$$
Dove X_i e X_j sono due punti nello spazio n-dimensionale, e X_{ik} e X_{jk} sono le loro coordinate(tale distanza è applicabile solo in spazi reali, quindi a parametri reali. Oltretutto la distanza euclidea utilizza il principio introdotto dal teorema di Pitagora [il teorema di Pitagora è un eccezione della formula precedente, precisamente solo per spazi bidimensionali]). Nonostante la distanza euclidea sia applicabile solo ad insiemi di parametri reali, essa è la più utilizzata, sopratutto per la sua similarità con la deviazione standard (guardare l'espessione matematica delle due grandezze)

2. **Distanza Euclidea Pesata**: La distanza euclidea pesata è una variante della distanza euclidea che permette di andare a pesare i vari parametri in base alla loro importanza. La formula è la seguente:
$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{n} a_k(X_{ik} - X_{jk})^2}$$
Dove a_k è il peso associato al parametro k. Tale distanza è utile quando si vuole dare più importanza ad alcuni parametri rispetto ad altri. Tale metrica è utilizzata spesso o quando i parametri non sono stati scalati o quando i parametri hanno livelli di importanza diversi.

3. **Distanza Euclidea Quadrata**: La distanza euclidea quadrata è una variante della distanza euclidea che evita di dover calcolare la radice quadrata, dato che tale operazione è computazionalmente costosa. La formula è la seguente:
$$d(X_i, X_j) = \sum_{k=1}^{n}(X_{ik} - X_{jk})^2$$
Tale distanza è utile quando si vuole evitare di effettuare la radice quadrata, dato che non cambia l'ordinamento delle distanze tra i punti. Tale metrica cerca di enfatizzare i valori più distanti (rendend distanze grandi ancora più grandi), e favorendo quelle piccole (rendendo distanze piccole ancora più piccole).

4. **Distanza di Chi-Quadrato (Chi-square distance)**: La distanza di Chi-Quadrato (Chi-square) è una metrica che viene utilizzata principalmente per dati categoriali. La formula è la seguente:
$$d(X_i, X_j) = \sum_{k=1}^{n} \frac{(X_{ik} - X_{jk})^2}{X_{ik} + X_{jk}}$$

Tale distanza è utile quando si vuole dare più importanza ai parametri che hanno valori più piccoli, dato che la somma al denominatore rende tale distanza più sensibile a piccole differenze tra i valori piccoli. Tale metrica, però, può essere applicata solo a dati che sono compatibili numericamente o che almeno appartengono allo stesso ordine di grandezza (se si avessero valori che hanno differenze grandi, il peso associato alla loro differenza sarebbe molto piccolo, e quindi non verrebbe considerata).

5. **Distanza di Hamming**: La distanza di Hamming è una metrica che viene utilizzata principalmente per dati categoriali. La formula è la seguente:

$$d(X_i, X_j) = \sum_{k=1}^{n} \delta(X_{ik}, X_{jk})$$

Dove $\delta(X_{ik}, X_{jk})$ è una funzione che vale 1 se $X_{ik} \neq X_{jk}$, altrimenti vale 0. Tale distanza è utile quando si vuole confrontare dati categoriali, dato che conta il numero di differenze tra i due vettori. Non è utile per dati numerici "continui", il che la rende difficilmente utilizzabile.

Info: Tutto le tecniche precedentemente presentate rispettano la definizione base di metrica. Una metrica si può definire formalmente come:

Dato uno spazio S, una funzione $d : S \times S \to \mathbb{R}$ è una metrica se per ogni $X_i, X_j, X_k \in S$ rispetta le seguenti proprietà:

- $d(X_i, X_j) \geq 0$ (Non negatività)
- $d(X_i, X_j) = 0 \iff X_i = X_j$ (Identità degli indiscernibili)
- $d(X_i, X_j) = d(X_j, X_i)$ (Simmetria)
- $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$ (Disuguaglianza triangolare)

Trattazione degli outliers

Gli **outliers** sono componenti del workload che sono molto distanti dalle altre componenti. La presenza di outliers può falsare il risultato finale del clustering, dato che tali componenti potrebbero essere considerate come cluster a se stanti, o potrebbero influenzare la posizione dei centroidi. Pertanto è utile andare a trattare gli outliers prima di effettuare il clustering. Le tecniche più utilizzate per trattare gli outliers sono:

- **Rimozione degli outliers**: Si va a rimuovere le componenti che sono molto distanti dalle altre componenti. Tale tecnica è utile quando si vuole evitare che gli outliers influenzino il risultato finale del clustering. Il problema principale di tale tecnica è che si potrebbe andare a rimuovere componenti che sono effettivamente parte del workload, e che potrebbero essere importanti per la caratterizzazione del workload stesso.

- **Trasformazione degli outliers**: Si va a trasformare le componenti che sono molto distanti dalle altre componenti. Tale tecnica è utile quando si vuole evitare che gli outliers influenzino il risultato finale del clustering, senza però rimuovere componenti che potrebbero essere importanti per la caratterizzazione del workload stesso. La trasformazione più comune è la normalizzazione

Data Scaling

Il **Data Scaling** è una tecnica che permette di andare a scalare i parametri delle componenti, in maniera tale che tutti i parametri abbiano lo stesso peso nella valutazione della distanza tra le componenti, quindi i dati che vengono trattati sono trattati secondo una distribuzione normale e quindi confrontabili anche guardando le unità di misura. Ci sono varie tecniche di normalizzazione che si possono applicare ai dati. Precisamente quelle utilizzate in questo corso sono:

- **Normalizzare ad una distribuzione normale (Z-score normalization)**: Per il calcolo di tale normalezzazione, si utilizza la formula:

  $$x_{ik}^{0} = \frac{x_{ik} - \overline{x_k}}{s_k}$$

  Dove x_{ik} è il valore del parametro x, $\overline{x_k}$ è la media del parametro x e s_{x_s} è la deviazione standard del parametro x. Tale tecnica permette di avere una distribuzione normale con media 0 e deviazione standard 1. Tale tecnica è molto efficiente quando i dati seguono una distribuzione normale.

- **Pesata**: Si va a pesare i parametri in base alla loro importanza. La formula è la seguente:

  $$x'_s = w_k * x_s$$

  Dove w_k è il peso associato al parametro x. Tale tecnica è utile quando si vuole dare più importanza ad alcuni parametri rispetto ad altri. w_k è un valore di peso relativo, e può essere calcolato anche mediante la formula: $w_k = \frac{1}{s_k}$

- **Range Normalizzation**: La range normalizzation (o in altri campi assimilabile all'equalizzazione dell'istogramma) è una tecnica che permette di andare a scalare i parametri in un intervallo specifico. Per comprendere meglio tale tecnica, si può vedere la formula come:

  $$x'_{ik} = \frac{x_{ik} - min(x_k)}{max(x_k) - min(x_k)}$$

  Il problema associato a tale tecnica di scaling è che è molto e fortemente influenzata dagli outliers, dato che tali valori andranno a definire il range di valori

- **Percentile Normalization**: La percentile normalization è una tecnica che permette di andare a scalare i parametri in base ai percentili. Tale tecnica è utile quando si vuole evitare che gli outliers influenzino il risultato finale del clustering. La formula è la seguente:

  $$x'_{ik} = \frac{x_{ik} - x_{2.5k}}{x_{97.5k} - x_{2.5k}}$$

  Tale formula cerca di andare a valutare il "massimo" e "minimo" della Range Normalization utilizzando i percentili, rendendo il sistema più resistente agli outliers.

3.2.6 Agglomerative Hierarchical Clustering

Le tecniche di clustering sono diverse e possono suddividersi in due macro-categorie: quelle **gerarchiche** e quelle **Non Gerarchiche**. Le tecniche gerarchiche permettono di costruire un sistema di cluster annidati tra loro (in base al livello del cluster), mentre le tecniche non gerarchiche cerca di non annidare in alcun modo i cluster. Nei nostri casi di studio la categoria di tecniche di clustering che ci interessa sono solo i clustering gerarchici agglomerativi (le tecniche di clustering gerarchico possono essere anche divisive, in base alla struttura della tecnica con cui si costruisce il sistema di cluster [3.4]).

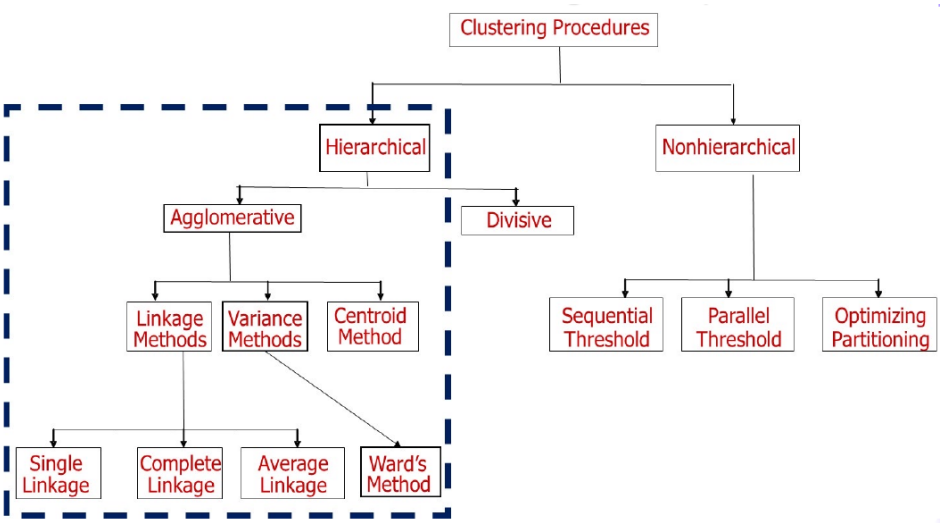


Figure 3.4: Esempio di clustering gerarchico

Fatta una paronamica generale sul clustering, ai fini del corso, andremo a concentrarci sul clustering gerarchico agglomerativo. Partendo dal considerare il **clustering gerarchico**, esso può essere di due tipi:

- **Agglomerativo**: Si parte dall'avere un cluster per ogni componente, e poi si prosegue agglomerando i cluster fino a che non si ottiene un unico cluster
- **Divisivo**: Si parte da un cluster in cui sono contenutre tutte le compoenti e si prosegue dividendo i vari cluster secondo particolari metodi.

Il clustering **gerarchico agglomerativo** è una tipologia di clustering che parte dall'avere tanti cluster (uno per ogni componente), e poi si occupa di andare ad agglomerare tali cluster in base a particolari metriche. Tali metriche si basano su concetti differenti di definizione di similarità, e per ognuno ci sono varie considerazioni da fare. Le tecniche principali di clustering agglomerativo sono:

- **Linkage Methods**: Vado a collegare i vari cluster in base alla metrica di distanza che vige tra di loro. Dato che non posso valutare la distanza di tutti i punti si può andare a considerare una metrica differente basandosi su diverse e particolari componenti del cluster:
  - **Single Linkage**: Vado a definire la metrica di distanza tra due cluster come la più piccola distanza tra due punti dei cluster (la distanza dei punti più vicini tra due cluster) [3.5a]

  - **Complete Linkage**: Vado a definire la metrica di distanza come la distanza maggiore tra due punti di un cluster [3.5b]
  - **Average Linkage**: Vado a definire la metrica di distanza come la distanza media dei punti tra due cluster [3.5c]

- **Centroid Methods**: Molto simile ai metodi linkage, ma al postro di considerare le istanze direttamente appartenenti al cluster, si va a considerare il centroide del cluster (la media dei punti appartenenti al cluster) e si va a valutare la distanza tra i centroidi dei vari cluster (il centroide non è detto che faccia parte degli elementi del cluster)

- **Variance Methods**: Tale metodo cerca di minimizzare la varianza intra-cluster e cerca di massimizzare la varianza inter-cluster. L'utilizzo effettivo di tale metodo ricade nell'utilizzo del metodo di Ward, che cerca di minimizzare la somma delle varianze all'interno di ogni cluster. A livello formale quello che si va a fare è calcolare la distranza tra due cluster mediante l'utilizzo della distanza euclidea quadrata tra i centroidi dei due cluster, pesata per il numero di elementi che compongono i due cluster. La formula è la seguente: Dato il cluster P ed il cluster Q, il calcolo della distanza mediante il **metodo di Ward** è:

$$d(P,Q) = 2\frac{|P||Q|}{|P|+|Q|}||(\overline{x}_P, \overline{x}_Q)||^2$$

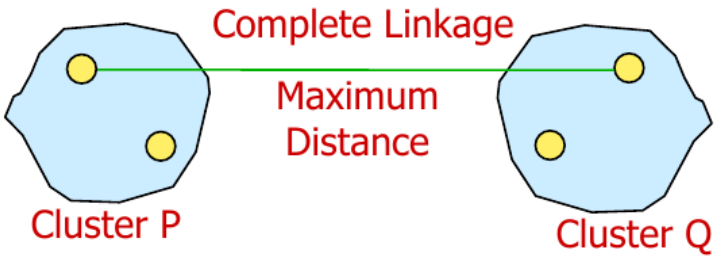
  Dove $|P|$ e $|Q|$ sono il numero di elementi che compongono i cluster P e Q, mentre $\overline{x}_P$ e $\overline{x}_Q$ sono i centroidi dei cluster P e Q.

In generale, quindi, un algoritmo di clustering gerarchico agglomerativo segue i seguenti passi:

1. Calcolo della matrice di distanza tra tutte le componenti date in input
2. Impostare ogni componente come un cluster a se stante
3. Ripetere fino a che non rimane un solo cluster:
   - Trovare i due cluster più vicini secondo la metrica di distanza scelta ed unirli in un unico cluster
   - Aggiornare la matrice di distanza per tenere conto del nuovo cluster
   - Ripeti fino a che non rimane un solo cluster

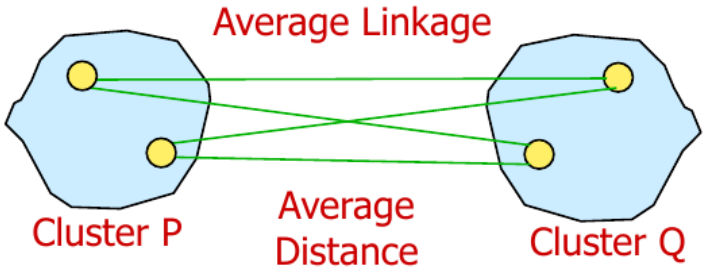
$$\min d(x,y), \qquad for\ x\ in\ P, y\ in\ Q$$

(a) Single Linkage



$$d(P,Q) = \max d(x,y), \qquad for\ x\ in\ P, y\ in\ Q$$

(b) Complete Linkage



$$d(P,Q) = mean\ d(x,y), \qquad for\ x\ in\ P, y\ in\ Q$$

(c) Average Linkage

Figure 3.5: Tipi di metriche per il Linkage Clustering

La parte in cui incidono di più le decisioni dei criteri di distanza (linkage, centroid, variance) è nella parte di calcolo della distanza dell'algoritmo. Ad ogni possibile metodo utilizzato, viene associato un diverso modo di calcolare la matrice di distanza, e di conseguenza, si avrà una diversa struttura del clustering finale. La scelta su quali cluster agglomerare rimane la stessa, ovvero si agglomerano i cluster con la distanza minore tra loro. (questo accade per tutti i metodi, tale metodo di scelta di agglomerazione prescinde dalla metrica utilizzata per calcolare la distanza tra i cluster [che invece contiene una serie di criteri di scelta precedentemente discussi]).

La parte più importante di un clustering gerarchico è il fatto che poi alla fine si va a ricavare un legame tra i vari cluster, che può essere rappresentato mediante un dendogramma. Un dendogramma è un grafo aciclico orientato, dove il nodo rappresenta il cluster "finale" (intero dataset), le foglie sono i singoli elementi, ed i nodi, sono i cluster che nel processo sono stati agglomerati. La comodità di avere un dendogramma è che si può andare a "tagliare" il grafo ad un certo livello, ottenendo così un numero di cluster desiderato. (guardare figura [3.6]). Quado vado a scegliere il livello a cui tagliare il dendogramma, mediante la larghezza del dendogramma, riesco a capire se due cluster se hanno una varianza più bassa o più alta. Più sono vicino alla radice, meno sono i cluster, ma la similarità intra-cluster si riduce. Mentre, più sono vicino alle foglie, più sono i cluster, e la similarità intra-cluster aumenta.

