SERVIZI INTEGRATI DI INGEGNERIA

Cp = 2.0
Cpk = 1.5
ppm = 3.4

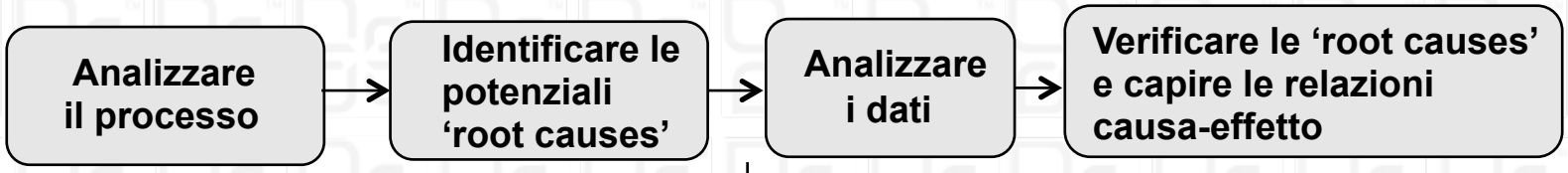
Graphical Tools
- Analizzare i dati graficamente

COCCO SERVIZI INTEGRATI DI INGEGNERIA SRL
VIA STAZIONE 2 - 10090 ROSTA (TO) ITALIA
TEL. +39.011.954.0208 - FAX +39.011.956.7863
INFO@COCCOINGEGNERIA.IT

SOCIETÀ DI INGEGNERIA UNIPERSONALE
CAPITALE SOCIALE INT. VERSATO 10.000 EURO
REG. IMP. DI TORINO R.E.A. N.1017155
PARTITA IVA E C.F. 09000840018

Ing. Pier Giorgio DELLA ROLE
<<Master Black Belt >>

L’analisi del processo e dei dati



Il processo

Le prime due fasi sono chiamate anche “process door” perchè hanno lo scopo di capire e ottenere informazioni direttamente dal processo stesso.

Gli strumenti sono indirizzati quindi ad analizzare e capire come il processo funziona.

I dati

Le ultime due fasi sono chiamate anche “data door” perchè hanno lo scopo di ottenere informazioni direttamente dai dati raccolti.

Gli strumenti comprendono quindi una serie Tools grafici e statistici per analizzare i dati.

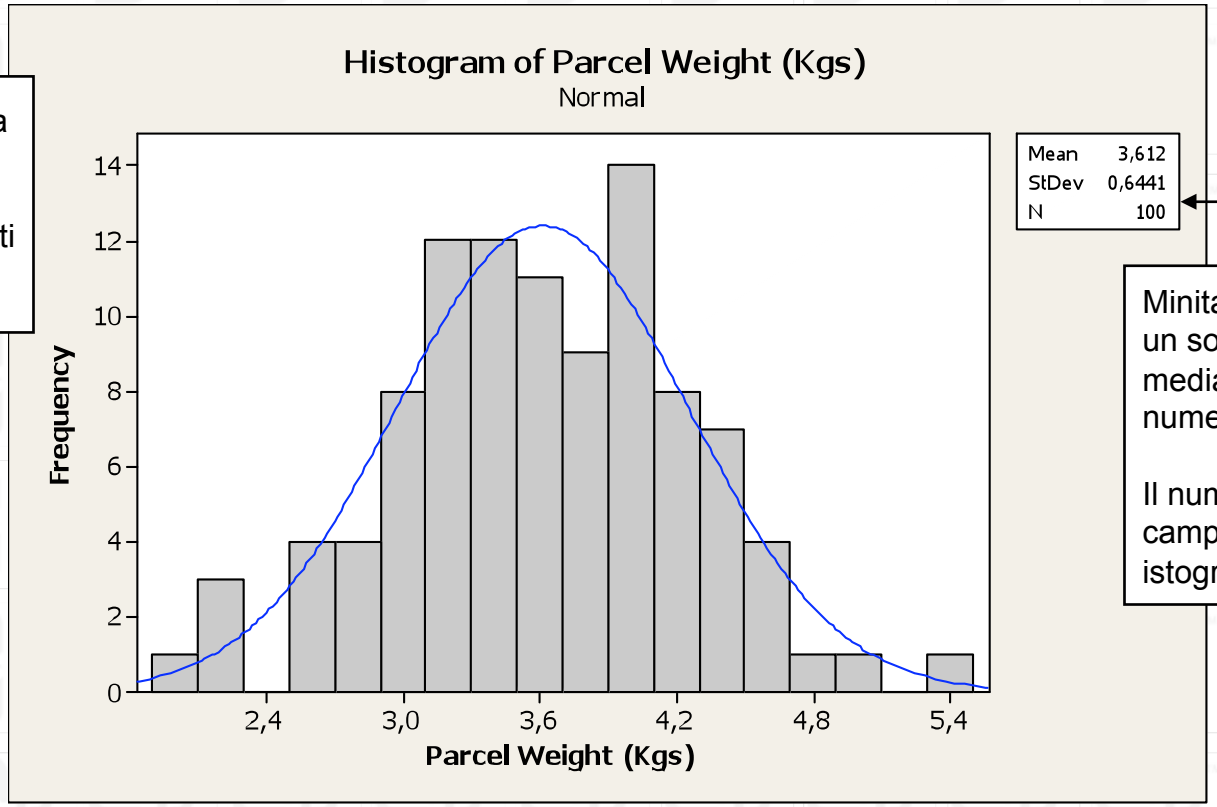
Introduzione

Gli strumenti grafici sono il punto di partenza per analizzare i dati. Teorie ed idee provenienti dall'analisi grafica saranno successivamente investigate con più sofisticati strumenti statistici.

1. Analizzare distribuzioni	- Histogram/Dot Plots - MINITAB's Graphical Summary - Probability Plot
2. Scoprire variazioni legate al tempo	- Time Series Plot (Run Charts)
3. Confrontare distribuzioni o gruppi di dati	- Box Plot - Individual Value Plot
4. Scoprire relazioni tra due set di dati	- Scatter Plot - Matrix Plot
5. Analizzare differenti categorie, sottogruppi, "root causes" di dati	- Pareto Chart

Istogrammi (MINITAB: Graph > Histogram)

Questo istogramma rappresenta la distribuzione dei pesi di pacchi gestiti da una Azienda di spedizioni.



Minitab 14 mostra un sommario per la media, sigma e numero di campioni

Il numero minimo di campioni per un istogramma è 25

Interpretare gli istogrammi: descrivere i risultati usando frasi del tipo:

"questo istogramma indica che I pesi dei pacchi variano da un minimo di 2 kg fino a 5,5 kg, mentre la maggior frequenza è compresa tra 3 e 4,5 kg"

"la distribuzione appare simmetrica attorno alla media di 3,6 kg e sembra essere normale"

Istogrammi

Alcune regole:

1. Servono almeno da 50 a 100 misure per valutare una distribuzione mediante istogramma.
   Le valutazioni riguardano la tendenza centrale (media), la variabilità (spread) e la forma (shape).

2. Il numero di classi è solitamente calcolato come:

$$k = \sqrt{n}$$

n = numero misure

3. La larghezza della classe è data dalla seguente formula:

$$H = \frac{R}{k}$$

R = range = Valore max – Valore min
K = numero di classi

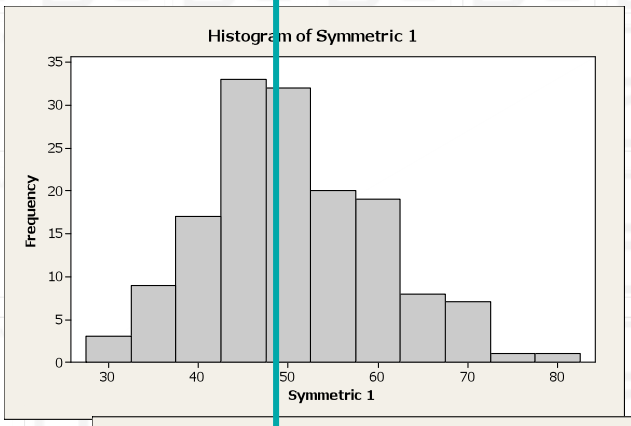
4. Determinare i confini della classe

   . Supponiamo che il valore minimo sia 9.00
   . H = 0.17 che viene arrotondato a 0.20
   . I confini della prima classe saranno 9.00 e 9.19 con punto centrale 9.10
   . I confini della seconda classe saranno 9.20 e 9.39 con punto centrale 9.3

Interpretare gli Istogrammi

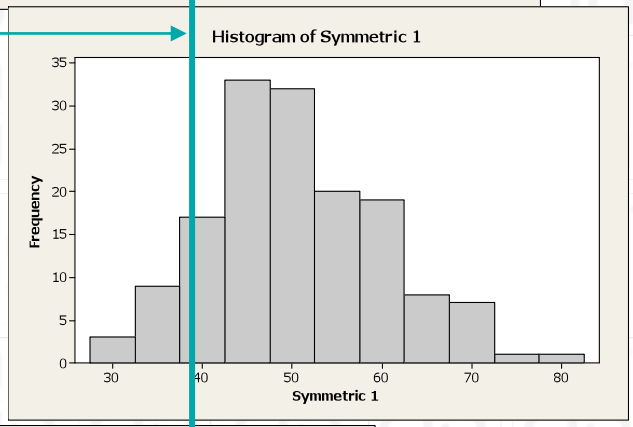
tendenza centrale (centering)

