POLITECNICO
MILANO 1863

# *PREDIZIONI MEGLIO DI COPPERFIELD*

**PROGETTO DI PROBABILITÀ E STATISTICA**

*Gruppo: Di Grandi Daniele, Hartog Alice*
*Anno scolastico 2017/18*

# Sommario:

# INTRODUZIONE

**Scopo del progetto**

Per il Progetto Integrativo del Laboratorio di Probabilità e Statistica dell'anno 2017/2018 ci siamo poste un problema riguardante la Statistica Predittiva: come può un'azienda ottimizzare il proprio stoccaggio di pezzi al fine di massimizzare la probabilità di soddisfare immediatamente un ordine da parte di un cliente?
Negli ultimi anni si tende sempre più a fare in modo che gli ordini siano caratterizzati da una notevole frammentazione delle quantità e da rapidi tempi di consegna, come si intuisce dall'e-commerce, ed è esattamente quello che vogliamo fare noi.

Per arrivare a questo scopo, è necessario comprendere la distribuzione degli ordini, in modo da gestirne al meglio l'entità quantitativa.
A questo proposito abbiamo individuato un articolo di particolare importanza dell'azienda CT Sistemi Plastici, commissionato dal loro cliente principale, e analizzato il modo e la frequenza con cui viene prodotto e ordinato.

L'obiettivo è quello di capire il modello di business che si adatta meglio a questa situazione, analizzando i dati fornitici dall'azienda stessa: iniziare la produzione non appena arriva l'ordine del cliente oppure prevedere ordini futuri per anticiparne la produzione?
Analizzeremo entrambe le soluzioni e valuteremo la migliore.

Provenienza e composizione dei dati: I dati provengono da un set di dati messi a disposizione dall'azienda CT Sistemi Plastici della provincia di Milano.
Il dataset contiene gli ordini staccati dal principale cliente e i dati relativi al magazzino scorte dell'articolo:

*-MANIFOLD NPM L4 PASTIGLIA DX, codice articolo: NPM400.*
D'ora in poi, ci riferiremo all'articolo con il suo codice.

Il nostro lavoro comprende:

i. Grafici (qqplot e box-plot) basati sui dati fornitici e suddivisi per anno;
ii. Tabelle relative a ordini e scorte;
iii. Attraverso uno stimatore e gli intervalli di confidenza abbiamo calcolato il valore medio che cercavamo, verificandolo con un test di ipotesi;
iv. Con due modelli di regressione lineare mostriamo la bontà delle nostre previsioni.

# STATISTICA DESCRITTIVA

*Foto dell'articolo NPM400, si tratta di un coperchio da inserire all'interno di batterie per automobili. Ha la funzione di far sfiatare i gas presenti internamente alla batteria, grazie a una pastiglia porosa inserita al suo interno. Il pezzo viene stampato tramite macchinari di stampaggio plastica a iniezione.*

Vediamo gli ordini che sono stati registrati mensilmente per i quattro anni e le scorte presenti a magazzino durante lo stesso periodo, per controllare se gli ordini siano stati soddisfatti con le sole scorte oppure no.

| Quantità | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|
| GEN | 22.000 | 20.000 | 19.000 | 25.120 |
| FEB | 7.750 | 25.250 | 48.540 | 67.000 |
| MAR | 5.000 | 4.000 | 6.000 | 14.715 |
| APR | 24.500 | 11.000 | 8.500 | 9.000 |
| MAG | 26.590 | 43.000 | 18.000 | 23.215 |
| GIU | 19.000 | 26.000 | 10.000 | 14.500 |
| LUG | 19.000 | 58.250 | 54.250 | 27.250 |
| AGO | 0 | 10.000 | 0 | 0 |
| SET | 8.250 | 48.000 | 21.500 | 20.250 |
| OTT | 15.000 | 21.500 | 20.250 | 44.750 |
| NOV | 17.000 | 22.000 | 11.380 | 17.750 |
| DIC | 30.000 | 25.000 | 17.250 | 3.000 |

