Metodi statistici per le ricerche di mercato

Prof.ssa Isabella Mingo
A.A. 2016-2017

SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Scienze Politiche, Sociologia, Comunicazione
Corso di laurea Magistrale in «Organizzazione e marketing per la comunicazione d'impresa»

Accettare o rifiutare l'ipotesi nulla: Alcune considerazioni

- ✓ I test di significatività sono test statistici che quantificano i dati in senso di probabilità: i livelli del 5% (0.05) e dell' 1% (0.01) sono livelli accettati come ***limiti del tutto convenzionali*** per stabilire la significatività di uno scarto dall'ipotesi nulla
- ✓ Ad esempio, il livello di 5% sta ad indicare che su 100 campioni estratti dalla popolazione 95 di essi produrranno una differenza tra la stima campionaria e il valore del parametro ipotizzato in H_0 piccola o nulla e solo 5 mostreranno una differenza molto alta

Errori nei test di ipotesi

- La procedura del test delle ipotesi è soggetta a due tipi di errore
  - *errore di I specie detto* alfa che consiste nel considerare *valide differenze che in realtà non esistono*
    - viene respinta H_0, mentre H_0 è vera
    - *Pr (errore I specie)* $= \alpha$

    Le tecniche per ridurre l'errore alfa consistono nel **ridurre** il livello di significatività del test

  - *errore di II specie detto Beta* non considera differenze che realmente sono presenti nella realtà
    - H_0 viene accettata, ma in realtà H_0 è falsa
    - *Pr (errore II specie)* $= \beta$

    Le tecniche per ridurre la probabilità dell'errore beta consistono nell'**aumentare** il livello di significatività

    Per ridurre simultaneamente i due tipi di errore occorre aumentare l'ampiezza campionaria

Confronto tra medie campionarie

Per stabilire se la differenza tra due valori medi campionari è significativa o è dovuta al caso si deve distinguere tra due situazioni:

- I due campioni sono indipendenti (sono tratti da popolazioni diverse)
- I due campioni sono dipendenti (sono estratti dalla stessa popolazione)

Si supponga ad esempio di voler confrontare :
- il consumo medio di prodotti solari tra un campione di uomini e uno di donne: si tratta di due popolazioni distinte o meglio due segmenti di una popolazione e dunque i due campioni sono indipendenti;
- la soddisfazione di un gruppo di clienti prima e dopo l'introduzione di una miglioria nel servizio di assistenza: stessa popolazione e stesso campione, ma in momenti diversi e dunque i due campioni sono dipendenti o appaiati.

Confronto tra medie campionarie : campioni indipendenti

Nel caso di campioni indipendenti di grandi dimensioni (n>30), di numerosità n_1 e n_2 la variabile differenza tra le due medie campionarie tende a una distribuzione normale con media :

$$\mu_{\overline{X_1}-\overline{X_2}} = \mu_1 - \mu_2$$

ed Errore standard:

$$\sigma_{\overline{X_1}-\overline{X_2}} = \sqrt{\sigma^2_{\overline{X_1}} + \sigma^2_{\overline{X_2}}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Il test sulle differenze medie , utilizzando la distribuzione Normale standardizzata, sarà:

$$Z = \frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

dove σ_1^2 e σ_2^2 sono le varianze delle due popolazioni, che possono essere stimate con gli scarti quadratici medi corretti dei rispettivi campioni.

Confronto tra medie campionarie : campioni indipendenti - Esercizio

Si supponga ad esempio di voler confrontare :
- il consumo medio di prodotti solari tra un campione di uomini n_1= 200 e uno di donne n_2=200.

Nel campione degli uomini risulta un consumo medio di 150 ml con uno scarto quadratico medio di 2,5 ml, mentre nel campione di donne risulta un consumo medio di 153 ml con uno scarto quadratico medio di 1,5 ml.
Con un livello di significatività del 5%, si può affermare che esistono differenza significative tra uomini e donne nel consumo dei prodotti solari o le differenza rilevate sono trascurabili?

