

CAPITOLO 2 – Disponibilità e produzione delle informazioni statistiche.

COSA CI SERVE SAPERE

E' utile ripassare i seguenti argomenti:

Media, varianza, distribuzione binomiale, distribuzione normale, stima puntuale, stima intervallare, verifica di ipotesi.



Argomenti del capitolo

Le informazioni statistiche per l'azienda: concetti generali e definizioni (par. 2.1)

Le principali fonti di dati per l'azienda (par. 2.2, par. 2.3, par 2.4)

Qualità della statistica e statistica ufficiale (par. 2.5)

La produzione di dati ad hoc: le indagini campionarie (par. 2.6)

Paragrafo 2.6

La produzione di dati ad hoc: le indagini campionarie

- L'indagine campionaria e le sue fasi
- Popolazione obiettivo, popolazione effettiva, popolazione d'indagine
- Selezione del campione
- La Stima di media e proporzione nel caso di campionamento casuale semplice e campionamento stratificato
- Tecniche di rilevazione
- Il questionario
- Valutazione dei risultati

L'indagine campionaria e le sue fasi

Quando i dati secondari non sono in grado di rispondere ad un quesito aziendale, è necessario che l'azienda realizzi in proprio una rilevazione (sondaggio di opinione o ricerca di mercato)

Esempio: per valutare il mercato potenziale delle lavastoviglie in provincia di Pisa è indispensabile conoscere quanti nuclei familiare possiedono una lavastoviglie. Le informazioni disponibili forniscono indicazioni dettagliate solamente per vasti territori (ripartizioni territoriali o regioni). In questo caso è necessario impostare una rilevazione ad hoc.

Due approcci



Indagine censuaria, totale o completa: il fenomeno di interesse è rilevato su **tutte le unità del collettivo**

Vantaggi

Non è presente l'errore campionario

Svantaggi

Costosa, lunghi tempi di realizzazione, alta probabilità di errori non campionari

Indagine parziale o campionaria: il fenomeno di interesse è rilevato su una parte (**campione**) delle unità del collettivo

Vantaggi:

Relativamente meno costosa, tempi di realizzazione più brevi, minore probabilità di errori non campionari.

Svantaggi

Presenza dell'errore campionario

Le fasi dell'indagine campionaria

- 1) definizione degli **obiettivi conoscitivi** della rilevazione;
- 2) identificazione della **popolazione di riferimento** o **popolazione obiettivo**;
- 3) scelta dei criteri di **selezione del campione** (piano di campionamento);
- 4) scelta della **metodologia di stima dei parametri di interesse**, nel caso di indagine campionaria con campionamento probabilistico;
- 5) scelta della **modalità di raccolta dei dati**;
- 6) messa a punto del **questionario**;
- 7) organizzazione della fase di **rilevazione dei dati**;
- 8) **valutazione dei costi** di realizzazione dell'intera indagine;

Alcuni concetti fondamentali

- **Popolazione obiettivo (universo, collettivo):** popolazione oggetto di indagine ovvero la popolazione a cui *vogliamo che siano generalizzate le informazioni rilevate sul campione.*

Supponiamo di voler condurre un'indagine campionaria sul fatturato medio annuo delle aziende che producono calzature in Italia nel 2011. Qual è la popolazione obiettivo?

- **Campione:** parte di popolazione sulla quale vengono raccolte le informazioni. Dimensione = n
- **Unità di rilevazione:** unità del campione
- **Unità di analisi:** individui/entità su cui vengono rilevate le informazioni

Dalla Popolazione di selezione alla Popolazione di indagine

- **Campione effettivo** e **Campione teorico** possono differire a causa delle “mancate risposte”
- Il campione effettivo fornisce evidenza soltanto per quella parte della popolazione (**Popolazione di indagine**) rappresentata dalle unità effettivamente osservate



Selezione del campione

- **Obiettivo di un'indagine campionaria:** la stima di parametri della popolazione (quelli più frequentemente richiesti sono medie, totali, proporzioni) sulla base della evidenza fornita dal campione.
- **Errore statistico:** differenza tra la stima del parametro e il valore del parametro nella popolazione. Nel caso di indagini campionarie, l'errore statistico comprende anche **l'errore campionario**
- **Selezione del campione:** soltanto se il campione è selezionato con meccanismo casuale (campione probabilistico), è possibile stimare l'errore campionario.

Campioni probabilistici e non probabilistici

- Nei **Campioni probabilistici**, le unità campionarie sono selezionate con meccanismo casuale. Le unità dei **Campioni non probabilistici** sono selezionate sulla base di scelte arbitrarie spesso dettate da considerazioni di ordine pratico.
- Per selezionare campioni probabilistici è necessario disporre della **lista di campionamento**. Quando è impossibile (oppure eccessivamente costoso ed oneroso), si ricorre a schemi di campionamento non probabilistici.

