

Facoltà di Giurisprudenza Università di Macerata

Corso di

Programmazione e valutazione dei servizi sociali

Il campionamento

docente: Cristina Davino

a.a.: 2009-2010

Le indagini statistiche

- Oggetto di ogni indagine statistica è la conoscenza di una popolazione.
 - ↳ L'insieme, l'aggregato di unità elementari in cui il fenomeno allo studio si manifesta.
- Una popolazione può essere:
 - Un insieme di soggetti → *i clienti di un'azienda*
 - Un insieme di unità amministrative → *i Comuni*
 - Un insieme di stabilimenti → *Le imprese manifatturiere*
 - Una superficie → *Il territorio di una regione*
 - Un insieme di eventi → *I fatti criminosi in un certo periodo*
 - ...

Le indagini statistiche

- Ai fini di una corretta comprensione del fenomeno analizzato, un universo statistico deve essere definito:

nei contenuti

nello spazio

nel tempo

Es.: Popolazione residente in Italia alla mezzanotte tra il 27 e il 28 ottobre 2001.

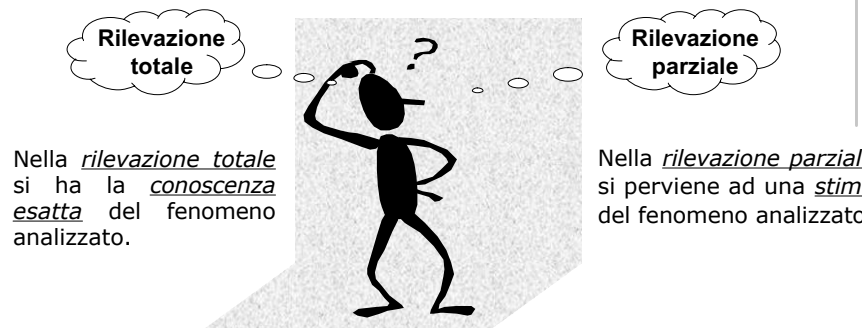
- Data una popolazione di N unità statistiche, un campione è un insieme di n unità selezionate tra le N della popolazione allo scopo di rappresentarla rispetto ai caratteri, o variabili, oggetto di studio.

Le indagini campionarie

Quindi...

- (a) Una parte delle unità della popolazione di riferimento viene selezionata per far parte del campione, seguendo un insieme interdipendente di regole che vengono denominate disegno di campionamento;
- (b) Le unità selezionate si sottopongono ad osservazione per:
 - b1) Ottenere informazioni su certe caratteristiche (statistiche) della popolazione;
 - b1) Analizzare le relazioni, semplici e complesse, che aiutino ad interpretare atteggiamenti o comportamenti dell'insieme oggetto di studio.

Il dilemma



D'altra parte, bisogna anche considerare:

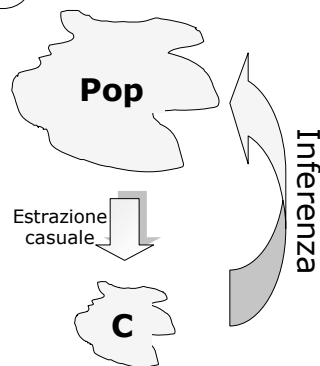
- I **tempi** della rilevazione;
- I **costi** della rilevazione;
- La **ricchezza di dettagli** della rilevazione;
- Gli **errori** associati alla rilevazione;

Riassumendo

- Le informazioni relative alla popolazione, cioè alle variabili che la caratterizzano, possono derivare da una:
 - (a) Rilevazione censuaria o totale
 - ➔ Si ha la conoscenza esatta del fenomeno analizzato.
 - Si preferisce:
 - ... per analisi a livello di micro-aree;
 - ... quando le unità da analizzare sono rare;
 - ... quando si vuole portare l'analisi ad un elevato livello di dettaglio.
 - (b) Rilevazione campionaria
 - ➔ Si perviene ad una stima del fenomeno.
 - Si preferisce:
 - ... quando è impossibile effettuare una rilevazione totale;
 - ... quando la rilevazione del carattere comporta la distruzione delle unità osservate;
 - ... quando si vogliono ridurre i costi e/o i tempi di un'indagine.

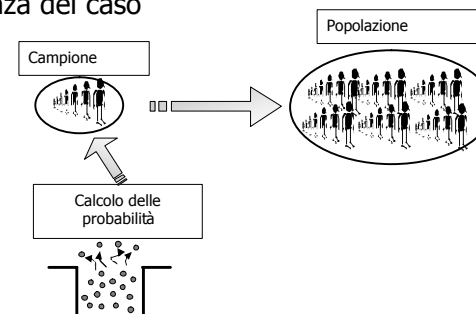
Il campionamento

Si definisce campionamento un procedimento attraverso il quale da un insieme di unità costituenti l'oggetto dello studio, si estrae un numero ridotto di casi scelti con criteri tali da consentire la generalizzazione all'intera popolazione dei risultati ottenuti.



Il campionamento e l'inferenza

- ✦ Il campione deve essere rappresentativo della popolazione
 - ➔ campionamento casuale
- ✦ Il calcolo delle probabilità esamina i risultati che si ottengono sotto l'influenza del caso



Il campionamento e l'inferenza



- Inferenza: utilizza statistiche del campione per effettuare la stima dei corrispondenti veri valori della popolazione
- In pratica, viene selezionato a caso dalla popolazione un campione unico di ampiezza predeterminata
- Campioni diversi forniranno stime diverse del parametro della popolazione

$$\begin{array}{ccccc} V & = & v & \pm & \varepsilon \\ \text{Parametro della} & & \text{Stima del} & & \text{Errore di} \\ \text{popolazione} & & \text{campione} & & \text{campionamento} \\ \text{(incognito)} & & & & \end{array}$$

L'errore di campionamento

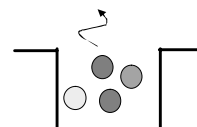
Si consideri la popolazione costituita da N=4 quattro ipermercati A, B, C, D. Le vendite effettuate da ciascuno di essi nel periodo 1/1/08-31/12/08 sono riportate nella seguente tabella:

Ipermercato	A	B	C	D
Vendite (in miliardi di euro)	4	1	3	2

$$Media = \frac{1}{4}(4 + 1 + 3 + 2) = 2,5$$

L'errore di campionamento

Numero del campione	Primo Elemento	Secondo Elemento	Media Campionaria
1	4	4	4,0
2	4	1	2,5
3	4	3	3,5
4	4	2	3,0
5	1	4	2,5
6	1	1	1,0
7	1	3	2,0
8	1	2	1,5
9	3	4	3,5
10	3	1	2,0
11	3	3	3,0
12	3	2	2,5
13	2	4	3,0
14	2	1	1,5
15	2	3	2,5
16	2	2	2,0



$$Media = \frac{40}{16} = 2,5$$

Le distribuzioni campionarie

- Inferenza: utilizza statistiche del campione per effettuare la stima dei corrispondenti veri valori della popolazione
- In pratica, viene selezionato a caso dalla popolazione un campione unico di ampiezza predeterminata
- Bisognerebbe prendere in esame ogni campione che avrebbe potuto manifestarsi

Distribuzioni campionarie

Parametri: valori caratteristici della popolazione

Statistiche: funzioni delle osservazioni campionarie

Statistica calcolata: numero ottenuto applicando la statistica al campione osservato

Distribuzione campionaria: valori che la statistica assume al variare del campione nell'universo campionario

Parametri e statistiche

Popolazione  Parametri → **Valori fissi**,
spesso non noti

Campione  Statistiche
o
Stimatori → **Variabili casuali**,
le cui determinazioni
dipendono dalle particolari
osservazioni scelte

Le distribuzioni campionarie

Valori che la statistica assume al variare
del campione nell'universo campionario

Distribuzioni Campionarie

Le conclusioni inferenziali, basate sull'unico campione osservato, devono essere giudicate sulla base della distribuzione di probabilità dei possibili campioni che potevano essere generati e dei quali quello osservato costituisce una realizzazione particolare.

V.C. Media Campionaria

- V.C. media campionaria: medie aritmetiche calcolate su tutti i campioni appartenenti allo spazio campionario
- Le medie variano al variare del campione estratto e, poiché i campioni sono estratti casualmente, i valori che può assumere la media campionaria sono realizzazioni di una v.c
- La distribuzione della v.c media campionaria dipende dalla distribuzione della popolazione X
- Quando la dimensione del campione è sufficientemente grande, la distribuzione della media campionaria può essere approssimata alla distribuzione normale qualunque sia la distribuzione della popolazione (*Teorema del Limite Centrale*)

Esempio sulla V.C. Media Campionaria

Si consideri la popolazione costituita da N=4 quattro ipermercati A, B, C, D. Le vendite effettuate da ciascuno di essi nel periodo 1/1/03-31/12/03 sono riportate nella seguente tabella:

Ipermercato	A	B	C	D
Vendite (in miliardi di euro)	4	1	3	2

1. Si calcolino la media μ e lo scarto quadratico medio σ della popolazione

$$\mu = \frac{1}{4}(4 + 1 + 3 + 2) = 2,5$$

$$\sigma = \sqrt{\frac{1}{N} \sum x_i^2 - \mu^2} = \sqrt{\frac{30}{4} - 2,5^2} = \sqrt{1,25} = 1,12$$

Esempio sulla V.C. Media Campionaria

2. Effettuando un campionamento **con ripetizione** si calcolino il valore atteso e lo scarto quadratico medio della v.c. media campionaria

- Universo dei campioni $n=2$ estratti con ripetizione (4^2) e relative medie campionarie

$$E(\bar{X}) = \frac{40}{16} = 2,5 = \mu$$

$$sqm(\bar{X}) = 0,79 = \frac{1,12}{\sqrt{2}}$$

Numero del campione	Primo Elemento	Secondo Elemento	Media Campionaria
1	4	4	4,0
2	4	1	2,5
3	4	3	3,5
4	4	2	3,0
5	1	4	2,5
6	1	1	1,0
7	1	3	2,0
8	1	2	1,5
9	3	4	3,5
10	3	1	2,0
11	3	3	3,0
12	3	2	2,5
13	2	4	3,0
14	2	1	1,5
15	2	3	2,5
16	2	2	2,0

Esempio sulla V.C. Media Campionaria

2. Effettuando un campionamento **senza ripetizione** si calcolino il valore atteso e lo scarto quadratico medio della v.c. media campionaria

- Universo dei campioni $n=2$ estratti senza ripetizione ($\frac{4!}{(4-2)!} = 12$) e relative medie campionarie

$$E(\bar{X}) = \frac{30}{12} = 2,5$$

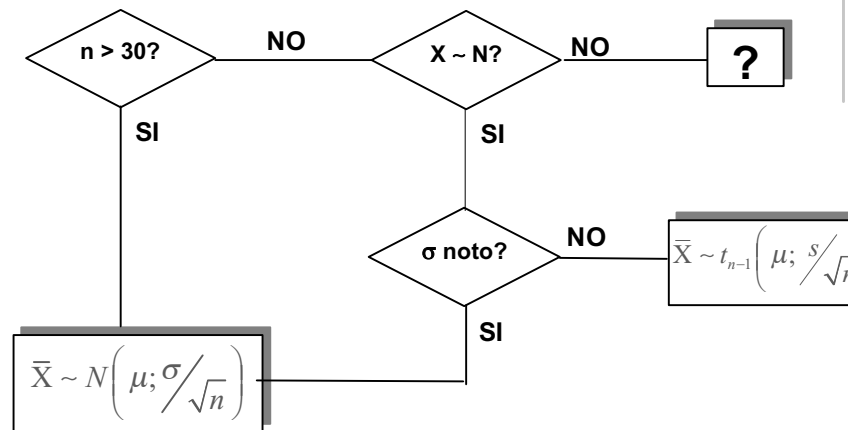
$$Var(\bar{X}) = 0,64 = \frac{1,12}{\sqrt{2}} \sqrt{\frac{2}{3}}$$

Numero del campione	Primo Elemento	Secondo Elemento	Media Campionaria
1	4	1	2,5
2	4	3	3,5
3	4	2	3,0
4	1	4	2,5
5	1	3	2,0
6	1	2	1,5
7	3	4	3,5
8	3	1	2,0
9	3	2	2,5
10	2	4	3,0
11	2	1	1,5
12	2	3	2,5

Esempio sulla V.C. Media Campionaria

	Campionamento con reintroduzione	Campionamento senza reintroduzione
Popolazione non finita	$E(\bar{X}) = \mu$ $\sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$	$E(\bar{X}) = \mu$ $\sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$
Popolazione finita	$E(\bar{X}) = \mu$ $\sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$	$E(\bar{X}) = \mu$ $\sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