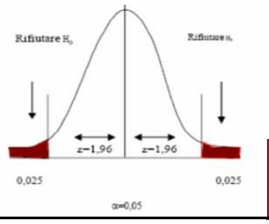
1) $H_0: \overline{X_1} = \overline{X_2}$ $\quad H_1 : \overline{X_1} \neq \overline{X_2}$

2) Distribuzione probabilistica: distribuzione normale Z (n>30)
Test a due code(si è interessati a verificare se il consumo medio è diverso
3) Livello di confidenza (1- α)=95%; livello di significatività : α/2=0,25.
4) Valore di Z critico in corrispondenza del livello di significatività prescelto
Z= ± 1,96

4) Calcolo del valore della statistica test

$$Z = \frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{s_1^2}{n_1-1}+\frac{s_2^2}{n_2-1}}} = \frac{150-153}{\sqrt{\frac{2.5^2}{199}+\frac{1.5^2}{199}}} = -14{,}52$$

Il valore Z è inferiore a -1,96, pertanto rifiutiamo l'ipotesi nulla: il consumo medio di prodotti solari tra uomini e donne è significativamente diverso, con un livello di significatività del 5%



Confronto tra medie campionarie : campioni indipendenti - Esercizio

- Da una ricerca di mercato è risultato che le vendite medie settimanali di quotidiani in un campione casuale di 110 edicole della zona Nord della città è di 2500 euro con uno scarto quadratico medio di 200, mentre su un campione di 88 edicole della zona Sud, le vendite medie ammontano a 2578 euro con uno scarto quadratico medio di 150euro.

Con un livello di significatività dell'1 %, si può affermare che esistono differenze significative tra le due zone?

1) $H_0: \overline{X_1} = \bar{X}_2$ $\quad H_1 : \overline{X_1} \neq \bar{X}_2$

2) Distribuzione probabilistica: distribuzione normale Z (n>30)
Test a due code
3) Livello di confidenza (1- α)=99%; livello di significatività : α/2=0,005.
4) Valore di Z critico in corrispondenza del livello di significatività prescelto
5) Z= ± 2,58

4) Calcolo del valore della statistica test

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}} = \frac{2500 - 2578}{\sqrt{\frac{200^2}{109} + \frac{150^2}{87}}} = -3{,}13$$

Il valore Z è inferiore a -2,58, rifiutiamo l'ipotesi nulla: le vendite medie sono significativamente diverse nelle due zone, con un livello di significatività del 5%.

Differenze tra proporzioni campionarie : campioni indipendenti

Nel caso di campioni indipendenti di grandi dimensioni (n>30), di numerosità n_1 e n_2 la variabile differenza tra le due proporzioni campionarie tende a una distribuzione normale con media :

$$P_{p_1 - p_2} = |p_1 - p_2|$$

ed Errore standard stimato:

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Test di verifica di ipotesi sul confronto tra proporzioni campionarie : campioni indipendenti

Nel caso di campioni indipendenti di grandi dimensioni (n>30), di numerosità n_1 e n_2 pe confrontare le proporzioni $p_1 e\, p_2$ il test stabilirà

$H_0 : p_1 - p_2 = 0$ $H_1 : p_1 - p_2 \neq 0$

Sotto l'assunzione $p_1 = p_2$ stimiamo il valore comune di $p_1 e\, p_2$ attraverso la proporzion campionaria per l'intero campione stima conglobata) :

$$\hat{P} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$

dove $\hat{P}$ indica la proporzione media calcolata sui due campioni

$$\sigma_{p_1 - p_2} = \sqrt{\frac{\hat{P}\,\hat{Q}}{n_1} + \frac{\hat{P}\,\hat{Q}}{n_2}}$$

Il test sulle differenze tra proporzioni, utilizzando la distribuzione Normale standardizzata, sarà:

$$Z = \frac{|p_1 - p_2|}{\sqrt{\frac{\hat{P}\,\hat{Q}}{n_1} + \frac{\hat{P}\,\hat{Q}}{n_2}}}$$

a.a. 20

Confronto tra proporzioni campionarie : campioni indipendenti - Esercizio

Su un campione di consumatori di n_1= 600 consumatori si è rilevato che la marca di Jeans preferita dal 28% è la ABC. Reiterando l'indagine dopo un anno su un altro campione n_2 di 600 intervistati , risulta che la marca ABC è preferita dal 25%.
Con un livello di significatività del 5%, si può affermare che esistono differenze significative tra le preferenze dei consumatori ?