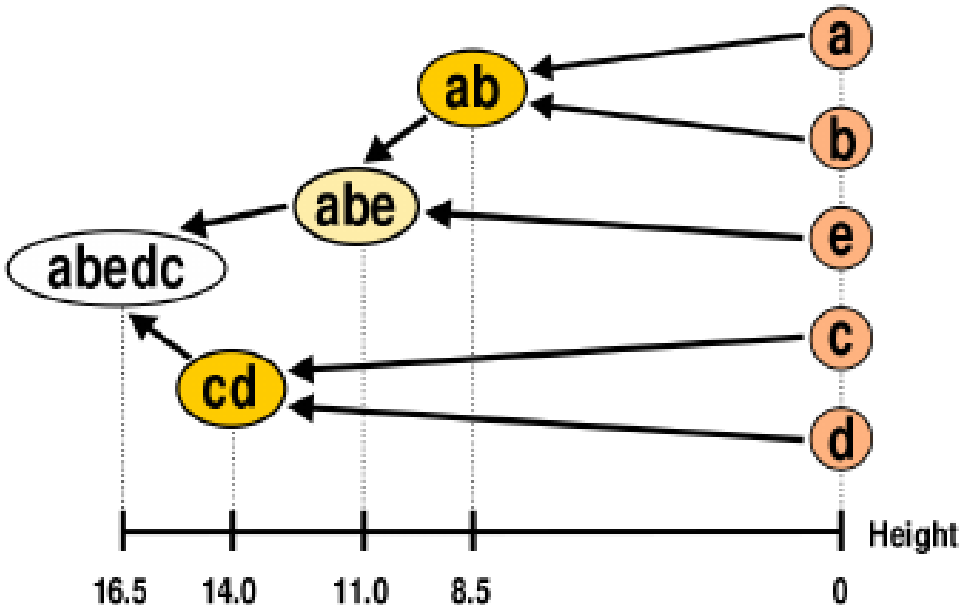


Figure 3.6: Esempio di dendogramma

Warning: In base al metodo utilizzato si avrà una clusterizzazione differente. Quindi la scelta del metodo di calcolo della distanza è fondamentale per ottenere un buon risultato finale.

Interpretazione del clustering

Alla fine del clustering, tutte le componenti facenti parte dell'insieme di dati in ingresso sono associati ad un singolo cluster. In generale posso eliminare i cluster poco influenti (che occupano poche risorse) e che hanno una dimensione molto piccola (poche componenti) [Attenzione, le regole precedentemente presentate per la selezione di un cluster non possono essere divisie, devono essere entrambe vere, altrimenti si potrebbe ricadere in errori di valutazione]. Ottenuti i cluster, si cerca di interpretarli in maniera funzionale (in baso alla tipologia di dato che si sta andando a valutare).

Una volta ottenuti e valutati i cluster si va a trovare un singolo valore rappresentativo per ogni cluster, ciò ci permetterà di poter rappresentare il workload mediante un minor numero di componenti.

I cluster tra loro devono essere quanto più dissimili possibile, mentre le componenti all'interno di un cluster devono essere quanto più simili possibile tra di loro. In statistica possiamo definire tale concetto parlando di varianza; più precisamente si enuncia che: *La varianza intra-cluster dev'essere quanto più piccola possibile, mentre la varianza inter-cluster dev'essere quanto più grande possibile.* L'unico problema legato all'utilizzo della varianza è che non rispetta la disuguaglianza triangolare, e pertanto non può essere utilizzata come metrica di distanza per il clustering. Pertanto si utilizza la devianza come metrica per valutare la bontà del clustering effettuato, questo perchè rispetta la disuguaglianza triangolare:

$$devianza\_totale = devianza\_intra\_cluster + devianza\_inter\_cluster$$

In maniera più formale, ci è richiesto un metodo per valutare tali devianze. Per valutare la devianza intra-cluster si utilizza la seguente formula:

Dati n oggetti in uno spazio vettoriale d-dimensionale
$X_i = (x_{i1}, x_{i2}, \ldots, x_{id})$, $\forall i \in [1, n]$, si definisce devianza intra-cluster:

$$intra\_cluster\_deviance = \sum_{k=1}^{K} \sum_{i=1}^{n_k} ||X_i - \overline{X}_k||^2$$

Dove:

- $\overline{X}_k = \frac{1}{n_k} \sum_{C(i)=k} X_i$: Centroide del cluster k
- n_k: Numero di oggetti nel cluster k
- $||X_i - X_j||^2 = \sum_{p=1}^{d} (x_{ip} - x_{jp})^2$
- K: Numero di cluster che si vanno a considerare

Mentre per valutare la devianza inter-cluster si utilizza la seguente formula:

Dati i K cluster e $\overline{X}$ media di tutti gli elementi. Si definisce la devianza inter-cluster come:

$$inter\_cluster\_deviance = \sum_{k=1}^{K} n_k ||\overline{X}_k - \overline{X}||^2$$

Dove (formalmente):

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$: Media di tutte le componenti date in ingresso all'algoritmo
- n: il numero di componenti dati in ingresso all'algoritmo

Si può dimostrare che la varianza intra-cluster e la varianza inter-cluster rispettano la disuguaglianza triangolare (vedi formula [3.2.6])

3.2.7 Valutazione della devianza persa

Quando vado ad effettuare la sintetizzazione del workload mediante l'utilizzo di PCA + clustering, devo valutare la quantità di devianza persa. La devianza persa è la quantità di devianza che non viene più rappresentata dalle componenti selezionate per rappresentare il workload. Per calcolarla, vado a valutare i seguenti valori:

- **Varianza persa PCA**: La devianza portata dai parametri che non sono stati selezionati dopo la PCA
- **Varianza persa Clustering**: La devianza intra-cluster, ovvero la devianza che non viene più rappresentata dalle componenti selezionate per rappresentare il workload

Per il calcolo della devianza totale persa vado a considerare la seguente formula:

$$devianza\_persa = devianza\_persa\_PCA + (devianza\_guadagnata\_PCA * devianza\_intra\_cluster)$$

Oppure in maniera identica posso utilizzare la formula:

$$devianza\_persa = 1 - (devianza\_guadagnata\_PCA * devianza\_inter\_cluster)$$

In questo secondo caso sono andato a valutare prima la quantità effettiva di varianza "guadagnata" tra i due processi e poi ho sottratto ad 1 (essendo tutto espresso in percentuali).

Esempio:
Una volta effettuata la PCA con la selezione dei parametri sul mio workload ho che la devianza considerata è del 95%.
Poi facendo il clustering, scopro che la mia devianza intra-cluster è pari al 10% e di conseguenza quella inter-cluster è pari al 95%.
Di conseguenza posso calcolare la devianza totale persa come:

$$devianza\_persa = devianza\_persa\_PCA + (devianza\_guadagnata\_PCA * devianza\_intra\_cluster) =$$

$$= 0.05 + (0.95 * 0.10) = 0.145 = 14.5\%$$

Se vado a calcolare tale valore con l'altra formula, ottengo:

$$devianza\_persa = 1 - (devianza\_guadagnata\_PCA * devianza\_inter\_cluster) =$$

$$= 1 - (0.95 * 0.90) = 1 - 0.855 = 0.145 = 14.5\%$$

Notiamo come i due valori siano esarramente uguali.

Info: *Tale dimostrazione è stata aggiunta solo a scopo didattico e non è richiesta ai fini dell'esame* Il collegamento tra i due calcoli è molto semplice e risulta dimostrabile. Per dimostrare tale collegamento si può procedere nel seguente modo (consideriamo la deviazione guadagnata dalla pca denotata come dgp, mentre la deviazione persa dalla pca dpp, mentre poi vado a considerare la deviazione intra-cluster come dic, mentre la deviazione inter-cluster dec). Posso dimostrare tale legame nel seguente modo:

$$deviazione\_persa = dpp + (dgp * dic) = 1 - dgp + (dgp * dic) =$$

$$= 1 - (1 - dic) * dgp = 1 - dec * dgp$$

Come possiamo vedere alla fine siamo arrivati alla seconda forma di calcolo della devianza persa totale

3.2.8 Modelli di Markov

Un **modello di Markov** è una tipologia di modello in cui uno stato successivo o prossimo, dipende solamente dallo stato corrente. Esso può essere rappresentato, quindi, mediante l'utilizzo di un grafo. Un modello di markov, quindi, ha un modo di decidere il prossimo stato (che dipenderà solo da quello attuale). La possibilità di raggiungere un dato stato considerato lo stato precendente viene chiamata **Probabilità di transizione**, mentre la tabella con tutte le probabilità di raggiungere uno stato dato lo stato corrente, è detta **Matrice delle transizioni** (in inglese sarebbe matrice di transizione, ma tale definizione collide con le conoscenze precedenti di algebra e geometria). Tale tipologia di modello permette di trovare dei "pattern" o path, che vengono seguiti durante "l'utilizzo" di un applicativo (workload). Permettendo di andare a stressare il sistema su quelle operazioni che si sanno sono più ripetute (probabili). La stima della tabella delle transizioni è fatta tramite un'approccio di tipo frequentista

Chapter 4

Caratterizzazione di Dati Misurati e Statistica Inferenziale

Resi noti una serie di dati misurati, quello che si vuole andare a fare è trovare un modello statistico che mi permetta di avere un'approssimazione valida di tali dati. Ci sono varie tipologie di tecniche e di approcci. Per grandi quantità di dati si farà un utilizzo massiccio della **Statistica inferenziale**, che comprende tutta una serie di metodi che permettono di andare a stimare un modello quantò più opportuno possibile ai dati che si stanno andando ad analizzare

4.1 Media, Mediana e Moda

Dato un insieme di dati, si possono andare a valutare 3 valori principali (e scalari), che ci permettono di approssimare (in maniera grossolana), la nostra distribuzione di dati (rappresentabile, anche, mediante un istogrammma). I valori a cui si fa riferimento sono:

- **Media**: Media statistica di tutti i valori che si sta andando a considerare. Se si ha un set di valoi $X = (x_1, x_2, \ldots, x_n)$, allora definiamo come media:

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

  La media, però, non tiene conto dell'asimmetria dei dati (skewness), quindi se ad esempio un sistema ha dei tempi di ripsosta sempre stabili, e poi, un singolo caso di tempo di risposta lungo, potrebbe avere la stessa media di un sistema che ha mediamente tempi di risposta più lunghi (molto sensibile ad eventuali comportamenti limite)

- **Mediana**: La mediana di un set di valori $X = (x_1, x_2, \ldots, x_n)$ è data dalla selezione del valore centrale del set di valori ordinati. Per capire, si ordina il set di valori X e poi si seleziona l'elemento presente in $\left\lfloor \frac{n}{2} \right\rfloor$. Ciò però presenta un problema, se il numrero di valori è dispari allora io seleziono l'unica posizione centrale presente (es. Se ho 3 elementi seleziono 1), mentre se ho un valore pari devo trovare un modo con cui scegliere quale elemento considerare, pertanto, essendo due i valori

centrali, se ne fa la media (es. se ho 4 elementi, farò la media del valore in posizione 1 ed in posizione 2). La mediana a differenza della media viene presa direttamente dai valori reali e non calcolata considerando tutti i valori, ciò gli permette di essere più resistente agli outlier e sopratutto a distribuzioni asimettriche (resistente alla skewness)

- **Moda**: La moda rappresenta il valore più probabile all'interno di una distribuzione (quello presentato più volte). Di conseguenza tiene conto del picco presente nell'istrogramma. Il problema della moda è che può non esistere (caso di distribuzioni uniformi), oppure può assumere più di un valore (immaginare una distribuzione bi-gaussiana). Nonostante le considerazioni precedenti, la media è totalmente immune agli outlier (vedo solo il più probabile), e mi permette di evitare anche le ambiguità di una particolare distribuzione

Info: Nella descrizione dei parametri precedenti si è parlato di asimettrie dei dati (skewness dei dati), tale valore per piccole quantità di campioni può essere approssimato con la formula:

$$skewness = \frac{y_{max}}{y_{min}}$$

Da tale formula comprendo che più il valore è alto e più i dati sono "asimettrici". Ma tale considerazione è più corretta quanto più è piccola la quantità di campioni che sto andando a considerare

Un semplice criterio per decidere quale tipologia di metrica utilizzare è quella mostrata in figura [4.1]

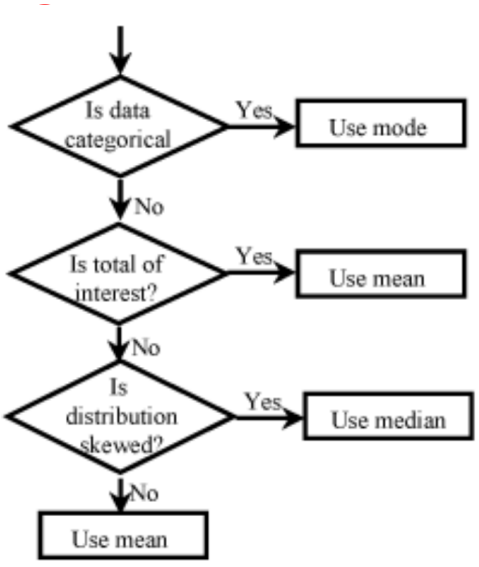


Figure 4.1: Semplice schema di un criterio per la scelta della metrica da utilizzare

I valori descritti, oltretutto, possono essere ottenuti osservando anche la distribuzione dei dati mediante degli istogrammi di occorrenza. In questo modo è anche più facile campire quale metrica sia meglio utilizzare.

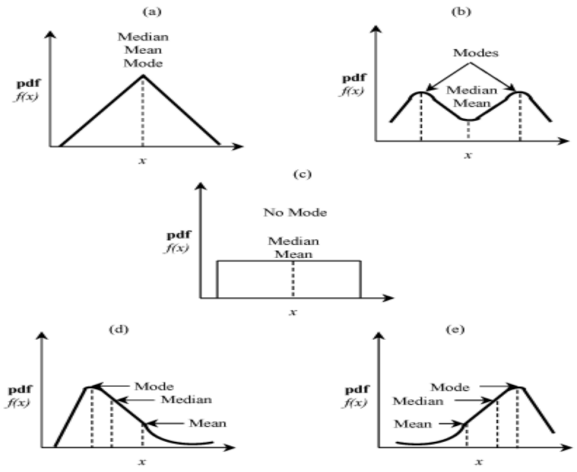


Figure 4.2: Grafici illustrativi dei parametri discussi

4.2 Indici di dispersione

Andare ad approssimare una distribuzione di valori con un singolo valore non basta, tale informazione non mi permette di considerare anche la **variabilità** che i dati possono avere nel tempo. Uno strumento utile per avere un'idea della variabilità, sono gli **indici di dispersione**. Tra le metriche più utilizzate possiamo trovare:

- **Range**: Il range è un parametro molto basilare che viene calcolato come: $v_{max} - Vmin$. Tale metrica, per quanto semplice, è anche molto poco resistente agli outlier
- **Varianza campionaria (o deviazione standard)**: Tale parametro è strettamente calcolato dai dait e dipende dalla loro distribuzione
- **10- e -90 Percentile**: Prendendo una distribuzione di dati si vanno a considerare due valori:
  - $P10$: Valore dei dati sotto il quale cade il 10% dei dati considerati
  - $P90$: Valore dei dati sotto il quale cade il 90% dei dati considerati (quindi solo il restante 10% è maggiore)

  Si definisce poi come indice di dispersione il range calcolato su tali valori: $range = P90 - P10$. Se vediamo tali valori in riferimento alla pdf dei dati (il loro istogramma) e alla CDF. Si ha che per calcolare il valore P90 e P10, vado a valutare il valore della CDF in base al percentile ricercato ($F(x_p) = 0.10$ o 0.90), oppure, nel caso della pdf vado a calutare l'integrale (guardare la relazione tra CDF e pdf) $\left[\int_{-\infty}^{x_p} f(x)dx = 0.10|0.90\right]$. Il percentile solitamente può essere chiamato anche **quantile**. La differenza sta nella notazione, per il quantile si dice 0.1-quantile (α-quantile), mentre il percentile si esprime come 10-percentile ($100 * \alpha$-percentile). Per stimare tali valori su un insieme di dati discreto (quindi non distribuzioni continue su cui effettuare integrali o derivate), posso andare ad effettuare, prima un ordinamento e poi a selezionare uno specifico valore in base alla α scelta. Precisamente si va a selezionare il valore in posizione: $[(n-1) * \alpha + 1]$

- **Semi Inter-Quantile Range**: Tale metrica va ad utilizzare la definizione di due particolari casi di quantile. Più precisamente si va a considerare i quartili: 0.75-quantile (75-percentile) [o terzo quantile] ed il 0.25-quantile (25-percentile) [o primo quantile]. Tali valori sono poi combinati e si va a calcolare il valore del Semi Inter-Quantile Range come:

$$SIQR = \frac{Q_3 - Q_1}{2}$$

- **Mean Absolute Deviation**: Vado a considerare la somma delle deviazioni, ma che sia sottoposta prima all'operazione di modulo, in modo da evitale l'azzeramento della somma delle deviazioni (come visto nel capitolo precedente). La formula è la seguente:

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

  Dove $\overline{x}$ è la media dei valori: $x_1, x_2, \ldots, x_n$.