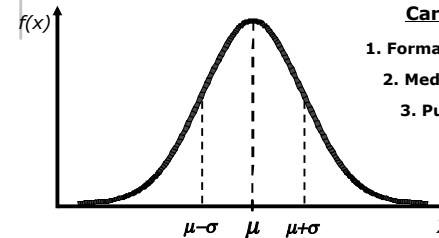
Distribuzione della V.C. Media Campionaria



La distribuzione normale

Una variabile casuale X segue una distribuzione Normale, con media μ e varianza σ^2 , se la sua funzione di densità di probabilità è data da:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Caratteristiche della distribuzione Normale

1. Forma campanulare e simmetrica
2. Media, mediana e moda coincidenti
3. Punto di flesso a distanza σ dalla media

La tavola della distribuzione normale standardizzata

Area sotto la curva di densità normale standardizzata calcolata tra 0 e Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1809	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3079	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3889	0.3907	0.3926	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4266	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4899	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4919	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986

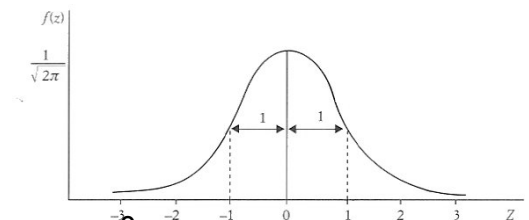
La distribuzione normale standardizzata

$$Z = \frac{X - \mu}{\sigma}$$

$$X \sim N(0, 1)$$

Proprietà:

- $\mu = 0$
- $\sigma = 1$
- il massimo di $f(x)$ si ha per $x=0$
- punti di flesso: $x=\pm 1$
- i valori della curva normale standardizzata sono tabulati



Stima per Intervalli

- ◆ Si cerca un intervallo che ha una particolare confidenza o probabilità di includere il parametro della popolazione

$$P(t_1 < \theta < t_2) = \underbrace{1 - \alpha}_{\text{Livello di confidenza}}$$

La stima per intervalli

La media della popolazione

Si supponga di aver estratto 10 campioni di 36 unità da una popolazione normale con media $\mu=10$ e varianza pari a 36. Per ognuno di questi campioni si è calcolata la media campionaria e l'intervallo di confidenza al 95%.

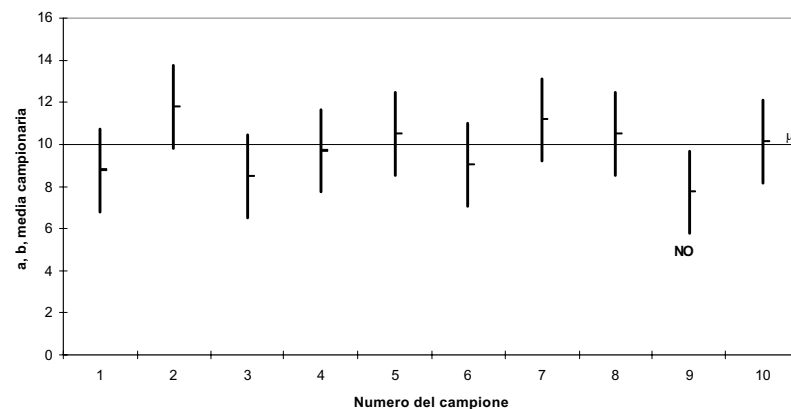
Estremi dell'intervallo:

$$a = \bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \quad b = \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}$$

Campione	$\bar{X}$	Estremo inferiore a	Estremo superiore b
1	8.75	6.79	10.71
2	11.75	9.79	13.71
3	8.45	6.49	10.41
4	9.70	7.74	11.66
5	10.50	8.54	12.46
6	9.00	7.04	10.96
7	11.15	9.19	13.11
8	10.50	8.54	12.46
9	7.75	5.79	9.71
10	10.10	8.14	12.06

La stima per intervalli

La media della popolazione



Stima per Intervalli

La media campionaria

- Quando il parametro μ della popolazione è incognito, il miglior modo per stimarlo è utilizzare la media campionaria.

- Quando la numerosità campionaria n è sufficientemente elevata si ha:

$$\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$$

- E' quindi possibile dire che, con probabilità $1-\alpha$, l'intervallo:

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

contiene il parametro incognito μ .



La numerosità campionaria

- Quando il parametro μ della popolazione è incognito, il miglior modo per stimarlo è utilizzare la media campionaria.

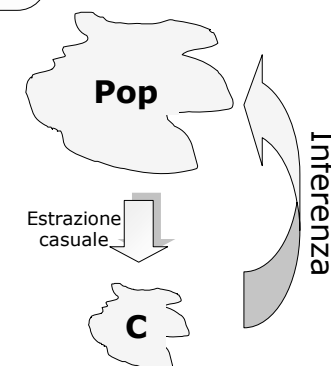
$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

V = v $\pm$ ϵ
 Parametro della popolazione (incognito) Stima del campione **Errore di campionamento**



Il campionamento

*Si definisce **campionamento** un procedimento attraverso il quale da un insieme di unità costituenti l'oggetto dello studio, si estrae un numero ridotto di casi scelti con criteri tali da consentire la generalizzazione all'intera popolazione dei risultati ottenuti.*



Le diverse tecniche di campionamento

- ◆ **Campionamento probabilistico**
 - Camp. casuale semplice
 - Camp. casuale stratificato
 - Camp. a due stadi
 - Camp. sistematico
- ◆ **Campionamento non probabilistico**
 - Camp. per quote
 - Disegno fattoriale
 - Camp. a scelta ragionata
 - Camp. bilanciato
 - Camp a valanga
 - Camp. telefonico

Il campionamento probabilistico

- ⇒ Le unità sono scelte in modo casuale (ma non "a casaccio").
- ⇒ La casualità interviene nella selezione delle unità e si ottiene attribuendo ad ogni unità della popolazione una probabilità nota e diversa da zero di essere selezionata.
- ⇒ Quando la probabilità di estrazione, oltre ad essere nota, è posta uguale per tutte le unità, si parla di **campionamento casuale semplice**.
- ⇒ In particolare, la casualità interviene nella selezione delle unità e si ottiene:
 - attribuendo ad ogni unità della popolazione una probabilità nota e diversa da zero di essere selezionata;
 - utilizzando in modo appropriato le tecniche per la selezione.

Il disegno di campionamento

Il disegno di campionamento è l'insieme delle decisioni prese per formare il campione.