*Questa tabella mostra la quantità di pezzi ordinati mensilmente.*

| | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|
| *Gennaio* | 15 | 15 | 20 | 15 |
| *Febbraio* | 10 | 20 | 30 | 40 |
| *Marzo* | 10 | 10 | 10 | 10 |
| *Aprile* | 0 | 10 | 7 | 5 |
| *Maggio* | 15 | 30 | 15 | 15 |
| *Giugno* | 15 | 30 | 10 | 15 |
| *Luglio* | 10 | 30 | 25 | 20 |
| *Agosto* | 0 | 5 | 0 | 0 |
| *Settembre* | 5 | 40 | 15 | 15 |
| *Ottobre* | 10 | 20 | 15 | 20 |
| *Novembre* | 10 | 20 | 10 | 10 |
| *Dicembre* | 20 | 30 | 10 | 0 |

*Questa tabella mostra l'inventario del magazzino del prodotto, qualche giorno prima di ciascun ordine. I dati sono in migliaia.*

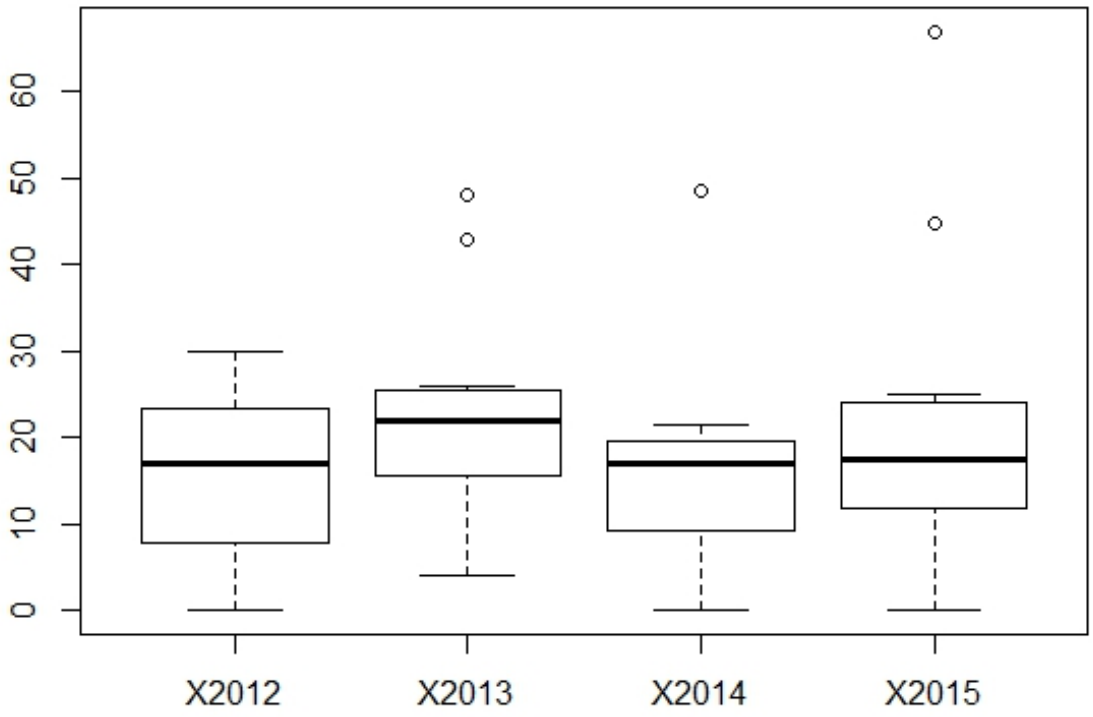
**I dati da ora in poi andranno considerati tutti come migliaia, per comodità di calcolo.**

Con questa disponibilità di magazzino calcolo la probabilità di soddisfare un ordine al suo arrivo, contando il numero di ordini arrivati in totale che sarebbero stati soddisfatti con la sola scorta. Facendo uso della legge forte dei grandi numeri ho il 20% di probabilità di soddisfare gli ordini (9 ordini soddisfatti su 45 ordini totali).