Campionamento probabilistico: alcuni concetti di base

Le unità sono selezionate con meccanismo casuale ed hanno tutte una probabilità nota e non nulla di essere selezionate

Y = variabile casuale che rappresenta il carattere osservato sulla popolazione di interesse

Θ = parametro da stimare

Immaginiamo di estrarre dalla popolazione il campione casuale:

$\{Y_1, Y_2, \dots, Y_n\}$
con realizzazioni campionarie $\{y_1, y_2, \dots, y_n\}$

Stimatore

$$T_n = f(Y_1, Y_2, \dots, Y_n)$$

Stima

$$t_n = f(y_1, y_2, \dots, y_n)$$

Altre definizioni fondamentali

$$f = \frac{n}{N} \quad \text{Frazione di campionamento}$$

π_i **Probabilità di inclusione** nel campione della singola unità i

ω_i **Peso base o fattore di espansione all'universo**: reciproco della probabilità di inclusione, può essere interpretato come il numero di elementi della popolazione rappresentati dalla unità i -esima

Lo stimatore di uno stesso parametro è diverso a seconda del **piano di campionamento** prescelto, cioè della regola con cui vengono selezionate le unità del campione.

Principali Schemi di campionamento PROBABILISTICO

Casuale semplice (con o senza ripetizione)

Campionamento Sistemático

Campionamento Stratificato

Campionamento a Grappoli

Campionamento a Stadi

CAMPIONAMENTO CASUALE SEMPLICE CON E SENZA RIPETIZIONE

- Il campionamento casuale semplice (CCS) rappresenta la base di riferimento per qualunque altro tipo di campionamento probabilistico.
- La procedura di selezione del campione è assimilabile all'estrazione di palline da un'urna, con o senza ripetizione.
- Il CCS attribuisce ad ogni singola unità della popolazione la stessa probabilità di essere estratta.

Probabilità di inclusione

$$\pi_i = \frac{n}{N}$$

Peso base ω_i (o fattore di espansione all'universo)

$$\omega_i = \frac{N}{n}$$

CAMPIONAMENTO SISTEMATICO

- Può essere utilizzato soltanto se le unità della lista di campionamento sono ordinabili secondo un carattere.
- Il campione selezionato è condizionato dall'ordinamento della lista. Ad esempio, nel caso di selezione sistematica da una lista di aziende ordinate per dimensione, il campione sistematico garantirà la presenza di aziende di dimensione diversa
- Il Campionamento sistematico per intervallo monetario (Monetary Unit Sampling), utilizzato nelle procedure di controllo contabile e di certificazione di bilancio, ha la caratteristica di attribuire una probabilità di inclusione nel campione maggiore per le voci contabili di importo più elevato [vedi Box 2.1].

Equivale allo schema di selezione utilizzato per estrarre un campione casuale semplice soltanto se l'estrazione è preceduta da un'operazione che dispone le unità della **lista in ordine casuale**.

REGOLA DI SELEZIONE DEL CAMPIONAMENTO SISTEMATICO

1. È selezionato casualmente un numero intero j compreso tra 1 e k , dove $k = \text{int}(N/n)$, ovvero k è il più grande intero minore od uguale a (N/n) . L'unità della lista corrispondente al numero j rappresenta la prima unità selezionata.
2. Le $(n-1)$ unità rimanenti sono individuate ogni k unità, seguendo l'ordine della lista. Per questo motivo k è detto **passo di campionamento**.

IL CAMPIONAMENTO STRATIFICATO

- Si basa sull'ipotesi è che esista una connessione tra **l'intensità del carattere di interesse** e una serie di **attributi noti** posseduti dalla popolazione (variabili ausiliarie o supplementari)
- Il primo passo consiste nel **classificare le unità della lista di campionamento in subpopolazioni omogenee** rispetto alle variabili ausiliarie. Si ottengono così sub-popolazioni o **strati**.
- Da ogni strato è **selezionato un campione casuale semplice** ottenendo un numero di campioni indipendenti pari al numero di strati. Infine, l'unione di tali campioni produce il **campione stratificato**.

