# Il caso dell'inferenza su attributi - 1

- Le stesse relazioni possono essere trovate nel caso di stima per attributi, ovvero nel caso in cui la variabile X sia qualitativa e si debba stimare la proporzione di unità che scelgono una delle sue categorie.
- Ad esempio, se si considera la votazione per un candidato sindaco, le categorie di riferimento sono i nomi dei vari candidati (A, B, C e così via); se puntiamo l'attenzione sul candidato A, gli associamo la variabile binaria che assumerà il valore "1" se si vota per questo candidato e "0" se non lo si vota.

# La variabile binaria «binomiale»

- Determinazioni:
- 0 per «non voto»
- 1 per «voto»
- $\pi$ probabilità di «voto»
- $1-\pi$ probabilità di «non voto»

- Qual è la media aritmetica?
- Qual è lo scarto quadratico medio?

# Il caso dell'inferenza su attributi - 2

- Questa variabile avrà come media nella popolazione il valore π, ovvero la proporzione di cittadini che lo votano, e $\sqrt{(\pi(1-\pi)}$ ) come scarto quadratico medio.
- Nel campione la proporzione degli elettori di A sarà p e l'intervallo di confidenza dell'induzione si modificherà in questo modo:

$$\text{Perc.}\{p-1{,}96\sqrt{(\pi\,(1-\pi)}\,)/\sqrt{n} \le \pi \le p+1{,}96\sqrt{(\pi(1-\pi)}\,)/\sqrt{n}\}=95\%$$

# Il caso dell'inferenza su attributi - 3

- Ovvero: se nel nostro campione di dimensione n una proporzione di p cittadini ha espresso l'intenzione di votare per il candidato A, nel 95% dei casi la proporzione di potenziali elettori di A nella popolazione sarà compresa nell'intervallo qui sopra enunciato.
- Anche in questo caso c'è il problema dell'inserimento nella formula (per lo scarto quadratico medio) di un valore incognito π, che è quello che stiamo cercando di stimare: possiamo, analogamente al caso precedente, sostituirlo con la proporzione p nel campione, oppure fare una scelta conservativa, mettendoci nel caso più sfavorevole possibile e sostituirlo col valore 0,5 (il valore cui corrisponde la varianza maggiore).

# La dimensione campionaria - 1

- In un campione di **900** unità, estratto da una popolazione di ampiezza grande, **225** intervistati hanno dichiarato di praticare almeno uno sport:
- la proporzione campionaria (la *statistica*) è quindi **.25**, ovvero il **25%**;
- i valori dell'intervallo di confidenza al 95% intorno a .25 saranno compresi fra **.25-ε** e **.25+ε**, ed ε è uguale a **$1.96\sqrt{(.25(1-.25))/900}$** ovvero **0.028**.

# La dimensione campionaria - 2

- Si può quindi affermare con un livello di fiducia del 95%, ovvero prevedendo di sbagliare solo 5 volte su 100 (1 su 20), che la proporzione degli sportivi nella popolazione di riferimento è compresa fra **.222** e **.278**, ovvero gli sportivi sono tra il **22,2%** e il **27,8%**.

# La dimensione campionaria - 3

- Si tratta di un risultato poco interessante, perché l'oscillazione dei risultati (la ***forchetta***) è molto ampia, oltre il **5%**.
- Se volessimo un risultato più accurato, ad esempio se richiedessimo che ε sia al massimo l'1%, dovremmo intervistare un numero "n" di unità statistiche tale da soddisfare la seguente disequazione:

$$1.96\sqrt{(.25(1-.25))}/\sqrt{n} \leq .01$$

- ovvero **n = 7203**, che corrisponde a un onere economico e di tempo molto sensibile.

# Il test di ipotesi

- In genere, più che cercare di stimare una *statistica* sulla popolazione, i metodi inferenziali tendono a verificare una qualche ipotesi sulla popolazione stessa: un'ipotesi in Statistica è proprio un'affermazione sulla popolazione, ossia la previsione che un parametro assuma o un particolare valore o ricada in un certo intervallo di valori.

# Il test di ipotesi

- Ad esempio, che un candidato abbia o meno la maggioranza di coloro che voteranno, che in un'azienda gli uomini siamo retribuiti meglio delle donne, che l'appartenenza a una certa categoria della popolazione influenzi il comportamento elettorale, che i risultati di un atleta migliorino dopo averlo sottoposto a un particolare tipo di allenamento e così via.

# Il test di ipotesi

- Una volta definita un'ipotesi sulla popolazione bisogna raccogliere i dati campionari e verificare se i risultati, sintetizzati in una statistica test (ossia una stima puntuale del parametro nella popolazione), ci permettono di rifiutare o meno la nostra ipotesi.

# Le ipotesi $H_0$ e $H_a$

- Per comodità costruiamo due ipotesi alla base del nostro ragionamento: un'**ipotesi nulla $H_0$** e un'**ipotesi alternativa Ha**.
- L'ipotesi nulla corrisponde, in genere, a una situazione di *assenza di effetto*, mentre quella alternativa presuppone un effetto, anche se non sarà possibile misurarlo col test.