Un elemento utile quando si sta guardando il grafico dei dati è il **boxplot**. Il boxplot ci da un informazione grafica iniziale su dove i valori di interesse per il calcolo delle diverse metriche si trovino. Principalmente ci permette di poter osservare le seguenti metriche [4.3]:

- **Valore massimo**
- **Valore minimo**
- **Mediana**(o 0.5-quantile/secondo quantile)
- **Primo Quantile**(0.25-quantile)
- **Terzo Quantile**(0.75-quantile)
- **Semi Inter-Quantile Range**

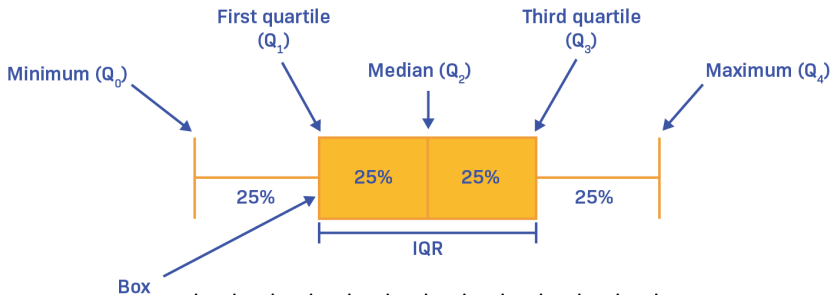


Figure 4.3: Struttura di un boxplot

La stima dell'indice di dispersione è qualcosa di complicato, sopratutto quando ci sono in mezzo gli outlier. Pertanto, solitamente, come metrica al posto della varianza (che non è completamente immune agli outlier), si preferisce utilizzare il Semi Iter-Quantile

Range (SIRQ), che è più resistente alla presenza di outlier (anche più di qualcuno). Mentre come metrica centrale si preferisce l'uso della mediana, questo poichè ci è assicurato che esista (a differenza della moda), e sopratutto che faccia parte dell'insieme di valori considerato (diversamente dalla media). Questo spiega anche la costruzione del boxplot e dei valori che va a mettere in evidenza.

4.3 Quantile-Quantile Plot

Il **Quantile-Quantile Plot** è un metodo che ci permette di confrontare se due popolazioni hanno una distribuzione simile o meno. Nel nostro caso lo si utilizzerà per confrontare la distribuzione della nostra popolazione rispetto alla distribuzione normale. La composizione di base di un grafico quantile-quantile generico è la seguente: Si va a definire un punto p come $p = (x_i, y_i)$, dove x_i è il valore teorico dell' i-esimo quantile, mentre y_i è il valore reale dell'i-esimo quantile. Per quanto riguarda i parametri reali non è complicato trovare i quartili (basta utilizzare la specifica formula in base ad α), mentre per i parametri teorici risulta più complesso dato che bisogna trovare il modo di invertire la CDF. Per ricavare i valori della normale (funzione gaussiana di media nulla e varianza unitaria). Vado a calcolare il quantile come: $q_i = \dfrac{i - 0.5}{2}$, da cui posso ricavare il valore teorico come: $x_i = 4.91 \left[q_i^{0.14} - (1 - q_i)^{0.14}\right]$. Quindi, in generale, data una popolazione posso plottare il grafico utilizzando le metriche esposte in precedenza, e confrontare la distribuzione dei dati reali con quelli di una normale. In questo modo cerco di capire quanto la distribuzione della popolazione sia simile ad una distribuzione gaussiana. Alcuni esempi sono mostrati alla figura: [4.4]

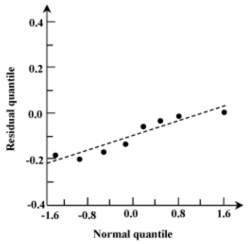
4.4 Intervalli di confidenza e dimensione del campionamento

Solitamente, andare a considerare tutta la popolazione, potrebbe essere oneroso, pertanto, si potrebbe considerare di fare delle valutazioni su un numero minore di componenti estratte in maniera randomica (campionamento della popolazione). La problematica principale risiede nelle assunzioni che si potrebbe fare sulle statistiche della popolazione intera, ovvero, calcolare qualche valore per l'intera popolazione a partire dalla popolazione intera. Tali metodoligie sono sotto il nome di **statistica inferenziale**. Per comprenderne bene il legame osservare l'immagine [4.5].

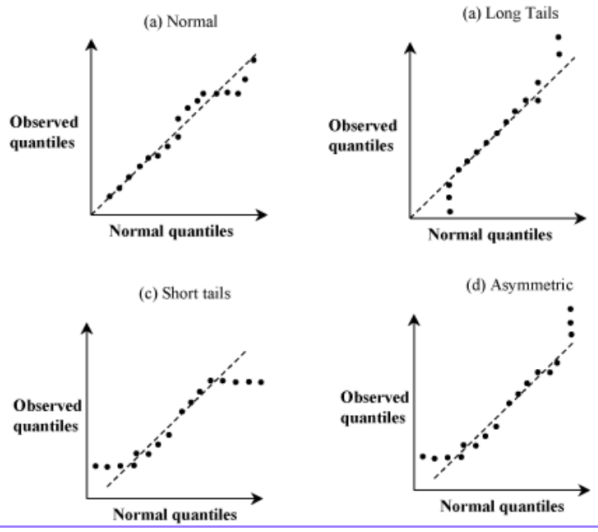
4.4.1 Campioni e Popolazione

È importante definire cosa si intende quando si sta parlando di popolazione o di campioni, ciò ci permetterà di poter definire tutta una serie di principi statistici, utili, per effettuare l'inferenza statistica. Formalmente si definisce **popolazione** l'insieme totale di tutte le componenti (o istanze), mentre si definisce **Campione**(O Sample), un insieme di n osservazioni provenienti dai campioni. In generale è importante, definire anche il significato di **parametri**, i dati associati alla popolazione, essi saranno indicati con le lettere greche (ad esempio la media e la varianza associate alla popolazione, tali valori

	y_i	q_i	x_i
1	-0.19	0.0625	-1.535
2	-0.14	0.1875	-0.885
3	-0.09	0.3125	-0.487
4	-0.04	0.4375	-0.157
5	0.04	0.5625	0.157
6	0.04	0.6875	0.487
7	0.14	0.8125	0.885
8	0.19	0.9375	1.535



(a) Esempio di Quantile-Quantile plot rispetto alla normale



(b) Diverse tipologie di rappresentazioni in base alla natura della popolazione

Figure 4.4: Applicazioni della Quanitile-Quantile plot (o rappresentazione Quantile-Quantile)

per qualunque saranno i campioni estratti, non varieranno [saranno costanti]); mentre si definiscono **statistiche** i valori associati ai sample della popolazione, esse saranno rappresentate da lettere comuni (tali valori variano da campione a campione e quindi non sono costanti).

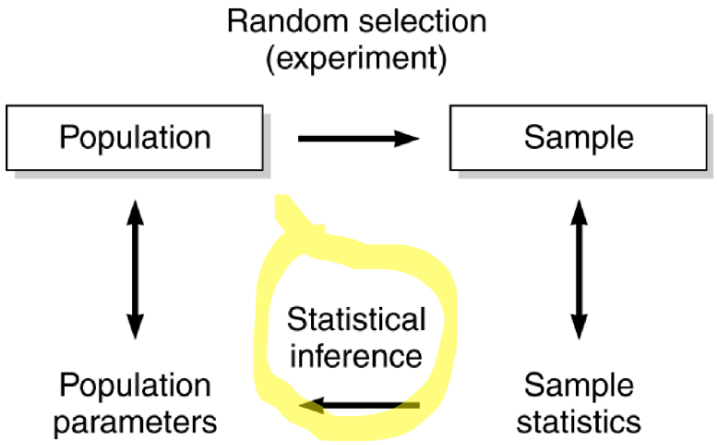


Figure 4.5: Collocazione logica della statistica inferenziale

Info: *Tale pezzo non è richiesto ai fini dell'esame ma potrebbe essere utile in una fase di approfondimento e studio della statistica inferenziale*
Distribuzioni Campionarie
Si definisce **Distribuzione Campionaria**, una distribuzione di probabilità associata alla variabile aleatoria $\overline{X}$, che è la composizione di diverse osservazioni $X_1, X_2, \ldots, X_n$ che sono ricavate da k-campioni della popolazione iniziale. $\overline{X}$ non è altro che una variabile aleatoria associata alla media delle osservazioni (media campionaria).

Warning: Quando parliamo di $\overline{X}$ stiamo parlando della variabile aleatoria (quindi un valore casuale), mentre la distribuzione campionaria è la funzione di distribuzione di probabilità(pdf) associata ad $\overline{X}$

Pertanto, possiamo definire $\overline{X}$ come:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Da cui, applicando il **teorema fondamentale della media**, si ottiene che:

$$E(\overline{X}) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n}$$

Andando a considerare un caso più semplice, dove la popolazione ha media nota e pari a μ e varianza σ^2, e dove le variabili aleatorie $X_1, X_2, \ldots, X_n$, associate alle osservazioni dei campioni, sono delle variabili aleatorie indipendenti ed identicamente distribuite (sono la stessa variabile aleatoria con media μ e varianza σ^2), allora si può dire che:

$$E(\overline{X}) = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$$

$$V(\overline{X}) = E((\overline{X} - \mu)^2) = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Da tali formule si sono ricavati dei valori associati alla distribuzione campionaria, dette statistiche. Com'è possibile notare, più la dimensione del sample size cresce e più la varianza diminuisce, ciò ci fa capire che l'errore fatto sulla media diminuisce man mano (base per la valutazione dello standard error). Andando ad invalidare le ipotesi fatte sulla struttura della popolazione (sulla sua distribuzione di probabilità), allora si può andare a valutare il **Teorema del limite centrale**.

4.4.2 Standard Error

In linea generale e formale, lo **Standard Error** non è altro che la deviazione standard della distribuzione campionaria. Quando vado a valutare le statistiche riferite ai vari campioni presi dalla popolazione, con sample size pari ad n, voglio capire quanto queste siano effettivamente vicine a quelle reali. La problematica risiede proprio nel calcolo dei parametri. Di base, non si possono fare delle assunzioni sulla distribuzione di probabilità della popolazione (solitamente non è gaussiana, ma ha una forma generica o

skewed (ambigua)). Pertanto non posso andare a valutare la varianza campionaria, questo poichè non posso fare assunzioni. Pertanto, in questi casi, ci viene in aiuto il **teorema del limite centrale**. Il **teorema del limite centrale** ci dice che:
Ipotesi
Siano $X_1, X_2, \ldots, X_n$, n-variabili aleatorie Indipendenti ed identicamente distribuite (hanno tutte la stessa distribuzione, poichè la loro base dei dati è la stessa)
Sia la sample size n, un valore orientativamente grande $n >= 30$
Tesi
Allora posso dire che la **distribuzione campionaria** della media (distribuzione della variabile aleatori $\overline{X}$), è riconducibile ad una funzione normale con:

- **Media** pari a μ (media della popolazione)
- **Varianza** pari a $\frac{\sigma^2}{n}$ (o deviazione standard $\frac{\sigma}{\sqrt{n}}$)

Se la sample size è relativamente bassa (empiricamente minore di 30), la distribuzione ha un'andamento differente chiamato **t-student**. (Questo accade poichè, solitamente, non conoscendo la varianza reale della distribuzione globale, si potrebbe andare ad utilizzare la varianza campionaria, ma tale varianza avrà anche lei un margine di errore [è una variabile aleatoria a n-1 gradi di libertà]. Di conseguenza quella relazione diventa una relazione assimilabile alla t-student).

Info: *Tale pezzo non è richieto per lo svolgimento dell'esame, ma solo al fine di approfondire alcuni concetti*
Seguendo quello che si è visto nel precedente approfondimento [4.4.1]; Il teorema del limite centrale non è altro che la conseguenza della composizione della variabile aleatoria $\overline{X}$, rispetto alle variabili aleatorie associate ai campioni. Se le variabili aleatorie sono indipendenti (ce lo assicurano le indipendenze tra le varie osservazioni) ed identicamente distribuite (il fatto che le varie variabili abbiano la stessa pdf(e quindi non lo stesso valore), ci assicura, che la media di tutte le distribuzioni sia μ), allora il teorema del limite centrale non è altro che la dimostrazione del caso particolare che si è mostrato in precedenza. Se si vuole avere una maggior formalità nell'esporre il teorema del limite centrale si può andare a definire, una nuova variabile aleatoria Z, come:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{n\to\infty} \mathcal{N}(0,1)$$

Tale Z, solitamente, viene assimilata ad una **t-student**, poichè non è sempre detto che si conosca la σ reale associata alla popolazione, pertanto si va ad utilizzare la varianza campionaria s. Data questa "sostituzione", come per la media, anche per la varianza bisognerebbe andare a definire degli intervalli di confidenza e degli errori rispetto al valore reale (pertanto viene trattata come una variabile aleatoria chi-quadrata), ciò ci permette di dire che la variabile aleatoria sopracitata Z, sia a tutti gli effetti una **t-student**.

Il **Teorema del limite centrale** non fa alcuna ipotesi sulla distribuzione della popolazione di partenza. Pertanto si mostra un esempio di applicazione mediante la figura [4.6]

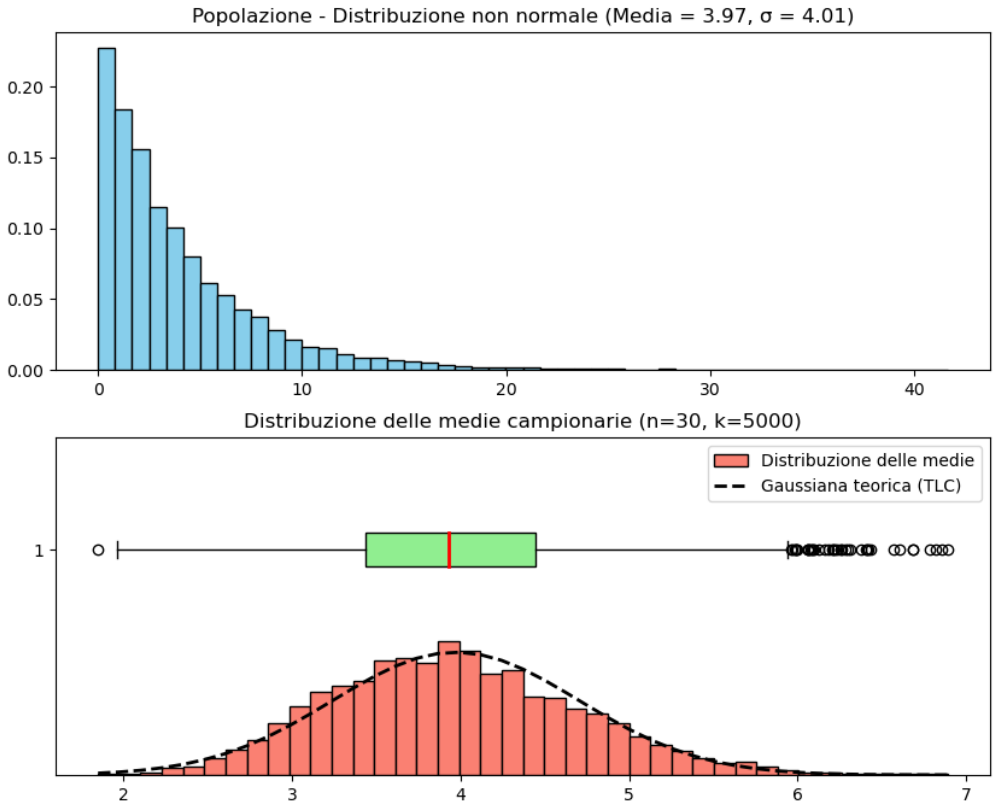


Figure 4.6: Applicazione del teorema del limite centrale

4.4.3 Intervalli di confidenza

Per comprendere appieno il significato di **intervallo di confidenza**, allora, bisogna esprimere in maniera migliore la variabile aleatoria $\overline{X}$. Quello che si vuole andare a fare è normalizzarla, per cercare di ricondurre la distribuzione ad una normale con varianza unitaria. Formalmente potremo dire che:
Siano $X_1, X_2, \ldots, X_n$ variabili aleatroie associate ai campioni di una popolazione che ha una distribuzione normale con una media non nota μ ed una varianza nota σ^2. Allora, la variabile aleatoria associata all'operazione di media $\overline{X}$ è, teoricamente, normale, centrata in μ e con varianza σ^2/n. Pertanto, posso andare a normalizzare tale distribuzione di probabilità ad una normale standard mediante l'applicazione della seguente formula (simile allo z-score):

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Warning: Per ora non si è ancora detto nulla sul come sia fatta la varianza, pertanto, se ci sono dei dubbi, risulta perfettamente in linea dato che ancora deve essere affrontata la questione. Essa sarà affrontata nel capitolo [4.4.4]

Un **Intervallo di confidenza**, quindi, non è altro che un intervallo numerico in cui la media ricade. In maniera formale, un intervallo di confidenza viene rappresentato mediante l'intervallo $[l, u]$, che ha la seguente caratteristica: $l \leq \mu \leq u$. l e u sono valori che sono stati calcolati partendo dalle osservazioni che sono state campionate (quini i valori sono associati al singolo campione). Pertanto, dato che si andranno a valutare per ogni campione, costituiranno anche loro delle vere e proprie variabili aleatorie L ed U. Supponendo di poterne valutare la distribuzione, si avrebbe che:

$$P(L \leq \mu \leq U) = 1 - \alpha$$

dove $0 \leq \alpha \leq 1$. Precisamente, l'intervallo delineato da l, u è l'intervallo di confidenza, mentre α è il livello di confidenza (la probabilità che la media non faccia parte di tale intervallo). Dal singolo campione, quindi, possiamo ricavare che la media potrebbe ricadere all'interno dell'intevallo: $l \leq \mu \leq u$.