Abbiamo quindi calcolato le probabilità di soddisfare gli ordini con quanto presente a magazzino, suddividendole mese per mese, utilizzando lo stesso procedimento:

| *Gen* | *Feb* | *Mar* | *Apr* | *Mag* | *Giu* | *Lug* | *Ago* | *Set* | *Ott* | *Nov* | *Dic* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.25 | 0.25 | 0.75 | 0 | 0 | 0.50 | 0 | 0 | 0.25 | 0 | 0 | 0.25 |

Attraverso un boxplot possiamo mostrare chiaramente lo squilibrio tra i dati:




In ciascun boxplot la mediana (il segmento più spesso) non è centrata ma tende verso l'alto, indicando una maggiore concentrazione dei dati nella zona superiore. Sono presenti anche dati dispersi (outlayer), e i baffi della scatola non sono assolutamente simmetrici: tutto ciò indica una forte asimmetria dei dati.

# TEST DI NORMALITÀ

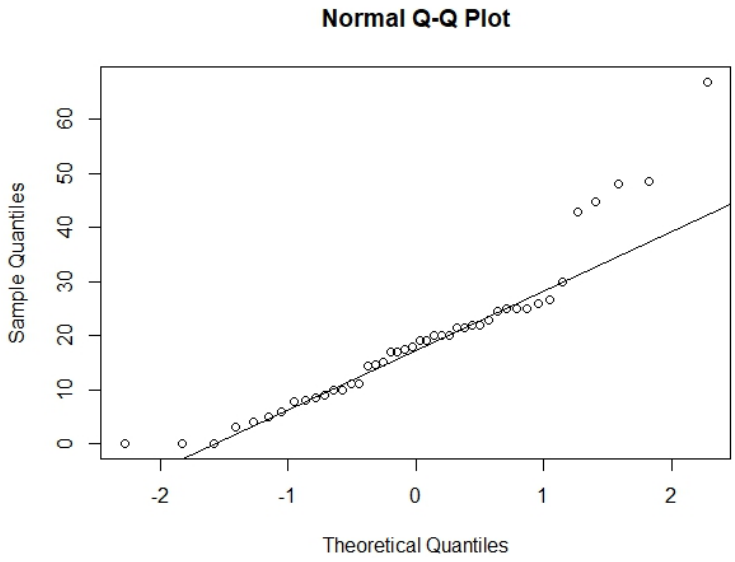
Per stabilire se i nostri dati seguono una distribuzione normale abbiamo effettuato lo Shapiro-Wilk test:

```
>shapiro.test(npm400)

Shapiro-Wilk normality test

data:  npm400
W = 0.89479, p-value = 0.0007555
```

Avendo un p-value molto basso, il mio campione non segue una distribuzione normale.
Il grafico QQ-Norm lo delinea chiaramente:




Il grafico mostra però che i dati al centro aderiscono abbastanza bene alla linea, e questo significa che ha una distribuzione simile a quella di una normale.

Inoltre, secondo il Teorema Centrale del Limite, con una numerosità sufficientemente grande (maggiore di 30), si può approssimare la distribuzione dei dati ad una normale. In questo caso i valori sono 48 e quindi possiamo applicare il teorema e approssimare i dati a una normale.
Inoltre abbiamo nello Shapiro-test anche un valore di statistica W pari a 0.89, valore che, insieme alla sua numerosità, è sufficiente per poter approssimare.

# INTERVALLI DI CONFIDENZA

Interessiamoci ora alla parte di predittiva chiedendoci quale sia l'intervallo di predizione entro cui giacciono gli ordini fatti per ogni mese.
Nel nostro caso purtroppo non è possibile calcolare questo intervallo, essendo i nostri dati asintoticamente normali, perciò l'alternativa migliore tra quelle studiate è usare gli intervalli di confidenza.

