www.sce.unimore.it

Scienze della Comunicazione
e dell'Economia

VERIFICA DI IPOTESI

Legacy Edition

Luca La Rocca
luca.larocca@unimore.it

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Ipotesi statistiche

Un'ipotesi statistica è una congettura sul valore di un parametro (nella popolazione di interesse per una certa indagine).

Per esempio (Borra & Di Ciaccio, 2008, Esempio 13.2.4) è un'ipotesi statistica la congettura che l'altezza media degli italiani nati nel 1980 sia pari a 175 cm:

$$\mu = 175$$

Un'ipotesi statistica è dunque individuata da un vincolo su un parametro: i valori che soddisfano il vincolo (qui uno solo) sono quelli per i quali la congettura è vera.

Per fare un altro esempio (Borra & Di Ciaccio, 2008, Esempio 13.2.1) è un'ipotesi statistica l'affermazione di un'azienda produttrice di batterie per autovetture secondo la quale la durata media di un certo modello di batteria è almeno pari a 3400 ore:

$$\mu \quad \geq \quad 3400$$

Qui il vincolo è un vincolo di disuguaglianza (invece che di uguaglianza) soddisfatto dagli infiniti valori della durata media

es. 3400, 3500, 4000, . . .

per i quali l'affermazione dell'azienda produttrice è vera.

Un terzo esempio di ipotesi statistica (Borra & Di Ciaccio, 2008, Esempio 13.2.6) è la congettura che la pressione sanguigna media dei soggetti che assumono un certo farmaco sia la stessa di quella dei soggetti che non lo assumono:

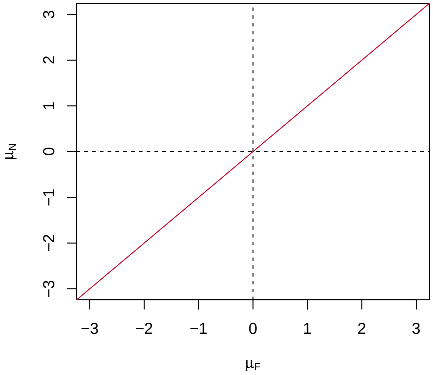
$$\mu_F = \mu_N$$

Qui il parametro è un vettore con due componenti:

$$\mu = (\mu_F, \mu_N)$$

L'ipotesi statistica è individuata, nel piano cartesiano, dalla bisettrice del primo e terzo quadrante…

$H_0 : \mu_F = \mu_N$



La verifica di un'ipotesi statistica consiste nello stabilire se un dato campione casuale (semplice) contiene "abbastanza" evidenza per rifiutare l'ipotesi in questione; per esempio

- si prendono a caso 40 italiani nati nel 1980 e se ne misurano le altezze: la loro media è "molto" diversa da 175?
- si prendono a caso 30 batterie e se ne osservano le durate: la loro media è "molto" minore di 3400?
- si somministra a 20 soggetti, presi a caso, il farmaco e ad altri 20 soggetti, sempre presi a caso, un placebo: le pressioni medie nei due gruppi sono "molto" diverse?

Se si, i dati forniscono una "chiara" indicazione contro l'ipotesi sottoposta a verifica e questa sarà rifiutata; altrimenti...

…sino a prova contraria l'ipotesi sottoposta a verifica sarà accettata: come mai questa asimmetria?

Le ipotesi sottoposte a verifica hanno tipicamente uno "status" privilegiato (per cui sono da considerarsi vere sino a prova contraria):

- nell'esempio dell'altezza degli italiani 175 cm potrebbe essere l'altezza dei nati nel 1950 (genitori dei soggetti osservati);
- nell'esempio delle batterie si presume la buona fede del produttore, finché non possiamo dimostrare che mente;
- nell'esempio della pressione sanguigna si presume l'inefficacia del farmaco, finché non possiamo dimostrarne l'efficacia.

In virtù del suo "status" privilegiato, l'ipotesi sottoposta a verifica si dice ipotesi nulla.