Notazione per il Campionamento stratificato

h = singolo strato

H = numero di strati (per cui $h=1, \dots, H$)

N_h = dimensione dell' h -esimo strato nella popolazione;

n_h = dimensione del campione estratto dallo strato h

$W_h = \frac{N_h}{N}$ = proporzione della popolazione nello strato h

$f_h = \frac{n_h}{N_h}$ = frazione di campionamento per lo strato h

Determinazione della numerosità campionaria per ciascun strato

Campione con stratificazione proporzionale: la frazione di campionamento del singolo strato è la stessa per ogni strato ed è pari alla frazione di campionamento definita per il campione nel suo complesso

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} \quad \text{Per ogni } h=1, \dots, H$$

Campione con stratificazione non proporzionale: la frazione di campionamento varia da strato a strato (assicurare maggior precisione alle stime degli strati poco numerosi o in cui la variabilità di Y è maggiore)

Vantaggi del Campionamento stratificato

- Rispetto al CCS assicura maggior precisione delle stime, a parità di numerosità campionaria.
- Consente di ottenere stime per strato (a condizione che la numerosità degli strati sia adeguata).

Esercizio

Sondaggio di opinione tra i clienti di un sito online specializzato nella vendita di libri e di prodotti multimediali. Il titolare del sito ha a disposizione l'elenco di 3000 clienti e di ciascuno di essi conosce l'età, il sesso e la cifra media per articolo spesa negli ultimi 12 mesi. Individuate il numero di unità da estrarre da ciascun strato, ipotizzando che $n = 350$ e che la popolazione risulti distribuita tra i 12 strati come riportato nella seguente tabella.

Genere	Fascia di spesa	Fascia di età			Totale
		18-39	40-59	60 e oltre	
Femmine	≤ 50 euro	300	360	240	900
	> 50	210	270	240	720
Maschi	≤ 50 euro	180	330	225	735
	> 50	162	300	183	645
Totale		852	1260	888	3000

Svolgimento e soluzione, libro di testo paragrafo 2.6

IL CAMPIONAMENTO A GRAPPOLI

- La lista degli N elementi è suddivisa in **grappoli**, ciascuno **rappresentativo della popolazione**, ovvero tale da riprodurre la variabilità del carattere di interesse nella popolazione.
- Selezione casuale di un numero di grappoli ed inclusione nel campione di tutti gli elementi ad essi appartenenti.
- Idealmente i grappoli devono essere individuati in modo che la variabilità del parametro da stimare sia **alta entro i grappoli** e **bassa tra grappoli**
- Nella prassi i grappoli corrispondono a raggruppamenti realmente esistenti (città, quartieri, edifici, famiglie).

IL CAMPIONAMENTO A STADI

- Può essere considerato come una variante del campionamento a grappoli.
- **Unità di primo stadio:** grappoli
- **Unità di secondo stadio:** unità elementari appartenenti ai grappoli

Selezione del campione: selezione di un campione di grappoli; selezione di campioni di unità elementari da ciascun grappolo

Il **Campionamento a due stadi** è utilizzato per rilevazioni campionarie di **grandi dimensioni**: ad es. molte indagini Istat con copertura nazionale, nelle quali i Comuni sono le unità di primo stadio e le Famiglie registrate nelle Anagrafi dei Comuni sono le unità di secondo stadio.

CAMPIONAMENTO NON PROBABILISTICO

Si ricorre al campionamento non probabilistico:

- per contenere i costi e i tempi di esecuzione dell'indagine;
- quando è impossibile reperire la lista delle unità della popolazione

Svantaggi:

- non offrono alcuna garanzia di rappresentatività della popolazione
- non consentono di fornire una stima dell'errore campionario

Principali Schemi di campionamento NON PROBABILISTICO

Campionamento di Comodo

Campionamento Ragionato

Campionamento per quote

Campionamento di comodo

L' intervistatore decide arbitrariamente chi intervistare

Nessuna garanzia di rappresentatività perché:

- i) l'insieme dei soggetti intervistati presenti nel luogo e nel momento scelti per l'intervista difficilmente rispecchia la popolazione di interesse;
- ii) L'intervistatore può consapevolmente o meno intervistare in prevalenza persone con certe caratteristiche;
- iii) non è prevista, in genere, alcuna correzione per la non risposta.

Esempi: interviste presso esercizi commerciali, televoto

Campionamento ragionato

Vengono selezionate specifiche unità in quanto ritenute depositarie delle informazioni rilevanti per il buon esito della ricerca

Esempi:

- campione di prodotti (detto paniere) sul quale i rilevatori Istat rilevano mensilmente i prezzi per misurare l'inflazione.
- testimoni privilegiati od opinion leaders.