Le fasi:

- ❶ definizione della struttura del campione
- ❷ selezione delle unità campionarie
- ❸ probabilità di inclusione delle singole unità
- ❹ determinazione della numerosità del campione



Il disegno di campionamento

❶ Definizione della struttura del campione

- ⇒ Richiede la definizione della lista delle unità che compongono l'universo che si intende osservare
- ⇒ Ad ogni unità deve essere attribuito un identificatore



PROBLEMI : Costi spesso eccessivi



SOLUZIONI : Campionamento su più livelli
Campionamento a grappoli

❷ Selezione delle unità campionarie

- ⇒ Selezione casuale con reinserimento
- ⇒ Selezione casuale senza reinserimento
- ⇒ Selezione casuale sistematica (passo: N/n)

Tavole dei numeri casuali

Le tecniche di selezione casuale

• Selezione casuale con reintroduzione (o bernoulliano)

Ogni elemento che viene estratto viene reintrodotta nella popolazione in modo tale che ad ogni estrazione successiva non venga alterata la composizione della popolazione ed ogni elemento estratto ha sempre la stessa probabilità di venire scelto.

- Probabilità di estrazione di ciascun elemento: $\frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N}$
- Universo campionario: N^n



- La numerosità della popolazione è, di fatto, considerata infinita;
- Una unità può essere estratta più volte;
- La probabilità di estrazione rimane costante.

Le tecniche di selezione casuale

• Selezione casuale senza reintroduzione

Ogni elemento, una volta estratto, non viene reimesso nella popolazione per cui, dopo ogni estrazione, la probabilità che gli elementi restanti entrino a far parte del campione viene modificata.

- Probabilità di estrazione di ciascun elemento: $\frac{1}{N}, \frac{1}{N-1}, \dots, \frac{1}{N-n+1}$
- Universo campionario: $N \cdot (N-1) \cdot \dots \cdot (N-n+1) = \frac{N!}{(N-n)!}$



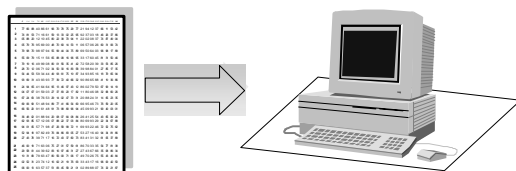
- La probabilità di estrazione varia ad ogni passo dell'estrazione

Il disegno di campionamento

② Selezione delle unità campionarie

Tavola dei numeri casuali

Generazione automatica di
 n numeri casuali



③ Probabilità di selezione delle unità campionarie

- costanti
- variabili

(generalmente in funzione della dimensione dell'unità)

La tavola dei numeri casuali

	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-16	17-18	19-20	21-22	23-24	25-26	27-28	29-30	31-32	33-34	35-36	37-38	39-40
1	77	66	88	40	86	61	96	70	78	75	29	77	21	94	12	37	66	11	53	42
2	74	81	53	71	16	61	59	13	33	02	25	95	92	37	03	18	46	26	37	86
3	05	88	20	12	10	45	80	22	38	70	94	11	22	02	08	37	74	87	49	04
4	05	79	76	95	69	00	48	70	60	14	53	11	06	57	06	26	60	31	06	74
5	79	98	70	98	97	94	55	99	44	04	75	89	69	50	64	03	96	98	17	89
6	55	09	79	15	11	56	65	88	08	16	96	95	33	17	60	45	81	31	50	46
7	79	19	16	49	99	08	80	01	56	35	41	42	72	58	20	39	33	53	85	26
8	28	70	12	06	71	02	34	50	30	16	83	58	39	98	84	01	27	85	17	35
9	54	44	53	59	34	44	49	93	61	75	19	87	34	93	85	16	18	79	65	94
10	93	69	31	43	93	93	77	39	72	40	66	32	90	86	65	88	41	19	36	86
11	24	94	65	41	64	64	95	13	46	97	43	12	86	02	79	50	67	90	14	19
12	04	07	67	01	59	03	27	37	83	20	17	82	11	80	46	08	32	68	60	26
13	67	24	63	38	76	53	29	14	02	47	70	31	20	88	24	31	14	65	23	35
14	69	06	90	51	48	94	89	77	41	66	54	60	66	95	46	73	76	59	20	05
15	66	56	20	91	61	48	91	73	98	80	96	94	45	09	93	21	90	40	03	01

④ La numerosità campionaria

Popolazione
 N

È l'insieme finito o infinito di unità, definito nei contenuti, nello spazio e nel tempo, oggetto dell'indagine statistica

Campione
 n

È costituito da un certo numero di unità, estratte con qualche procedimento da una popolazione, al fine di rappresentarla quanto ai caratteri oggetto di studio

$$V = v \pm \varepsilon$$

Parametro della popolazione (incognito) Stima del campione **Errore di campionamento**

"La *numerosità ottima* di un campione è quella che consente di ottenere gli obiettivi dell'indagine al *minimo costo* e sarà il numero minimo in base al quale le stime raggiungeranno il livello di attendibilità atteso."
(L. Fabbris: *L'indagine campionaria - NIS*)

Determinazione della numerosità ottimale

- Fissare la quantità di errore che si è disposti ad accettare nell'uso del campione per stimare il parametro della popolazione (*errore di campionamento ammesso*, ε)
- Stimare lo scarto quadratico medio se non sono disponibili dati del passato
- Fissare il livello di confidenza desiderato

► Intervallo della stima per la media:

- a. Con n grande e schema di campionamento con reintroduzione:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \frac{z_{\alpha/2}^2 \cdot \sigma^2}{\varepsilon^2}$$

- b. Con n grande e schema di campionamento senza reintroduzione:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \Rightarrow \quad n = \frac{\left(\frac{z_{\alpha/2} \cdot \sigma}{\varepsilon} \right)^2}{1 + \frac{1}{N} \cdot \left(\frac{z_{\alpha/2} \cdot \sigma}{\varepsilon} \right)^2}$$

Determinazione della numerosità ottimale

Intervallo della stima per la media:

Metodo empirico

- Si determina la numerosità n_0 seguendo lo schema A;
- Se il valore di n_0 così calcolato risulta più piccolo del 5% di N , si utilizza il valore di n_0 ;
- Se n_0 risulta superiore al 5% di N , si introduce un fattore di correzione che calcola il valore corretto con la formula:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Esempio

Un'azienda che imbottiglia una bibita gassata vuole indagare sulla forza della pressione interna della bibita presente in una lattina. Supponendo che la forza della pressione sia distribuita normalmente ed abbia uno s.q.m. 28psi, quante lattine bisogna controllare se si vuole stimare la pressione media delle lattine prodotte dall'azienda entro l'intervallo di ± 3 psi? Si consideri un livello di confidenza del 95%.