centrato

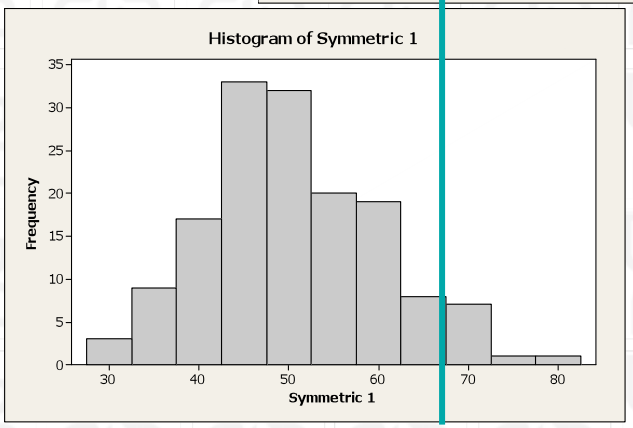


Specifiche cliente

Processo troppo alto



Processo troppo basso

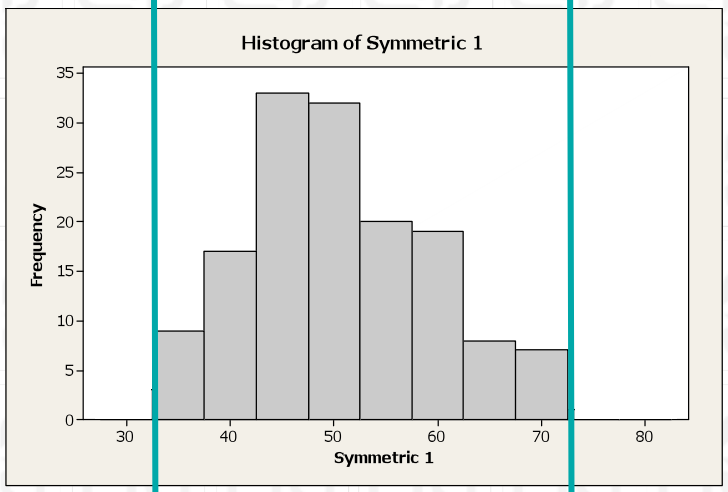


Interpretare gli Istogrammi

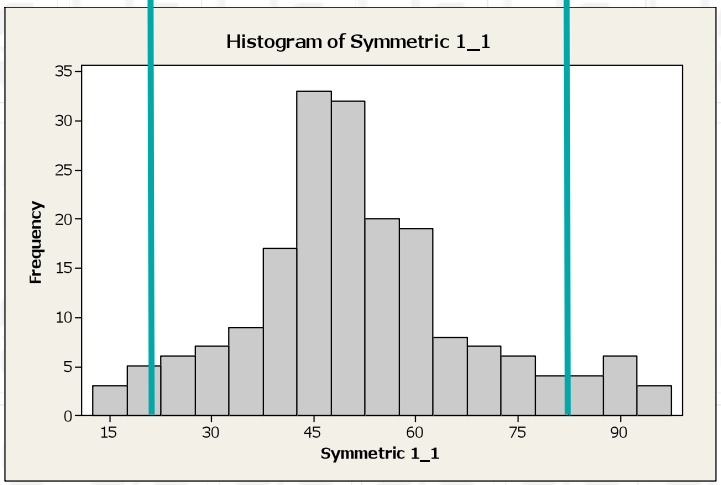
Variabilità (spread)

Specifiche cliente

Processo entro le specifiche cliente

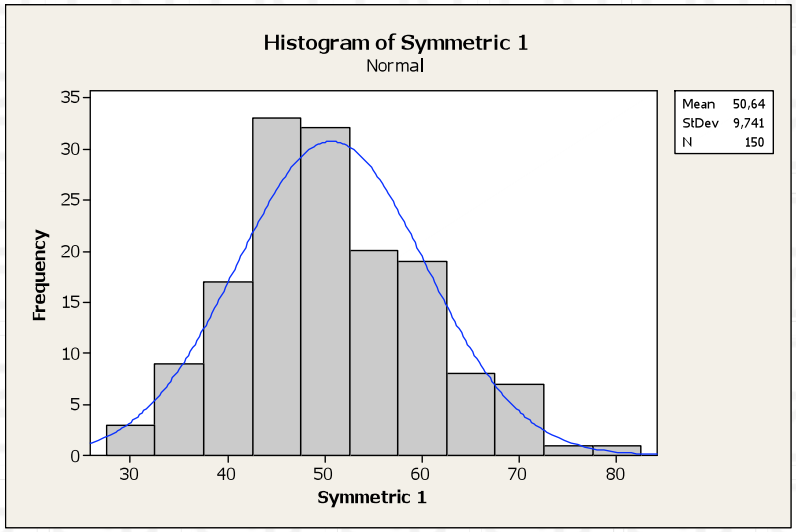


Processo troppo variabile

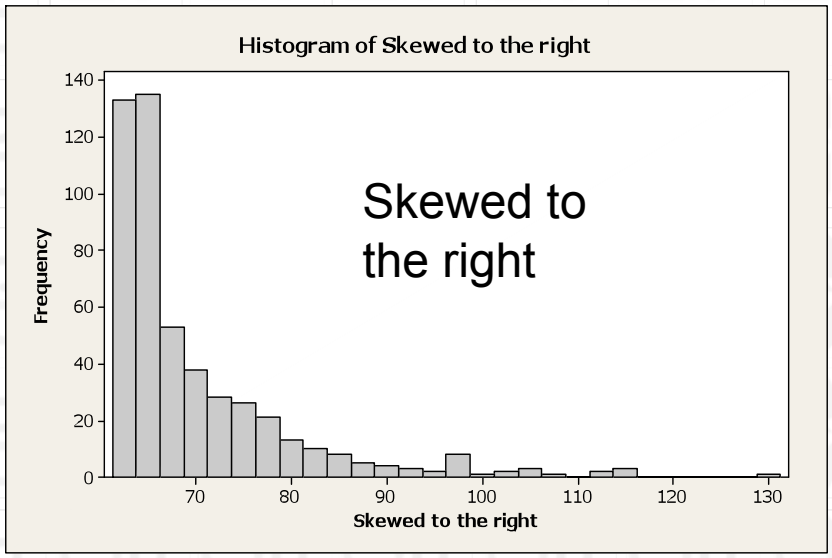


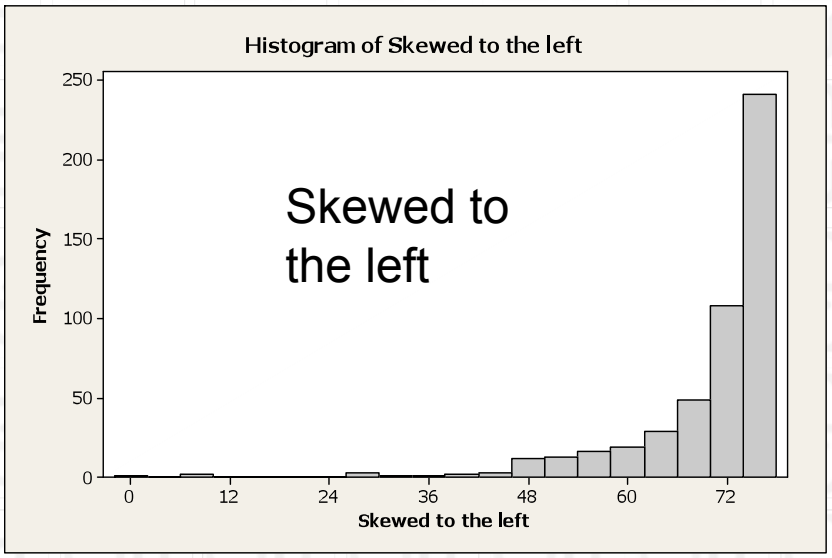
Interpretare gli Istogrammi

Forma (shape)



Distribuzione normale



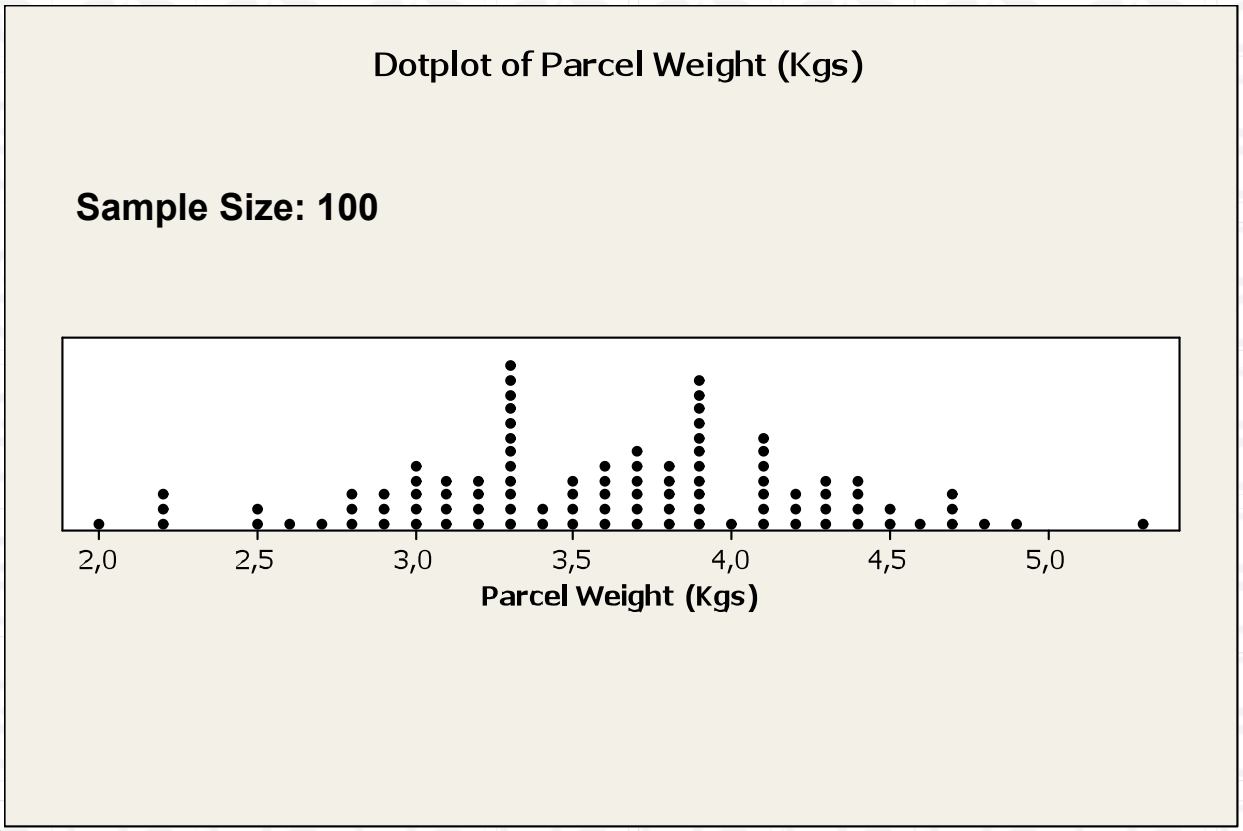


Dot Plots

(MINITAB: Graph > Dotplot)

Questo Dot Plot mostra gli stessi dati sui pesi dei pacchi visti prima, però qui ogni punto rappresenta un valore misurato

Il Dot Plot usa un differente numero di colonne rispetto all'istogramma e quindi la rappresentazione è diversa.



