

**Statistica Applicata**  
**Corso di Laurea in Scienze Naturali**  
**a. a. 2018/2019**

prof. Federico Plazzi

19 e 26 Luglio 2019

Nome: \_\_\_\_\_

Cognome: \_\_\_\_\_

Matricola: \_\_\_\_\_

**Alcune indicazioni:**

- La prova è costituita da cinque esercizi; dopo ogni esercizio c'è lo spazio in cui scrivere la risposta o le risposte. In caso questo spazio non sia sufficiente, si può continuare a rispondere sul retro del foglio, avendo cura di indicare il numero dell'esercizio a fianco della continuazione della risposta.
- Alcuni esercizi richiedono semplici calcoli, per i quali è consentito l'uso di una calcolatrice ed eventualmente la consultazione di una o più delle tabelle allegate.
- Altri esercizi richiedono invece la lettura dei dati: verrà valutata in questo caso l'argomentazione che giustifica l'interpretazione fornita.
- La durata massima della prova è di 60 minuti.
- Si prega di non scrivere nulla sulle tabelle allegate.

# 1 Dati

Una grande catena di supermercati intende lanciare una linea di prodotti biologici. Per esplorare il possibile gradimento del pubblico, sceglie uno dei prodotti in questione e lo fa assaggiare a un *panel* di 15 assaggiatori insieme allo stesso prodotto ottenuto però con le tecniche industriali destinate alla produzione su larga scala, ovviamente senza indicare l'origine dei due prodotti. Gli assaggiatori vengono scelti tra i potenziali clienti e non si tratta perciò di assaggiatori professionisti: viene chiesto loro di esprimere, per entrambi i prodotti (“biologico” e “industriale”), un voto in decimi in base al sapore. La tabella 1 mostra i risultati dell'indagine.

Tabella 1: Voti espressi dal *panel* di assaggiatori

Assaggiatore	Biologico	Industriale
A	9	6
B	8	10
C	7	7
D	7	6
E	8	5
F	8	7
G	8	8
H	9	7
I	6	10
J	9	9
K	8	9
L	9	10
M	7	8
N	8	7
O	6	5

# 2 Esercizi

## 2.1 Statistiche di base

Calcolare media, devianza, varianza e deviazione standard dei voti conseguiti dal prodotto biologico.

<i>Media</i>	<i>Devianza</i>	<i>Varianza</i>	<i>Deviazione standard</i>
7,8	14,4	0,96	0,9798

## 2.2 Distribuzione dei dati

La figura 1 mostra la distribuzione di tutti i 30 voti. Quale potrebbe essere il risultato del relativo test di Shapiro e Wilk tra i tre elencati di seguito? Perché? Cosa possiamo dire della distribuzione dei dati?

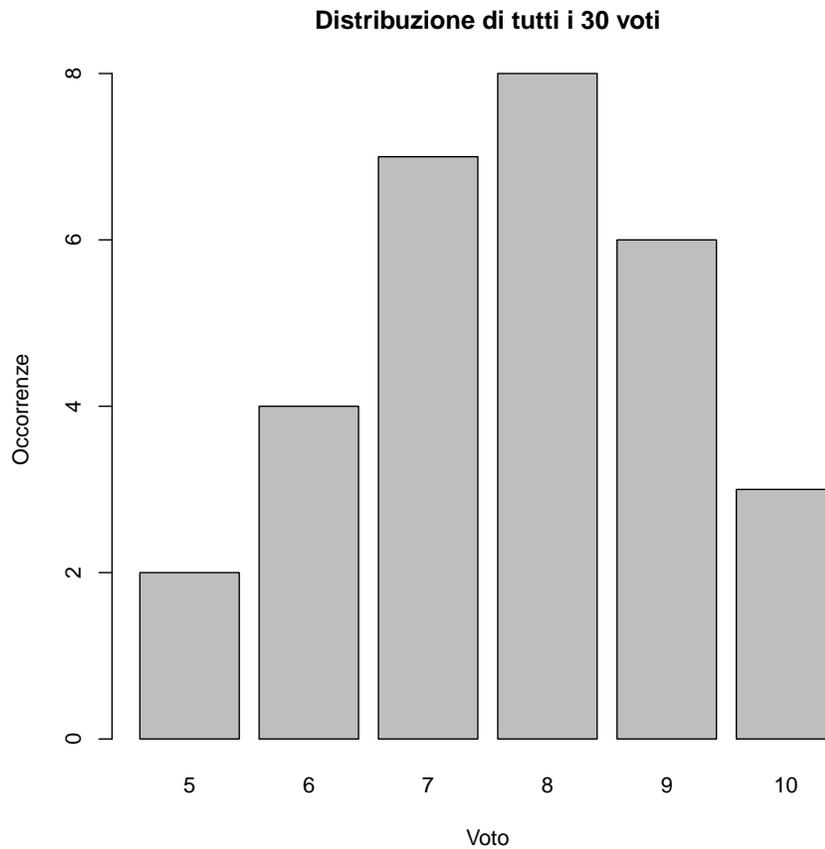


Figura 1: Distribuzione dei voti.