Campionamento per quote

- Ha le caratteristiche di un campionamento di comodo ma popolazione e campione **condividono esattamente la medesima composizione** rispetto a determinate caratteristiche delle unità osservate.
- Si differenzia dal campionamento stratificato (con stratificazione proporzionale) per **il metodo di selezione delle unità da ciascun strato della popolazione**: esse sono scelte dall'intervistatore che liberamente decide chi coinvolgere nell'indagine fino a **saturare ogni quota**.

La Stima di media e proporzione nel caso di campionamento casuale semplice e campionamento stratificato

Y è la caratteristica o variabile oggetto di indagine nella popolazione obiettivo.

I parametri da stimare sono:

μ = media della variabile Y nella popolazione

π = proporzione di casi che presentano il carattere Y nella popolazione

Esempi: stima del fatturato medio annuo delle imprese del Nord-est Italia; stima della proporzione di imprese che nel 2011 ha registrato un fatturato inferiore a quello dell'anno precedente

CCS: stima puntuale della media (μ)

Uno stimatore corretto della media è la media campionaria:

$$T_{\bar{Y}} = \frac{1}{n} \sum_{i=1}^n Y_i$$

Nel caso di **CCS con ripetizione**, l'errore standard è pari a:

$$ES(T_{\bar{Y}}) = \sqrt{\frac{\sigma_Y^2}{n}}$$

Ma σ^2 spesso non è nota, per cui è necessario ottenerne una stima a partire dai dati campionari

La varianza campionaria S_Y^2 è uno stimatore corretto della varianza

$$S_Y^2 = \frac{1}{n-1} (Y_i - \bar{Y})^2$$

Nel caso di **CCS senza reimmissione**, l'errore standard della media campionaria risulta essere

$$ES(T_{\bar{Y}}) = \sqrt{(1-f) \cdot \frac{S_Y^2}{n}}$$

Fattore di correzione per popolazioni finite - Tende a 1 quando f tende a 0

Esercizio

Un'impresa specializzata nella fornitura di assistenza idraulica 24 ore su 24, vuole stimare il "tempo medio di attesa" per gli interventi realizzati nell'ultimo mese, intervistando un campione di clienti. Dalla lista dei 115 clienti, viene estratto un campione casuale semplice di $n = 25$ clienti. I risultati delle osservazioni sono riportati nella tabella che segue.

Fornite la stima puntuale della media di Y e dell'errore standard

Y = numero di ore di attesa prima dell'intervento

n.	Etichette	Y	n.	Etichette	Y
1	209	1	13	15	2
2	295	4	14	46	3
3	178	3	15	300	2
4	253	4	16	232	4
5	24	0	17	282	3
6	357	3	18	167	3
7	159	3	19	173	1
8	271	0	20	219	1
9	279	4	21	13	4
10	244	1	22	85	0
11	239	2	23	35	3
12	20	0	24	239	1
			25	123	4

Soluzioni

$$\bar{y} = \frac{1}{25} (1 + 4 + 3 + \dots + 4) = 2,24$$

$$s_Y^2 = \frac{1}{25 - 1} \cdot \left[(1 - 2,24)^2 + (4 - 2,24)^2 \dots + ((4 - 2,24)^2) \right] = 2,11$$

$$(1 - f) = \left(1 - \frac{n}{N} \right) = 1 - \frac{25}{115} = 0,78$$

$$es(T_{\bar{y}}) = \sqrt{0,78 \cdot \frac{2,11}{25}} = 0,26$$

CCS: stima puntuale della proporzione (π)

Definiamo la variabile dicotomica Z che rileva la presenza/assenza del carattere osservato negli elementi della popolazione. In particolare la variabile Z assume valore 1 se il carattere è presente, 0 se assente.

Lo stimatore della proporzione T_p è uguale alla media campionaria della variabile Z , che a sua volta corrisponde alla proporzione dei casi del campione (proporzione campionaria) che presenta la caratteristica in esame.

$$T_P = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{n_k}{n}$$

dove n_k è il numero di unità del campione che possiedono l'attributo di interesse

L'errore standard della proporzione campionaria, nel caso di CCS senza ripetizione è dato da:

$$ES(T_P) = \sqrt{(1-f) \cdot \frac{\pi(1-\pi)}{n-1}}$$

Il valore di π (proporzione di casi nella popolazione) è ovviamente **ignoto**

Il problema viene affrontato nei seguenti modi:

- 1) si utilizza la proporzione campionaria P ;
- 2) sulla base di conoscenze a priori sul fenomeno in esame si indica una stima approssimativa della proporzione;
- 3) si pone $p=0,5$ in via precauzionale. Si dimostra infatti che l'errore standard raggiunge il suo massimo proprio quando $p = 0,5$. Di conseguenza l'errore standard che otterremo sarà eventualmente sovrastimato.