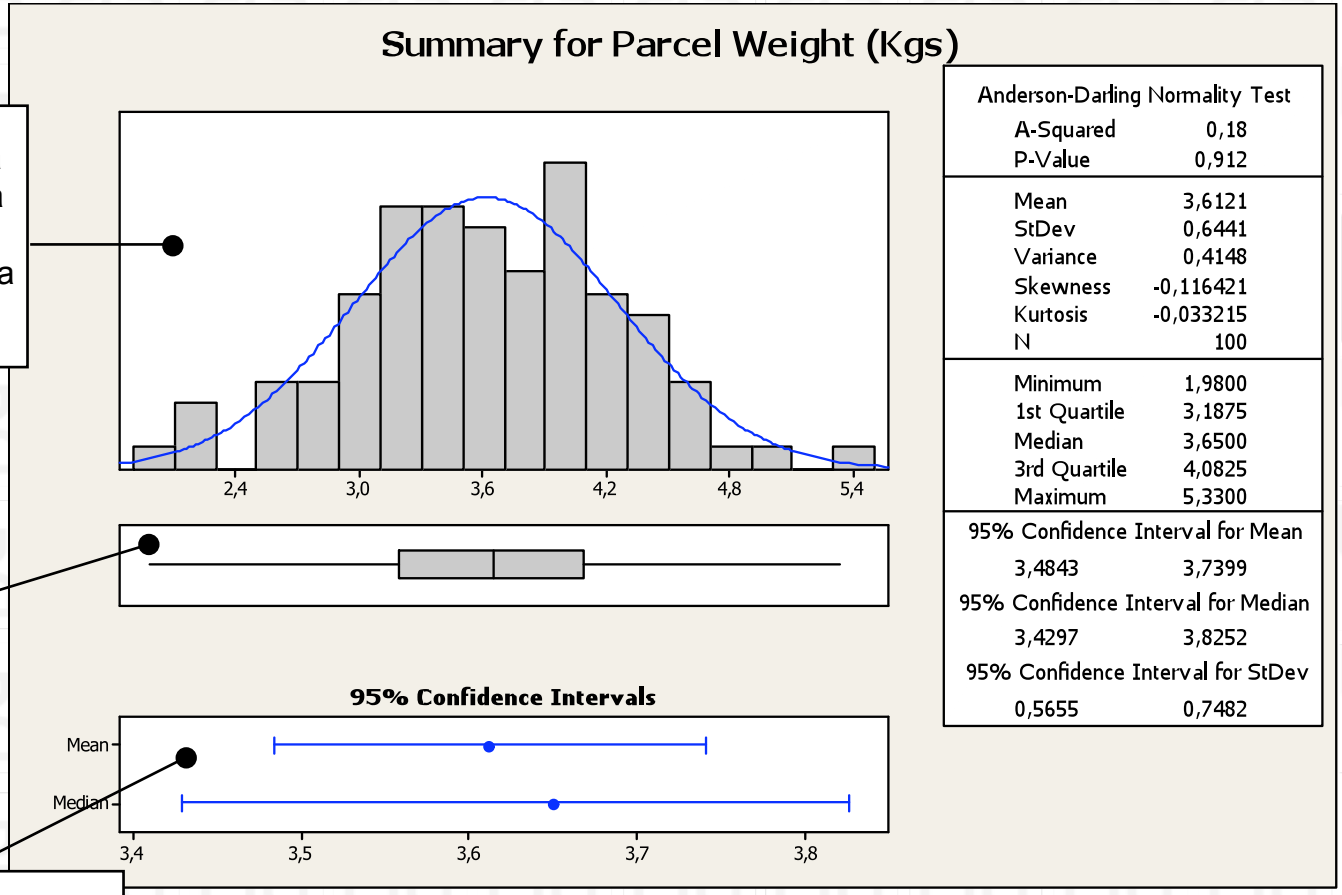
Dot Plot o Istogramma?

Sono entrambi distribuzioni di frequenza ed il loro uso dipende dalle applicazioni:

. L'Istogramma riassume bene i dati ed utile per un primo esame
. Il Dot Plot mostra un punto per ogni dato (più dettaglio) e permette l'uso della funzione "brush" di MINITAB per indagare punti anomali

MINITAB's Graphical Summary

(MINITAB: Basic Statistics > Graphical Summary



L'**Istogramma** mostra un processo che varia tra 2 e 5,5 kg e la cui distribuzione è vicina a quella normale.

Il **box plot** mostra la distribuzione dei dati ed usa le stesse scale dell'istogramma.

Gli **intervalli di confidenza** per la media e la mediana hanno invece una propria scala.

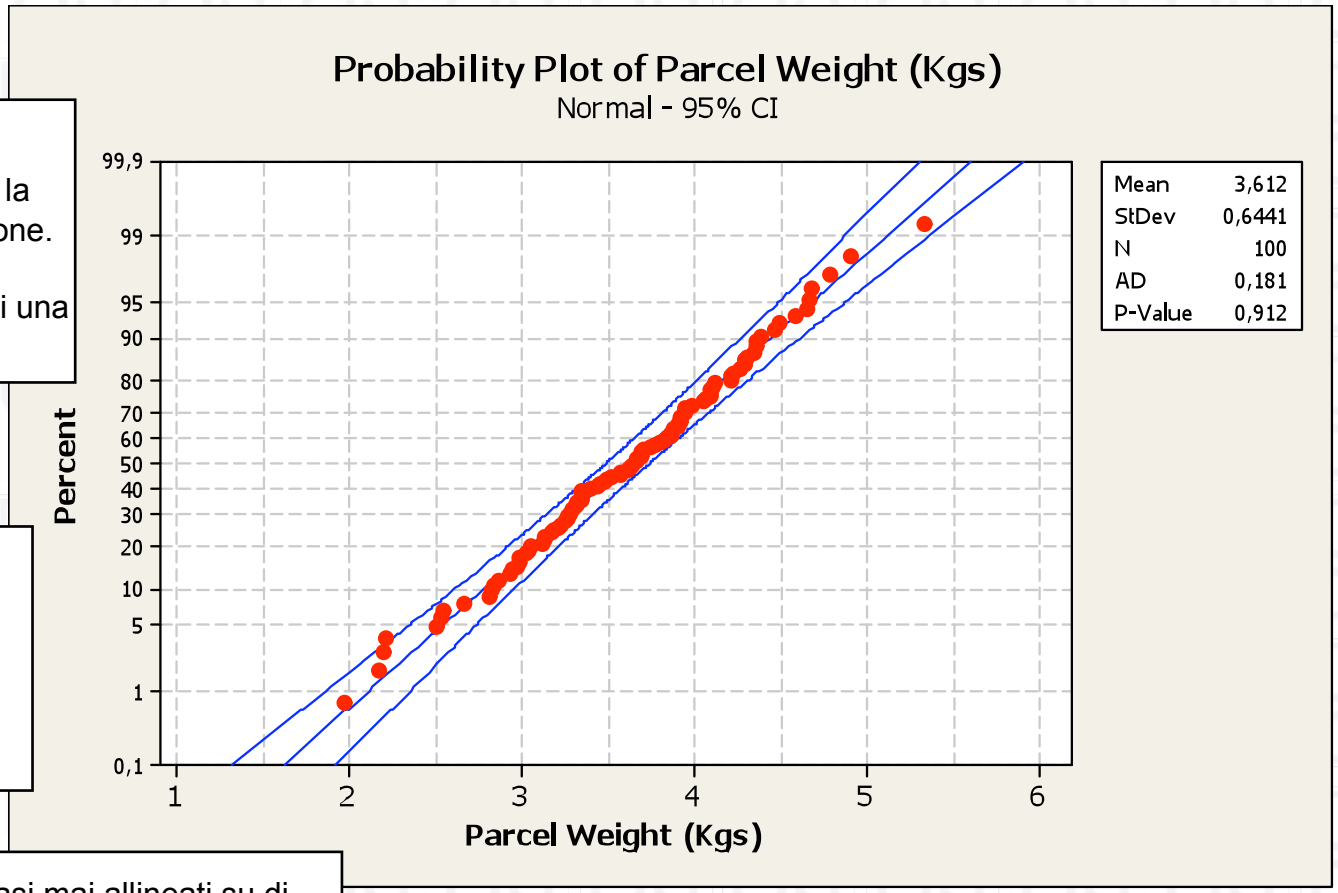
Gli Intervalli di Confidenza (CI) hanno il seguente significato:

- Possiamo essere 95% confidenti che la media dei pesi dei pacchi cade all'interno dei limiti dell'intervallo di confidenza qui indicati

Probability Plots

(MINITAB: Graph > Probability Plot)

Probability Plots sono grafici per decidere circa la normalità della distribuzione. I dati sono normalmente distribuiti se cadono su di una linea retta.

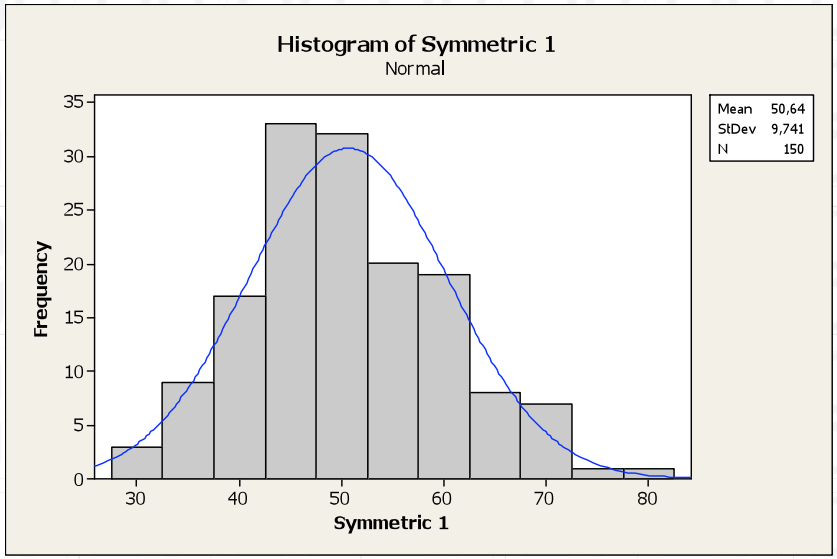


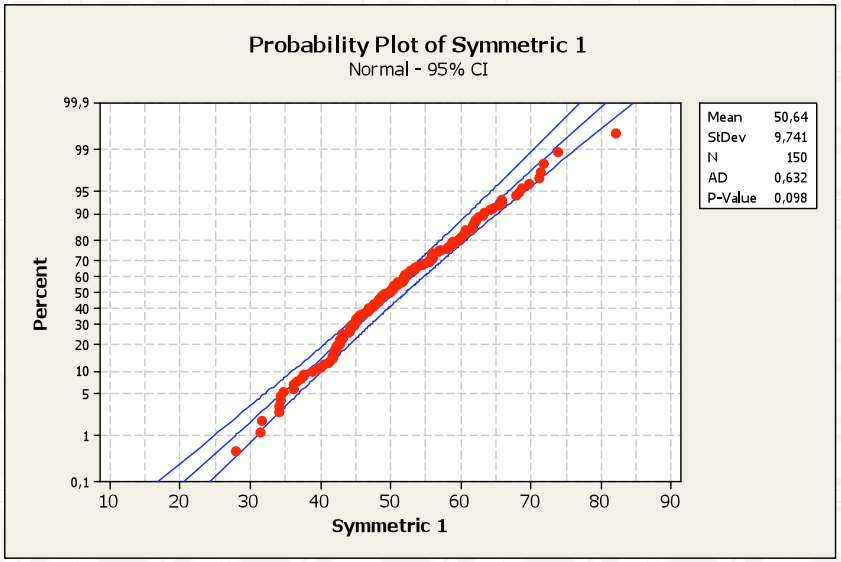
L'asse verticale rappresenta l'area sotto la curva normale (cumulative probability) e la sua scala è logaritmica.