Per calcolare l'intervallo di confidenza a partire dai dati presenti, si va a considerare il valore che può assumere la precedente probabilità $(1-\alpha)$, andando a fissare α, si possono andare a delimitare due punti simmetrici al punto della media della media campionaria, che delimitano due aree esterne la cui somma è proprio α (di conseguenza due aree laterali che misurano $\alpha/2$ l'una [4.7])

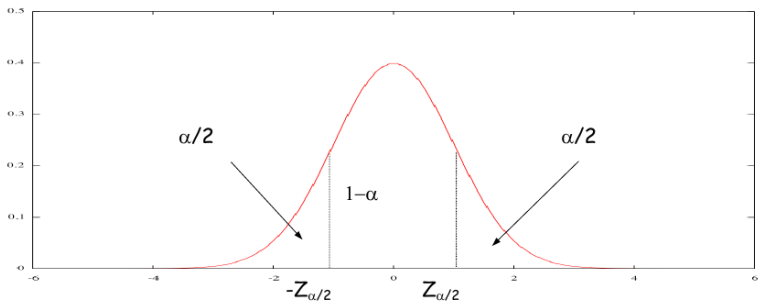


Figure 4.7: Intervallo di confidenza rispetto alla normale

Pertanto, essendo i valori della normale standard, noti, possiamo andare a scrivere la seguente probabilità:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

Andando a sostituire con il valore che ci siamo calcolati a partire dalla variabile aleatoria associata alla distribuzione campionaria, si ha:

$$P(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

da cui:

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Da tale conclusione possiamo ricavare l'intervallo di confidenza (quindi i valori l e u). Considerando quindi, $\overline{x}$ la media della distribuzione campionaria, σ^2 la varianza della popolazione iniziale (tale parametro non è banale) e sample size pari ad n, allora posso definire un **intervallo di confidenza** al $100 * (1 - \alpha)\%$ di **livello di confidenza**, come:

$$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Con $z_{\alpha/2}$ il punto di $\frac{\alpha}{2}$-quaritile o 100*($\frac{\alpha}{2}$)-percentile
La lunghezza dell'intervallo di confidenza ci permette di misurare la **precisione** di stima, o meglio, ci dice il range in cui la media originale della popolazione possa ricadere con un certo livello di confidenza (solitamente 95%, ma dipende dall α utilizzata per in calcolo dell'intervallo). Oltretutto, si può calcolare anche il margine di errore che si effettua con la media reale mediante la valutazione del seguente valore:

$$E = error = |\overline{x} - \mu|$$

Warning: Fare attenzione quando leggete tali appunti. $\overline{X}$ il valore che ha X grande indica la variabile aleatoria a cui la distribuzione campionaria funge da pdf. Mentre $\overline{x}$ è il valore della media calcolata per il singolo campione (o insieme di osservazione estratto dalla popolazione)

I fattori che principalmente incidono sulle dimensioni dell'intervallo di confidenza sono:

- **Sample size (n)**: Il sample size incide profondamente sulla dimensione dell'intervallo di confidenza. Più esso è grande e più il valore di media calcolato sul singolo campione risulta "preciso". (per intenderci, matematicamente, si va a ridurre il range a parità di livello di confidenza, in cui la media originale può ricadere)
- **Livello di confidenza**: Più il livello di confidenza è alto e più il range tende ad aumentare, mentre se si richiede dei livelli di confidenza più bassi, allora il range tende a diminuire. (si va ad agire, matematicamenre, sempre sul termine che va ad infierire sulla media nel calcolo del range. Per capirci, il livello di confidenza infierisce sul calcolo del parametro $z_{\alpha/2}$)
- **Variabilità della popolazione**: Altro parametro fondamentale per il calcolo dell'intervallo di confidenza è la deviazione standard della popolazione iniziale, più la popolazione iniziale ha una deviazione standard grande, più è difficile dedurre la media a partire dai campioni

4.4.4 Iperparametri per gli intervalli di confidenza

Nel precedente paragrafo si è affrontato la discussione riguardante gli intervalli di confidenza. Quello di cui però non si è discusso è il come gli iperparametri associati alla determinazione dell'intervallo di confidenza (n,σ), vengano valutati o scelti. Pertanto, si sono definite delle regole, che ne hanno permesso la corretta valutazione.

Sample size

La **Sample size** è un parametro molto importante è va valutato con cura (date tutte le implicazioni che si hanno su di esso). Pertanto si è cercato un metodo di calcolo per valutarne al meglio il valore. Formalmente:

Definito $\overline{x}$ il valore approssimativo (valutato sul singolo campione) della media μ e σ la deviazione standard della popolazione e scelti: il livello di confidenza $100 * (1 - \alpha)\%$ e la stima superiore dell'errore desiderato E(definito come $E = |\overline{x} - \mu|$). Allora si può calcolare la sample size come:

$$n = \left(\frac{\sigma * z_{\alpha/2}}{E}\right)^2$$

Tale valore sarà un valore reale, dato che n dev'essere un'intero, allora si approssima sempre tale valore per eccesso, questo per garantire che i valori che sono stati dati in ingresso (E, α) siano rispettati.

Info: *Tale pezzo è stato inserito sono a scopo di approfondimento, non è richiesto ai fini dell'esame*

Il calcolo della **Sample size**, deriva direttamente dalla definizione di intervallo di confidenza, combinata con la definizione di errore. Precisamente la dimostrazione è composta dai seguenti passaggi: Partendo dalla definizione dell'intervallo di confidenza:

$$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Spostando la media al centro si ottiene:

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu - \overline{x} \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Se al centro poniamo un modulo, otteniamo la definizione di $E = |\overline{x} - \mu| = |\mu - \overline{x}|$, ma, essendo che i valori saranno sicuramente maggiori di 0 (dato l'inserimento del modulo), allora posso ignorare la parte di sinistra della disequazione, ottenendo:

$$E \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Dato che cerchiamo il valore minimo, si può ignorare il minore e considerare solo l'uguaglianza, da cui:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \iff n = \left(\frac{\sigma * z_{\alpha/2}}{E}\right)^2$$

Da cui la tesi

Varianza

Fino ad ora abbiamo considerato la deviazione standard della popolazione σ, nota. Solitamente, però, non si ha a che fare con popolazioni caratterizzate (non si conosce ne la media ne la varianza). Per risolvere tale problema si fa utilizzo del **Teorema del limite centrale**, che ci permette di dire che, date le variabili aleatorie ($X_1, X_2, \ldots, X_n$)

legate alla variabile aleatoria della media campionaria ($\overline{X}$), e dato un valore di n abbastanza grande, allora, scalando con lo z-score la $\overline{X}$, otteniamo una distribuzione normale standard, più formalmente:

$$\overline{X}' = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

dato che però σ non lo conosciamo, esso può essere sostituito con la deviazione standard campionaria, che per grandi valori di n risulta eguale alla deviazione standard normale. Formalmente si può dire che:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

Dove, per n molto grande (idealmente $n \to \infty$) il termine $\frac{n}{n-1} \to 1$, e che quindi ci permette di trovare una buona approssimazione della varianza per la varianza campionaria. Quindi, da ora in poi, per n sufficientemente grandi si potrà dire che $s \approx \sigma$. Da tale considerazione, quindi, si farà riferimento alla distribuzione **t-student**, che sarà definita come:

$$Z = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

Tale distribuzione, per valori grandi di n [per via del teorema del limite centrale] tende ad essere una distribuzione normale standard, il che ci permette di poter definire l'**intervallo di confidenza per grandi campioni**(large sample confidence interval), come:

$$\overline{X} - \frac{s * z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{s * z_{\alpha/2}}{\sqrt{n}}$$

il quale avrà un certo **livello di confidenza** di approssimativamente: $100 * (1-\alpha)\%$.

4.4.5 t-student

Come accennato precedentemente, mediante un'operazione di z-socre sulla variabile aleatoria $\overline{X}$, e considerando la varianza $\sigma \approx s$, allora si avrà una variabile aleatoria detta **t-student**. Essa viene caratterizzata da due principali parametri (solitamente), ovvero, il parametro α ed i suoi gradi di libertà (nel nostro caso $n-1$). La variabile aleatoria T viene definita come:

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

data tale definizione, cambia anche la definizione dell'**intervallo di confidenza**. Precisamente, la definizione si basa sulla considerazione che si sta avendo a che fare con una t-student, e che i suoi valori sono tabulati secondo sia la variabile α che rispetto ai suoi gradi di libertà $n-1$. Infatti, formalmente:

$$\overline{x} - \frac{t_{(1-\alpha/2),n-1} * s}{\sqrt{n}} \leq \mu \leq \overline{x} + \frac{t_{1-\alpha/2,n-1} * s}{\sqrt{n}}$$

il livello di confidenza associato a tale intervallo è pari a $100 * (1-\alpha)\%$.

4.5 Test di ipotesi

Un **test di ipotesi** è una serie di tecniche utilizzate per decidere se un **ipotesi** fatta sui dati sia veritiera o meno. Con le conoscenze pregresse (delle precedenti sezioni), si possono affrontare tali test. Uno dei più semplici in quest'ambito è lo **zero mean test**, tale test va a verificare semplicemente se l'intervallo di confidenza calcolato include 0 o meno [4.8].

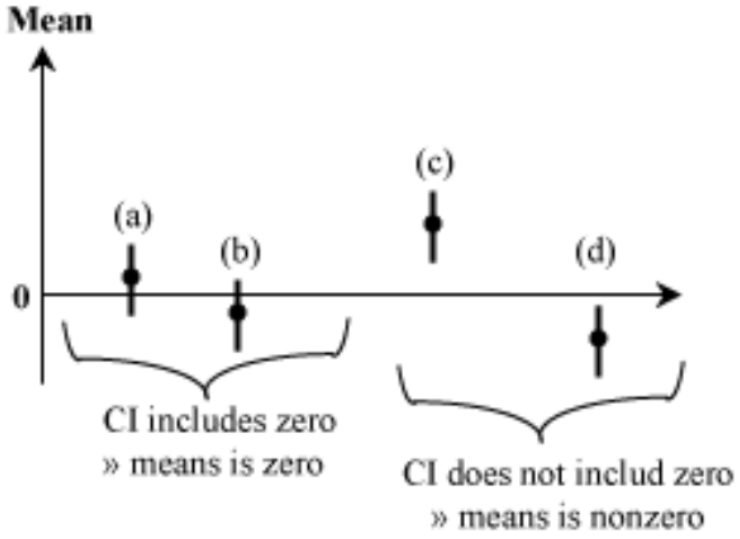


Figure 4.8: Zero Mean Test

I testi di ipotesi sono molto utili sopratutto quando si vogliono andare a confrontare diverse alternative (o diversi sistemi). Entrando nel merito, esistono due modalità principali con cui è possibile confrontare due diversi sistemi (A e B), ovvero;

- **Paired observations**: Si vanno ad effettuare n esperimenti speculari su entrambi i sistemi, quindi entrambi avranno un set di espetimenti con cardinalità uguale. Pertanto si cerca di trovare una corrispondenza tra l'i-esimo test effettuato sul sistema A e l'i-esimo test effettuato sul sistema B. Una volta effettuati gli esperimenti si va a costruire la distribuzione delle differenze ($d_i = A_i - B_i$), e nel caso l'**intervallo di confidenza** include lo 0, allora i due sistemi sono statisticamente equivalenti. es. si hanno due sistemi e vengono sottoposti entrambi agli stessi 6 workload, voglio campire con un livello di confidenza del 90% se un sistema è "migliore" dell'altro. Una volta valutati i valori (magari del response time), vado a valutarne le varie differenze per coppie paired (vado ad effettuare la differenza tra valori appartenenti allo stesso workload), una volta ottenuto il vettore delle differenze vado a valutare l'intervallo di confidenza, se questo contiene 0 allora i due sistemi sono equivalenti atatisticamente, altrimenti sono statisticamente differenti.

- **Unpaired observations**: Si vanno ad effettuare le valutazioni su ogni sistema in maniera indipendente dall'altro (quindi non è detto siano gli stessi stress). Un volta ottenuti i risultati si vanno a costruire gli intervalli di confidenza, ognuno per se, e poi si può ricadere in tre casi:

- **Nessuma sovrapposizione**: I due intervalli di confidenza non si intersecano in alcun punto, il che predilige che un sistema sia prevalente all'altro
- **Medie incluse negli intervalli**: Quando i due intervalli di confidenza non solo si sovrappongono, ma la media dei due sistemi è inclusa nel range dell'altro e vicecersa (indice di equivalenza statistica)
- **Intervalli intersecati ma senza media**: Quando l'intersezione tra i due intervalli di confidenza è non nulla, ma le medie sono entrambe esterne all'altro intervallo di confidenza. In questi casi non è deducibile niente e bisogna effettuare un **t-test** (vedere [**??**])

Per una migliore comprensione di quali di questi concetti osservare l'immagine [4.9]

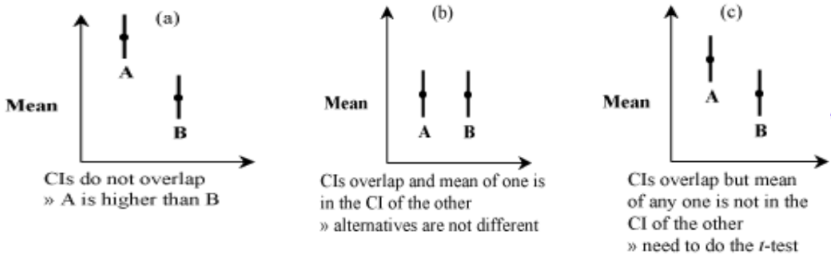


Figure 4.9: Casi di unpaired observations

Compreso cosa sia un **test di ipotesi**, allora, possiamo passare a definire quella che è un ipotesi:
Un **ipotesi** non è che un affermazione che viene fatta sui parametri di una o più popolazioni. Ed il **test di ipotesi** non è altro che la procedura per confermare o meno le ipotesi. Si possono definire le ipotesi partendo dalla **Null Hypothesis**, essa racchiude quello che è il concetto che si vuole verificare (es. Ipotesi sulla media della popolazione $H_0 : \mu = \mu_0$) [solitamente rappresenta lo stato corrente o quello senza effetto], e poi vi è la controparte detta **Alternative Hypothesis** [Afferma che c'è un effetto reale e solitamente quando si raccolgono i dati se cerca il modo di favorire H_1 e di rigettare H_0]. Sulla alternative hypothesis si possono definire:

- **one-sided alternative hyphotesis**: si va a definire una confutazione dell'ipotesi nulla evincendo una relazione d'ordine (es. $H_1 : \mu < \mu_0$)
- La **two-sided alternative hypothesis**: si va a definire una confutazione dell'ipotesi nulla senza andare a mostrare particolari relazioni d'ordine ($H_1 : \mu \neq \mu_0$)

La suddivisione di tali ipotesi alternative risiede nel come poi si andranno a valutare le regioni critiche, principalmente, nella one-sided alternative hypothesis si va a definire la regione critica andando a valutare una sola delle code della gaussiana, mentre nel caso della two-sided alternative hypothesis si vanno a valutare entrambe le code della gaussiana (definite mediante quartili).

Fare dei test di ipotesi implica l'assunzione di decisioni che, in alcuni casi, possono risultare errate. Per questo motivo è fondamentale definire delle **metriche di valutazione del rischio**, in modo da quantificare e controllare la probabilità di giungere a

una conclusione sbagliata. Nella pratica, raramente è possibile analizzare l'intera popolazione: si lavora quindi con campioni casuali, dai quali si ricavano informazioni utili per trarre conclusioni sulla popolazione di interesse. Attraverso l'**inferenza statistica**, si stima il comportamento dei parametri e si eseguono test di ipotesi per decidere se accettare o rifiutare l'ipotesi nulla H_0 a favore dell'ipotesi alternativa H_1. Tali procedure consentono di confrontare le formulazioni su basi oggettive, mantenendo sempre consapevolezza dei rischi associati e della probabilità di commettere errori nelle decisioni.