Esercizio

Riprendiamo l'esercizio svolto in precedenza e definiamo Z come segue:

$Z = 0$ se le ore di attesa sono minori o uguali a 3;

$Z = 1$ se le ore di attesa sono maggiori di 3

Stimate la proporzione di clienti che hanno atteso un numero di ore maggiore di 3 e il relativo errore standard, ponendo $p=0,5$

SOLUZIONE

Stima puntuale proporzione campionaria

$$p = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{1}{25} (6) = \frac{6}{25} = 0,24$$

Errore standard

$$es(T_{\pi}) = \sqrt{0,78 \cdot \frac{0,5 \cdot 0,5}{24}} = 0,09$$

Campionamento Stratificato: stima puntuale della media (μ)

media ponderata degli stimatori
ottenuti per ciascun strato.

Stimatore della media

$$T_{\bar{Y}_{ST}} = \sum_{h=1}^H W_h \bar{Y}_h = \sum_{h=1}^H W_h \cdot \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$$

N_h = dimensione dell' h -esimo strato nella popolazione;

$W_h = \frac{N_h}{N}$ = proporzione della popolazione nello strato h ;

\bar{Y}_h = stimatore della media dello strato h .

n_h = dimensione del campione estratto dall' h -esimo strato

Errore standard della media campionaria

$$ES(T_{\bar{Y}_{ST}}) = \sqrt{\sum_{h=1}^H W_h^2 \cdot (1 - f_h) \cdot \frac{S_h^2}{n_h}}$$

$$f_h = \frac{n_h}{N_h} = \text{frazione di campionamento relativa allo strato } h$$

$$S_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{ih} - \bar{Y}_h)^2 = \text{stimatore corretto della varianza nello strato } h$$

Campionamento Stratificato: stima puntuale della proporzione (π)

Stimatore della media

$$T_{P_{ST}} = \sum_{h=1}^H W_h P_h = \sum_{h=1}^H W_h \frac{r_h}{n_h}$$

Dove r_h indica il numero di elementi del campione dello strato h che presentano la caratteristica in esame

Errore standard della proporzione campionaria

$$es(T_{P_{ST}}) = \sqrt{\sum_{h=1}^H W_h^2 \cdot (1 - f_h) \cdot \frac{P_h \cdot (1 - P_h)}{n_h \cdot (n_h - 1)}}$$

Osservazione importante

Nel caso di campionamento con stratificazione proporzionale, la frazione di campionamento di ciascuno strato risulta uguale a quella relativa al campione nel suo complesso, cioè :

$$f_h = \frac{n_h}{N_h} = \frac{n}{N}$$

In questo caso, si può dimostrare che le formule utilizzate per la stima della media e della proporzione coincidono con quelle applicate per il campionamento casuale semplice.

Esercizio:

riprendiamo l'esercizio precedente ma anziché un CCS supponiamo di estrarre una campione stratificato (con **stratificazione proporzionale**) di dimensione $n = 25$. Gli strati sono definiti considerando la fascia oraria della chiamata di intervento. In particolare:

STRATO = 1 Include i clienti che hanno contattato l'impresa tra le 20 di sera e le 8 di mattina. Nella popolazione esso include **45** clienti.

STRATO = 2 Raggruppa i clienti che hanno contattato l'impresa tra le 8 di mattina e le 20 di sera. Nella popolazione lo strato contiene **70** clienti

Stimate media e proporzione e relativi errori standard

Campione stratificato $Y =$ ore di attesa

n.	Etichette	strato	Y	n.	Etichette	strato	Y
1	8	1	0	13	68	1	4
2	49	2	5	14	157	2	3
3	241	2	2	15	254	2	4
4	164	1	4	16	257	1	2
5	118	1	2	17	68	2	2
6	209	1	1	18	282	2	3
7	222	2	5	19	28	1	2
8	99	2	4	20	203	1	0
9	195	2	2	21	181	2	1
10	93	2	5	22	246	2	2
11	76	1	1	23	269	2	3
12	105	2	3	24	193	1	3
				25	109	2	1

Riorganizziamo le informazioni nella seguente tabella:

	N_h	W_h	n_h	y_h	\bar{y}_h	s_h^2	r_h	P_h
strato 1	45	0,39	10	19	1,90	2,10	2	0,20
strato 2	70	0,61	15	45	3,00	1,86	5	0,33

Nelle colonne 1 e 2 sono registrate le dimensioni, assoluta e relativa, degli strati nella popolazione; le colonne 3, 4, 5 e 6 contengono rispettivamente la dimensione campionaria, i totali, le media e le varianze campionarie elementari relative a ciascuno strato; la colonne 7 e 8, infine, riportano il numero e la proporzione di pronto intervento realizzati dopo oltre 3 ore di attesa.