La negazione dell'ipotesi nulla si dice ipotesi alternativa:

- nell'esempio dell'altezza degli italiani l'ipotesi alternativa è che la media dei nati nel 1980 sia diversa da quella dei nati nel 1950...

$$\mu \neq 175$$

- nell'esempio delle batterie l'ipotesi alternativa è che la durata media sia minore di 3400 ore...

$$\mu < 3400$$

- nell'esempio della pressione sanguigna l'ipotesi alternativa è che la pressione media dei soggetti che assumono il farmaco sia diversa da quella dei soggetti che non lo assumono...

$$\mu_F \neq \mu_N$$

Si dice sistema di ipotesi la bipartizione dello spazio parametrico in ipotesi nulla e ipotesi alternativa:

- nell'esempio dell'altezza degli italiani

$$\begin{aligned} H_0 &: \mu = 175 \\ H_1 &: \mu \neq 175 \end{aligned}$$

- nell'esempio delle batterie

$$\begin{aligned} H_0 &: \mu \geq 3400 \\ H_1 &: \mu < 3400 \end{aligned}$$

- nell'esempio della pressione sanguigna

$$\begin{aligned} H_0 &: \mu_F = \mu_N \\ H_1 &: \mu_F \neq \mu_N \end{aligned}$$

Diremo che un'ipotesi è semplice quando è definita da un vincolo di uguaglianza:

- $\mu = 175$ nell'esempio dell'altezza dei nati nel 1980;
- $\mu_F = \mu_N$ ($\mu_F - \mu_N = 0$) nell'esempio della pressione sanguigna.

Diremo che un'ipotesi è composita quando è definita da un vincolo di disuguaglianza:

- $\mu \geq 3400$ nell'esempio delle batterie;
- l'ipotesi alternativa in tutti e tre gli esempi ($\mu \neq 175$, $\mu_F \neq \mu_N$ e $\mu < 3400$).

Un'ipotesi composita, tipicalmente l'ipotesi alternativa, si dice unidirezionale (unilatera) se è definita da un vincolo del tipo

$$\mu \quad < \quad \mu_0,$$

es. $\mu_0 = 3400$ nel caso delle batterie, oppure

$$\mu \quad > \quad \mu_0,$$

mentre si dice bidirezionale (bilatera) se è definita da un vincolo del tipo

$$\mu \quad \neq \quad \mu_0,$$

es. $\mu_0 = 175$ nel caso dell'altezza dei nati nel 1980 e $\mu_F - \mu_N \neq 0$ nell'esempio della pressione sanguigna.

Ipotesi statistiche

Test statistici

Il caso di medie e proporzioni

Confronto tra due gruppi

Il caso della deviazione standard

Come stabiliamo quando un campione fornisce una "chiara" indicazione contro l'ipotesi nulla?

Un test statistico è una regola per discriminare i campioni che, se osservati, portano al rifiuto di H_0 (da quelli che, se osservati, portano ad accettare H_0... sino a prova contraria).

La regione di rifiuto di un test statistico è formata dai campioni che contengono "abbastanza" evidenza contro H_0.

La regione di accettazione di un test statistico è formata dai campioni che <u>non</u> contengono "abbastanza" evidenza contro H_0.

La regione di rifiuto e quella di accettazione bipartiscono lo spazio campionario, consentendoci di prendere una decisione (rifiutare o meno l'ipotesi nulla) qualunque sia il campione osservato.

La regione di rifiuto di un test statistico è tipicamente definita per mezzo di una statistica test (funzione dei dati e dell'ipotesi nulla).

Per esempio, se μ è l'altezza media degli italiani nati nel 1980, per il sistema di ipotesi

$$\begin{aligned} H_0 : \quad \mu &= 175 \\ H_1 : \quad \mu &\neq 175 \end{aligned}$$

si può definire la regione di rifiuto come l'insieme dei campioni per i quali

$$|T| > t_\alpha \qquad \text{cioè} \qquad T < -t_\alpha \quad \text{o} \quad T > +t_\alpha$$

dove

$$T = \frac{\bar{X} - \mu_0}{\widehat{\mathcal{SE}}}$$

Si ricordi che $\widehat{\mathcal{SE}} = S/\sqrt{n}$ cosicché la statistica test T è funzione della media campionaria $\bar{X}$, della deviazione standard "corretta" S, della numerosità campionaria $n = 40$ e del valore del parametro sotto l'ipotesi nulla (semplice) $\mu_0 = 175$.

Il valore critico t_α è scelto in modo che

$$\mathbb{P}(T > t_\alpha | H_0) = \alpha$$

per un "opportuno" α (es. $\alpha = 0.05$) sulla base della distribuzione campionaria di T sotto H_0 (t di Student con $n - 1$ gradi di libertà): la regione di rifiuto consisterà di valori "estremi" sotto l'ipotesi nulla.

Perché scegliamo in questo modo la regione di rifiuto?

La verifica di H_0 darà un risultato errato in due casi:

- H_0 è vera e noi la rifiutiamo, nel qual caso parleremo di errore di prima specie;
- H_0 è falsa e noi la accettiamo, nel qual caso pareremo di errore di seconda specie.