L'intervallo di confidenza per la media μ al 95% è stato calcolato nel seguente modo:

$$\mu \in \bar{X} \pm t_{\alpha/2} \sqrt{\frac{S^2}{n}}$$

L' intervallo di confidenza della distribuzione della totalità degli ordini è $\mu \in (16; 22.6)$.
I quantili sono delle t di Student, in quanto la varianza utilizzata è quella campionaria ($S^2$).
$\bar{X}$ è invece la media campionaria e n il numero del campione.

Se vogliamo ottimizzare la produzione, è meglio calcolare un intervallo di confidenza mensile:

| | MEDIA | IC inf | IC sup |
|---|---|---|---|
| *Gennaio* | 21,5 | 17,368 | 25,632 |
| *Febbraio* | 37 | 32,918 | 41,182 |
| *Marzo* | 7,429 | 3,293 | 11,556 |
| *Aprile* | 13,25 | 9,118 | 17,382 |
| *Maggio* | 27,65 | 23,518 | 31,782 |
| *Giugno* | 17,375 | 13,243 | 21,507 |
| *Luglio* | 39,69 | 25,980 | 53,39 |
| *Agosto* | 2,5 | 0 | 6,632 |
| *Settembre* | 24,375 | 20,243 | 28,507 |
| *Ottobre* | 25,3 | 21,168 | 29,432 |
| *Novembre* | 16,875 | 12,743 | 21,01 |
| *Dicembre* | 18,75 | 14,618 | 22,882 |

*L'intervallo di confidenza inferiore di Agosto sarebbe un numero negativo, ma per ovvi motivi un ordine non può essere minore di 0, perciò è stato ridimensionato a zero.*

Per migliorare il mio sistema di produzione devo stimare un valore medio per ciascun mese in modo che, non appena la mia scorta scende al di sotto di esso, comincio a produrre con un macchinario libero per raggiungere nuovamente la soglia.

Per stimare il valore utilizzo il metodo di massima verosimiglianza.
Da teoria sappiamo che la stima di massima verosimiglianza per la media di una popolazione normale è $\bar{X}$, la media campionaria, che avevamo già inserito nella tabella sovrastante.

L'ultimo passaggio di questo discorso sarebbe quello di verificare se il valore stimato giace all'interno dell'intervallo di confidenza, ma essendo l'intervallo centrato proprio sulla media campionaria è già verificato.

Quindi, se le scorte scendessero sotto questo valore, dovrei iniziare la produzione per raggiungerlo nuovamente.
In questo modo, all'arrivo dell'ordine, se il quantitativo richiesto sarà superiore alla media prevista, basterà preparare gli ultimi pezzi necessari, e la spedizione sarà notevolmente più veloce.

Per un'ottimizzazione maggiore, la quantità di pezzi da tenere a scorta dev'essere all'incirca compresa tra la media campionaria e l'intervallo superiore di confidenza.
La probabilità di soddisfare l'ordine avendo in magazzino un quantitativo compreso tra la media campionaria e l'intervallo superiore è stata ricalcolata con il metodo utilizzato a pag.5.
In questo modo le probabilità sono comprese tra 60% e 72%.

# TEST DI IPOTESI

Abbiamo deciso di fare un test di ipotesi sulla media di ciascun mese, per avere un maggior controllo e poter affermare che la media campionaria è esattamente il miglior stimatore della media.
La decisione di svolgere questo test proprio sulla media è stata presa in quanto, dopo l'ottimizzazione fatta nella pagina precedente, la media è diventata il minimo quantitativo da tenere a magazzino, e risulta evidente che sia quindi il valore più importante da analizzare.

Il test di ipotesi è strutturato nel seguente modo:

$$H0: \mu = \mu_0 \text{ contro } H1: \mu < \mu_0$$

Con significatività di $\alpha$ pari al 5%.
Il test è stato fatto per ciascun mese, qui riportiamo per questioni di sintesi quello per il mese di gennaio.

```
One Sample t-test

data:  p
t = 0, df = 3, p-value = 0.5
alternative hypothesis: true mean is less than 21.5
95 percent confidence interval:
     -Inf 24.61321
sample estimates:
mean of x
     21.5
```

Dove p è il campione per ciascun mese, con rispettiva media e varianza campionarie.
Il risultato mi dice, per ciascun mese, che non ho sufficiente evidenza statistica per rifiutare l'ipotesi nulla, e perciò la media è corretta.