- Shapiro-Wilk test  
 $W = 0.99831$ ,  $p\text{-value} = 0.0003$
- Shapiro-Wilk test  
 $W = 0.94294$ ,  $p\text{-value} = 0.1091$
- Shapiro-Wilk test  
 $W = 0.73902$ ,  $p\text{-value} = 0.1574$

*L'ipotesi nulla del test di Shapiro e Wilk è la normalità della distribuzione dei dati. Il grafico mostra una certa normalità nella distribuzione, sebbene non appaia perfettamente simmetrico; d'altra parte, non si tratta né di dati continui né di dati privi di limiti superiore e inferiore. Il primo risultato proposto per il test di Shapiro e Wilk ha un p-value inferiore a 0,05 (peraltro a fronte di un valore di  $W$  vicinissimo a 1) e come tale indurrebbe a rifiutare l'ipotesi nulla; tra gli altri due, l'ultimo non è molto verosimile, proponendo di non rifiutare l'ipotesi nulla di normalità ( $P = 0,1574$ ) a fronte di un valore di  $W$  molto lontano da 1. Il risultato corretto è il quindi il secondo e la distribuzione è da considerarsi normale.*

### 2.3 Prodotto biologico e prodotto industriale

Per effettuare il confronto tra il gradimento del prodotto biologico e il gradimento del prodotto industriale, vengono effettuati otto diversi test. Qual è quello corretto? Cosa possiamo concludere?

- Test  $t$  a campioni indipendenti a due code:  
Two Sample t-test  
 $t = 0.3873$ ,  $df = 22.655$ ,  $p\text{-value} = 0.0326$   
alternative hypothesis: true difference in means is not equal to 0
- Test  $t$  a campioni appaiati a due code:  
Paired t-test  
 $t = 0.4166$ ,  $df = 14$ ,  $p\text{-value} = 0.6833$   
alternative hypothesis: true difference in means is not equal to 0
- Test  $t$  a campioni indipendenti a una coda:  
Two Sample t-test  
 $t = 0.3873$ ,  $df = 22.655$ ,  $p\text{-value} = 0.3511$   
alternative hypothesis: true difference in means is greater than 0
- Test  $t$  a campioni appaiati a una coda:  
Paired t-test  
 $t = 0.4166$ ,  $df = 14$ ,  $p\text{-value} = 0.0004$   
alternative hypothesis: true difference in means is greater than 0
- Test di Mann e Whitney a due code:  
Wilcoxon rank sum test with continuity correction  
 $W = 122$ ,  $p\text{-value} = 0.7029$   
alternative hypothesis: true location shift is not equal to 0
- Test di Wilcoxon a due code:  
Wilcoxon signed rank test with continuity correction  
 $V = 45.5$ ,  $p\text{-value} = 2.2436e-7$   
alternative hypothesis: true location shift is not equal to 0

- Test di Mann e Whitney a una coda:  
 Wilcoxon rank sum test with continuity correction  
 $W = 122$ , p-value = 0.3514  
 alternative hypothesis: true location shift is greater than 0
- Test di Wilcoxon a una coda:  
 Wilcoxon signed rank test with continuity correction  
 $V = 45.5$ , p-value = 0.0103  
 alternative hypothesis: true location shift is greater than 0

La variabile è a distribuzione normale. I campioni sono da considerarsi appaiati, in quanto è preferibile confrontare il parere di ogni assaggiatore per il prodotto biologico con il parere dello stesso assaggiatore per il prodotto industriale, a maggior ragione perché si tratta di assaggiatori non professionisti e quindi, si suppone, influenzati da variabili personali. La direzione della preferenza, inoltre, non è nota a priori. Tutto questo porta a decidere per il test  $t$  a campioni appaiati a due code. Il p-value è superiore a 0,05 (0,6833) e quindi le differenze tra i due campioni sono da ritenersi non significative.

## 2.4 Correlazioni interne

Il fatto che si tratti di assaggiatori non professionisti potrebbe portare a pensare che ci siano degli “errori di misura” nel gradimento: un assaggiatore meglio disposto, quindi, potrebbe dare un voto più alto a entrambi i prodotti di uno più pretenzioso. Si decide di indagare questo punto attraverso un modello di correlazione lineare tra i voti del prodotto biologico e quelli di quello industriale: se venisse individuata una correlazione, significherebbe che un assaggiatore che dia un voto (per esempio) alto al prodotto biologico tenderebbe a darlo anche a quello industriale. Di seguito (tabella 2) vengono mostrati i risultati: cosa se ne può concludere?

Tabella 2: Modello di correlazione lineare.

	Stima	p-value	$r$	$R^2$
Intercetta	5,5417			
Pendenza	0,2639	0,5806	0,15526	0,02411

La correlazione è totalmente assente ( $P = 0,5806$ ,  $r = 0,1553$ ,  $R^2 = 0,0241$ ). Non sembra quindi che ci siano evidenti “errori di misura” nel gradimento del panel, per quanto composto da assaggiatori non professionisti.

## 2.5 Interessi commerciali

La catena di supermercati cerca anche di capire se i voti pari a 9 o 10 ottenuti dal prodotto industriale siano indice di un migliore gradimento di questo prodotto rispetto a quello biologico o se siano oscillazioni casuali. Per farlo, conta il numero di voti pari a 9 e 10 per i due prodotti e li mette a confronto. Quale test statistico bisogna adoperare per stabilire se c'è una differenza nei numeri di voti pari a 9 e 10 nei due prodotti? Cosa se ne può concludere?

*Il test corretto è il test del  $\chi^2$ , perché il numero di voti di un certo tipo è una variabile qualitativa. Si ottiene una tabella a doppia entrata:*

	$\leq 8$	$\geq 9$
Prodotto biologico	11	4
Prodotto industriale	10	5

*Il valore di  $\chi^2$  con un grado di libertà e correzione di continuità è 0, che è ovviamente inferiore ai valori tabulati ( $P = 1$ ). Concludiamo che non c'è la benché minima differenza nella distribuzione dei voti pari a 9 e 10.*