La probabilità di commettere un errore (di prima o seconda specie) è controllata dalla funzione potenza del test

$$\mathbb{P}(|T| > t_\alpha | \mu)$$

cioè dalla probabilità di rifiutare H_0 al variare di μ in $H_0 \cup H_1 \ldots$

…in particolare la probabilità di un errore di prima specie sarà (sfruttando la simmetria della distribuzione t di Student)

$$\mathbb{P}(|T| > t_\alpha | H_0) \quad = \quad 2 \times \mathbb{P}(T > t_\alpha | H_0) \quad = \quad 2\alpha \quad = \quad 10\%$$

e sarà detta livello di significatività del test (probabilità di rifiutare l'ipotesi nulla se questa è vera).

La scelta di α (determinando il valore critico t_α) permette di fissare un certo livello di significatività (livello al quale l'ipotesi nulla sarà verificata); nell'esempio stabiliamo che per noi è <u>accettabile</u> commettere un errore di prima specie in un campione su dieci (supponendo la nulla vera).

D'altra parte la probabilità di un errore di seconda specie sarà

$$\mathbb{P}(|T| \leq t_\alpha | H_1) \quad = \quad 1 - \mathbb{P}(|T| > t_\alpha | H_1)$$

e per renderla piccola converrà scegliere una regione di rifiuto con elevata probabilità sotto H_1 (a rigore occorrerebbe specificare $\mu \in H_1$, visto che l'alternativa non è semplice come la nulla…).

La ricerca di un test potente sotto l'alternativa (quanto più possibile, compatibilmente con il livello di significatività prescelto) determina la forma della regione di rifiuto e ci porta a rifiutare per valori "estremi" di T.

Borra & Di Ciaccio (2008, p. 370) discutono la scelta della numerosità campionaria n per ottenere una certa potenza sotto l'alternativa.

In caso di rifiuto diremo che l'ipotesi alternativa è statisticamente significativa al livello di significatività fissato...

...come fissarlo?

Il contesto applicativo può suggerire maggiore o minore prudenza (tanto più è basso il livello di significatività, tanto più l'ipotesi nulla sarà vera sino a prova contraria); tipicamente, in assenza di esigenze applicative specifiche, interessano due valori convenzionali:

- il 5% quando non diversamente specificato ($\alpha = 0.025$);
- l'1% nel qual caso si parla di alta significatività ($\alpha = 0.005$).

Se per esempio (campione casuale semplice di $n = 40$ altezze) si trova

$$
\begin{aligned}
\bar{x} &= 178 \\
s &= 8
\end{aligned}
$$

il valore osservato della statistica test sarà

$$
t = \frac{178 - 175}{\frac{8}{\sqrt{40}}} \simeq 2.37
$$

da confrontare <u>tipicamente</u> con i valori critici (39 gradi di libertà... praticamente la normale standard)

$$
\begin{aligned}
t_{0.025} &= 2.02 \\
t_{0.005} &= 2.71
\end{aligned}
$$

Dato che il valore osservato della statistica test ($t = 2.37$) è maggiore del valore critico $t_{0.025} = 2.02$, rifiutiamo al livello 5% l'ipotesi che l'altezza media dei nati nel 1980 sia la stessa dei nati nel 1950: l'incremento di altezza osservato è significativo, cioè può generalizzarsi dal campione alla popolazione (accettando di sbagliare, se H_0 è vera, nel 5% dei campioni).

Dato che il valore osservato della statistica test ($t = 2.37$) è minore del valore critico $t_{0.005} = 2.71$, accettiamo al livello 1% l'ipotesi che l'altezza media dei nati nel 1980 sia la stessa dei nati nel 1950: l'incremento osservato non è altamente significativo, cioè non può generalizzarsi dal campione alla popolazione (accettando di sbagliare, se H_0 è vera, nell'1% dei campioni).

Poiché le conclusioni dipendono dal livello di significatività, piuttosto che scegliere preventivamente un livello al quale verificare H_0, nella pratica si preferisce spesso calcolare il livello di significatività osservato

$$p \quad = \quad \mathbb{P}(|T| > t | H_0)$$

noto come p-value o, in traduzione italiana, come “valore p”.

Si tratta della proporzione di campioni in cui, se è vera l'ipotesi nulla, la statistica test T assume un valore più “estremo” del valore osservato t: una misura (inversa) dell'evidenza fornita dai dati contro H_0 (valore “piccoli” di p indicano “molta” evidenza contro H_0).

ATTENZIONE: il p-value non è la probabilità che l'ipotesi nulla sia vera alla luce dei dati (per valutare la quale occorrerebbe avventurarsi nel reame della statistica bayesiana. . . cosa che qui non facciamo).