Livello di fiducia=95%

$$\sigma=28 \quad n=50 \quad \bar{x} = 235$$

$$\bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

Idf=90%	⇒	$z=1,64$
Idf=95%	⇒	$z=1,96$
Idf=99%	⇒	$z=2,33$

$$n = \frac{z^2 \sigma^2}{\varepsilon^2} = \frac{1.96^2 \cdot 28^2}{3^2} = 334.6 \approx 335$$

Determinazione della numerosità ottimale

Intervallo della stima per la proporzione:

- a. Con n grande e schema di campionamento con reintroduzione:
- $$p \pm z_{\alpha/2} \cdot \sqrt{\frac{\pi \times (1-\pi)}{n}} \Rightarrow n = \frac{z_{\alpha/2}^2 \times \pi(1-\pi)}{\varepsilon^2}$$

- b. Con n grande e schema di campionamento senza reintroduzione:

$$p \pm z_{\alpha/2} \cdot \sqrt{\frac{\pi \times (1-\pi)}{n}} \cdot \sqrt{\frac{N-n}{N-1}} \Rightarrow n = \frac{z_{\alpha/2}^2 \cdot \pi(1-\pi)}{\varepsilon^2} \cdot \frac{N}{1 + \frac{1}{N} \cdot \frac{z_{\alpha/2}^2 \times \pi(1-\pi)}{\varepsilon^2}}$$

Metodo empirico

- Nel caso di massima variabilità ($\pi=0,5$), si può porre $z=2$.

Si ha allora:
$$n = \frac{z_{\alpha/2}^2 \times \pi(1-\pi)}{\varepsilon^2} = \frac{2^2 \times \frac{1}{2} \cdot \frac{1}{2}}{\varepsilon^2} = \frac{1}{\varepsilon^2}$$

Determinazione della numerosità ottimale

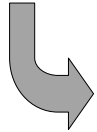
Stima per la proporzione:

5%		2%		1%	
N	n	N	n	N	n
100	80	100	96	100	99
300	170	300	270	300	296
500	220	500	415	500	475
1000	285	1000	715	1000	910
5000	370	5000	1660	5000	3330
> 8000	400 (n_0)	10000	2000	10000	5000
		>50000	2500 (n_0)	20000	6350
				>200000	10000 (n_0)

(livello di confidenza = 95%)

Determinazione della numerosità ottimale

- Stima dei parametri di una sola variabile
- Stima dei parametri di una pluralità di variabili



- Determinazione della numerosità campionaria per ciascuna variabile
- Assumere come ampiezza campionaria l' n più elevato

- Obiettivo dell'analisi

L'errore nella ricerca sociale

- ◆ Errore di selezione
 - 1. Errore di copertura
 - 2. Errore di non-risposta
 - 3. Errore di campionamento
- ◆ Errore di osservazione
- ◆ Errore di trattamento dati

1. Errore di copertura

- Lista della popolazione
- Aggiornamento
- Duplicazioni
- Incompletezza

Soluzioni



- Ridefinire la popolazione
- Trascurare gli esclusi
- Integrare il campione

Errore di non-risposta

"Il concetto di estrazione casuale è in teoria semplicissimo [...]; questa semplicità si rivela però illusoria [...] gli esseri umani differiscono dalle palline dell'urna per due aspetti essenziali: non sono a portata di mano del ricercatore [...] e sono pienamente liberi di non rispondere" (Marradi, 1989)

Le cause dell'errore di non-risposta:

- **Mancato contatto con i soggetti estratti**
 - Difficoltà a raggiungere i soggetti
 - Irreperibilità dei soggetti campionati
- **Rifiuti a rispondere**
 - Diffidenza nei confronti dell'estraneo
 - Insicurezza nei confronti di una prova
 - Rifiuto di carattere ideologico

Come affrontare l'errore di non-risposta:

- **Ripetuti ritorni sulle persone non raggiunte dall'intervista**
- **Tecniche di ponderazione**

Errore di campionamento

L'errore di campionamento è direttamente proporzionale al livello di fiducia che si vuole avere nella stima ed alla variabilità del fenomeno studiato ed inversamente proporzionale all'ampiezza del campione

A. Stima di una media μ

$$\varepsilon = z \frac{s}{\sqrt{n}} \sqrt{1-f}$$

dove

- z = coefficiente dipendente dal livello di fiducia della stima
- s = deviazione standard campionaria
- n = ampiezza del campione
- $1-f$ = fattore di correzione per popolazioni finite ($f=n/N$)

B. Stima di una proporzione π

$$\varepsilon = z \sqrt{\frac{\pi(1-\pi)}{n-1}} \sqrt{1-f}$$

dove

- p = proporzione campionaria
- $q = 1-p$

Errore di campionamento

Esempio

Per stimare il reddito medio di una popolazione di 10.000 soggetti si costruiscono due campioni rispettivamente di 1.000 e 100 casi. Dai dati di questi campioni si ottiene, per la variabile reddito mensile, la media aritmetica e la deviazione standard:

n	media	s
1.000	1.253.000	311.000
100	1.250.000	308.000

Ad un livello di confidenza del 95%, l'errore di campionamento nei due casi risulta:

$$\text{Campione di 1.000 casi } e = 1,96 \frac{311.000}{\sqrt{1.000}} \sqrt{1-0,10} \cong 18.700$$

$$\text{Campione di 100 casi } e = 1,96 \frac{308.000}{\sqrt{100}} \cong 61.600$$

Errore di campionamento

- ◆ La formula per il calcolo della numerosità campionaria si riferisce ad analisi monovariate
- ◆ Raramente la stima di singole variabili esaurisce l'interesse del ricercatore sociale
- ◆ Il ricercatore sociale è soprattutto interessato alle relazioni tra le variabili
- ◆ La dimensione del campione dipende:
 - Dalla distribuzione delle variabili studiate
 - Dal tipo di analisi che si intende effettuare

Errore di campionamento

Analisi monovariata

Praticanti	25,7 ± 4,2	istruz.superiore	63,1 ± 4,6
Non praticanti	74,3 ± 4,2	Istruz.inferiore	36,9 ± 4,6
n	420		420

Analisi bivariata

	Istr.sup.	Istr. Inf.
Praticanti	22,6 ± 5,0	30,9 ± 7,3
Non praticanti	77,4 ± 5,0	69,1 ± 7,3
n	265	155

Analisi trivariata

	Giovani		Adulti		Anziani	
	Istr.sup.	Istr. Inf.	Istr.sup.	Istr. Inf.	Istr.sup.	Istr. Inf.
Praticanti	19,4	27,8	17,0	28,3	24,2	43,9
Non praticanti	80,6	72,2	83,0	71,7	75,9	56,1
n	72	36	94	53	99	66

Errore	±9,2	±14,8	±7,6	±12,2	±8,5	±12,1
---------------	------	-------	------	-------	------	-------

Una prima riflessione

Campione casuale ➡ E' un campione estratto da una popolazione in cui tutte le unità hanno probabilità non nulla di essere estratte.