Per effettuare il **test di ipotesi**, solitamente, si effettuano i seguenti passi:

1. **Dichiarazione dell'ipotesi nulla**: Effettuare un affermazione sui parametri della popolazione (ATTENZIONE non il campione ma la popolazione), ad es. ($H_0 : \mu = \mu_0$)

2. **Livello di confidenza**: definizione di un parametro α che indicherà la probabilità di commettere errori di **primo tipo**

3. **Decidere la tipologia di test**: Scegliere quale test andare ad effettuare sui dati (es. t-test, z-test, ecc.)

4. **Calcolo delle statistiche per il test**

5. **Valutazione delle ipotesi**: la valutazione delle ipotesi differisce in base alla tipologia di test che si vuole andare ad effettuare

4.5.1 P-value

Il **P-value approach** va a valutare il **p-value**, un valore che mi permette di capire quanto i miei dati siano compatibili con la mia ipotesi nulla. Per comprendere meglio il significato di tale valore facciamo l'esempio di una moneta:

Info: *Esempio moneta p-value*
Si vuole valutare se una moneta sia equilibrata o meno. Pertanto si avranno le seguenti ipotesi:

$$H_0 : p(testa) = 0.5$$

$$H_1 : p(testa) \neq 0.5$$

Si effettuano 10 tiri e si ottiene testa 8 volte. Allora si va a valutare il p-value come la probabilità che esca 8 volte testa **considerando l'ipotesi nulla come vera**. Pertanto si va a valutare che:

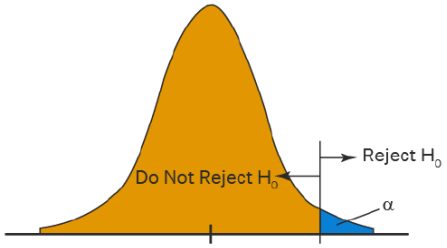
$$P(testa\ per\ 9\ volte) = \binom{10}{8}\left(\frac{1}{2}\right)^8 * \left(\frac{1}{2}\right)^2 = 45\left(\frac{1}{2^10}\right) = 0.04394 = 4.39\%$$

Se ho considerato un livello di confidenza al 90%, il p-value, essendo minore del valore di $alpha = 0.5$, allora l'ipotesi H_0 è certamente falsa. Mentre nel caso fosse stato maggiore di tale valore, l'ipotesi H_1 poteva essere confermata. Più è grande il p-value e maggiore è la "giusta" scelta dell'ipotesi H_0

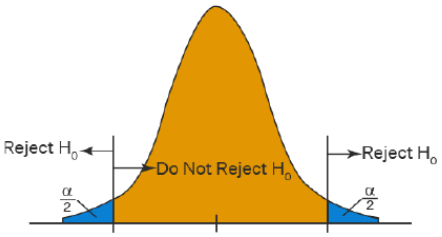
In generale, la determinazione del p-value viene effettuata mediante l'integrale della distribuzione, che sottostà al valore effettivamente osservato (solitamente si lavora sui valori limite dell'ipotesi H_0, quindi quei valori che sono al confine della scelta per tale ipotesi). Un'altro modo per interpretare il p-value è associandolo al **rischio** di un errato rigetto dell'ipotesi H_0.

4.5.2 Tipologie di errori

Nel precedente capitolo si fa riferimento a **regione critica** e **valori critici**. Per capire meglio a cosa si fa riferimento, si va a considerare uno specifico intervallo di confidenza, in cui l'ipotesi H_0 è vera. La **regione critica** è quella parte del dominio che rigetta l'ipotesi H_0 mentre i **valori critici** sono i valori limite per la regione critica. La definizione di quali sia la regione critica ed i relativi valori critici viene definita a partire dalla tipologia di alternative hypothesis fatta, per comprendere al meglio cosa si intende per regione critica e per valori critici, osservare la figura [4.10], dove le regioni in blu sono le regioni critiche, ed i valori evidenziati con delle linee nere, invece, sono i valori critici.



(a) One-Sided alternative hypothesis critical region



(b) Two-Sided alternative hypothesis critical regions

(c) Two-Sided alternative hypothesis critical regions (esempio logico)

Figure 4.10: Critical region per le differenti tipologie di alternative hypothesis

Errore di tipo I

L'**errore di tipo I** è definito come: Rigettare l'ipotesi nulla H_0 quando essa è in realta vera. La probabilità associata all'errore di tipo I, viene chiamata α**-error** (o significance level o dimensione del test. Sono varie le definizioni)

$$\alpha = P(errore\ di\ tipo\ I) = P(H_0\ sia\ rigettato\ quando\ H_0\ e'\ vero)$$

Per definire tale parametro si hanno due strade:

- Decido il parametro α e poi vado a valutarmi i valori critici e le regioni critiche
- La regione prefissata ed i valori critici sono stati fissati in precedenza, da tali valori vado a ricavarmi il mio parametro α

Per comprendere al meglio come calcolare il parametro α a partire dai valori critici, si fa il seguente esempio:
Si considera un'intervallo di confidenza come [48.5, 51.5]. L'errore di tipo I occorre quando ho un valore di $\overline{x}$ che è fuori dal range, quando nella realtà la media $\mu = 50$. Supponendo che la popolazione sia una gaussiana con media $mu = 50$ e varianza $\sigma^2 = 2.5^2$, andando a valutare la distribuzione delle medie campionarie, si ottiene una normale (per il teorema del limite centrale), con media $\mu = 50$ e varianza $\sigma^2/n = \sigma/\sqrt{n} = 0.79$ = . Pertanto, si può andare a definire α come:

$$\alpha = P(\overline{X} < 48.5\ con\ \mu = 50) + P(\overline{X} > 51.5\ con\ \mu = 50)$$

Andando a valutare gli z_values, si ottiene:

$$z_1 = \frac{lower\_critical\_value - \mu}{\sigma/\sqrt{n}} = \frac{48.5 - 50}{0.79} = -\frac{1.5}{0.79} \approx -1.90$$

$$z_2 = \frac{upper\_critical\_value - \mu}{\sigma/\sqrt{n}} = \frac{51.5 - 50}{0.79} = \frac{1.5}{0.79} \approx 1.90$$

e quindi posso andare a definire α come:

$$\alpha = P(Z < -1.90) + P(Z > 1.90) \approx 0.028717 + 0.028717 = 0.057434$$

Questo lo possiamo fare poichè mediante la "normalizzazione" della variabile aleatoria, possiamo accedere ai valori di probabilità andando a prelevarli all'interno delle tabelle inerenti alla funzione t-student. Il valore trovato, ci permette di capire che rigettare l'ipotesi $H_0 : \mu = 50$ quando la media è effettivamente $\mu = 50$ può accadere con una probabilità pari al 5.74%. La cosa da notare è che, se anche non cambiassi i valori critici, ma aumentassi la sample size, la probabilità di errore di tipo I diminuirebbe, il che aggiunge un ulteriore peso nella scelta della sample size effettiva. Per capirci, aumenterebbe il valore associato alla varianza della distribuzione campionaria $\sigma/\sqrt{n}$ Il che, indirettamente fa aumentare anche i valori delle z_1 e z_2, che aumentando in modulo, riducono l'area sottostante, e quindi la probabilità di avere un errore di tipo I

$$n = 16$$

$$\sigma/\sqrt{n} = 2.5/\sqrt{16} = 0.625$$

$$z_1 = \frac{48.5 - 50}{0.625} = 2.4$$

$$z_2 = \frac{51.5 - 50}{0.625} = 2.4$$

$$\alpha = P(Z < 2.4) + P(Z > 2.4) = 0.0082 + 0.0082 = 0.0164 = 1.64\%$$

La probabilità di avere un errore di tipo I ora si è ridotta drasticamente, a parità di valori critici (con probabilmente un livello di confidenza diverso), ma ciò ci basta per capire come la sample size è importante anche per la definizione dell'errore di tipo I

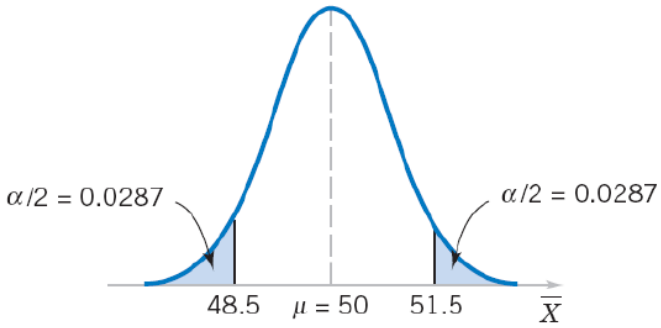


Figure 4.11: Interpretazione grafica dell'errore di tipo I

Errore di tipo II

L'**errore di tipo II** è definito come: Non Rigettare l'ipotesi nulla H_0 quando essa è in realta falsa. La probabilità associata all'errore di tipo II, viene chiamata β**-error**.

$$\beta = P(errore\ di\ tipo\ II) = P(Non\ rigetto\ di\ H_0\ quando\ H_0\ falso)$$

Per comprendere in magliera ottimale la suddivisione tra le due tipologie di errori, si può far riferimento alla tabella [4.12]

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	no error	type II error
Reject H_0	type I error	no error

Figure 4.12: Tabella per comprendere la posizione e la definizione degli errori di tipo I e II

Per capire appieno come interpretare l'**errore di tipo II**, si va a svolgere il seguente esempio:
Consideriamo sempre l'esempio precedente dove si ha l'intervallo di confidenza [48.5, 51.5], e la varianza della popolazione nota $\sigma = 2.5$. Andare a valutare l'errore di tipo II vuol dire:

$$\beta = P(errore\ di\ tipo\ II) = P(Non\ rigettare\ H_0\ quando\ H_0\ e'\ falsa) =$$

$$P(48.5 \leq \overline{X} \leq 51.5\ con\ \mu = 52)$$

andando a valutare i valori delle z_1 e z_2, si avrà:

$$z_1 = \frac{lower\_critical\_value - \mu}{\sigma/\sqrt{n}} = \frac{48.5 - 52}{0.79} = -\frac{3.5}{0.79} \approx -4.43$$

$$z_2 = \frac{upper\_critical\_value - \mu}{\sigma/\sqrt{n}} = \frac{51.5 - 52}{0.79} = -\frac{0.5}{0.79} \approx -0.63$$

da cui andando a sostituire mediante la variabile normalizzata, si avrà che:

$$P(-4.43 \leq Z \leq -0.63) = P(Z \leq -0.63) - P(Z \leq -4.43) = 0.2643 - 0.0000 = 0.2643 = 26.43\%$$

Tale percentuale è il valore di β che indica la probabilità di commettere l'errore di non rigettare l'ipotesi H_0 a discapito del reale valore assunto dalla media. Si vuole far notare che l'errore di tipo II aumenta quanto più diminuisce la distanza tra la media effettiva e quella che si vuole testare. Per capire, se prima la media reale fosse stata 50.5 al posto di 52, si sarebbe avuta la seguente probabilità di errore di tipo II:

$$z_1 = \frac{48.5 - 50.5}{0.79} = -\frac{2}{0.79} \approx -2.53$$

$$z_2 = \frac{51.5 - 50.5}{0.79} = \frac{1}{0.79} \approx 1.27$$

Valutando la probabilità tramite la formula scomposta della Z:

$$P(Z \leq 1.27) - P(Z \leq -2.53) = 0.8980 - 0.0057 = 0.8923 = 89.23\%$$

Notiamo come con la riduzione della "distanza" tra le due medie, aumenta la probabilità di errore di tipo II. A livello grafico, viene rappresentato come l'area sottostante la distribuzione reale per cui si darebbe l'ipotesi sbagliata (compresa tra i valori critici). Per capire meglio tale concetto osservare la figura [4.13].

Come anche per l'errore di tipo I, l'errore di tipo II è fortemente legato alla sample size, più essa è alta e più l'errore di tipo II diminuisce, questo poichè un aumento della sample size fa sì che aumenti anche il valore in modulo delle z_i, che, più sono grandi e minore saranno le loro probabilità, di conseguenza, sarà minore anche l'errore di tipo II (β).

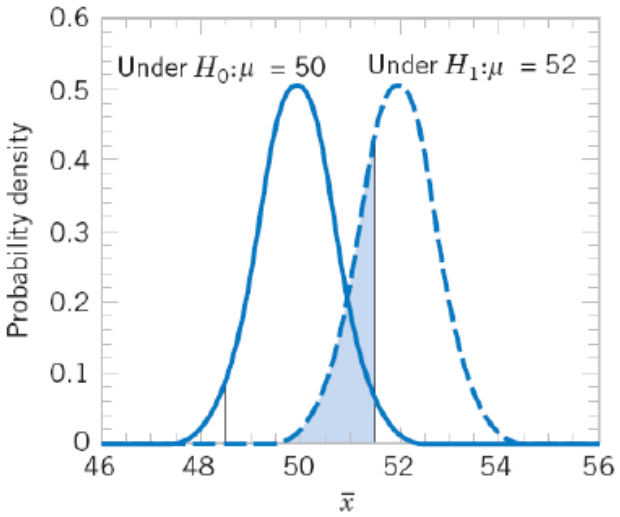


Figure 4.13: Area a cui è associata la probabilità di tipo II (caso test per media 50 con media effettiva 52)

4.5.3 Potenza di un test statistico

Gli errori che si sono definiti in precedenza non bastano per definire quanto un test valuti correttamente le sue ipotesi. Pertanto di definisce la **potenza** di un test statistico, come la probabilità di rigettare l'ipotesi nulla H_0 quando l'ipotesi alternativa è vera. (tale definizione viene data, dato che gli errori parlano di "falle" che il test può avere, il che ci fa capire che non ci sono abbastanza informazioni per avere la risposta corretta [con una certa probabilità], ma non si va a valutare, invece, quanto il test di ipotesi sia tollerante rispetto ad altre implicazioni). La potenza di un test statistico valuta ed esprime la **sensitività** del test statistico, dove per sensitività si intende l'abilità di un test di riconoscere le differenze. La definizione formale di tale parametro è:

$$power = 1 - \beta$$

che sta ad indicare la probabilità di rifiuto di un ipotesi H_0 che è effettivamente falsa. A livello grafico corrisponde all'area sottesa alla curva originale (quella con l'ipotesi di media vera, che risiede al di fuori del critical value, che sta ad indicare quando l'ipotesi H_0 viene rigettata per un motivo valido dato dalla distribuzione valida)[4.14].

4.5.4 One Sample Hypothesis Test

Si vanno ad effettuare delle ipotesi a partire dal singolo campione (attenzione singolo campione e non distribuzione campionaria). Un esempio è l'ipotesi basata sulla media dove:

$$H_0 : \mu = \mu_0$$

dove μ_0 può essere visto come generico valore che la media può assumere (se $\mu_0 = 0$, si sta parlando dello **zero mean test**).

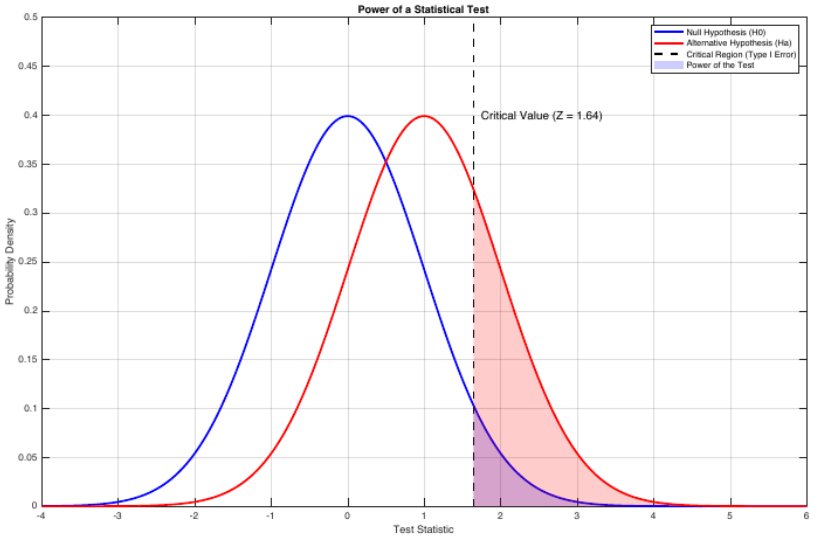


Figure 4.14: Definizione grafica della potenza di un test statistico

Zero Mean Test

Lo zero mean test permette di andare a valutare se, dato un campione, la sua media sia 0.

$$\overline{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \rightarrow z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Andando a valutare, poi, sulla base di z, l'intervallo di confidenza, posso verificare se quest'ultimo contiene 0 o meno, e sopratutto quanto è "distante" da 0. (Il che indirettamente definisce la distanza dalla media μ_0, e quindi verificare se la struttura è **statisticamente insignificante** rispetto a μ_0).

4.5.5 Comparing two alternatives

Quando si vogliono andare ad effettuare delle ipotesi basate sui parametri di due distribuzioni differenti, quali ad esempio media o varianza, bisogna, in un primo momento, prefissare la null hypothesis, poi si va a valutare il modo con cui andare a calcolare le statistiche associate ai sample (o con che criterio andarle a calcolare).