I dati non saranno quasi mai allineati su di una retta, per questo motivo MINITAB mostra anche i limiti dell'Intervallo di Confidenza (95%) e **se tutti I punti cadono all'interno della banda** possiamo assumere che la distribuzione è normale.

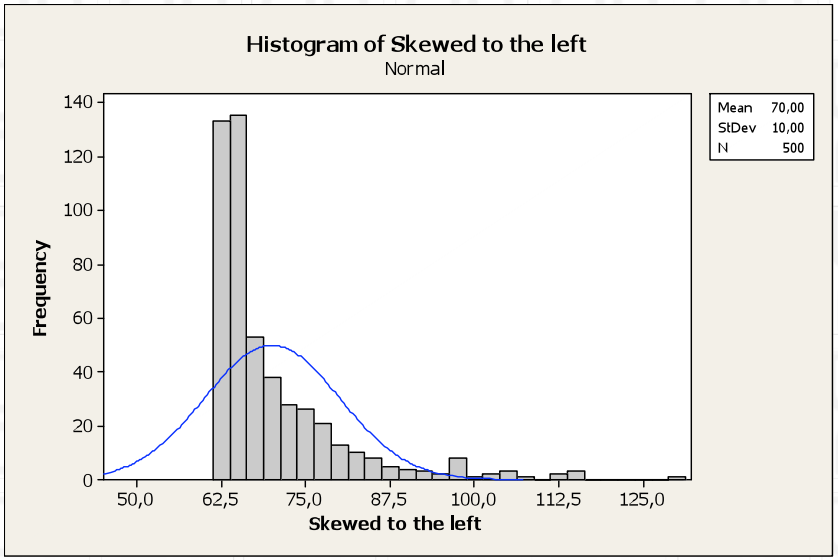
Probability Plots

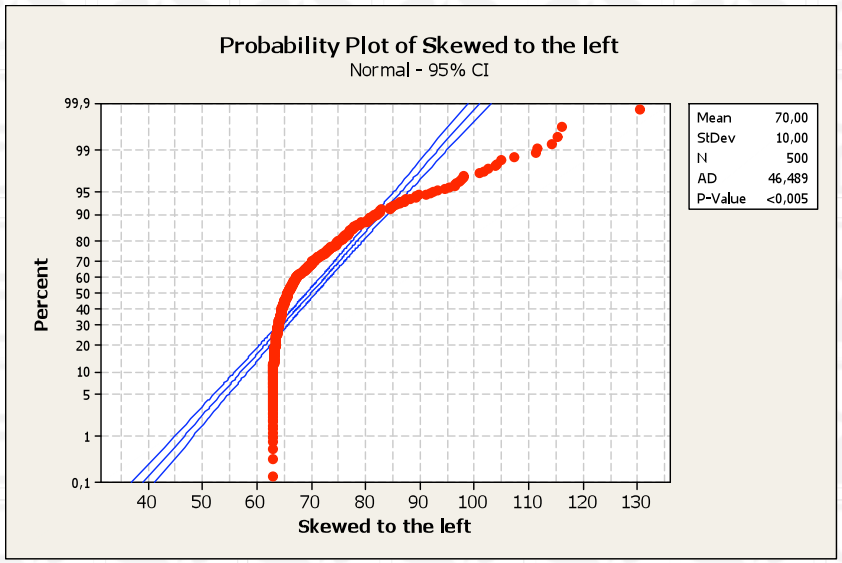
Distribuzione Normale





Distribuzione non normale





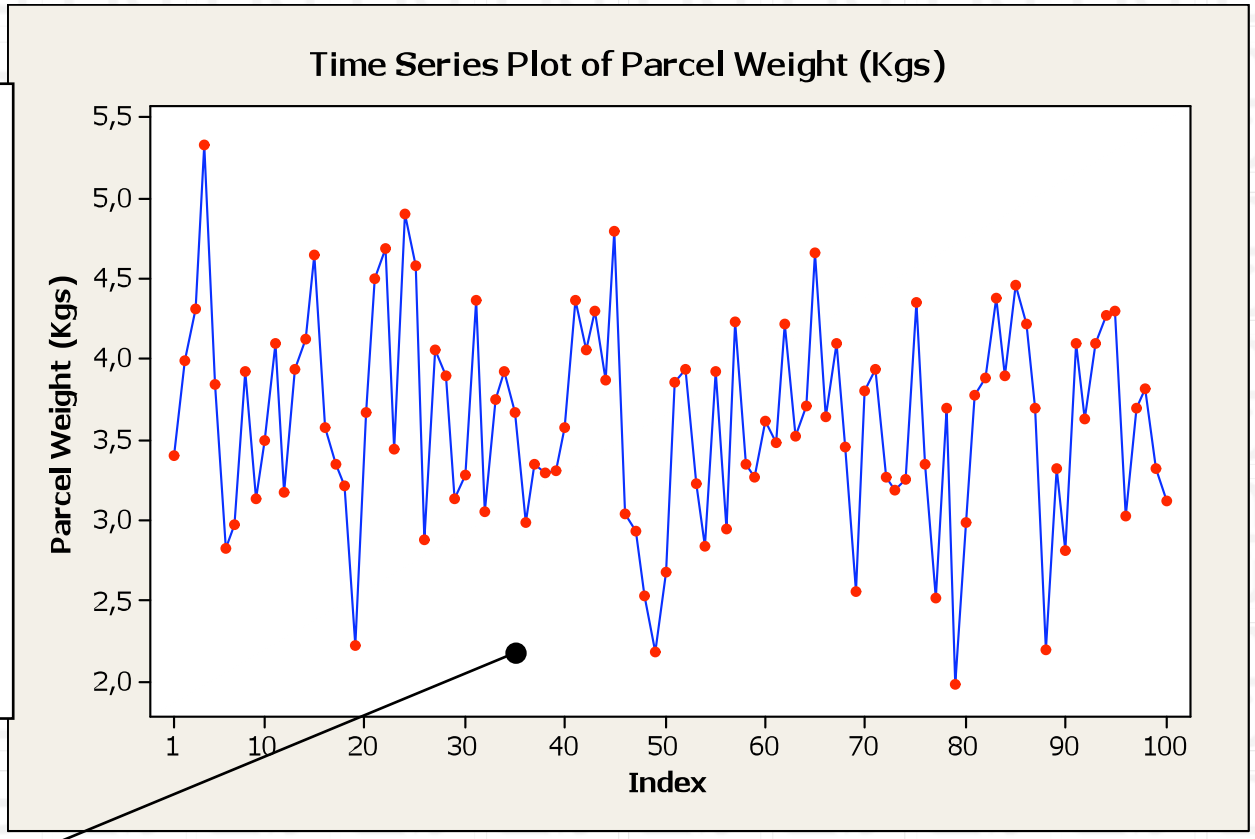
Time Series Plots

(MINITAB: Graph > Time Series Plot)

Time Series Plots può evidenziare variazioni e trends legati al fattore tempo.

Vengono evidenziati:

. Trends crescenti o decrescenti
. Cambi di variabilità
. Differenze tra short-term e long-term
. Dati con andamento ciclico
. Tutto quello che appare non essere casuale



Questo plot mostra il peso dei pacchi in funzione del tempo.
Sembra ragionevolmente casuale (non ci sono trends o cicli) e stabile.

Se abbiamo necessità di una tecnica più avanzata per valutare la stabilità, si dovrà usare una **SPC chart** – che è essenzialmente una Time Series Plot più sofisticata.

Analisi di Pareto

(MINITAB: Stat > Quality Tools > Pareto Charts)

Pareto Charts sono diagrammi di frequenza per dati classificati in categorie dove i dati più frequenti sono visualizzati in ordine da sinistra verso destra.

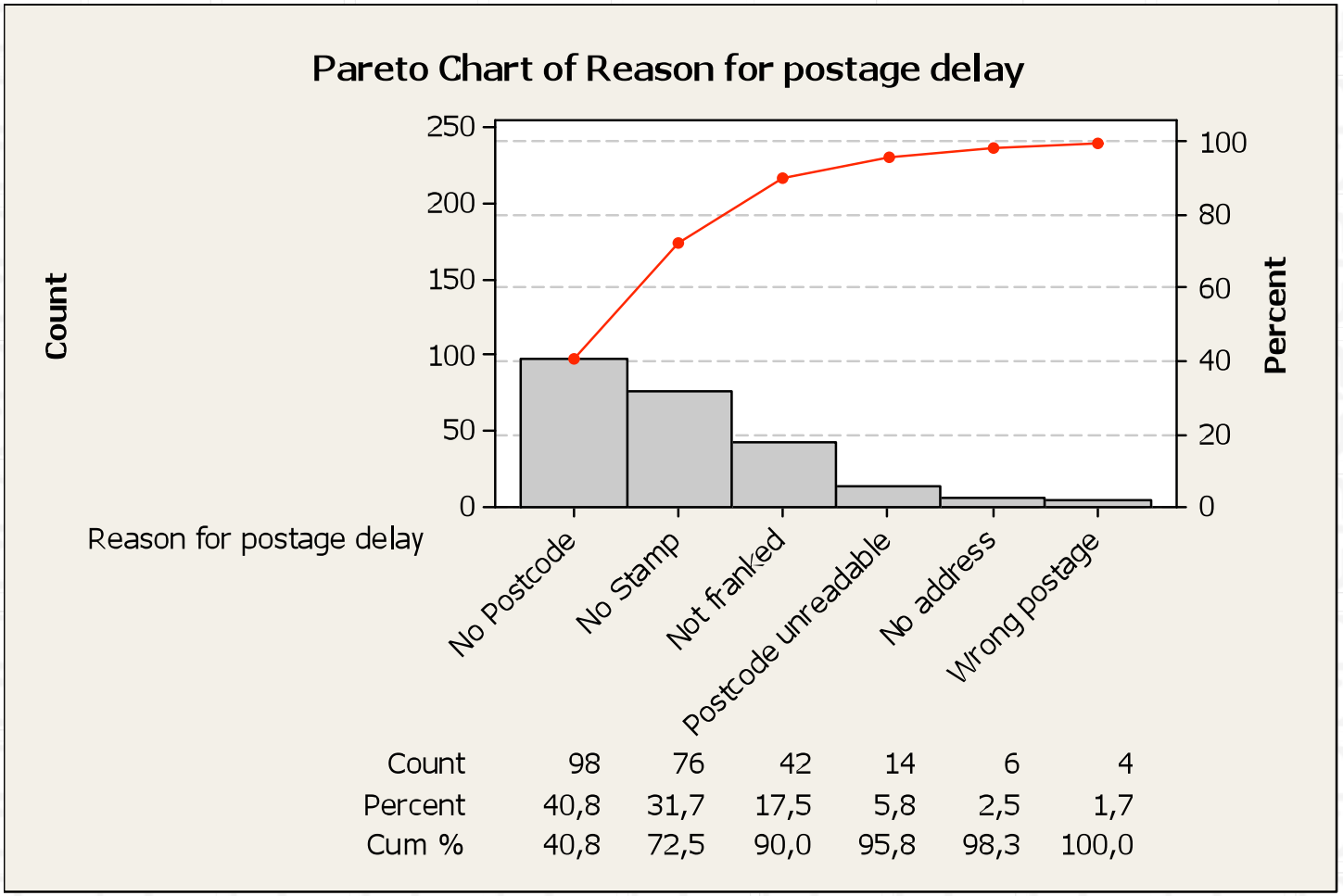
Reason for postage delay	Location
Not franked	urban
No Postcode	urban
No Stamp	rural
No Stamp	urban
Postcode unreadable	rural
No Postcode	urban
Not franked	urban
No Stamp	rural
No Stamp	rural
Not franked	urban