# Il test

- Il test, infatti, valuta l'evidenza campionaria dell'ipotesi $H_0$, ossia investiga se i dati contraddicano l'ipotesi nulla in maniera da suggerire che $H_a$ sia vera.
- In altre parole, si suppone che $H_0$ sia vera. Quindi, se si trova che i dati riscontrati nel campione molto difficilmente possono essere fatti risalire a quella ipotesi (perché la probabilità del test è molto bassa), allora si propende per l'ipotesi alternativa.

# Il valore di probabilità (*p-value*)

- Una volta che si è calcolato il test statistico, ossia la stima puntuale campionaria del parametro della popolazione, conoscendo la sua distribuzione campionaria si può individuare quale sarebbe la probabilità di verificarsi di un tale valore, o di uno più grande, qualora fosse vera l'ipotesi nulla.
- Questa probabilità è il ***p-value***, che viene fornito per tutti i test nei principali software statistici disponibili.

# Come trovare il *p-value*

- La conoscenza del *p-value* ci evita di andare a consultare tavole differenti a seconda di test differenti: bisogna ricordare che il test è significativo (ossia si rifiuta l'ipotesi nulla) quando il *p-value* è inferiore a un livello di probabilità da noi scelto (0,01; 0,05; 0,001 e così via), oppure quando è superiore ai valori sulle tavole corrispondenti ai livelli di probabilità scelti.

# Sintesi - 1

Possiamo quindi riassumere i vari passi di un test di ipotesi:

a- si formulano l'ipotesi nulla e quella alternativa, relativamente al parametro nella popolazione;

b- a seconda del tipo di dati a disposizione si calcola il test statistico nel campione;

c- utilizzando le informazioni sulla distribuzione campionaria del test, qualora sia vera l'ipotesi nulla, si calcola il *p-value*;

# Sintesi - 2

d- confrontando il *p-value* con il valore di probabilità con il quale assegniamo il livello di fiducia nella nostra decisione, rifiutiamo o non rifiutiamo l'ipotesi nulla;

e- il procedimento può non finire qui, in quanto quando rifiutiamo l'ipotesi nulla con un p-value significativo allo 0,05, abbiamo sempre un rischio - nel 5% dei casi - di aver rifiutato un ipotesi vera; così quando non la rifiutiamo abbiamo sempre il rischio di non aver rifiutato un'ipotesi falsa[1].

[1] Questi ulteriori passi fanno parte della Teoria delle decisioni statistiche, che non affrontiamo oggi.

# Test di ipotesi nel caso di una media

- Chiudiamo questa parte con l'esempio relativo a una variabile quantitativa: in questo caso il parametro è la media nella popolazione e il test statistico è la media campionaria. Abbiamo un campione di anziani maschi dai 65 ai 70 anni che sono pensionati in Case di riposo della Regione Lazio.
- Da studi geriatrici sappiamo che il peso medio $\mu_x$ in quella fascia d'età è di 70 chilogrammi, con un $\sigma_x$ di 10 chilogrammi. Vogliamo vedere se i ricoverati sono più o meno ben nutriti dei loro coetanei. Il campione è di 49 anziani e il peso medio $M_x$ è uguale a 68 chilogrammi con $s_x$=6 chilogrammi.

## ....

- L’ipotesi nulla è che i ricoverati siano nutriti altrettanto bene dei loro coetanei che vivono a casa: $H_0$ è che $\mu_x=70$; $H_a$ è $\mu_x\neq70$ (bidirezionale).
- L’ipotesi nulla equivale a dire che non c’è nessuna differenza fra gli anziani nelle due situazioni, quella alternativa che il trattamento, ossia il soggiorno nelle Case di riposo, ha un qualche effetto, positivo o negativo che sia sulla nutrizione degli anziani.

## ....

- Il test statistico è la media campionaria (68), che standardizziamo rispetto alla distribuzione delle medie campionarie, che è normale con media 70 e s.q.m. pari a $10/\sqrt{49}$.
- Il,valore standardizzato ($z_x = (x-\mu_x)/\sigma_x/\sqrt{n}$) è -1,4, che ha un *p-value* pari a 0,0808, ossia 8,08%, ben superiore al 2,5% che sta nella coda di un intervallo di confidenza al 95%.
- Il valore è quindi dentro l'intervallo e noi non possiamo rifiutare l'ipotesi nulla. Il peso medio più basso sarà dovuto alla variabilità campionaria e non a una situazione oggettivamente diversa.

## .....

- In questo test abbiamo usato al posto di $\sigma_x$ un dato fornito da studi geriatrici. Se lo avessimo, invece, stimato tramite $s_x$ (=5), dato che la variabilità in appartenenti allo stesso gruppo sembra essere molto più ridotta e quindi il test più accurato, avremmo dovuto considerare il valore della *t di Student*

$$t_x = (x-\mu_x)/s_x/\sqrt{n} = -2,8.$$

- Questo risultato ci fornisce un *p-value* pari a 0,0048 che ci spinge a rifiutare l'ipotesi nulla!