# AGGIORNAMENTO E CONTROLLI

Dopo tutte le ipotesi fatte, stimiamo che con la media calcolata per ciascun mese dovremmo riuscire con una buona percentuale a soddisfare gli ordini degli anni futuri.
Perciò, abbiamo preso il numero di pezzi ordinati negli anni 2016 e 2017 per verificare:

| | 2016 | 2017 |
|---|---|---|
| *Gennaio* | 19,00 | 19,00 |
| *Febbraio* | 35,00 | 31,50 |
| *Marzo* | 6,75 | 5,00 |
| *Aprile* | 10,00 | 10,50 |
| *Maggio* | 25,00 | 22,75 |
| *Giugno* | 20,25 | 15,00 |
| *Luglio* | 18,00 | 30,50 |
| *Agosto* | 3,75 | 2,50 |
| *Settembre* | 21,50 | 21,50 |
| *Ottobre* | 22,25 | 22,50 |
| *Novembre* | 15,75 | 15,00 |
| *Dicembre* | 14,00 | 12,00 |

Per affinare il modello di mantenimento a scorta è necessario calcolare una nuova media campionaria tenendo conto anche di questi due anni.
Sebbene la media campionaria iniziale avrebbe soddisfatto un'altissima percentuale degli ordini nei sei anni, abbiamo deciso di ricalcolarla, avendo notato un leggero ribassamento del quantitativo degli ordini: sarebbe stato inutile tenere a scorta un valore di troppo superiore al necessario, dato che per l'azienda ha un costo anche lo stoccaggio.

Le nuove medie campionarie $\bar{X}$, ovvero le nuove quantità minime da tenere a scorta (con rispettivo intervallo di confidenza) sono le seguenti:

| | $\bar{X}$ | IC inf | IC sup |
|---|---|---|---|
| *Gennaio* | 20,69 | 18,10 | 23,27 |
| *Febbraio* | 35,51 | 14,13 | 56,88 |
| *Marzo* | 5,15 | 3,87 | 6,43 |
| *Aprile* | 12,12 | 5,69 | 18,55 |
| *Maggio* | 26,41 | 17,36 | 35,47 |
| *Giugno* | 18,71 | 12,75 | 24,67 |
| *Luglio* | 32,21 | 12,12 | 52,30 |
| *Agosto* | 2,71 | 0,00 | 6,81 |
| *Settembre* | 23,33 | 9,57 | 37,10 |
| *Ottobre* | 23,67 | 12,49 | 34,85 |
| *Novembre* | 16,06 | 12,03 | 20,09 |
| *Dicembre* | 16,66 | 6,41 | 26,91 |

*Anche qui l'intervallo inferiore di Agosto è stato ridimensionato a 0.*

Notiamo quindi che la probabilità di soddisfare un ordine immediatamente al suo arrivo, calcolata (come a pag.5) dopo tutte le analisi effettuate è la seguente:

PRIMA:
Probabilità di soddisfare un ordine immediatamente al suo arrivo: 20%.
Mensilmente:

| *Gen* | *Feb* | *Mar* | *Apr* | *Mag* | *Giu* | *Lug* | *Ago* | *Set* | *Ott* | *Nov* | *Dic* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.25 | 0.25 | 0.75 | 0 | 0 | 0.50 | 0 | 0 | 0.25 | 0 | 0 | 0.25 |

DOPO:
Calcolando la probabilità utilizzando come quantitativo un intervallo (dalla media campionaria all'intervallo di confidenza superiore) otteniamo:
Probabilità di soddisfare un ordine immediatamente al suo arrivo: 64% - 80%.
Mensilmente:

| *Gen* | *Feb* | *Mar* | *Apr* | *Mag* | *Giu* | *Lug* | *Ago* | *Set* | *Ott* | *Nov* | *Dic* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.66-0.83 | 0.66-0.83 | 0.5-0.66 | 0.83 | 0.66-0.83 | 0.33-0.83 | 0.66 | 0.66-0.83 | 0.83 | 0.83 | 0.5-0.83 | 0.5-0.83 |

Osserviamo che la probabilità di soddisfare un ordine generico si è alzata notevolmente, così come sono aumentate anche quelle mensili.