1) $H_0: p_1 = p_2$ $H_1 : p_1 \neq p_2$

2) Distribuzione probabilistica: distribuzione normale Z (n>30)
Test a due code
3) Livello di confidenza (1- α)=95%; livello di significatività : α/2=0,025.
4) Valore di Z critico in corrispondenza del livello di significatività prescelto : Z= ± 1,96

4) Calcolo del valore della statistica test

$Z = \frac{|p_1 - p_2|}{\sqrt{\frac{\hat{P}\,\hat{Q}}{n_1} + \frac{\hat{P}\,\hat{Q}}{n_2}}}$ $\hat{P} = \frac{0{,}28 \cdot 600 + 0{,}25 \cdot 600}{600 + 600} = 0{,}265$ $\hat{Q} = = (1 - 0{,}265) = 0{,}735$

$$Z = \frac{|0{,}28 - 0{,}25|}{\sqrt{\frac{0{,}265 \cdot 0{,}735}{600} + \frac{0{,}265 \cdot 0{,}735}{600}}} = 1{,}178$$

a.a

Il valore Z è compreso nell'intervallo -1,96 ---- + 1,96 pertanto accettiamo l'ipotesi nulla: la preferenza dei consumatori per la marca ABC non è cambiata nel tempo.

Esercizio

Da una indagine riguardante l'efficacia di un nuovo spot pubblicitario è risultato che, su un campione di 450 giovani da 25 a 35 anni, il 40% ha dichiarato di aver acquistato il prodotto a seguito dello spot. Su un altro campione di 380 adulti da 36 a 45 anni invece la percentuale è risultata del 31%. Con un livello di significatività dell'1%, si può affermare che lo spot ha una diversa efficacia sui due target?

1) $H_0: p_1 = p_2$ $\quad H_1: p_1 \neq p_2$

2) Distribuzione probabilistica: distribuzione normale Z (n>30)
Il valore Z dei nostro test è maggiore di 2,58 pertanto rifiutiamo l'ipotesi nulla: l'efficacia dello spot sui due target è significativamente differente con un livello di significatività dell'1%.

Confronto tra proporzioni interdipendenti

Nelle ricerche di mercato a volte le proporzioni messe a confronto sono calcolate sullo stesso campione e dunque non sono indipendenti tra loro.

Si voglia ad esempio verificare l'ipotesi che non vi siano differenze significative tra le percentuali dei clienti che, in un campione di 480 individui, hanno scelto il marchio ALFA per la qualità e per il costo (tabella seguente)

Motivi della scelta del marchio ALFA	%
Qualità dei prodotti	28
Prezzi dei prodotti	31
Facilità di reperire i prodotti	18
Altri motivi	23
Totale	100

In tal caso le modalità di risposta sono state rilevate sullo stesso campione e si escludono tra loro, pertanto nel calcolare l'errore standard, occorre tenerne conto utilizzando la seguente relazione:

$$\sigma_{p_1 - p_2} = \sqrt{\frac{1}{n}(p_1 q_1 + p_2 q_2 + 2 p_1 p_2)}$$

$$\sigma_{p_1 - p_2} = \sqrt{\frac{1}{480}(0{,}28 \cdot 0{,}72) + (0{,}31 \cdot 0{,}69) + 2(0{,}28 \cdot 0{,}31)} = 0{,}035$$

Confronto tra proporzioni interdipendenti (segue)

Il valore del test per il nostro campione di 480 clienti sarà:

$$Z = \frac{|0{,}28-0{,}31|}{0{,}035} = 0{,}86$$

Se scegliamo un livello di significatività di 0,01 , il valore Z critico è $\pm$ 2,58 e il valore del nostro test cade nell'intervallo - 2,58 ---- + 2,58 , pertanto accettiamo l'ipotesi nulla: le differenze tra le percentuali di risposte alle modalità « Qualità» e «Prezzi» non sono significative; le due motivazioni nel campione considerato si equivalgono.

Esercizio

Con un livello di significatività del 5%, si verifichi l'ipotesi che non vi siano differenze significative tra le percentuali dei clienti che, in un campione di 350 individui, hanno scelto il gestore WINDOFON per la trasparenza delle offerte e per la copertura (tabella seguente)

Motivazioni scelta gestore	%
Copertura	26
Costi	28
Trasparenza offerte	35
Servizio assistenza	11
Totale	100

1) $H_0: p_1 = p_2$ $\quad H_1: p_1 \neq p_2$

2) Distribuzione probabilistica: distribuzione normale Z (n>30); Test a due code

Livello di confidenza (1- α)=95%; livello di significatività : α/2=0,025.

3) Valore di Z critico in corrispondenza del livello di significatività prescelto : Z= $\pm$ 1,96

4) Calcolo del valore della statistica test

$$\sigma_{p_1-p_2} = \sqrt{\frac{1}{350}(0{,}35 \cdot 0{,}65) + (0{,}26 \cdot 0{,}74) + 2(0{,}35 \cdot 0{,}26)} = 0{,}042$$

$$Z = \frac{|0{,}35-0{,}26|}{0{,}042} = 2{,}14$$

il valore del nostro test è maggiore a 1,96, pertanto cade nella zona di rifiuto dell'ipotesi nulla. Le differenze tra le percentuali di risposte alle modalità Trasparenza delle offerte e Copertura sono significative.