Una volta calcolato p le conclusioni sulla significatività o meno della differenza osservata si ottengono confrontandolo con uno o più livelli di interesse: p è il minimo livello al quale rifiutiamo l'ipotesi nulla.

Nell'esempio con l'aiuto di un calcolatore troviamo

$$p = \mathbb{P}(|T| > 2.37|H_0) = 2 \times \mathbb{P}(T > 2.37|H_0) = 2.28\%$$

di modo che la differenza osservata è significativa al 3%,
non significativa al 2%; dalla tabella di Borra & Di Ciaccio (2008, p. 493) si evince solo che

$$2\% < p < 5\%$$

in quanto $t_{0.025} < t < t_{0.010}$, ma tanto basta per concludere (usando le tipiche soglie convenzionali).

Si osservi che, poiché l'alternativa è bilatera, le stesse conclusioni varrebbero se avessimo osservato $t = -2.37$ ($\bar{x} = 172$) eccetto che in questo caso avremmo osservato un decremento significativo (ma non altamente significativo).

In alternativa (prima di effettuare il test o meglio di raccogliere i dati) potremmo convincerci (es. considerando i mutamenti nell'alimentazione) che un eventuale cambiamento sia necessariamente stato un aumento (non possa essere stato una diminuzione) dell'altezza media e quindi formulare il sistema di ipotesi

$$\begin{array}{lcl} H_0: & \mu & = & 175 \\ H_1: & \mu & > & 175 \end{array}$$

In questo caso (alternativa unilatera destra) rifiuteremo solo per valori “grandi” di T: valori “piccoli” di T non forniscono alcuna evidenza contro H_0, semmai forniscono evidenza in favore di H_0...

...pertanto il p-value, calcolato su una coda, varrà

$$p = \mathbb{P}(T > 2.37 | H_0) = 1.14\%$$

esattamente la metà di prima, cosicché si rifiuta al 2% e non all'1%.

Si noti come con l'alternativa unilatera sia più facile rifiutare (p-value dimezzato): nel dubbio è dunque più prudente optare per l'alternativa bilatera (altrimenti si avvantaggia H_1, mentre vogliamo considerare H_0 vera sino a prova contraria).

Se, avendo optato per l'alternativa unilatera destra, osservassimo

$$\bar{x} \quad = \quad 172 \quad < \quad 175 \quad = \quad \mu_0$$

avremmo $t = -2.37$, il p-value (calcolato sulla sola coda destra) varrebbe

$$p = \mathbb{P}(T > -2.37 | H_0) = 1 - \mathbb{P}(T > 2.37 | H_0) = 98.86\% > 50\%$$

e non vi sarebbe alcun motivo di rifiutare: la media campionaria costituirebbe evidenza in favore dell'ipotesi nulla, avendo noi escluso a priori l'eventualità $\mu < 175$; attenzione a scegliere l'alternativa unilatera prima di vedere i dati.

Nell’esempio delle batterie abbiamo il sistema di ipotesi

$$\begin{array}{llll} H_0: & \mu & \geq & 3400 \\ H_1: & \mu & < & 3400 \end{array}$$

dove l’alternativa unilatera (sinistra) nasce naturalmente in contrapposizione a una nulla composita...

...in questo caso, dunque, non abbiamo necessità di scegliere l’alternativa; si pone tuttavia un nuovo problema:

> *quale valore di μ usiamo per calcolare t (e determinare la distribuzione di T sotto H_0)?*

Da questa scelta dipenderà il p-value (coda sinistra)...

...infatti, per $\bar{x}$, s ed n fissate, troviamo che

$$p = \mathbb{P}(T < t|\mu) = \mathbb{P}(T < (\bar{x} - \mu)/(s/\sqrt{n})|\mu)$$

al variare di $\mu \geq \mu_0$ definisce una funzione decrescente di μ: tanto più grande prendiamo μ, tanto più piccolo sarà p...

...e quindi più facile rifiutare H_0: poiché consideriamo la nulla vera sino a prova contraria, prenderemo μ il più piccolo possibile e cioè pari al valore di soglia: $\mu = \mu_0$.