Possiamo utilizzare questi nuovi dati come fonte per la regressione lineare.

# REGRESSIONE

La domanda che ci poniamo è la seguente: che correlazione esiste tra media campionaria mensile e ordini futuri?

Strutturiamo la regressione in questo modo:

Come variabile indipendente scegliamo le medie campionarie e come variabile dipendente il quantitativo futuro di ordini.
In seguito consideriamo due diverse regressioni:

1. Relazioniamo le medie campionarie ottenute fino al 2015 con gli ordini arrivati nel 2016.

   Quindi:
   $X_1$ = media campionaria di gennaio fino al 2015
   $Y_1$ = ordine effettuato a gennaio 2016
   .
   .
   .
   $X_{12}$ = media campionaria di dicembre fino al 2015
   $Y_{12}$ = ordine effettuato a dicembre 2016

Da R otteniamo il seguente output:

```
> modello1 = lm(formula = y ~ regr, x = TRUE, y = TRUE)
> summary(modello1)

Call:
lm(formula = y ~ regr, x = TRUE, y = TRUE)

Residuals:
   Min     1Q  Median     3Q    Max
-14.269  -2.736   1.063   2.718   8.300

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1215     3.9106   1.310   0.2196
regr          0.5832     0.1670   3.492   0.0058 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.024 on 10 degrees of freedom
Multiple R-squared:  0.5495,     Adjusted R-squared:  0.5044
F-statistic:  12.2 on 1 and 10 DF,  p-value: 0.005799
```

Analizzando l'output notiamo che l'intercetta ha un p-value abbastanza alto (0,2196) e quindi al fine di migliorare il modello potrebbe essere eliminata, mentre il coefficiente di regressione ha un p-value di 0,0058, relativamente basso, e quindi ci da informazioni importanti.
Guardando invece la statistica $R^2$ notiamo un valore di 0,5495, mentre di 0,5044 per quella aggiustata. Il valore ottenuto è abbastanza basso, perciò la bontà del modello non è sufficiente a garantire la sua correttezza.

2. Affiniamo la media campionaria includendo anche gli ordini del 2016 e facciamo il medesimo procedimento del punto 1, questa volta utilizzando gli ordini del 2017 come variabile dipendente e osserviamo se il modello presenta un miglioramento.

Da R otteniamo il seguente output:

```
> modello2 = lm(formula = j ~ regr2, x = TRUE, y = TRUE)
> summary(modello2)

Call:
lm(formula = j ~ regr2, x = TRUE, y = TRUE)

Residuals:
   Min     1Q  Median     3Q    Max
-5.8253 -1.1180 -0.8796  1.7793  6.9801

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.45475    2.23954   0.650    0.531
regr2        0.71984    0.09991   7.205 2.91e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.408 on 10 degrees of freedom
Multiple R-squared:  0.8385,    Adjusted R-squared:  0.8223
F-statistic: 51.91 on 1 and 10 DF,  p-value: 2.91e-05
```

Analizzando il modello aggiornato, notiamo che il p-value dell'intercetta si è alzato ulteriormente, a conferma che sarebbe un regressore da eliminare. Il p-value del coefficiente di regressione si è abbassato fino a un valore di 2,91e-05, segno che è un regressore che ha molto da dire sulla distribuzione.
Il coefficiente $R^2$ ha un valore di 0,8385 (prossimo a 1) perciò, superando il valore di 0,7, ha rivelato che il modello è molto buono.