Un campione è **rappresentativo**... ...quando è estratto in modo casuale (e **non** quando è grande!).

Un campione **grande**... ...è associato ad un minore errore delle stime.

Quindi... ...la cosa migliore è avere un campione grande scelto in modo casuale ;



...è molto meglio avere un campione piccolo estratto in modo casuale che un campione grande estratto "a casaccio".

Il campionamento casuale semplice

“Il campionamento casuale semplice è raramente applicato nelle indagini statistiche, sia perché la selezione è completamente affidata al caso e non considera le informazioni note a priori sulla popolazione, sia perché nelle indagini su vasta scala comporta un piano di rilevazione costoso e di difficile realizzazione dal punto di vista organizzativo, necessitando inoltre della lista completa della popolazione che spesso non è disponibile” (Corbetta, 1999) .

Altri campioni probabilistici

Campionamento sistematico

Le unità campionarie non vengono estratte mediante sorteggio ma selezionandone sistematicamente una ogni dato intervallo (ad es. $k=N/n$).

Il campionamento sistematico consente di ottenere campioni casuali anche nella situazione in cui manchi la lista della popolazione e N sia sconosciuto (per es. un cliente ogni *tot* che escono dal negozio)

- ✓ Deve essere rispettato il requisito che tutte le unità abbiano la stessa probabilità di essere incluse
- ✓ Deve essere evitata ogni forma di scelta diversa da quella predeterminata dall'intervallo di campionamento

Altri campioni probabilistici

Campionamento stratificato (proporzionale o non proporzionale)

- (a) Suddividere la popolazione in sottopopolazioni (*strati*) il più possibile omogenee rispetto alla variabile da stimare, utilizzando una variabile ad essa correlata;
- (b) Estrarre un campione casuale semplice da ogni strato
- (c) Unire i campioni dei singoli strati per ottenere il campione globale.

A parità di ampiezza del campione, assicura un minore errore di campionamento rispetto al campionamento casuale semplice

Es.: Stima del Reddito Variabile correlata: Professione

1. Si estrae un campione da ciascuno strato mediante un processo di campionamento casuale semplice;
 2. Si calcolano le medie dei vari strati;
 3. Si stima la media attraverso la media ponderata delle medie campionarie, con pesi dati dalle numerosità relative dei vari strati
- Operaio
 - Impiegato
 - Dirigente
 - Libero prof.

Quando si stratifica

La stratificazione si usa quando si vuole...

- evidenziare insiemi di unità significative per la ricerca;
- separare sottopopolazioni con caratteristiche speciali;
- utilizzare informazioni note, mantenendo la casualità dell'estrazione;
- individuare sottopopolazioni omogenee rispetto alla variabile da studiare e ottenere stime più efficienti (maggiore precisione a parità di ampiezza) di quelle ottenibili con un campione casuale semplice.

La stratificazione può essere “forzata” ...

- Quando le sottopopolazioni si trovano su liste distinte;

↳ Es.: Campione estratto dalle liste elettorali, con schedine di diverso colore tra maschi e femmine.

I diversi tipi di stratificazione

• Il campione stratificato proporzionale

↳ Riproduce la stessa composizione degli strati nella popolazione

Es.: Popolazione occupati

- Operaio 35%
- Impiegato 45%
- Dirigente 15%
- Libero prof. 5%

$n=3000$

La numerosità dei singoli strati si ottiene moltiplicando n per la frequenza relativa (il peso) del singolo strato:

- Operaio: $3000 \times 0,35 = 1050$
- Impiegato: $3000 \times 0,45 = 1350$
- Dirigente: $3000 \times 0,15 = 450$
- Libero prof.: $3000 \times 0,05 = 150$

I diversi tipi di stratificazione

• Il campione stratificato non proporzionale

↳ Si usa quando si decide di sovrarappresentare alcuni strati (e quindi di sottorappresentarne altri).

Tipicamente, gli strati sovrarappresentati sono quelli meno numerosi.

Es.: Popolazione occupati

- Operaio: ~~1050~~ → 1000
- Impiegato: ~~1350~~ → 1200
- Dirigente: ~~450~~ → 500
- Libero prof.: ~~150~~ → 300

Il campione, quindi, non riproduce la composizione della popolazione, e nelle analisi andrà dunque effettuata una operazione di riponderazione.

Le variabili di stratificazione

Regola n° 1

Non esistono criteri assoluti o oggettivi per la scelta delle variabili di stratificazione ma solo indicazioni di massima.

Suggerimenti

- Le variabili scelte per la stratificazione devono essere correlate con la variabile, o le variabili, osservate e tra loro indipendenti;
- Una buona variabile di stratificazione è, normalmente, la suddivisione territoriale;
- Un'altra è la dimensione dell'unità.
- Nelle indagini multiscopo, la scelta delle variabili di stratificazione non è più finalizzata alla massima efficienza ma ad una migliore suddivisione della popolazione sulla base delle conoscenze che si hanno sul fenomeno;

Il numero di strati

Regola n° 2

Non esistono criteri assoluti o oggettivi per la scelta del numero di strati ma solo indicazioni di massima.

Suggerimenti

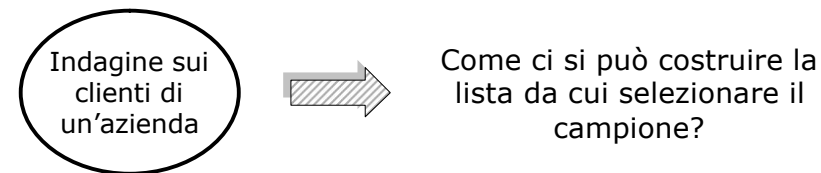
- L'efficienza delle stime aumenta con il numero di strati;
- Tuttavia, in linea di tendenza, dopo un certo numero di suddivisioni della popolazione il beneficio in termini di efficienza è modesto;
- Inoltre, all'aumentare del numero di strati crescono i costi della stratificazione e della selezione del campione;
- Un numero elevato di strati è auspicabile quando il campionamento è sul base territoriale, poiché si controlla la dispersione delle unità e si rende più agevole l'organizzazione e l'esecuzione del lavoro sul campo.