Paired Observations

Quando si parla di paired observations, si intende che il numerodi osservazioni valutate tra i due sistemi è uguale e che ogni osservazione non è casuale, ma è frutto dello stesso "workload" (o sistema di valutazione). Pertanto, quello che si può andare a fare con una coppia di campioni prelevata dai due sistemi è quella di valutare la differenza tra i due campioni per poi effettuare uno **zero mean approach**, che ci permetterà di capire se le medie sono uguali. Formalmente si vede che:

Fatta l'ipotesi:

$$H_0 : \mu_1 = \mu_2 \iff \mu_1 - \mu_2 = 0$$

Vado quindi a valutare la differenza delle medie come:

$$\overline{x}_1 - \overline{x}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_d^2}{n}) \implies \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sigma_d/\sqrt{n}} = z \sim \mathcal{N}(0,1)$$

Warning: La σ_d non è valutata come la somma delle deviazioni standard delle due distribuzioni. Bisogna andare a valutare la deviazione standard della differenza tra le due variabili aleatorie.

Quindi la valutazione del capire se due sistemi siano statisticamente coerenti o meno viene delegata all'analisi della "singola" variabile aleatoria d, che quindi ne permette di trattare tale valore come singolo sample.

Unpaired Observations

Nel caso di osservazioni unpaired, allora il discorso diventa più complesso. Di base non possiamo usare la tecnica della differenza dato che il numero di campioni che si vuole andare a confrontare è differente e di diversa natura. Pertanto quello che si va a fare è valutare in primis le proprie statistiche sui campioni in maniera isolata (quindi si calcolano gli intervalli di confidenza sui singolari sample), e poi, tramite dei test "visivi" si vanno a valutare le casistiche in cui ci si può trovare [4.9]. Altra metodologia che invece può essere d'aiuto è utilizzare le tecniche per le **two-sample hipothesis testing**, che cercano di utilizzare delle "relazioni" tra le distribuzioni per ricavarne delle valutazioni statistiche. Esempi di tali tecniche sono:

- **z-test**: Si va a costruire l'intervallo di confidenza per poi andare a valutare le ipotesi (esempio visto fin'ora). Viene utilizzata principalmente quando si hanno a disposizione campioni con un sample size (n) molto grande e quando la varianza della popolazione è nota. Dato che si è nel caso di considerazione di unpaired observations, sì va a definire la struttura della funzione z, conoscendo le varianze delle due distribuzioni σ_1 e σ_2, si ricava z come:

  $$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  Per andare a definire l'intervallo di confidenza tramite lo z-test, vado a definire il seguente intervallo:

  $$\left[\overline{x} - \frac{\sigma * z_{1-\alpha/2}}{\sqrt{n}}, \overline{x} + \frac{\sigma * z_{1-\alpha/2}}{\sqrt{n}}\right]$$

- **t-test**: Al posto di approssimare il valore ad una normale standard, si vanno ad utilizzare i valori di una funzione t-student che dipenderà sia dal parametro α, che dai suoi gradi di libertà. Viene utilizzata soprattutto quando la varianza della popolazione non è nota e quindi ci si trova costretti ad utilizzare la varianza campionaria. In generale la varianza campionaria viene vista come una funzione chi-squared $\mathcal{X}_k^2$ che non è altro che una composizione di quadrati di variabili aleatorie gaussiane standard, più formalmente, siano $Z_1, Z_2, \ldots, Z_k$, k variabili aleatorie

normali standard ed indipendenti (IID), si definisce formalmente distribuzione chi-squared come:

$$\mathcal{X}_k^2 = \sum_{i=1}^{k} Z_i^2$$

Tale funzione ha un solo parametro che è k che sta ad indicare il numero di gradi di libertà. Andiamo ad enunciare questo poichè alla base del t-test vi è la definizione della t-student, come:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \implies t_{n-1} = \frac{\overline{x} - \mu}{S/\sqrt{n}}$$

dove S è definita come:

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

Con qualche passaggio algebrico, andiamo a definire la distribuzione t_{n-1} in maniera dipendente da z, come:

$$t_{n-1} = \frac{\overline{x} - \mu}{S/\sqrt{n}} = \frac{\overline{x} - \mu}{\sqrt{\frac{S^2\sigma^2(n-1)}{n\sigma^2(n-1)}}} = \frac{z}{\sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}}} = \frac{z}{\sqrt{\frac{k}{(n-1)}}}$$

facendo in questo modo ho ottenuto la distribuzione $k = \dfrac{S^2(n-1)}{\sigma^2}$, che è una funzione chi-quadrata con gradi di libertà pari a $n-1$. Ciò sta a dimostrare come la distribuzione t_{n-1} sia una t-student con n-1 gradi di libertà. In questo caso, invece, si va a definire l'intervallo di confidenza come:

$$\left[\overline{x} - \frac{S * t_{1-\frac{\alpha}{2},\, n-1}}{\sqrt{n}},\ \overline{x} + \frac{S * t_{1-\frac{\alpha}{2},\, n-1}}{\sqrt{n}}\right]$$

Non si può definire una di queste tecniche come dominante rispetto all'altra, in generale la loro applicazione dipende fortemente dal contesto in cui ci si trova e la situazione in cui ci si trova. In generale il "filtro" per il metodo da scegliere viene fatto in base a diversi parametri, come la conoscenza della varianza della/delle popolazioni, se l'ipotesi che viene fatta richiede il confronto di due popolazioni o di una singola, la dimensione della sample size, ecc. Diciamo che i parametri sono vari. In maniera piuttosto empirica si sono trovate delle corrispondenze nelle metodologie di utilizzo e la loro condizione associata. Tale relazione è evidenziata all'interno dell'immagine [4.15]

Z -test		
One-sample z-test	$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	(Normal population **or** *n* large) **and** σ known.
Two-sample z-test	$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	(Normal population or *n* large) **and** independent observations **and** σ_1 and σ_2 are known
One-sample t-test	$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$	(Normal population **or** *n* large) **and** σ unknown
Paired t-test	$t = \frac{\overline{x_1} - \overline{x_2}}{S_d/\sqrt{n}}$	(Normal population of differences **or** *n* large) **and** σ unknown
Two-sample t-test (pooled)	$t = \frac{\overline{x_1} - \overline{x_2}}{S_p\sqrt{1/n_1 + 1/n_2}}$	(Normal pop. **or** $n_1 + n_2 > 40$) **and** independent observations **and** $\sigma_1 = \sigma_2$ unknown *S_p is the pooled std*
Two-sample t-test (unpooled)	$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	(Normal pop. **or** $n_1 + n_2 > 40$) **and** independent observations **and** $\sigma_1 \neq \sigma_2$ both unknown

Figure 4.15: Tabella di scelta della tipologia di test

4.5.6 Esempio di utilizzo dello Z-test

Considera di essere un data analyst per la tua compagnia, e di dover valutare il **tempo di risposta**, che è la chiave del servizio, dato il **nuovo rilascio del software**. Si vuole quindi comparare se il nuovo update ha **cambiato significativamente** la response time rispetto al **sistema precedente**.

Verifica delle assunzioni

Nel nostro caso:

- **Data (nel nostro caso la media campionaria)**: segue una distribuzione normale (ciò ci viene garantito dal fatto che si avrà una sample size molto grande e che quindi, applicando il teorema del limire centrale, si potrà dire che i dati seguono una normale)
- **La deviazione standard**: si conosce in maniera approssimativa andandola a calcolare mediante la **sample standard Deviation**, che è una buona stima della deviazione standard della popolazione data la grande dimensione del campione
- **Independent samples**: Ogni misurazione fatta della response time non è dipendente in alcun modo dalle altre

Definire le ipotesi

La definizione delle ipotesi è la seguente:

- H_0: l'aggiornamento non ha alcun effetto sul response time (le medie sono uguali):

$$H_0 : \mu_{before} = \mu_{after}$$

- H_1: l'aggiornamento ha ridotto il response time, quindi:

$$H_1 : \mu_{before} > \mu_{after}$$

Valutazione dei dati

Vado a valutare, a partire dai dati a disposizione, i seguenti parametri:

- **Sample Size Before Update**(n_{before}): 100
- **Sample Size after Update**(n_{after}): 100
- **Mean Response Time Before Update** ($\overline{X}_{before}$): 130 ms
- **Mean Response Time After Update** ($\overline{X}_{after}$): 110 ms
- **Sample Standard Deviation Before Update**(s_{before}): 15 ms
- **Sample Standard Deviation After Update**(s_{after}): 12 ms

Effettuazione dello Z-test

Dato che si vuole effettuare il **two-sample z-test**, che ha come sua formula la seguente:

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Andando a valutare la deviazione standard, si ottiene anche lo **standard error**, che è possibile calcolare come:

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{s_{before}^2}{n_{before}} + \frac{s_{after}^2}{n_{after}}} = \sqrt{\frac{15^2}{100} + \frac{12^2}{100}} = \sqrt{3.69} \approx 1.92$$

Calcolato lo standard error vado a valutare la Z-statistic, come:

$$z = \frac{\overline{X}_{before} - \overline{X}_{after}}{SE} = \frac{130 - 110}{1.92} \approx 10.42$$

Una volta valutata la statistica, si vanno a valutare i valori critici e la regione critica. Per fare ciò si parte dal prefissare un **livello di confidenza**, per l'esempio si sceglierà del 5% ($\alpha = 0.05$). Dato che nel nostro caso siamo in una one-tailed test (consideriamo solo une delle due code, da come è impostata l'ipotesi), allora si va a calcolare il valore critico come:

$$Z_\alpha = 1.645$$

(*Tale valore è stato prelevato da una tabella specifica*)
Ma, dato che la z-statistic è 10.42, che è maggiore di 1.645, allora si va a rifiutare l'ipotesi nulla.

Conclusion

Alla fine il risultato ci suggerisce che l'aggiornamento che è stato effettuato al software ha ridotto in maniera significativa il response time, supportando l'ipotesi che la nuova versione ha portato ad un incremento delle performance.

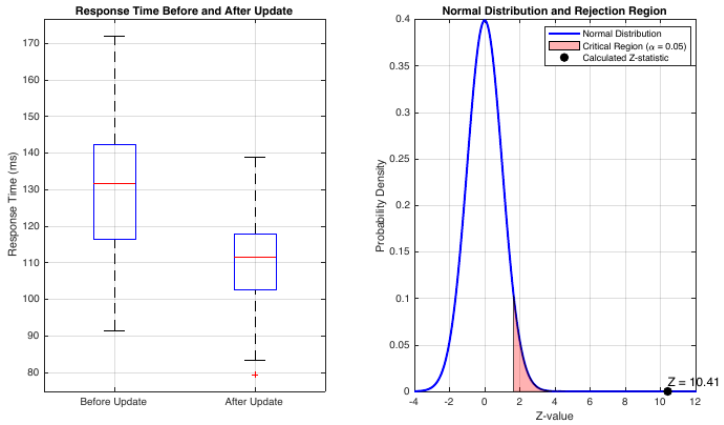


Figure 4.16: Grafici e boxplots associati all'esempio svolto

Chapter 5

Modelli di Regressione

Un **Modello di regressione**, permette di poter valutare il valore di una variabile come funzione di ulteriori altre variabili. Per essere precisi, si definiscono:

- **Response variables**: Le variabili che vengono stimate dal modello
- **Predictor variables**(o predictors o factors): Variabili utilizzare per andare a valutare le Response variables

I modelli regressivi possono essere di due principali tipologie:

- **Lineare**: Modelli più utilizzati nella pratica. In generale vanno ad utilizzare tutti i parametri, ma molto spesso (ed anche in questo corso), si possono considerare dei modelli che si basano su un singolo predictor variable, tali modelli sono più semplici e vengono chiamati **simple linear regression models**
- **Non Lineare**: Modelli non molto utilizzati che fanno uso delle relazioni non lineari tra le predictor variables

Oltre ad andare a definire i modelli e la loro struttura, bisogna definire quanto un modello sia qualitativamente buono rispetto ai dati utilizzati, quindi bisogna definire delle metriche di **qualità del modello**. Tali tecniche permettono di capire se il modello trovato è affine ai dati utilizzati per calcolarlo. Più precisamente, possiamo avere due tipologie di approcci al seguente problema:

- **Approccio visuale**: Si va a valutare il modello in base a quanto questo sia "chiuso" rispetto ai valori considerati (quanto questo approssimi bene l'andamento delle mie osservazioni). Per comprendere bene tale concetto si può far riferimento alla figure [5.1]
- **Valutazione degli errori**: In primis devo capire come andare a definire il mio errore. Il modo più semplice per andarlo a definire è mediante la differenza tra il valore modellato su x (quindi la y calcolata) ed il valore reale rispetto alla variabile x. In genereale, l'errore, in questa modalità, viene definito come la distanza verticale tra il valore osservato e quello calcolato

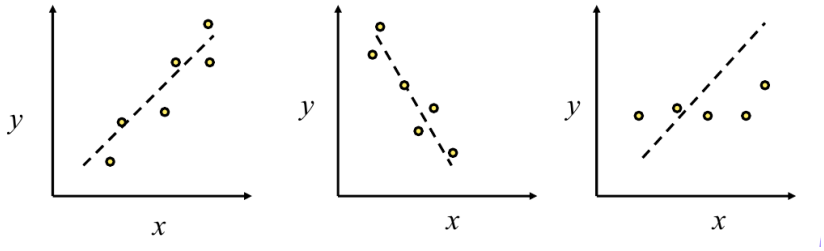


Figure 5.1: Valutazione visiva dei modelli di regressione lineari

5.1 Modelli di regressione lineari semplici

I **simple linear regression model**(modelli di regressione lineare semplici), sono dei modelli che hanno un singolo parameter predictor. Queste presupposizione rende più semplice la trattazione di tali modelli, in particolare, il loro calcolo. Per comprendere meglio come si vanno a stimare i parametri del semplice modello lineare, bisogna passare per i **parametri di qualità del modello**. Più precisamente dalla valutazione dell'errore. Andando ad aggiungere più formalismo:
Considerando le n osservazioni:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

si va a costruire un modello lineare come:

$$\hat{y} = b_0 + b_1 * x$$

basato su tali osservazioni. Per valutare l'errore si prosegue nel calcolo del residuo come:

$$e_i = y_i - \hat{y}_i$$

Con tale definizione, si può proseguire andando ad imporre la condizione, che la somma di tutti i residui sia pari a 0:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - b_0 - b_1 * x_i) = 0$$

Andando a risolvere tale equazione non si troverebbe una **soluzione univoca**, poichè si avrebbe una equazione in 2 incognite, che quindi ha infinito alla 1 soluzioni. Pertanto vi è richiesto di definire un'ulteriore equazione, ovvero quella che fa riferimento all'**SSE**(Sum of Squared Error), che dev'essere minimizzato, ed è definito come:

$$SSE = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - b_0 - b_1 * x_i)^2$$

Questo però è un problema di minimizzazione vincolata, che quindi viene risolto in certi modi utilizzando varie tecniche. Noi utilizzeremo una tecnica pervenuta dall'esame di analisi 1, ovvero andremo a derivare la funzione quadrata e a porla uguale a 1.

Info: *Il prof non ha dato molto peso a questa parte, banalizzandola particolarmente. Di seguito, però, si riporta la dimostrazione per il calcolo dei coefficienti b1 e b0*
Partendo dalla media dei residui e ponendola uguale a 0, si ottiene che:

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 * x_i) = \overline{y} - b_0 - b_1 * \overline{x} = 0$$

Da cui:

$$b_0 = \overline{y} - b_1 * \overline{x}$$

Andando a sostituire tale definizione nella formula di calcolo dell'errore si avrebbe:

$$e_i = y_i - \overline{y} + b_1 * \overline{x} - b_1 * x_i = (y_i - \overline{y}) - b_1 * (x_i - \overline{x})$$

Qui mi fermo e vado a valutare l'errore quadratico, come:

$$\frac{SSE}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left((y_i - \overline{y}) - b_1 * (x_i - \overline{x})\right)^2 =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 + \frac{b_1^2}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{2\ b_1}{n-1}\sum_{i=1}^{n}((y_i - \overline{y}) * (x_i - \overline{x}))$$

Andando a denominare $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}$, $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ e $s_{xy} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}((y_i - \overline{y}) * (x_i - \overline{x}))}$, si ottiene:

$$\frac{SSE}{n-1} = s_y^2 + b_1^2 s_x^2 - 2b_1 s_{xy}^2$$

Andando a derivare rispetto a b_1 si ottiene:

$$\frac{1}{n-1}\frac{dSSE}{db_1} = 2b_1 s_x^2 - 2s_{xy}^2 = 0$$

da cui, si può calcolare b_1 come:

$$b_1 = \frac{s_{xy}^2}{s_x^2} = \frac{\sum_{i=1}^{n}((y_i - \overline{y}) * (x_i - \overline{x}))}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Affiancando a tale equazione quella calcolata in precedenza per b_0, allora si ha la coppia di equazioni adatta per minimizzare l'SSE e valutare un modello di regressione quanto più possibile di qualità

In particolare, la valutazione che vado a fare del parametro b_1 è legata fortemente al **trend** che i miei dati potrebbero avere. Quando b_1 è un valore diverso da 0, allora ciò vuol dire che si è trovato un trend all'interno dei dati. Questo è causato dal fatto che la

valutazione del valore b_1 sia dipendente da un operazione di "correlazione" tra i valori predittori (predictor variables) ed i parametri di risposta (Response variables).