Opzione 1: **“Chart defects data in”**

Deve essere usata quando abbiamo I dati originali (ontenuti in una colonna).

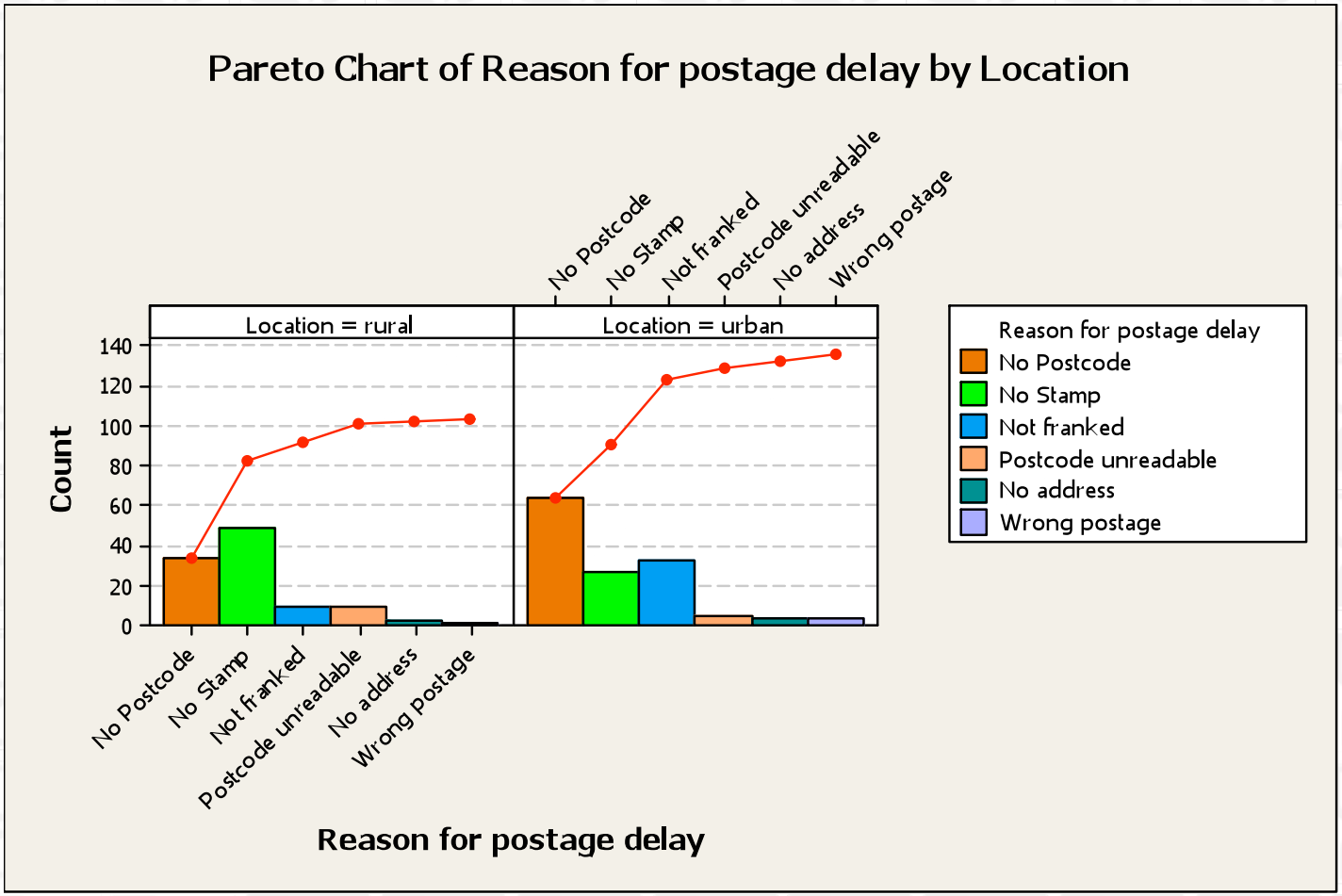
Per questo esempio, se si vogliono due Pareto Charts (una per ciascuna location – rural/urban) allora occorre attivare l’opzione “by variable”.
In tutti gli altri casi lasciarla bianca.

Per questo esempio, “no Postcode” è il motivo più frequente per il ritardo, con 98 casi (40,8%) e “no stamp” è il secondo motivo con 76 76 (31,7%).
Insieme (cumulativo) essi rappresentano il 72,5% di casi in si è avuto ritardo.



Count	98	76	42	14	6	4
Percent	40,8	31,7	17,5	5,8	2,5	1,7
Cum %	40,8	72,5	90,0	95,8	98,3	100,0

Come anticipato nella pagina precedente, l'opzione "by variable" può essere usata con lo scopo di produrre due Pareto Charts in MINITAB relative a due categorie diverse (location può essere rural o urban).

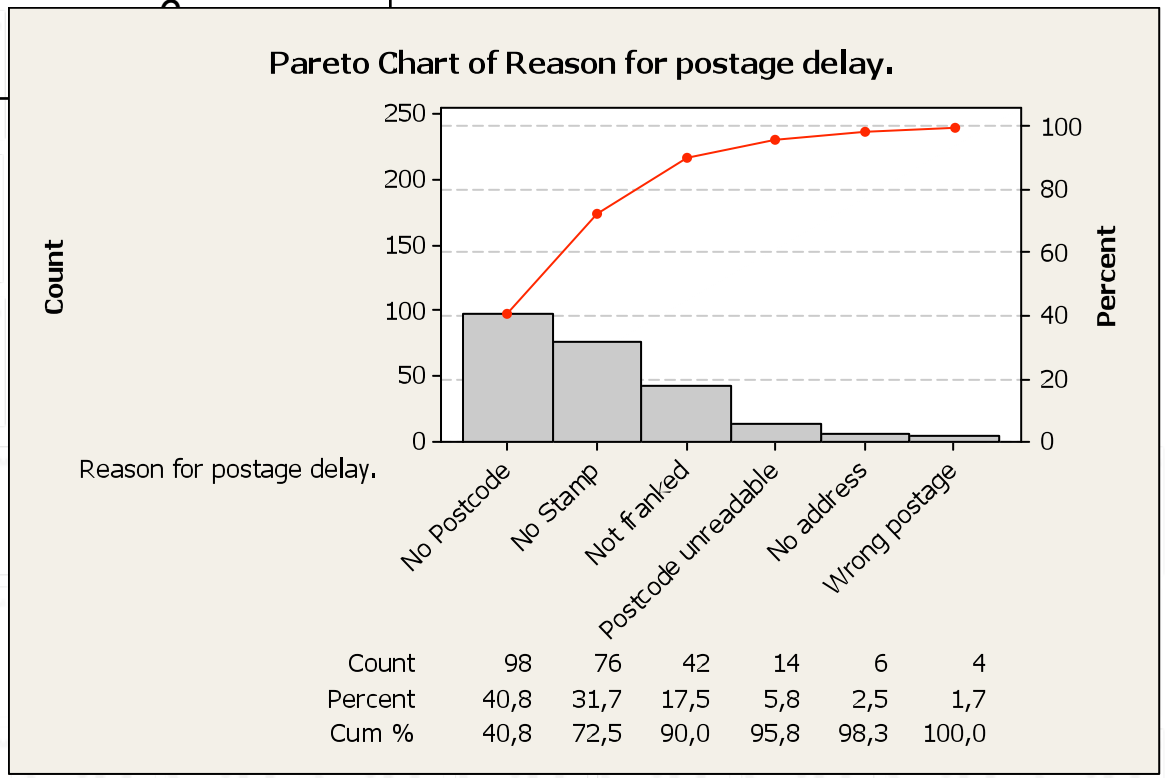


Reason for postage delay.	Location.
Not franked	42
No Stamp	76
No Postcode	98
Postcode unreadable	14
No address	
Wrong postage	

Opzione 2: **“Chart defect table”**

Deve essere usata quando i dati sono disponibili nel formato “table” come l’esempio di sinistra.
Usando questa opzione, non possiamo avere l’ulteriore suddivisione “by variable”.

I risultati sono gli stessi di pagina 10



Stratificare i dati

Analizzare le differenze tra sotto gruppi di dati può fornire utili informazioni sul processo in esame. Il metodo di dividere i dati in sotto-gruppi è chiamato “stratificazione”.

Parcel Weight (Kgs)	Location	Business Type	Service Used
3,40	England	Commercial	Normal
3,99	England	Commercial	Express
4,31	England	Commercial	Normal
5,33	France	Commercial	Overnight
3,84	England	Commercial	Normal
2,82	France	Residential	
2,97	England	Residential	Express
3,92	England	Residential	Express
3,75	France	Residential	Normal

Le base-dati spesso contengono molte informazioni sui dati suddivisi in categorie relative all’ambiente da cui i dati stessi provengono. Con esse si possono sviluppare molte teorie e idee utili per la fase di analisi.
I dati dell’esempio provengono dall’Azienda di consegne. Informazioni su **location, business e service** sono state raccolte unitamente al **peso** dei pacchi e tali informazioni possono fornire notizie utili sul processo di consegna.