Altri campioni probabilistici

Campionamento a stadi

- ✓ Non comporta un aumento di efficienza rispetto al CCS ma una semplificazione della procedura di estrazione ed una diminuzione dei costi di rilevazione.
- ✓ Rappresenta una scelta forzata quando manca la lista completa delle unità della popolazione.
- ✓ Nel caso più semplice (*due stadi*) le unità vengono divise in unità primarie e unità secondarie, e solo per queste ultime sarà necessario disporre della liste.

Campionamento su più stadi

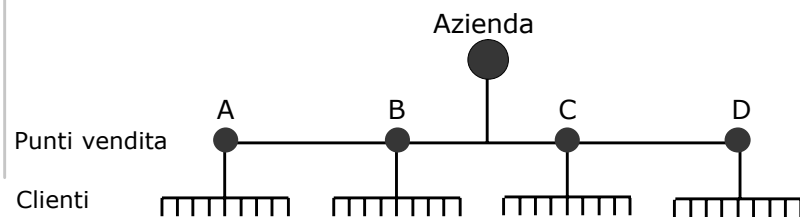


- Si considera cliente chi, in un giorno determinato, effettua un acquisto presso un punto vendita;
- I punti vendita fungono, quindi, da contenitori dei clienti che vi si trovano al momento della rilevazione;
- Le popolazioni che si considerano sono, di fatto, due:

punti vendita
clienti

 i primi ad un livello gerarchicamente superiore ai secondi;

Campionamento su più stadi



Un campionamento su *più stadi*, o su *più livelli*, prevede:

1. Una selezione dei punti vendita; Unità di primo stadio
2. L'estrazione di un campione da ciascuno dei punti vendita selezionati Unità di secondo stadio

Campionamento su più stadi

- Sono popolazioni gerarchiche quelle per le quali la popolazione finale di unità è contenuta in un insieme di unità di livello superiore;
- Per selezionare un campione è necessaria la lista delle unità;
- Ad ogni stadio sono necessarie le sole liste delle sub-popolazioni contenute nelle unità selezionate a livello superiore;
- Il campionamento a più stadi è quindi tipico delle situazioni in cui le liste della popolazione da sottoporre a indagine *non sono disponibili* o sono *costose* da reperire;

passi

- a. Determinare il numero di stadi su cui effettuare la selezione;
 - Accessibilità delle liste;
 - Costi;
 - Reperibilità delle informazioni.
- b. Individuare le caratteristiche per stratificare le unità di primo stadio;
 - Dimensioni
- c. Decidere quante unità selezionare al primo e ai successivi stadi;
- d. Decidere come selezionare le unità.

Vantaggi e svantaggi



Flessibilità e adattabilità

- L'estrazione si può effettuare con criteri differenti a ogni stadio;

Riduzione dei costi

- La rilevazione dei dati è concentrata sui punti selezionati al primo stadio;
- L'organizzazione del lavoro (formazione delle liste, selezione del campione, reclutamento del personale, esecuzione della rilevazione, supervisione sul campo, ...) risulta quindi facilitata;



Complessità della metodologia di stima

Rischio di stime inefficienti

- Le unità appartenenti ad un insieme coeso tendono ad assomigliarsi e quindi le risposte risultano penalizzate nella loro variabilità

Altri campioni probabilistici

Campionamento a grappoli

- ✓ E' una procedura molto simile a quella del C. a stadi e viene utilizzata quando la popolazione è naturalmente suddivisa in gruppi di unità spazialmente contigue (famiglie, classi scolastiche, viaggiatori di un aereo, ecc.).
- ✓ E' utile quando manca la lista delle unità elementari
- ✓ Si scelgono casualmente dei grappoli di unità e si considerano tutte le unità appartenenti a tali grappoli

Campioni *non* probabilistici

Campionamento per quote

E' un campionamento stratificato con scelta rimessa all'intervistatore

1. Si suddivide la popolazione in sottogruppi sulla base di variabili di cui si conosce la distribuzione
2. Si determinano le quote del campione
3. All'interno di ciascuna quota, l'intervistatore è libero di scegliere a sua discrezione i soggetti da intervistare



Libertà concessa all'intervistatore



Risparmi di costo

Un esempio di camp. per quote

Distribuzione % della popolazione negli strati

	Istr. Sup.		Istr. Inf.	
	M	F	M	F
Giovani	9%	8%	4%	4%
Adulti	12%	10%	6%	7%
Anziani	14%	10%	7%	9%

Numerosità degli strati nel campione (**n=240**)

	Istr. Sup.		Istr. Inf.	
	M	F	M	F
Giovani	37	35	18	18
Adulti	50	44	24	29
Anziani	59	40	28	38

Campioni *non* probabilistici

◆ Campionamento a scelta ragionata

- Precede storicamente il campione casuale (o statistico) di cui non può utilizzare le proprietà derivanti dalla teoria della probabilità
- E' costituito da unità scelte in modo da somigliare nell'insieme alla popolazione da cui sono tratte
- E' tanto più rappresentativo quanto più sono vere le informazioni su cui si basa la scelta
- Non sono applicabili gli schemi di calcolo delle probabilità
- Dipende molto dalle scelte degli operatori

◆ Campionamento bilanciato

- E' un camp. ragionato nel quale si selezionano le unità in modo che la media del campione, per determinate variabili, sia prossima alla media della popolazione

Campioni *non* probabilistici

◆ Campionamento a valanga

- E' utile in caso di popolazioni "clandestine"
- I soggetti da inserire nel campione vengono individuati attraverso gli stessi soggetti intervistati
- Con il procedere della rilevazione il numero dei nominativi cresce esponenzialmente ("a valanga")
- Si rischia di selezionare le persone socialmente più attive

Le origini delle indagini campionarie



1936:

Elezioni Presidenziali U.S.A.