Soluzione

$$\bar{y}_{ST} = \sum_{h=1}^H W_h \bar{y}_h = (0,39 \cdot 1,9) + (0,41 \cdot 3) = 2,57$$

$$es(\bar{y}_{ST}) = \sqrt{\left(0,39^2 \cdot 0,78 \cdot \frac{2,10}{10}\right) + \left(0,61^2 \cdot 0,78 \cdot \frac{1,86}{15}\right)} = 0,25$$

$$P_{ST} = \sum_{h=1}^H W_h P_h = (0,39 \cdot 0,20) + (0,61 \cdot 0,33) = 0,28$$

$$es(P_{ST}) = \sqrt{\left(0,39^2 \cdot 0,78 \cdot \frac{0,20 \cdot 0,80}{10 \cdot 9}\right) + \left(0,61^2 \cdot 0,78 \cdot \frac{0,33 \cdot 0,67}{15 \cdot 14}\right)} = 0,08$$

La stima per intervallo nel CCS

Quanto è precisa la nostra stima?

Costruiamo intorno alla stima puntuale un **intervallo di confidenza**, entro il quale, con un **livello di fiducia pari a $(1-\alpha)100$** sarà compreso il valore del parametro nella popolazione. α è detto **livello di significatività** o di rischio.

La stima dell'intervallo $I_{(1-\alpha)}$ richiede che sia **nota** la **distribuzione campionaria dello stimatore**.

Distribuzione della media e proporzione campionarie – Grandi campioni.

Media: la distribuzione della media campionaria può essere approssimata alla **distribuzione Normale** quando $n \geq 30$.

Proporzione: la distribuzione della proporzione campionaria può essere approssimata alla **Normale** quando la dimensione campionaria n è tale assicurare che $pn \geq 5$ e $n(1-p) \geq 5$ dove p è la proporzione campionaria.

Nel caso dunque di approssimazione alla Normale, gli estremi dell'intervallo di confidenza per la media e la proporzione saranno individuati rispettivamente da:

$$\bar{y} \pm z_{\frac{\alpha}{2}} \cdot es(T_{\bar{Y}})$$

$$p \pm z_{\frac{\alpha}{2}} \cdot es(T_P)$$

Distribuzione della media e proporzione campionarie – Piccoli campioni.

Nel caso di ***piccoli campioni*** estratti da popolazione Normali, la distribuzione dello stimatore della media campionaria sarà a sua volta Normale, a condizione che sia **nota la varianza del carattere di interesse** nella popolazione.

Nel caso, invece, in cui la varianza sia ignota e si utilizzi la varianza campionaria corretta come sua stima, lo stimatore avrà **distribuzione t di Student con $n-1$** gradi di libertà, dove n indica la dimensione del campione. In questo caso gli estremi dell'intervallo saranno calcolati sostituendo ai centili della Normale, i centili della t di Student.

Vedi Esempio 2.5 del libro di testo

Determinazione della numerosità campionaria

L'accuratezza della stima può essere migliorata aumentando la **dimensione campionaria**. Ma i costi crescono con la numerosità campionaria. Trade-off accuratezza e contenimento costi

Qual è la soglia minima della dimensione campionaria in grado di assicurare stime sufficientemente precise?

ERRORE CAMPIONARIO = MISURA DI IMPRECISIONE

$$e = z_{\frac{\alpha}{2}} \cdot es(T_n)$$

e = errore campionario,
“semiampiezza” dell’intervallo di
confidenza
 T_n è lo stimatore di un generico
parametro della popolazione.

Nel caso della stima della media

Se assumiamo che $(1-f)$ sia approssimabile a 1 (grandi popolazioni o estrazione con ripetizione), l'errore campionario è pari a:

$$e = z_{\frac{\alpha}{2}} \cdot \frac{\sigma_Y}{\sqrt{n}}$$

Da cui:

$$n = \frac{z_{\frac{\alpha}{2}}^2 \sigma_Y^2}{e^2}$$

Una volta **prefissato il livello di precisione** desiderato, è possibile **calcolare la dimensione campionaria** in grado di assicurarlo, a condizione che sia **nota la varianza di Y** nella popolazione o che sia possibile stimarla.

Tecniche di rilevazione

- Indicano il metodo di raccolta delle informazioni presso le unità selezionate
- La scelta dipende degli obiettivi della ricerca ma anche da considerazioni sui tempi e costi della rilevazione.