Stratificare i dati – Perchè suddividere i dati

. Le differenze tra sotto-gruppi possono essere una causa importante di variazione del processo
. Identificare i sotto-gruppi che sono differenti può aiutare nella fase di analisi a capire perchè sono differenti:

- I sottogruppi con le migliori performance possono essere presi come “benchmark”
- Le cause dei sottogruppi con le peggiori performance possono essere analizzate usando metodi quali i “5 why’s” o Ishikawa, etc..

Donade utili per stratificare i dati……

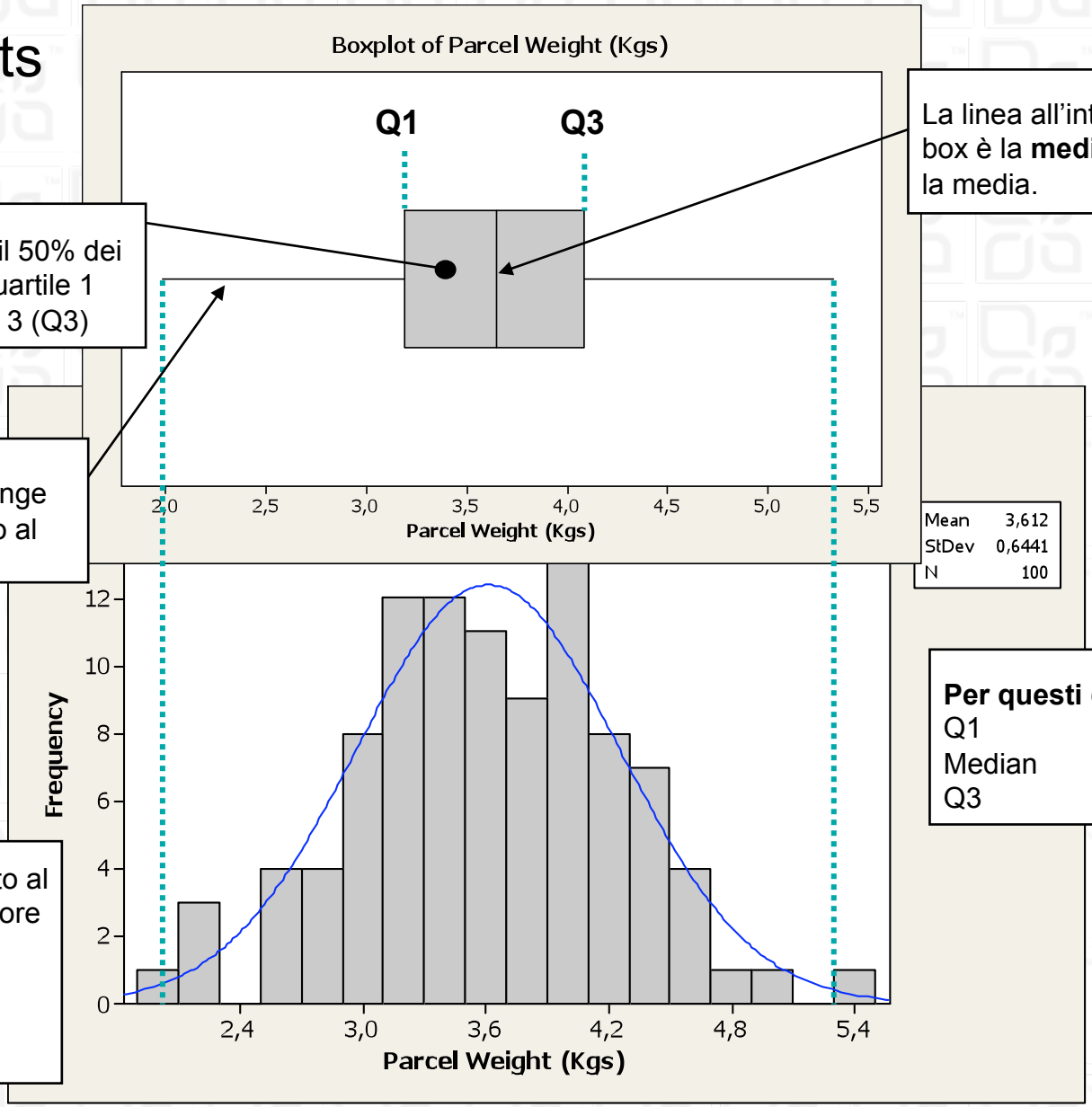
. Ci sono differenze nel peso dei pacchi tra le varie regioni?
. I clienti commerciali spediscono pacchi più pesanti?
. I clienti privati hanno maggior varietà di pacchi?
. Ilpeso dei pacchi influenza il servizio usato?

Graphical tools per analizzare e confrontare sotto-gruppi:

. Box Plots – utile per confrontare sotto-gruppi con almeno 25 dati

. Individual Value Plots – utile per confrontare sotto-gruppi con meno di 25 dati

Box Plots



Il **box** rappresenta il 50% dei dati, a partire dal quartile 1 (Q1) fino al quartile 3 (Q3)

La linea all'interno del box è la **mediana** e non la media.

I **whiskers** rappresentano il range dei dati: dal minimo al massimo.

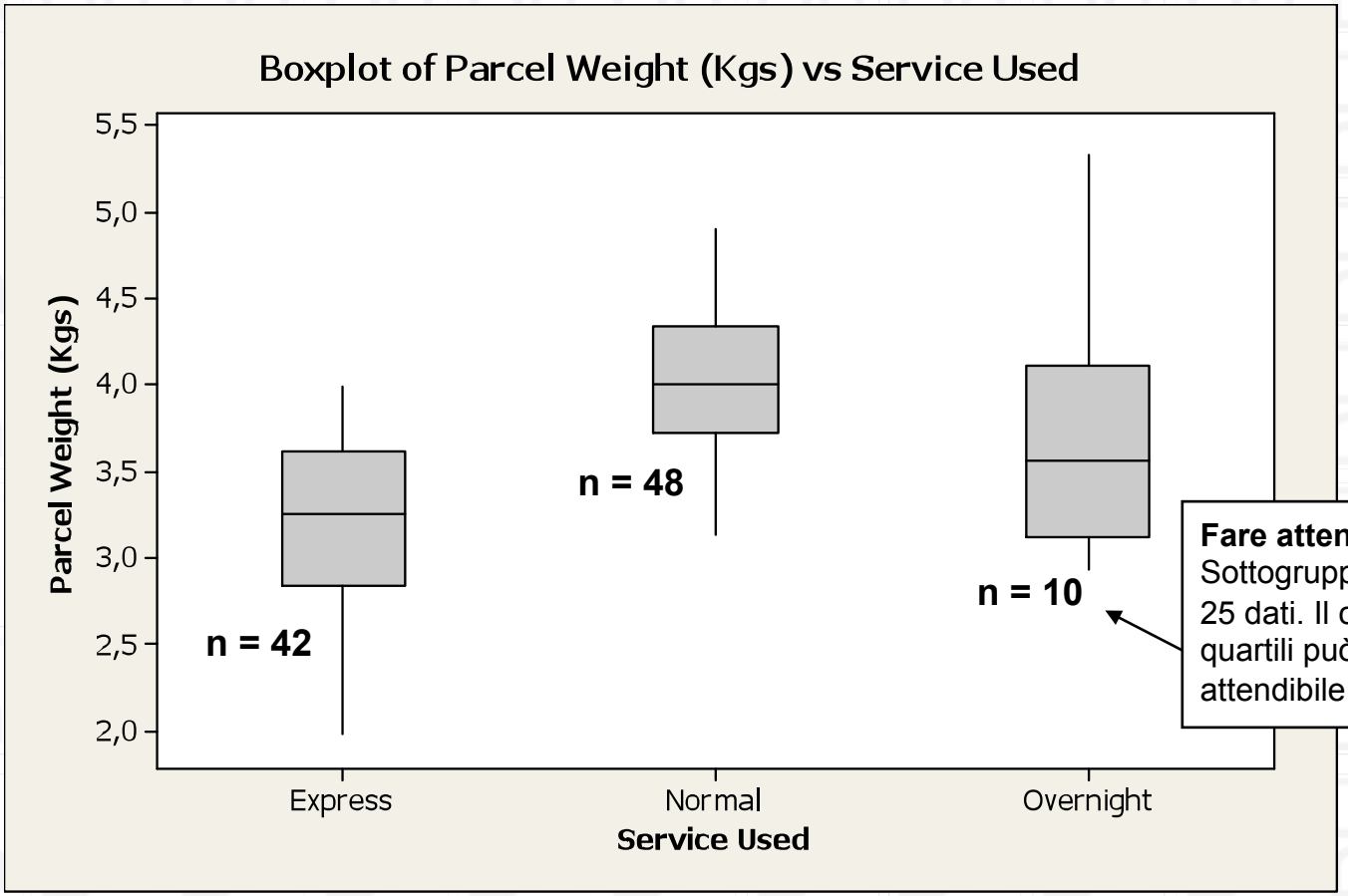
Per questi dati:

Q1	= 3.20
Median	= 3.65
Q3	= 4.09

Outliers: ogni punto al di là di un certo valore sono considerati "outliers" e sono segnalati con un asterisco (*).

Box Plots

(MINITAB: Graph > Box Plot)