CONFRONTO:
Innanzitutto notiamo che l'errore residuo standard del secondo modello si è dimezzato rispetto al primo. Inoltre, come già osservato, il secondo modello presenta un coefficiente $R^2$ più alto del primo e un p-value del coefficiente di regressione molto più basso.
Possiamo dunque concludere che il secondo modello si è migliorato e che nel corso degli anni, continuando a collezionare dati e aggiornandolo, si affinerà ulteriormente.

# DATI ECONOMICI

Per concludere indichiamo qualche costo, per sostenere maggiormente la nostra ipotesi iniziale.

Prima dell'analisi avevamo i costi di produzione dei singoli pezzi e una mora se la consegna non partiva entro 7 giorni (contratto stipulato con l'azienda).
Costo pezzo: 0,10€
Mora: 0,01€ per ciascun pezzo per ogni giorno di ritardo **(°)**

Dopo l'analisi abbiamo i costi di produzione dei singoli pezzi e i costi di stoccaggio.
Costo pezzo: 0,10€
Costi stoccaggio: 0,10€ (costo pezzo)*0,57 € (prezzo di vendita)*numero di pezzi stoccati*2/12 (numero mesi **(°°)**)*1/4 (numero settimane al mese)

Per i giorni di ritardo **(°)** bisogna sapere che ogni giorno vengono prodotti 2000 pezzi.
Per produrne 20690 pezzi (media dei pezzi prodotti a gennaio) servono più di 10 giorni, perciò in questo caso la mora sarà calcolata su 10-7 = 3 giorni.

Per il numero di mesi **(°°)** ne considero due perché a scorta i pezzi rimangono l'ultima settimana del mese precedente e la prima del mese in cui arriverà l'ordine (poiché gli ordini arrivano generalmente dopo la prima settimana del mese).
Per il numero di settimane tengo conto di una settimana per ogni mese. In definitiva, non sto considerando il costo mensile, ma il costo per ciascun ordine, indipendentemente dal fatto che lo stoccaggio sia distribuito su due mesi.

Calcoliamo quindi i costi medi totali per i pezzi da produrre per gennaio per comprenderne il confronto:

Prima dell'analisi: 0,10 €*20690 + 0,01€*20690*3 = 2689,7€
Dopo l'analisi: 0,10€*20690 + 0,10€*0,57€*20690*2/12*1/4 = 2118,14€

Abbiamo una diminuzione dei costi medi di 571,56€.

# CONCLUSIONE

Negli ultimi anni lo scopo di tutte le aziende è quello di velocizzare le spedizioni, in certi casi addirittura anticipando gli ordini. Così facendo i clienti saranno più propensi ad acquistare nuovamente ed in breve tempo, mentre l'azienda avrà la possibilità di soddisfare più ordini contemporaneamente.
Iniziare la fase di produzione solo all'arrivo dell'ordine può portare ad un impiego scorretto e non ottimale dei macchinari, che potrebbero essere utilizzati per anticipare la produzione durante i periodi di fermo, calibrando il quantitativo da produrre: l'obiettivo sarà rimanere negli intervalli calcolati.

Non rispettando la quantità minima, quando capita che più ordini si sovrappongano e non c'è disponibilità di macchinari per iniziare lo stampaggio dei pezzi, si accumulano ritardi nell'avvio della produzione con conseguenti slittamenti anche nelle spedizioni e pagamento di more.
Inoltre, il personale lavorativo dell'azienda addetto allo stampaggio dei pezzi sarà occupato a colmare i vari ritardi e gestire le sovrapposizioni, aumentando il rischio di commettere errori, lavorando sotto pressione e di fretta, con conseguente perdita di denaro e tempo per l'azienda.

Abbiamo quindi testato la nostra soluzione, e a valle dei calcoli effettuati possiamo dire che il modello di previsione creato, da un punto di vista statistico, ha funzionato: siamo passati da una probabilità molto bassa del 20%, dovuta ad una minima quantità di pezzi a scorta, ad una compresa tra il 64% e l'80%, che si affina a mano a mano che gli anni passano e si collezioneranno più dati.