5.1.1 Devianze

Come illustrato in precedenza, l'errore che viene commesso viene valutato come la somma dei quadrati delle deviazioni, il che quindi ci fa pensare alla definizione delle **devianze**. La devianza, in particolare, viene utilizzata all'interno della valutazione dell'errore, perchè rispetta la relazione triangolare, ovvero, la somma delle devianze ottenute da una separazione dei dati, è uguale alla devianza totale. Per applicare tale concetto al caso dei modelli lineari è utile andare a definire i seguenti concetti:

- **SSE**(Sum of Squares for Errors): Si va ad effettuare la somma dei quadrati degli errori (deviazioni) commesse dal modello. Spesso può anche essere chiamato **Residual Sum of Squares**.

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

- **SST**(Total Sum of Squares): Questa è la devianza associata alla variabile y e va a valutare tutta la sua variabilità.

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- **SSR**(Sum of Squares for regression): Devianza associata alla variabile $\hat{y}$ e va a valutare la sua variabilità.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

Tali valori sono legati tra loro mediante la seguente relazione:

$$SST = SSR + SSE$$

che è dimostrabile sia algebricamente che matematicamente, ma ciò prescinde dagli obbiettivi del corso (in gnerale bisognerebbe andare au utilizzare un approccio algebrico matriciale non proprio semplice, ma come detto, tale dimostrazione prescinde dagli obbiettivi del corso).

Info: *Tale dimostrazione prescinde dalle conoscenze del corso. Nonostante sia presente nel materiale fornito a lezione, tale dimostrazione non è richiesta ai fini dell'esame*

Dimostrazione della relazione triangolare

Per capire come effettuare tale dimostrazione, si parte dal ridefinire il seguente problema:

Ipotesi

Si è utilizzato un modello di regressione lineare che calcola b_0 e b_1 partendo dalle seguenti presupposizioni:

$$\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

$$\frac{dSSE}{db_1} = 0 \implies \frac{d(\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2)}{db_1} = -2\sum_{i=1}^{n}(e_i x_i) = 0$$

Dimostrazione

A partire dalla definizione di SST, si effettuano i seguenti passaggi algebrici:

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})]^2 =$$

$$= \sum_{i=1}^{n}[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})] = SSE + SSR + 2\sum_{i=1}^{n} e_i(\hat{y}_i - \overline{y}) =$$

$$= SSE + SSR + 2(\sum_{i=1}^{n} e_i\hat{y}_i - \sum_{i=1}^{n} e_i\overline{y})$$

Da questo risultato, date le ipotesi, si può dedurre che il secondo termine: $\left[\sum_{i=1}^{n} e_i\overline{y}\right]$ è 0, dato che sarebbe il calcolo della somma degli errori per una costante (che annulla tale somma). Per quanto riguarda invece l'altro termine, questo va analizzato nel seguente modo:

$$\sum_{i=1}^{n} e_i\hat{y}_i = \sum_{i=1}^{n} e_i(b_0 + b_1 x_i) = \sum_{i=1}^{n} b_0 e_i + \sum_{i=1}^{n} b_1 x_i e_i = 0$$

Il primo termine di questa equazione, si annulla per lo stesso motivo del precedente, mentre il secondo termine si annulla per via che b_1 è una costante e che il termine $\sum_{i=1}^{n} e_i x_i$, date le ipotesi, è nullo.

Pertanto, considerate tali "annullamenti" si ricava la tesi, ovvero la relazione triangolare:

$$SST = SSE + SSR$$

Coefficiente di determinazione

Per valutare la qualità di un modello lineare, si possono utilizzare i tre parametri precedentemente presentati, ovvero: SSE, SSR ed SST. Da tali valori possiamo calcolare il **coefficiente di determinazione**, che ci dice la percentuale rispetto alla devianza totale, della devianza presa dall'algoritmo di regressione (la devianza della regressione se è maggiore di quella dell'errore è buon segno, dato che si ha una buona rappresentabilità dei dati che si ha in ingresso). Per il calcolo di tale coefficente, quindi, si effettua il seguente rapporto:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

Più R^2 è alto e più il modello lineare è di qualità. Vi è però un problema, R^2 cresce al crescere degli elementi considerati, il che potrebbe sembrare un bene, ma non è così, dato che non è la cardinalità della popolazione che mi dice che c'è un trend trai i dati. Pertanto, si è definita una versione normalizzata della R^2, detta **adjusted $\mathbf{R^2}$**.

Adjusted $\mathbf{R^2}$

L'**Adjusted $\mathbf{R^2}$** va a valutare la percentuale di devianza che è stata coperta, essa, però, tiene conto anche del numero di parametri indipendenti all'interno del mio dataset. Più parametri dipendenti ho e maggiore potrebbe essere la varianza catturata dal trend. Quindi, considerando k il numero di elementi indipendenti del mio dataset ed n il numero totale di elementi. Sì definisce adjusted R^2 come:

$$\overline{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

Dove non tengo solo conto della quantità di dati, ma anche della loro indipendenza. Più i dati sono dipendenti e più il trend che si va a trovare è conforme ai dati da rappresentare, mentre, più questi sono indipendenti, e più risulta difficile andarli a valutare. Ciò, quindi va ad aggiungere un "peso specifico" alla variabile aggiunta rispetto alle altre.

Deviazione standard degli errori

Una volta ottenuta la somma di tutti gli errori, ci è richiesto di andare a calcolare anche il possibile intervallo di confidenza associato a tali valori. Per farlo, in primis, bisogna definire come calcolare la deviazione standard. Per farlo si prosegue con la seguente formula:

$$s_e = \sqrt{\frac{SSE}{n - 2}}$$

Si divide l'SSE per $n - 2$ poichè questi sono i gradi di libertà associati all'SSE. Ciò è dovuto principalmente al fatto che si sono utilizzati 2 gradi di libertà per determinare i valori di b_0 e b_1. Oltretutto, si può definire anche l'**MSE**(Mean Squared Error) come: $MSE = \frac{SSE}{n - 2}$.

Una cosa curiosa da notare è come possiamo andare a valutare i gradi di libertà sommando i vari gradi di libertà associati ad ogni varianza. $n - 1 = (n - 1) + 1$ che sarebbero i gradi di libertà di $SST = SSE + SSR$

5.1.2 Parameters and Statistics

Fino ad ora abbiamo ragionato con le statistiche b_0 e b_1 (quindi questi valori fanno riferimenti ai risultati ottenuti su un singolo campione di osservazioni). In generale il modello associato alla popolazione, quindi la valutazione dei parametri, viene effettuata in altri modi. I parametri associati alla popolazione, in particolare, vengono denotati come:

$$y = \beta_0 + \beta_1 x$$

dati tali valori, ho bisogno di trovare un modo per correlare le statistiche con i parametri del mio problema. Pertanto, come per altri problemi di statistica inferenziale, quello che si va a fare è sostenzialmente, valutare un intervallo di confidenza con un certo livello di confidenza, e sulla base di tali intervalli, trovare un possibile valore per i parametri associati alla popolazione.

Gli intervalli di confidenza vengono gestiti mediante una t-student, dato che non si conosce la "varianza" dei dati originali (Non ho ancora trovato il modello originale). Pertanto, vado a calcolarmi le devizioni standard associate al parametro b_0 e b_1 come:

$$s_{b_0} = s_e \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \right]^{1/2}$$

$$s_{b_1} = \frac{s_e}{[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2]^{1/2}}$$

Da cui, prefissando un livello di confidenza pari a $100*(1-\alpha)\%$, posso ricavare l'intervallo di confidenza per entrambi i parametri. Per farlo vado a considerare in primis il valore della t-student per $1 - \alpha/2$ con $n - 2$ gradi di libertà. ($t_{1-\alpha/2;n-2}$). Da cui ricavo gli intervalli di confidenza come:

$$b_0 \mp t s_{b_0}$$

$$b_1 \mp t s_{b_1}$$

Bisogna fare attenzione a tali intervalli, sopratutto per il parametro b_1, poichè, se l'intervallo di confidenza contiene 0, nulla lo esclude come valore, il che potrebbe pregiudicare la mancanza di un trend tra i dati.

Info: *Questa parte il prof in aula l'ha skippata. Viene aggiunta solo a scopo di completezza e conoscenza*
Precedentemente si è valutato l'intervallo di confidenza dei parametri. Ma si potrebbe andare a valutare anche l'intervallo di confidenza della media dei valori che possono essere assunti da un algoritmo di regressione. Pertanto. Sì definisce m il numero di osservazioni totali nella popolazione, n il numero di osservazioni del singolo campione, definito tutto ciò, si definisce la deviazione standard associata alla media dei valori y come:

$$s_{\hat{y}_{mp}} = s_e \left[\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \right]^{1/2}$$

5.1.3 Visual Tests for Assumptions

Andare ad effettuare la regressione richiede he siano verificate una serie di assunzioni rispetto alle variabili che si stanno andando a considerare. In generale le assunzioni che si vanno ad effettuare sono:

- **Relazione lineare**: La relazione tra la variabile y e la variabile x è lineare
- **Predictor non stocastico**: La variabile x dev'essere non-stocastica (il che vuol dire che non ha un comportamento casuale) e non deve avere alcun tipo di errore di misurazione
- **Indipendenza degli Errori**: Il modello degli errori dev'essere indipendente
- **Omoschedasticità degli errori**(Homoschedasticity of errors): Gli errori devono essere normalmente distribuiti, la cui media è nulla e la deviazione standard costante.

Info: *Tale pezzo è stato inserito solo a scopo informativo, ma non viene richiesto ai fini dello svolgimento dell'esame*
Omoschedasticità
La omoschedasticità si verifica quando la varianza del termine d'errore ϵ_i è costante per tutte le osservazioni, cioè:

$$Var(\epsilon_i) = \sigma$$

In termini intuitivi, significa che la dispersione dei punti osservati attorno alla retta di regressione è uniforme: gli errori non diventano né più grandi né più piccoli al crescere (o al variare) di x_i. Se invece la varianza degli errori cresce (o decresce) con x_i, si ha **eteroschedasticità**, indice di una modellazione dei parametri non adeguata o di un modello che non cattura correttamente la relazione tra le variabili.

Definite le assunzioni che bisogna fare rispetto ai dati, è utile, andarsi a plottare i valori che si voglione mettere in relazione (x ed y), per valutare se effettuare una regressione lineare o meno. Per effettuare tali test visuali vi sono tecniche per ogni tipologia di assunzione che si vuole verificare.

Relazione Lineare

In prima battuta mi conviene andare a plottare i dati di cui voglio valutare un'algoritmo di regressione, in modo da capire se tra essi sussiste una relazione d'ordine o meno. Per degli esempi fare riferimento alla figura [5.2]

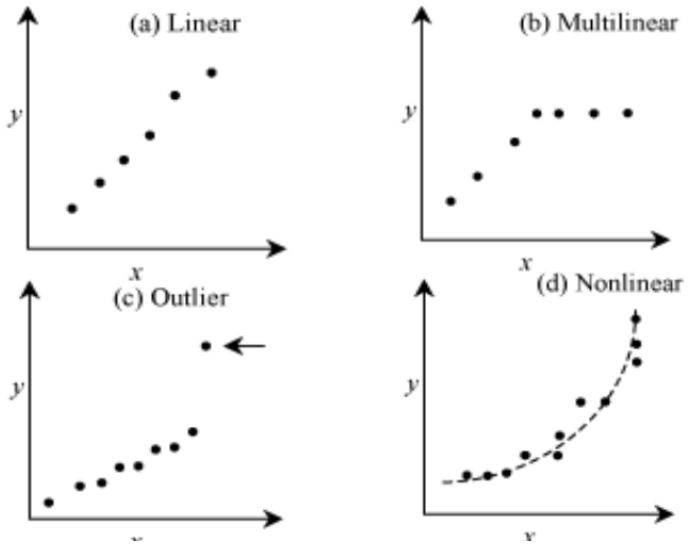


Figure 5.2: Esempi di dati in relazione tra loro

Indipendenza degli errori

Per verificare l'indipendenza degli errori a livello grafico, si va a plottare la distribuzione degli errori e_i rispetto alle corrispettive variabili predette $\hat{y}_i$, ciò viene fatto per verificare se ci sia dipendenza in base alla variabile predetta, e quindi "scagionare" dei trend indesiderati all'interno degli errori. Per comprendere tale concetto, visualizzare l'immagine [5.3], che mostra degli esempi di errori sia indipendenti, che dipendenti in diverse forme.

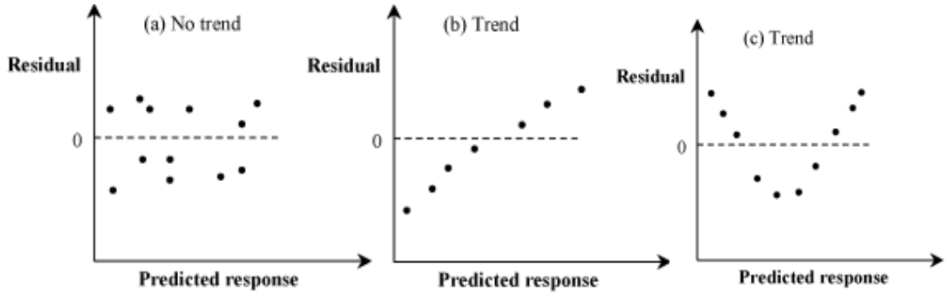


Figure 5.3: Esempi di distribuzioni degli errori rispetto alla variabile $\hat{y}_i$

A volte, è conveniente andare a plottare gli errori e_i in riferimento al numero dell'esperimento stesso i, in modo da verificare se magari ci sono effetti nascosti (errata inizializzazione, procedure sperimentali) o delle condizioni dell'ambiente (temperatura ed umidità) che

vanno ad incidere su tali errori (per fare ciò vado sempre a verificare i possibili trend presenti all'interno di tale relazione). Esempi di tali plot sono quelli presenti in figura [5.4]

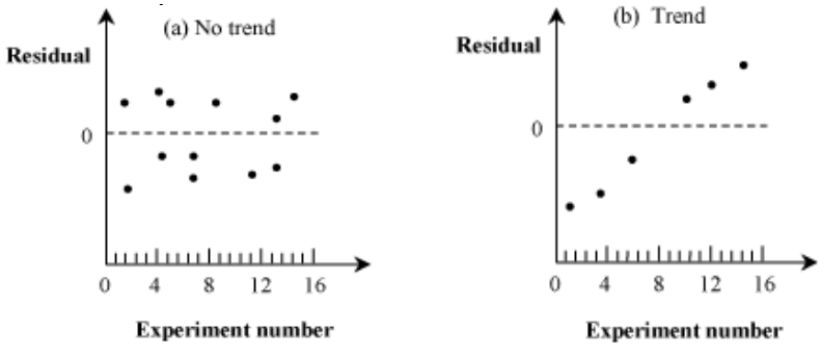


Figure 5.4: Plotting degli errori rispetto alla posizione di valutazione

Errori distribuiti normalmente

Per verificare se gli errori seguono una distribuzione normale, ci viene in aiuto il **Quantile-Quantile plotting**, che mette in relazione la distribuzione ottenuta dagli errori, rispetto ad una distribuzione normale. Esempi di applicazione della QQ-Plot sono rappresentati in figura [5.5]

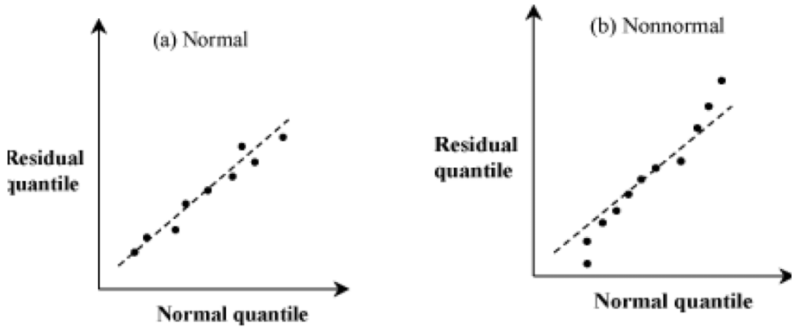


Figure 5.5: QQ-Plot per le distribuzioni degli errori

Deviazione standard costante

Quello che si va a fare, si va a plottare gli errori in riferimento alla risposta predetta (che dipende dalle predictor variables), andando a valutare se ci sono relazioni. Tale test visivo ci permette di verificare l'**Omoschedasticità**(Deviazione standard costante e media nulla) [Questa è la definizione mostrata a lezione e dalle dispense, anche se sui testi e in rete la definizione comprende solo la parte di deviazione standard (solo ipotesi di costanza di tale valore)]. Guardando il grafico, è importante stare attenti ad osservare se c'è un trend degli errori rispetto alla risposta predetta. Poichè in tal caso, non si possono fare ipotesi sulla costanza della varianza. Esempi di tali grafici sono quelli mostrati alla figure [5.6]