Tipologie di tecniche di rilevazione

- Intervista diretta
- Autocompilazione del questionario (indagini postali)
- Intervista telefonica
- Indagine web

INTERVISTA DIRETTA

- Interazione diretta tra intervistato e intervistatore
- Strumenti di supporto: libretto istruzioni, lettera preavviso
- Tecnologia CAPI (Computer Assisted Personal Interview)

Vantaggi	Svantaggi
la risposta proviene dall'individuo effettivamente selezionato	l'intervistatore può influenzare le risposte
l'intervistatore può riuscire ad instaurare un rapporto di fiducia	l'intervistatore potrebbe cambiare la formulazione delle domande
l'intervistato può chiedere spiegazioni	l'intervistato potrebbe sentirsi a disagio
	Alti costi

AUTOCOMPILAZIONE

La tecnica dell'auto-compilazione prevede che il rispondente compili da solo il questionario con l'aiuto di un libretto di istruzioni.

Vantaggi	Svantaggi
Economicità	Alta percentuale di mancati ritorni
Minori difficoltà organizzative	Questionari incompleti
L'intervistato ha minor difficoltà a rispondere su argomenti "sensibili"	Non corretta comprensione dei quesiti

INTERVISTA TELEFONICA

- In genere è effettuata con tecnologia CATI (Computer Assisted Telephone Interviewing)
- Utilizza l'elenco degli abbonati alla rete telefonica fissa come lista di campionamento

Vantaggi	Svantaggi
Minori costi rispetto all'intervista diretta	Sotto-copertura della lista di campionamento
Velocità di esecuzione ed elaborazione (se CATI)	
Migliore qualità dei dati tramite controlli di coerenza (se CATI)	

INDAGINE WEB

Vantaggi	Svantaggi
costi limitati	stessi limiti evidenziati per l'auto-compilazione
possibilità di aver accesso ad una sterminata platea di rispondenti	sotto-copertura della popolazione (accesso limitato ad internet)
Velocità di esecuzione ed elaborazione dei dati raccolti	alto tasso di non risposta
	E' difficile valutare la qualità dei dati

Il Questionario

- Il **questionario** è il supporto con cui vengono raccolte le informazioni di interesse per gli scopi conoscitivi dell'indagine.
- La fase di redazione del questionario viene concettualmente dopo la definizione di obiettivi conoscitivi, popolazione obiettivo, piano di campionamento e tecnica di somministrazione

Obiettivi di un buon questionario

- Rilevare effettivamente le informazioni di interesse (cosa voglio sapere?)
- Rilevare le informazioni sul numero maggiore di unità campionate.

Alcuni accorgimenti necessari

l'intervista deve essere percepita come **semplice, diretta, motivata, non eccessivamente impegnativa**

Per evitare "mancate risposte"

- presentazione del titolare e/o del committente dell'indagine e breve descrizione degli scopi della ricerca (lettera di preavviso o prima parte del questionario)
- fornire garanzie riservatezza
- porre le domande personali (o comunque su attributi demografici e sociali) in fondo al questionario nel caso di indagini telefoniche.
- Curare il linguaggio (**wording**)
- Formulare domande brevi, chiare e focalizzate in modo preciso sull'argomento di interesse.

Tipi di domande

La formulazione delle domande è diversa a seconda del tipo di **risposta attesa**.

Domanda aperta: non prevede modalità di risposta ma lascia il rispondente libero di esprimere il proprio pensiero, nella forma a lui più congeniale.

Come definirebbe la situazione economica della sua famiglia?

.....

Domanda chiusa: l'intervistato deve scegliere una (risposta unica) o più risposte (risposta multipla) tra un numero limitato di possibilità

Come definirebbe la condizione economica della sua famiglia?

peggiore rispetto allo scorso anno

equivalente a quella dello scorso anno

migliore di quella dello scorso anno



Domanda filtro:

- E' un tipo particolare di domanda chiusa che ha la funzione di convogliare soltanto una parte dei rispondenti verso domande successive.
- La domanda filtro è utilizzata quando le unità di analisi dell'indagine non sono omogenee, per cui determinate informazioni devono (o possono) essere rilevate solo su un sottoinsieme.

La condizione economica della sua famiglia è peggiorata rispetto allo scorso anno?

Si

No

Tale domanda, ad esempio, potrebbe convogliare chi ha risposto SI ad una sezione del questionario dedicata a rilevare le principali cause di peggioramento della condizione economica familiare.