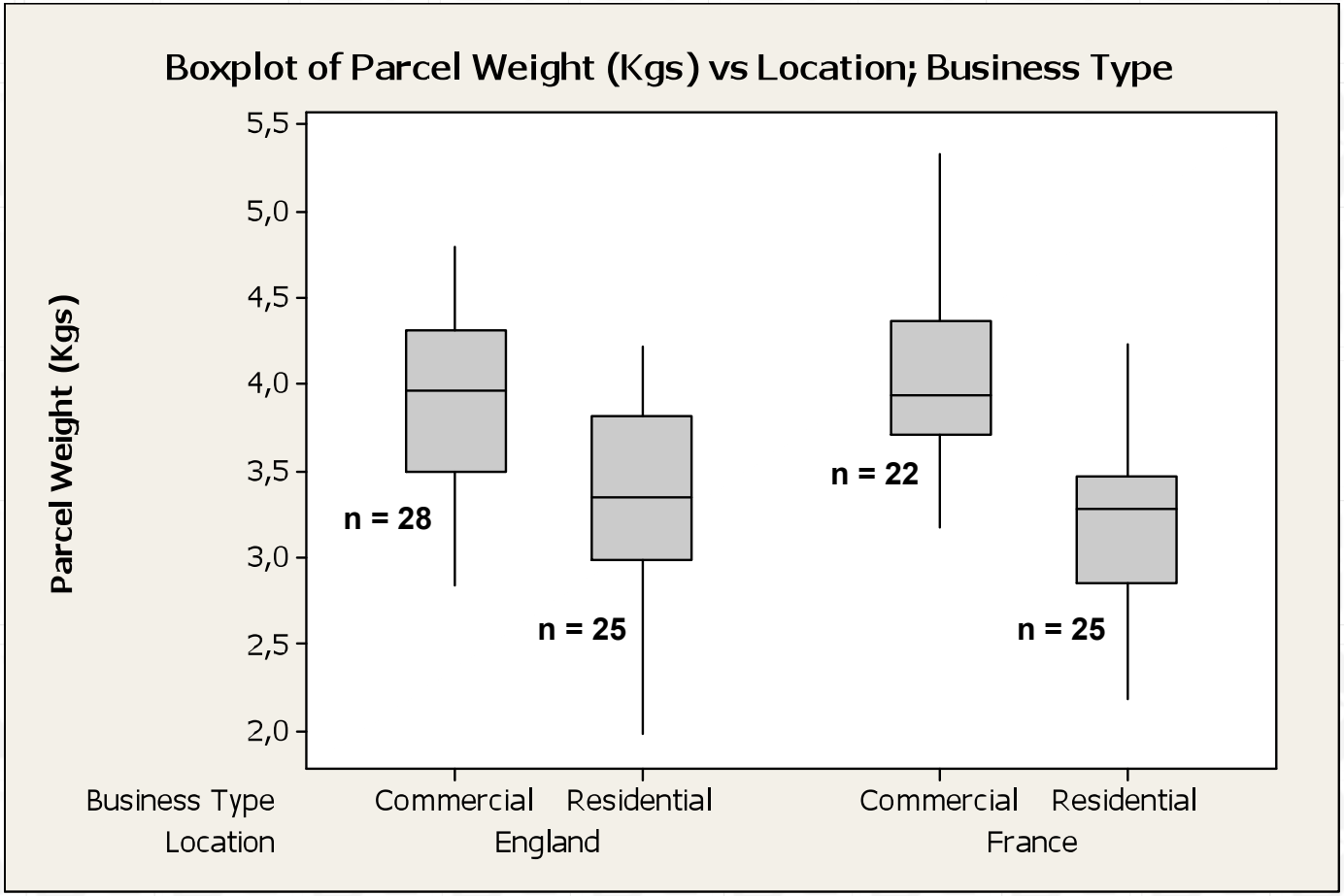
Alcune conclusioni:

. Il servizio “Express” sembra essere stato usato per pacchi con peso inferiore rispetto al servizio “Normal”

. La variazione (range) del peso dei pacchi tra Espress e Normal sembra essere circa la stessa

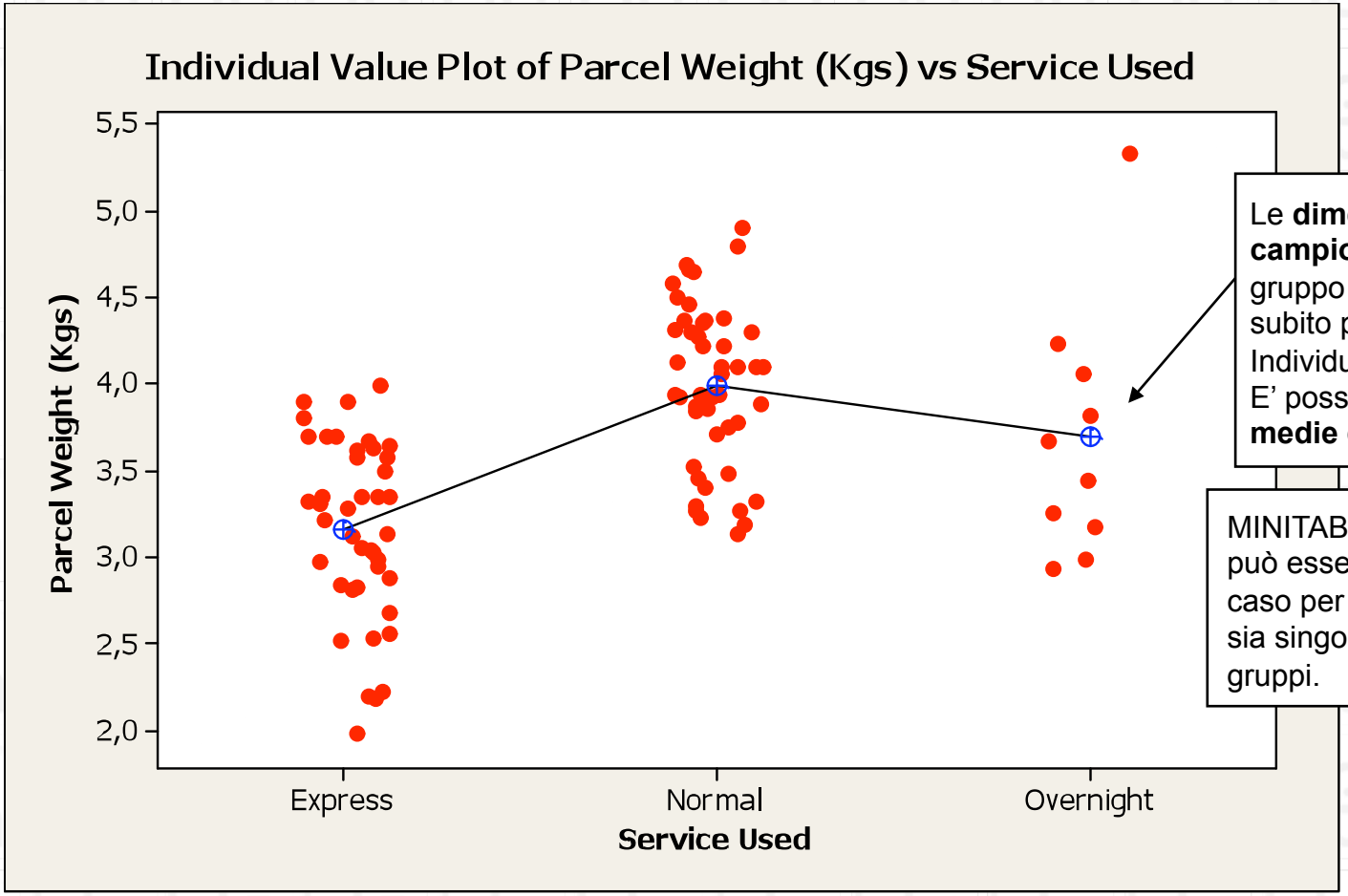
Box Plots

Introducendo due colonne nella casella “categorical variable” di MINITAB, i dati sono stratificati per entrambe le variabili.



Individual Value Plots

Individual Value Plots sono da preferire ai Box Plots quando le dimensioni del campione sono piccole (< 25)



Le **dimensioni del campione** del sotto-gruppo "overnight" sono subito percepibili con il Individual Value Plot.
E' possibile tracciare le **medie** dei sotto-gruppi.

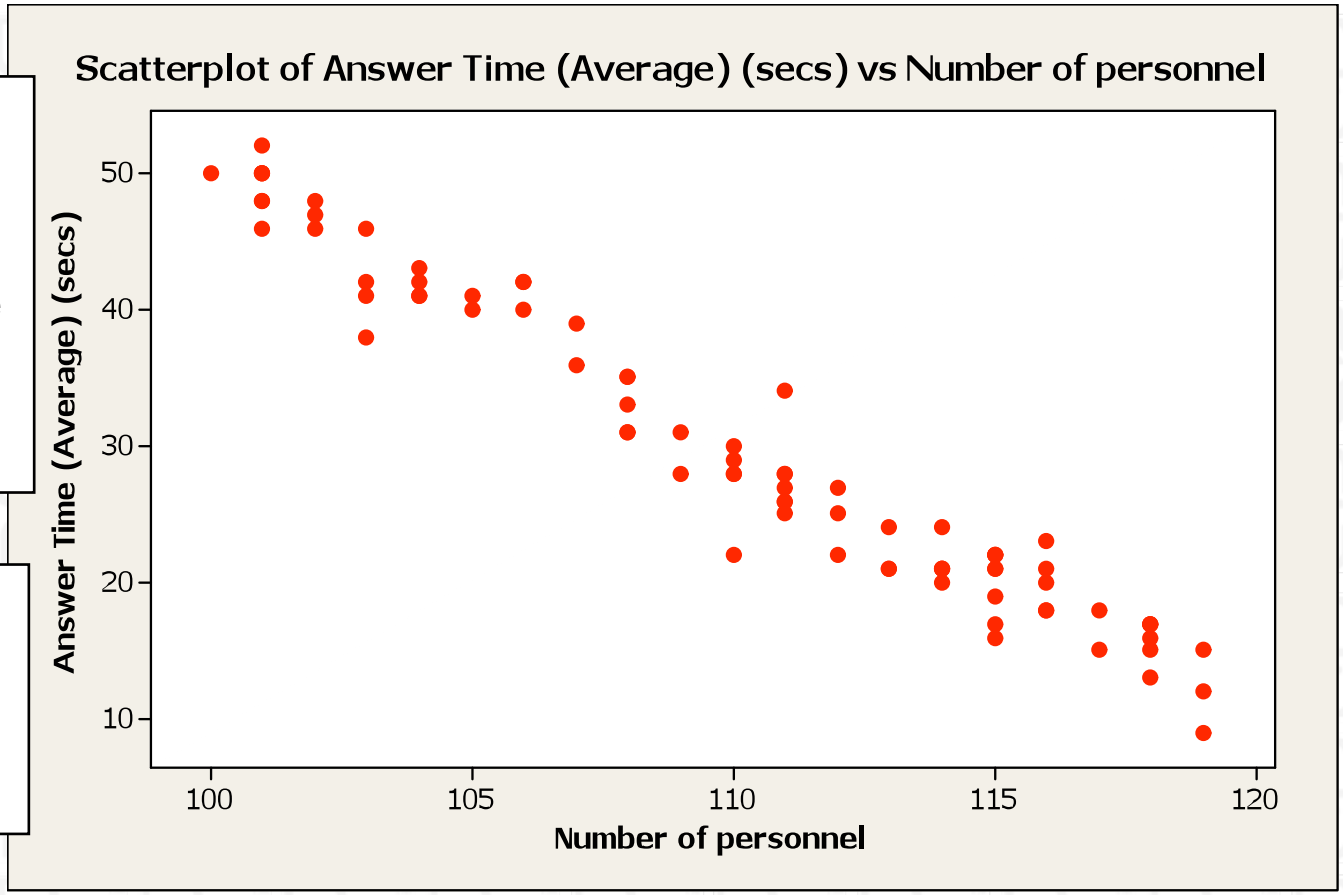
MINITAB's **Brush function** può essere usata in questo caso per analizzare I dati sia singolarmente che a gruppi.