Confronto tra proporzioni interdipendenti nel caso di modalità di risposta che non si escludono tra di loro

Nelle ricerche di mercato a volte le proporzioni messe a confronto sono calcolate sullo stesso campione e rigurdano modalità di risposta multiple che non sono escludono tra loro.

Si voglia ad esempio verificare l'ipotesi al livello di significatività α=0,05 che non vi siano differenze significative tra le percentuali dei clienti che, in un campione di 600 individui, hanno scelto il marchio ALFA per la qualità e per i prezzi , sapendo che le risposte alla domanda erano multiple e che il 4% dei rispondenti hanno scelto sia la prima che la seconda modalità (tabella seguente)

Perché ha scelto il marchio ALFA (più risposte)

Motivi	%
Qualità dei prodotti	13
Prezzi	10
Facilità di reperire i prodotti	60
Assistenza post-vendità	24
Altri motivi	3

In tal caso le modalità di risposta sono state rilevate sullo stesso campione e non si escludono tra loro, pertanto nel calcolare l'errore standard, occorre tenerne conto utilizzando la seguente relazione:

$$\sigma_{p_1 - p_2} = \sqrt{\frac{1}{n}(p_1 q_1 + p_2 q_2 + 2(p_1\, p_2 - p_{12})}$$

Dove p_{12} è la proporzione di rispondenti che hanno scelto le due modalità

$$\sigma_{p_1 - p_2} = \sqrt{\frac{1}{600}(0{,}13 \cdot 0{,}87) + (0{,}10 \cdot 0{,}90) + 2(0{,}13 \cdot 0{,}10) - 0{,}04} = 0{,}0158$$

$$Z = \frac{|0{,}13 - 0{,}10|}{0{,}0158} = 1{,}9$$

il valore del nostro test è inferiore a |1,96|, pertanto cade nella zona di accettazione dell'ipotesi nulla. Le differenze tra le percentuali di risposte alle modalità qualità e prezzi non sono significative, posto un livello di significatività del 5%.

Esercizio

Con un livello di significatività del 5%, si verifichi l'ipotesi che non vi siano differenze significative tra le percentuali dei clienti che, in un campione di 350 individui, hanno scelto il gestore WINDOFON per la trasparenza delle offerte e per la copertura sapendo che il 3% ha scelto entrambe le risposte (tabella seguente)

Motivazione scelta gestore(risposte multiple)	%
Copertura	36
Costi	38
Trasparenza offerte	35
Servizio Assistenza	24

Test del Chi Quadrato

E' adatto alla soluzione di problemi riguardanti l'analisi delle contingenze e dunque lo studio delle relazioni tra mutabili.

Nella prima parte di questo corso abbiamo imparato, a partire da una tabella a doppia entrata , a calcolare il χ^2 :

$$\chi^2 = \sum_{i=1}^{H}\sum_{j=1}^{K}\frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}}$$

Il problema fondamentale è stabilire se le differenze fra le frequenze osservate n_{ij} e le frequenze teoriche n^*_{ij} , ossia le frequenze attese nel caso di indipendenza tra le due mutabili, sono statisticamente significative o dovute al caso.

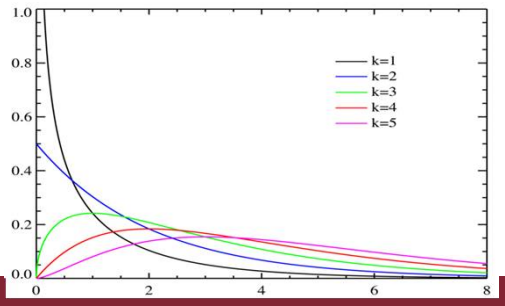
Si tratta dunque di verificare una ipotesi statistica seguendo le stesse procedure viste precedentemente. Questa volta però ci riferiamo ad un'altra distribuzione di probabilità: quella del Chi quadrato.

La distribuzione del Chi Quadrato

Ha le seguenti caratteristiche:

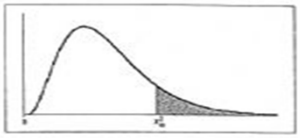
- Varia tra 0 e +∞
- È asintotica all'asse delle ascisse
- L'area sottostante esprime la probabilità corrispondente a ciascun valore di χ^2
- La forma dipende dai gradi di libertà k.
  In una tabella a doppia entrata i gradi di libertà sono dati da:
  k= (c-1)(r-1)

- la probabilità corrispondente a ciascun valore di χ^2 per k gradi di libertà si può trovare in apposite tavole.