Domande aperte e domande chiuse: vantaggi e svantaggi

DOMANDE APERTE:

V consentono di cogliere fedelmente il pensiero del rispondente e di ottenere risposte difficilmente prevedibili

S L'elaborazione ed analisi delle risposte, oltre ad essere complessa, richiede molto tempo

DOMANDE CHIUSE:

V l'elaborazione ed analisi delle risposte è più semplice e veloce

S rischio di "forzare" il rispondente verso risposte già stabilite, che potrebbero non riflettere la posizione dell'intervistato (problema della modalità "Altro").

Misurazione delle modalità di risposta

- Una delle difficoltà maggiori consiste nel predisporre i quesiti in modo che le risposte siano **traducibili in misure**.
- In alcuni casi l'unità di misura è chiara: età, peso, numero di figli
- La rilevazione di percezioni, atteggiamenti, opinioni crea maggiori problemi perché non esiste una unità di misura condivisa.



Definizione di SCALE AD HOC che forniscono **ordinamenti** più che misurazioni vere e proprie.

Le Scale

Scala nominale: assegna un codice (solitamente numerico) al fine di identificare l'unità di analisi rispetto al possesso di un determinato carattere. Sono utilizzabili nel caso di **caratteri qualitativi sconnessi**.

Es: 1=maschio, 2= femmina

Scale ordinali: assegnano l'unità a categorie diverse, tra le quali esiste un ordinamento naturale. Possono essere derivati **indici di posizione (mediana, quartili, ecc)**.

Es: ordine delle preferenze espresse da un consumatore rispetto a più marche di uno stesso prodotto.

Scala a intervallo: si assegnano valori numerici alle modalità ed è possibile quantificare la distanza tra di esse (ad es. la differenza tra la modalità 2 e 3 è equivalente a quella tra 4 e 5). Si possono calcolare **valori medi, indici di variabilità, coefficienti di correlazione lineare e applicare tecniche di analisi multivariata.**

Qual è il suo giudizio sulla qualità del biscotto in promozione?

1 (molto negat.) 2(negat.) 3 (né negat., né posit.) 4 (posit.) 5(molto posit.)

Scala a rapporti: assegnano valori numerici alle risposte in maniera tale che esista uno zero assoluto. I dati rilevati con questo tipo di scala possono essere utilizzati per calcolare **indici statistici di qualsiasi tipo.**

Esprimete il vostro giudizio sulla qualità del caffè in promozione, ponendo un segno sul seguente segmento

Non mi piace | _____ | Mi piace moltissimo

Esempi di scale a intervallo

Scala Likert: si utilizza per valutare l'atteggiamento di un intervistato nei confronti di una affermazione. Le modalità di risposta sono 5 o 7 ed esprimono accordo o disaccordo con l'affermazione in questione.

Indichi in che misura condivide l'opinione: "consumare abitualmente cibi biologici migliora lo stato di salute"

1 molto in disaccordo	2 in disaccordo	3 né d' accordo, né in disaccordo	4 d' accordo	5 molto d' accordo
-----------------------------	--------------------	--	-----------------	--------------------------

Non sempre conviene prevedere la modalità neutra in quanto essa può accentrare su di sé buona parte delle risposte, per indifferenza, per pigrizia, per desiderio di non prendere una posizione

Scala del differenziale semantico: si definiscono aggettivi o proposizioni di significato opposto (utile/inutile, vero/falso, corretto/scorretto).

L'intervistato indica la posizione che meglio riflette il proprio modo di sentire, lungo una scala che prevede 5 o 7 modalità. A tali modalità sono assegnati i punteggi -3,-2,-1,0,1,2,3 (nel caso di 7 modalità).

Indicate la vostra opinione sulla utilità del data warehouse della vostra azienda

INUTILE	molto	Abba- stanza	poco	né utile né inutile	poco	Abba- stanza	molto	UTILE
----------------	--------------	-------------------------	-------------	------------------------------------	-------------	-------------------------	--------------	--------------

Valutazione dei risultati

Per valutare se siamo in presenza o meno di una indagine seria e credibile dobbiamo verificare:

- la possibilità di accedere a tutte le informazioni relative agli strumenti di indagine utilizzati (definizioni, classificazioni, questionario, piano di campionamento, questionario ecc).
- l'accuratezza o precisione delle stime. Essa è misurata dall'errore statistico o **errore totale**

Errore totale = Errore campionario + Errore non campionario

Gli **errori NON campionari** sono generati da imperfezioni negli aspetti organizzativi della rilevazione e nella fase di registrazione dei risultati. Interessano sia le indagini campionarie sia quelle censuarie.

Errore statistico o
errore totale

Tendono ad aumentare al crescere della dimensione campionaria, per questo una indagine campionaria può essere considerata più attendibile di una completa

Errore campionario

Errori non campionari:

- errori di copertura
- errori da mancate risposte
- errori di misurazione