Candidati: F.D. Roosevelt, A. Landon



Indagine
Literary Digest

- 10 milioni di fac-simile di schede elettorali inviate a nominativi estratti dagli elenchi telefonici e dai registri automobilistici;
- 2 milioni di risposte.
- Risultato previsto: Roosevelt 41% Landon 59%

Indagine
Gallup

- Alcune migliaia di interviste ad elettori estratti casualmente dall'intera popolazione.
- Risultato previsto: Roosevelt 60% Landon 40%

Le origini delle indagini campionarie



1936:

Elezioni Presidenziali U.S.A.

Candidati: F.D. Roosevelt, A. Landon



Risultato

Roosevelt 61%



Gli errori del Literary Digest

1. Errore di copertura

- Le liste utilizzate non erano complete;
- Gli elenchi di proprietari di auto e di intestatari di telefono non erano rappresentativi dell'intera nazione ma solo dei ceti più abbienti;
- Questi ceti tendevano a votare prevalentemente repubblicano.

2. Autoselezione del campione

- Le caratteristiche socio-demografiche dei cittadini che risposero al sondaggio erano presumibilmente diverse da quelle di coloro che non risposero (*istruzione, reddito, ecc.*);
- Tali caratteristiche erano tutte correlate al voto.

Le possibili soluzioni

1. Errore di copertura

- └ Ridefinizione della popolazione;
- └ Integrazione del campione con unità estratte dalla parte di popolazione non considerata;

2. Autoselezione del campione — Riponderazione

- E' una procedura mediante la quale la composizione del campione viene modificata artificialmente (in sede di elaborazione) per renderla più vicina alla distribuzione della popolazione;
- Nella sua forma più semplice, si realizza attribuendo dei *pesi* alle singole unità campionate.

La riponderazione

- E' una procedura mediante la quale la composizione del campione viene modificata artificialmente (in sede di elaborazione) per renderla più vicina alla distribuzione della popolazione;
- Nella sua forma più semplice, si realizza attribuendo dei *pesi* alle singole unità campionate.
- La riponderazione si può effettuare:
 - a) Partendo dalle probabilità di inclusione dei soggetti nel campione
 - b) Sulla base di conoscenze che si hanno sulla popolazione (post-stratificazione)
 - c) Sulla base della conoscenza di caratteristiche dei non rispondenti

La riponderazione

La riponderazione effettuata partendo dalle probabilità di inclusione dei soggetti nel campione

- Spesso nei campioni utilizzati nella ricerca sociale la probabilità di inclusione non è uguale per tutti i soggetti
- I pesi della ponderazione vengono calcolati a partire dalle probabilità note di inclusione dei soggetti

La riponderazione effettuata sulla base di conoscenze che si hanno sulla popolazione

- A posteriori, dopo la rilevazione, si confronta la distribuzione di alcune variabili di interesse nella popolazione con quella risultante nel campione
- I dati campionari vengono corretti moltiplicando ogni caso del campione per un coefficiente di ponderazione pari al rapporto *quota teorica (nella popolazione) / quota rilevata (nel campione)*

La riponderazione effettuata sulla base della conoscenza di caratteristiche dei non rispondenti

- Le interviste raccolte vengono riponderate in modo tale da tenere conto dei rifiuti a rispondere

Un esempio di riponderazione

- Se è nota la distribuzione di alcune variabili nella popolazione, è possibile confrontare questa distribuzione con quella risultante dal campione, correggendo i dati campionari in modo da farli corrispondere, per queste variabili, ai dati nella popolazione;
- L'operazione si effettua moltiplicando ogni unità del campione per un coefficiente di ponderazione (peso) pari al rapporto quota teorica/quota rilevata della categoria di appartenenza.

Es.: Variabile "Genere" $\begin{cases} \text{Quota teorica Maschi: 49\%} \\ \text{Maschi nel campione: 58\%} \end{cases}$

- Si moltiplica ogni soggetto maschio per il peso: $\frac{49}{58} = 0,84$
- Si moltiplica ogni soggetto femmina per il peso: $\frac{51}{42} = 1,21$

La riponderazione

La riponderazione va usata con estrema cautela poiché, pur consentendo il rispetto delle proporzioni cercate, riproduce le caratteristiche delle unità già presenti, non aggiungendo, quindi, variabilità.

Riassumendo

Il campionamento offre molteplici vantaggi in termini di *costi* della rilevazione, di *tempo* richiesto per la raccolta dati e elaborazione, di *organizzazione*, di *approfondimento* e di *accuratezza*.

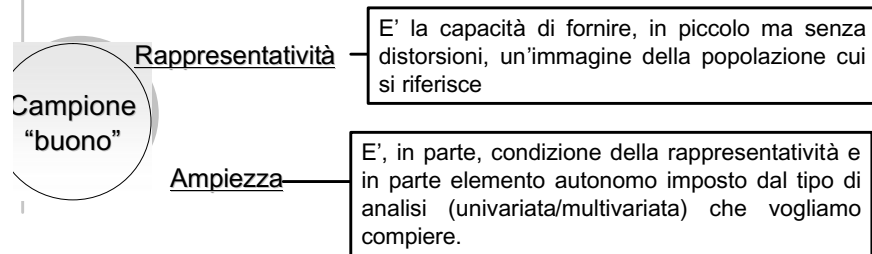
Lo studio sull'intera popolazione conduce al *valore esatto* del parametro studiato, il campione porta solo ad una *stima* di esso. Le stime basate su un campione sono affette da un *errore di campionamento*. Se il campione è stato costruito sulla base di una procedura probabilistica, l'entità di tale errore può essere determinata dalla teoria statistica.

Riassumendo

L'ampiezza del campione è direttamente proporzionale al livello di fiducia della stima, alla variabilità del fenomeno studiato, ed inversamente proporzionale all'errore che il ricercatore è disposto ad accettare

L'errore di campionamento dipende solo in piccola misura dalla dimensione della popolazione mentre è fortemente influenzato dall'ampiezza del campione.

I problemi legati al campionamento



- ✓ E' meglio intervistare 300 soggetti scelti casualmente dalla lista della popolazione che 1000 raggiunti in modo arbitrario.
- ✓ E' sempre auspicabile un confronto fra le caratteristiche del campione e quelle note delle popolazione.

Dove ripetere questa parte?...

P. Corbetta (1999)

Metodologia e tecniche della ricerca sociale
(il Mulino)

Cap. 8

-
- oppure

P. Corbetta (2003)

La ricerca sociale: metodologia e tecniche —
◦ Vol. IV. Le tecniche quantitative — Il Mulino

◦ Cap. 1