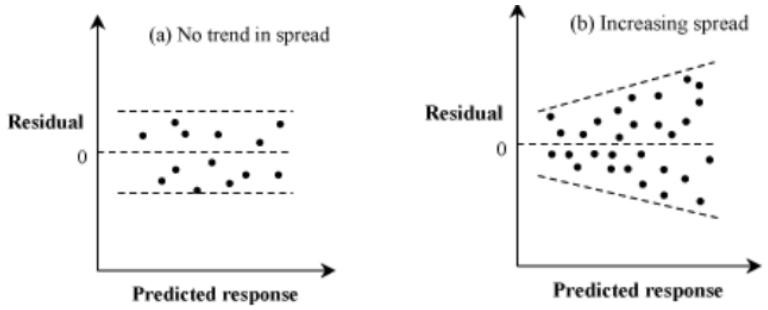


Figure 5.6: Distribuzione degli errori

L'Omoschedasticità si può verificare (o confermare) anche mediante l'utilizzo di uno specifico test statistico. Vi sono varie tecniche e modelli con cui tale test può essere effettuato, tali tecniche vanno sotto il nome di ANOVA (Analysis Of Variance) in letteratura. Ad esempio:

- **p-value:** È la probabilità di osservare un risultato uguale o più estremo di quello ottenuto, assumendo vera l'ipotesi nulla. Nei test di uguaglianza delle varianze indica quanto è plausibile che le varianze siano effettivamente uguali. Se $p < \alpha$ (es. 0.05), si rifiuta l'ipotesi nulla: le varianze non sono uguali.
- **Test di Levene:** Verifica se le varianze di più campioni sono uguali, basandosi sulle deviazioni assolute dalla media (o mediana) di ciascun campione. È più robusto del test di Bartlett rispetto alla non-normalità. Ipotesi nulla: le varianze dei campioni sono uguali.
- **Test di Brown–Forsythe:** Variante del test di Levene che utilizza la mediana invece della media, rendendolo ancora più robusto a outlier e distribuzioni asimmetriche. Ipotesi nulla: le varianze dei campioni sono uguali. È consigliato quando la normalità non può essere assunta.
- **Test di Bartlett:** Test classico per verificare l'uguaglianza delle varianze basato su una statistica χ^2. È potente se i dati sono normali, ma molto sensibile alle deviazioni dalla normalità. Ipotesi nulla: tutte le varianze sono uguali; se $p < \alpha$, si conclude che almeno una varianza differisce.
- Ce ne sono altri ma prescindono da tale corso

Nel caso in cui, invece, i dati siano normali ma Eteroschedatici, allora è conveniente utilizzare il **Welch's t-test**.

Warning: Le specifiche fatte sulle varie tecniche sopraelencate sono state aggiunte per un leggero grado di comprensione in più. Ma non sono richieste al fine di effettuare l'esame

5.1.4 Regressione Lineare Non parametrica

Quando si parla di **Regressione lineare non parametrica**, si intende quell'insieme di tecniche che cerca un trend di un certo tipo, senza fare alcuna ipotesi sulla forma della

popolazione o sulla distribuzione che i dati devono avere, richiede solo che ci sia indipendenza sui dati presenti. Per tali tipologie di test si utilizza il **test di Mann-Kendal**, che va a valutare la tendenza di una serie, e verifica se questa sia monotona o meno. Formalizzando tale concetto, andiamo a definirne le specifiche:
Dato un insieme di una coppia di valori, di cui si vuole verificare un eventuale trend (insieme di punti nel piano), $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$, si definisce un valore S che sarà utilizzato per la verifica. Sì definisce il valore S come:

$$S = \sum_{i<j} sign((x_j - x_i)(y_j - y_i))$$

Quindi si va ad effettuare una somma della valutazione dei parametri (se concordi si somma, se discordi si sottrae). Più precisamente si va a controllare se i punti siano **concordanti** (Concordant, danno un contributo positivo) o **discordanti** (discordant, danno un contributo negativo). Per comprendere tale relazione guardare l'immagine [5.7].

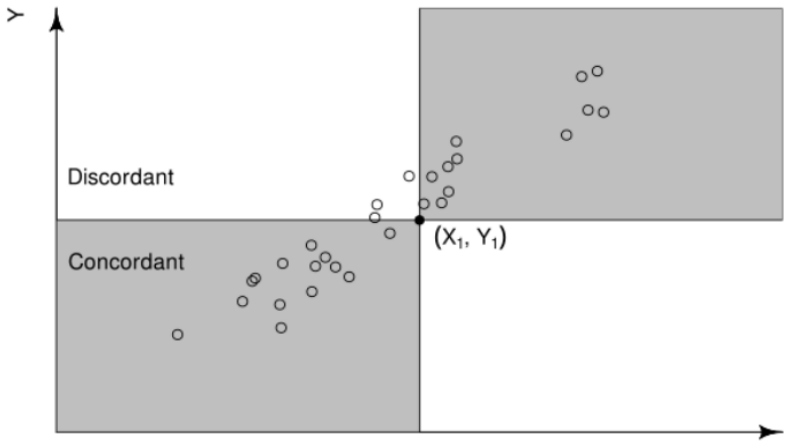


Figure 5.7: Punti concordanti e discordanti

Il valore di S ci permette di capire, in base al segno ed al valore che assume, se i dati hanno una certa tendenza. Per capire al meglio tale tendenza, ci calcola un'altro parametro τ_a che indica il **coefficente di correlazione**. Più S e τ_a sono grandi e più con forza posso rigettare l'ipotesi nulla (quella associata alla mancanza di trend nei dati). Per calcolare la τ_a si effettua il seguente calcolo

$$\tau_a = \frac{S}{\binom{n}{2}} = \frac{S}{\frac{n(n-1)}{2}}$$

Con tale valore, vado a riscalare il valore di S ad un range $-1 \leq \tau_a \leq 1$, poichè è come se stessi andando ad aggiungere un peso per ogni valore "valutato".

Dato che solitamente, si potrebbe incappare nel problema dei "pareggi" (ovvero valori nulli). Si può andare a valutare un'altro parametro che è:

$$\tau_b = \frac{S}{\sqrt{(\binom{n}{2} - n_1)(\binom{n}{2} - n_2)}}$$

Dove n_1 ed n_2 sono il numero di pareggi per ogni campione che si va a considerare (x_i ed y_i).

Procedura di Sen

Un modo per essere più robusti agli outlier e trovare un eventuale trend, è la seguente procedura. Considerato che i dati mostrati **posseggono tra loro un trend**, si vanno a valutare le pendenze, e si seleziona il valore della mediana tra queste. Per la valutazione delle pendenze si prosegue nel seguente calcolo:

$$slope_{ij} = \frac{y_j - y_i}{x_j - x_i}$$

5.2 Altri modelli di regressione

A volte la **Simple linear regression** non riesce a compensare tutti i casi richiesti o leggermente più complessi. Pertanto è utile andare ad esplorare diverse tecniche che ne permettono una buona stima

5.2.1 Regressione Lineare Multipla

A differenza della regressione lineare semplice, dove si cerca una relazione tra una singola variabile predictor e una singola predictable variable. Talvolta, utilizzare tale metodologia può essere limitante. Pertanto si definisce la **Multiple Linear Regression**(Regressione Lineare Multipla), che va a considerare una relazione **lineare** tra diverse variabili predictor. Per ogni campione si effettua una predizione rispetto al numero di parametri che si sono andari a utilizzare (predictor variables). Formalmente, definendo:

- $\mathbf{x_{ij}}$: Valore dell'attributo j-esimo per il campione i-esimo
- $\mathbf{y_j}$: j-esimo valore predetto
- $\mathbf{b_i}$: coefficente associato al parametro i-esimo
- $\mathbf{e_j}$: Errore associato al j-esimo valore predetto

si può scrivere la relazione tra i predictor variables e le predictable variables come:

$$y_j = b_0 + b_1 x_{j1} + b_2 x_{j2} + \cdots + b_k x_{jk} + e_j$$

Oppure in maniera estesa come:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$$

Con la notazione vettoriale

$$\underline{\mathbf{y}} = \underline{\underline{\mathbf{X}}}\, \underline{\mathbf{b}} + \underline{\mathbf{e}}$$

Per il calcolo dei parametri, si cerca di andare a minimizzare la formula dell'errore, ovvero:

$$e_i = y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik}$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2$$

Scrivendo $\sum_{i=1}^{n} e_i^2 = (\underline{y} - \underline{\underline{X}}\ \underline{b})^T (\underline{y} - \underline{\underline{X}}\ \underline{b})$, da cui, derivando rispetto a b, si ottiene:

$$\underline{b} = (\underline{\underline{X}}^t\ \underline{\underline{X}})^{-1} \underline{\underline{X}}^T\ \underline{y}$$

Compreso come effettuare la stima dei valori per effettuare la regressione lineare multipla, si prosegue nell'effettuazione della valutazione degli "errori", delle veriazioni e delle devianze. Precisamente, partendo dalle variazione e dalle devianze, si ha che:

$$SSY = \sum_{i=1}^{n} y_i^2, SS0 = n\overline{y}^2$$

$$SST = SSY - SS0$$

$$SSR = SST - SSE$$

Dove i gradi di libertà della SST sono sempre $n-1$ mentre quelli della SSE sono $n-k-1$. Si valutano, anche, il coefficiente di determinazione e la deviazione standard, come:

$$R^2 = \frac{SSR}{SST}$$

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

Per il calcolo del **MSE**(Mean squared error), si effettua la seguente formula:

$$MSE = \frac{SSE}{n-k-1}$$

Mentre la deviazione standard dei parametri, viene valutata come:

$$s_{b_j} = s_e \sqrt{c_{jj}}$$

dove c_{jj} è il j-esimo elemento diagonale della matrice $\underline{\underline{C}} = (\underline{\underline{X}}^T \underline{\underline{X}})^{-1}$

5.2.2 Multicolinearità

La **Multicolinearità** vuol dire effettuare la regressione con una dipendenza lineare tra le variabili predictor (tale situazione è detta colineare). Tale situazione, ci permette di capire che si avrebbe una variabile in più su cui si sta effettuando la regressione (quindi di per se, inutile). Quindi, solitamente, è utile capire se due variabili sono colineari o meno (correlate o meno). Per effettuare tale valutazione si va a calcolare la correlazione tra i due elementi, che poi viene valutata. La correlazione tra i due elementi viene calcolata come:

$$Correlation(x_1, x_2) = R_{x_1 x_2} =$$

$$= \frac{\sum_{i=1}^{n} x_{1i}x_{2i} - \frac{1}{n}\left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} x_{2i}\right)}{\left[\sum_{i=1}^{n} x_{1i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} x_{1i}\right)\right]^{1/2}\left[\sum_{i=1}^{n} x_{2i}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_{2i}\right)\left(\sum_{i=1}^{n} x_{2i}\right)\right]^{1/2}}$$

Quindi, aggiungere una variabile predictor non è detto che sia sempre una buona idea, poichè, la correlazione tra due predittori, riduce l'accuracy e va ad amplificare la varianza (Valori della matrice di correlazione vista in precedenza).

5.2.3 Regressione Curvilinea

Talvolta, può capitare che la relazione presente tra y e x non sia lineare, ma che comunque racchiude una funzione non lineare. Quello che sostanzialmente si può andare a fare, è cercare una relazione lineare rispetto a funzioni non lineari. Per capire, immaginiamo che y ed x siano legate dalla legge $y = bx^{\alpha}$, allore quello che si potrebbe andare a fare, è valutare il modello logaritmico di tale funzione come:

$$ln(y) = ln(b) + \alpha ln(x)$$

Il che quindi mi consente di utilizzare le tecniche affrontate per la regressione lineare, per trovare i valori di b e α. Oltre a tale legame ce ne sono altri, che sono mostrati nella figura [5.8]

Nonlinear	Linear
$y = a + b/x$	$y = a + b(1/x)$
$y = 1 / (a+bx)$	$(1/y) = a + bx$
$y = x(a+bx)$	$(x/y) = a + bx$
$y = ab^x$	$\ln(y) = \ln(a) + (\ln(b))x$

Figure 5.8: Forme di trasformazione di funzioni non lineari in relazioni lineari

5.2.4 Outliers

Bisogna fare attenzione a come andare a trattare gli **outlier**, poiché incidono fortemente sulla regressione. Non si possono escludere automaticamente, perché potrebbero contenere un'informazione importante sul comportamento reale del sistema; eliminarli senza una verifica accurata potrebbe portare a **risultati distorti o conclusioni errate**.

Una cosa che si potrebbe andare a fare è **formulare delle assunzioni sulla distribuzione** dei dati e verificare se l'outlier rispetta tali ipotesi. In caso contrario, potrebbe effettivamente trattarsi di un errore sperimentale o di misura.

Inoltre, è utile **visualizzare i dati con uno scatter plot**, in modo da individuare graficamente eventuali punti anomali. Se si sospetta la presenza di un errore, si può **ripetere l'esperimento o l'analisi** per confermare il risultato.

Un approccio prudente consiste anche nel **ripetere la regressione sia con che senza l'outlier**, confrontando i risultati ottenuti: se le differenze sono significative, l'outlier ha un'influenza importante e va studiato con maggiore attenzione. In alternativa, si possono **suddividere i dati in sottoinsiemi** (ad esempio, diverse regioni operative) e costruire **modelli separati** per ciascun gruppo, in modo da descrivere meglio il comportamento complessivo del sistema.

5.2.5 Errori comuni nella Regressione

Quando si applicano gli algoritmi di regressione, si può cadere in diversi problimi che sono provocati da errori durante la fase di "valutazione" della regressione. Gli errori che

più comunemente vengono fatti sono:

- **Non effettuare verifica visuale**: Se non si vanno ad osservare i dati mentre si applicano delle soluzioni di regressione, si potrebbero avere delle R^2 che sono comunque molto alte, a discapito di una regressione non proprio ottimale. Vedere figura [5.9]
- **Ignorare parametri importanti**: Nonostante la visualizzazione di un grafico sia fondamentale, è altrettanto importante andare a valutare anche gli **intervalli di confidenza** e il **coefficente di determinazione**R^2
- **Errata selezione dei predittori**: Sì può ricadere nell'errore di andare a valutare troppi predittori che aumentano la complessità totale del sistema e, se correlati, potrebbero essere ridondanti per l'algoritmo di regressione
- **Utilizzo improprio dei valori**: Si vanno a considerare dei valori solo nel loro range di misurazione ma non nella totalità del range delle operazioni

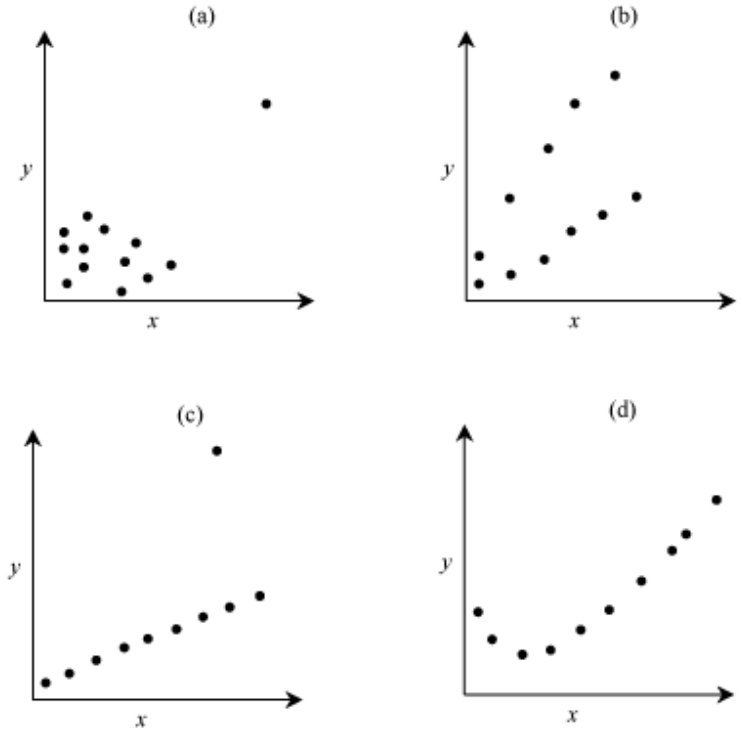


Figure 5.9: Curve con alto R^2 ma non opportunamente "regressibili"

Part II

Esercitazioni

Chapter 1

Web Server

FNS