Scatter Plots

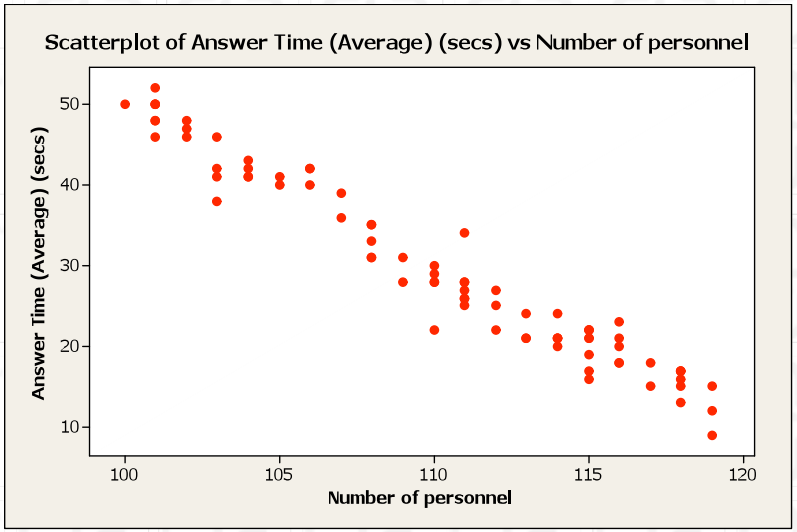
Tutte le volte che dobbiamo investigare se esiste una “relazione” tra due variabili continue, lo “scatter plot” è lo strumento migliore per una prima analisi.

Nell’esempio riportato a destra, si vuole analizzare il tempo di attesa di un Call Center. Sono stati presi in considerazione 80 turni e per ciascuno di essi è stato misurato il tempo di attesa in funzione del N° di personale.

Il diagramma segnala che esiste una relazione tra le due variabili: più alto è il numero di personale e minore è il tempo di attesa.



Scatter Plots



Causa & Effetto:

Fare attenzione prima di concludere che esiste una “relazione causa-effetto” **diretta**. A volte esiste una terza variabile che le influenza entrambe!
In questo caso sembra razionale affermare che con più personale i tempi di attesa diminuiscono.

Tipi di dati:

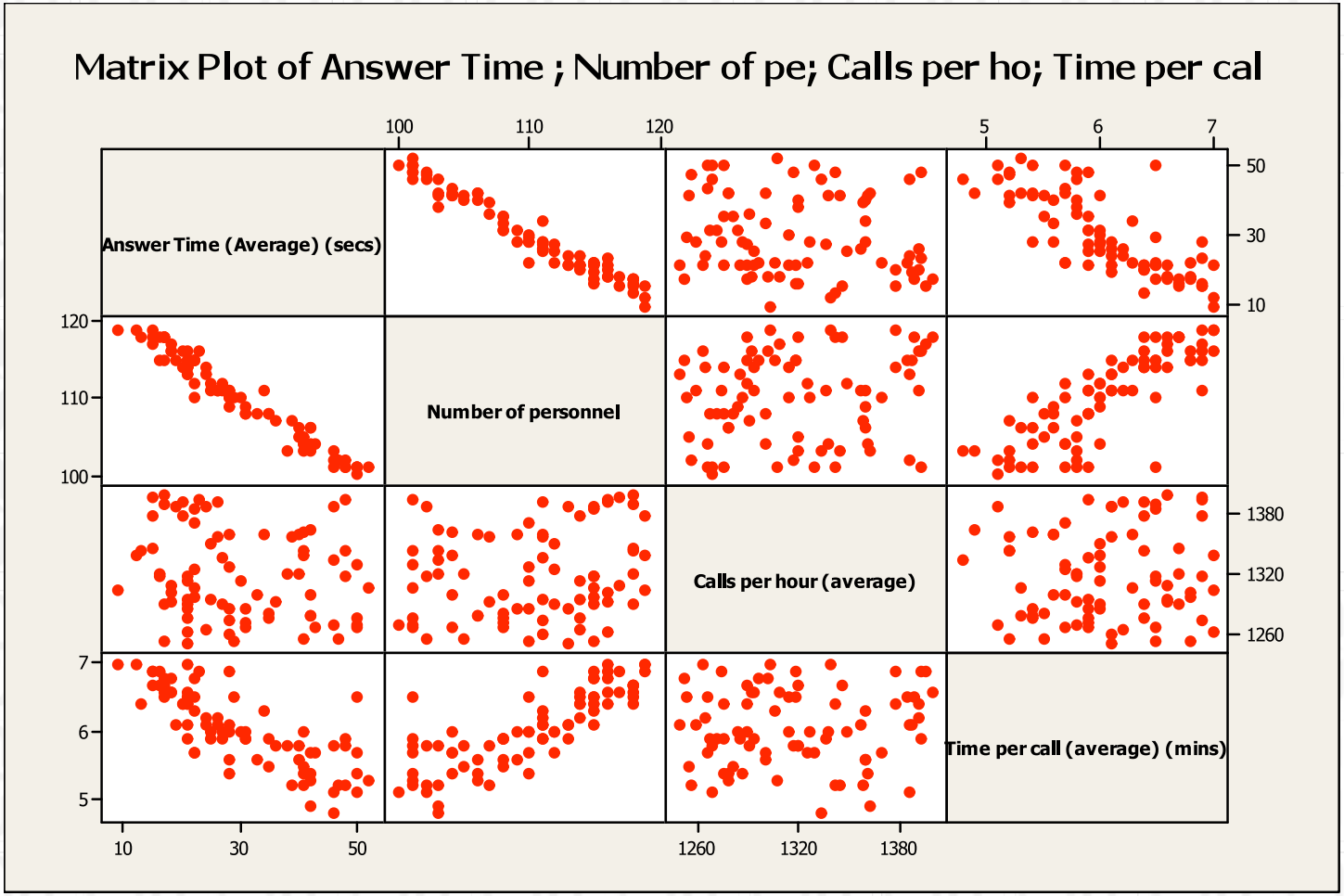
Scatter plots sono adatti per dati di tipo continuo e dati “count”. Nell’esempio il tempo di attesa E di tipo continuo, mentre il numero di personale è un dato “count”.

Lo scatter plot è il primo passo per definire se tra due variabili esiste una **correlazione**.
Nei casi in cui la relazione non è chiara, la correlazione (coefficiente di) aiuta a decidere se la relazione esiste.

La Regressione è una tecnica ancora più avanzata in quanto definisce la relazione (correlazione) in formato matematico (equazione).
La più semplice delle tecniche di regressione è la **Fitted Line Plot**.

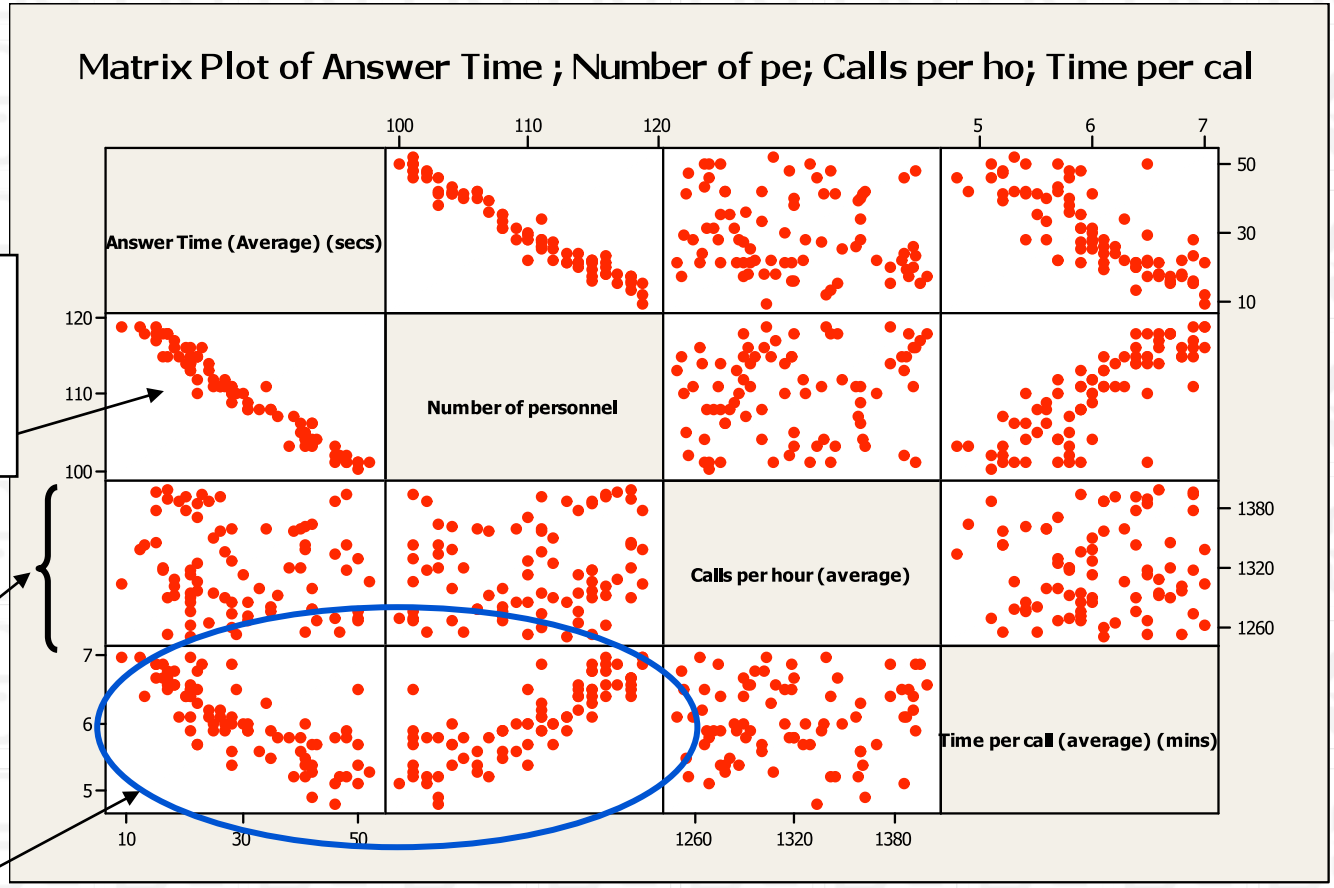
Matrix Plots

Matrix Plots forniscono un quadro di Scatter Plots per più variabili.
Le potenziali correlazioni possono essere identificate e analizzate.



Matrix Plots

MINITAB crea due immagini di scatter plots (bottom left e top right)



Come abbiamo visto **answer time** e **number of personnel** sono correlate e lo ritroviamo di nuovo.

Il volume di **incoming calls per hour** non sembra essere correlato con **answer time**, con **number of personnel** o con **time per call.**

Tuttavia, sembra ci sia una relazione tra **time per call** sia con **answer time** che con **number of personnel.**