Tavole statistiche – DISTRIBUZIONE CHI-QUADRATO

Valori della variabile χ^2 (χ^2_α) in corrispondenza di aree α sotto la coda della distribuzione chi-quadrato e al variare dei gradi di libertà ν



ν	α=0,005	0,01	0,025	0,05	0,100	0,200	0,250	0,500
1	7,88	6,63	5,02	3,84	2,71	1,64	1,32	0,45
2	10,60	9,21	7,38	5,99	4,61	3,22	2,77	1,39
3	12,84	11,34	9,35	7,81	6,25	4,64	4,11	2,37
4	14,86	13,28	11,14	9,49	7,78	5,99	5,39	3,36
5	16,75	15,09	12,83	11,07	9,24	7,29	6,63	4,35
6	18,55	16,81	14,45	12,59	10,64	8,56	7,84	5,35
7	20,28	18,48	16,01	14,07	12,02	9,80	9,04	6,35
8	21,95	20,09	17,53	15,51	13,36	11,03	10,22	7,34
9	23,59	21,67	19,02	16,92	14,68	12,24	11,39	8,34
10	25,19	23,21	20,48	18,31	15,99	13,44	12,55	9,34
11	26,76	24,73	21,92	19,68	17,28	14,63	13,70	10,34
12	28,30	26,22	23,34	21,03	18,55	15,81	14,85	11,34
13	29,82	27,69	24,74	22,36	19,81	16,98	15,98	12,34
14	31,32	29,14	26,12	23,68	21,06	18,15	17,12	13,34
15	32,80	30,58	27,49	25,00	22,31	19,31	18,25	14,34
16	34,27	32,00	28,85	26,30	23,54	20,47	19,37	15,34
17	35,72	33,41	30,19	27,59	24,77	21,61	20,49	16,34
18	37,16	34,81	31,53	28,87	25,99	22,76	21,60	17,34
19	38,58	36,19	32,85	30,14	27,20	23,90	22,72	18,34
20	40,00	37,57	34,17	31,41	28,41	25,04	23,83	19,34
21	41,40	38,93	35,48	32,67	29,62	26,17	24,93	20,34
22	42,80	40,29	36,78	33,92	30,81	27,30	26,04	21,34
23	44,18	41,64	38,08	35,17	32,01	28,43	27,14	22,34
24	45,56	42,98	39,36	36,42	33,20	29,55	28,24	23,34
25	46,93	44,31	40,65	37,65	34,38	30,68	29,34	24,34
26	48,29	45,64	41,92	38,89	35,56	31,79	30,43	25,34
27	49,65	46,96	43,19	40,11	36,74	32,91	31,53	26,34
28	50,99	48,28	44,46	41,34	37,92	34,03	32,62	27,34

Test del Chi quadrato: esempio

In base ai dati tratti da un campione casuale, riportati nella tabella seguente, e sapendo che il valore del Chi quadrato è:
$\chi^2 = 10,89$, possiamo affermare con un livello di significatività del 5% che la soddisfazione per la qualità del prodotto è significativamente differente tra maschi e femmine nella popolazione di riferimento?

	Soddisfazione per la qualità del prodotto				
	Molto	Abbastanz	Poco	per niente	Totale
Maschio	62	212	110	87	471
Femmina	58	218	167	74	517
Totale	120	430	277	161	988

Test del Chi quadrato: esempio

1) H_0: $n_{ij} = n^*_{ij}$ (le variabili sono indipendenti)
   H_a: $n_{ij} \neq n^*_{ij}$ (le variabili sono dipendenti)

2) Distribuzione di probabilità: Chi quadrato
   livello di significatività α=0,05
   Gradi di libertà: k= (c-1)(r-1)= (4-1)(2-1)=3
   Sulle tavole del Chi quadrato si individua il valore di χ^2 critico in corrispondenza del livello di significatività prescelto e dei gradi di libertà.
   χ^2= 7,81

3) Il valore nella nostra tabella è
χ^2 = 10, 89 > 7,81
Cade nell'area di rifiuto.
Pertanto possiamo affermare, con un livello di significatività del 5%, che le variabili non sono indipendenti, pertanto c'è relazione tra sesso degli intervistati e soddisfazione per la qualità del prodotto.