In pratica opereremo come se il sistema di ipotesi fosse

$$\begin{aligned} H_0 &: \mu = 3400 \\ H_1 &: \mu < 3400 \end{aligned}$$

Se per esempio si provano $n = 30$ batterie e si trova

$$\begin{aligned} \bar{x} &= 3351 \quad \text{ore} \\ s &= 242 \quad \text{ore} \end{aligned}$$

l'errore standard stimato sarà

$$\widehat{se} = \frac{242}{\sqrt{30}} = 44.18$$

e la statistica test osservata varrà

$$t = \frac{3351 - 3400}{44.18} = -\frac{49.00}{44.18} = -1.11;$$

troveremo allora (con l'aiuto di un calcolatore)

$$p = \mathbb{P}\{T < -1.11\} = 13.8\%$$

dove T segue la distribuzione t di Student con $n - 1 = 29$ gradi di libertà, ovvero (avvalendoci della tabella di Borra & Di Ciaccio, 2008, p. 493)

$$p > 10\%$$

in quanto per simmetria

$$\mathbb{P}\{T < -1.11\} = \mathbb{P}\{T > 1.11\}$$

e $|t| = 1.11 < 1.31 = t_{0.100}$.

In questo caso non abbiamo abbastanza evidenza per confutare l'affermazione dell'azienda produttrice: dobbiamo accettarla.

In altre parole: è troppo plausibile che la media campionaria osservata sia inferiore alla media di popolazione per un puro "caso", cioè perché abbiamo selezionato un campione "sfortunato", per potere scartare (con ragionevole certezza) questa eventualità.

Se non siamo convinti, possiamo (investire tempo e denaro per) mettere alla prova un campione più numeroso di batterie (resta infatti vero che le durate osservate forniscono evidenza contro H_0, sebbene tale evidenza non sia sufficiente a farcela rifiutare).

Ipotesi statistiche

Test statistici

Il caso di medie e proporzioni

- Test per la media di un carattere non dicotomico
- Test per una proporzione (carattere dicotomico)

Confronto tra due gruppi

Il caso della deviazione standard

Abbiamo visto che la verifica di un'ipotesi si articola nei seguenti passi:

1. formalizzazione di un sistema di ipotesi (nulla/alternativa);
2. scelta di una statistica test (regione di rifiuto);
3. calcolo del *p*-value (distribuzione campionaria);
4. confronto del *p*-value con uno o più valori di soglia (conclusioni).

La distribuzione campionaria della statistica test dipende, in generale, da alcuni assunti (es. popolazione normale, campione "grande") che determinano l'ambito di applicabilità di una data procedura per la verifica di ipotesi (test statistico).

Nel seguito si formalizzano i test statistici comunemente impiegati per la verifica di ipotesi su medie e proporzioni (medie di caratteri dicotomici).

Sia $x_1, \ldots, x_n$ un campione casuale semplice per il carattere X con media di popolazione μ; se X ha distribuzione di popolazione normale o se $x_1, \ldots, x_n$ è un campione "grande" (diciamo $n \geq 30$) e si vuole verificare l'ipotesi nulla

$$H_0 : \mu = \mu_0$$

contro l'alternativa

- bilatera

$$H_1 : \mu \neq \mu_0$$

- unilatera destra

$$H_1 : \mu > \mu_0$$

- unilatera sinistra

$$H_1 : \mu < \mu_0$$

ci si può basare sulla statistica test osservata

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

dove $\bar{x}$ e s sono media e deviazione standard “corretta” di $x_1, \ldots x_n$.

Il p-value sarà

- $p = 2 \times \mathbb{P}\{T > |t|\}$ se l’alternativa è bilatera
- $p = \mathbb{P}\{T > t\}$ se l’alternativa è unilatera destra
- $p = \mathbb{P}\{T < t\}$ se l’alternativa è unilatera sinistra

dove T ha distribuzione t di Student con $n - 1$ gradi di libertà
(al crescere di n questa si avvicina sempre più alla normale standard).

Quella appena formalizzata è la procedura che abbiamo applicato nell'esempio dell'altezza degli italiani e nell'esempio della durata delle batterie per autovetture; ci sarà di nuovo utile quando tratteremo il caso del confronto tra due gruppi (campioni appaiati).

Nel caso particolare (per la verità raro) in cui sia nota la deviazione standard di popolazione σ, ci si può basare sulla statistica test

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

la cui distribuzione è la normale standard (anche per n "piccolo", se la distribuzione di popolazione è normale).

Cosa cambia?

Si usa σ al posto di s nel calcolo della statistica osservata e si denota quest'ultima con z, invece che con t, per ricordare la distribuzione da cui proviene (se l'ipotesi nulla è vera)...

...quindi si calcola il p-value sulla base della distribuzione normale standard (questo è possibile con una qualsiasi tavola della normale standard, oltre che mediante un calcolatore).

Un'applicazione di questa procedura è presentata, per esempio, da Borra & Di Ciaccio (2008, Esempio 14.2.3); l'uso della distribuzione normale standard nella verifica di ipotesi sarà esemplificato quando applicheremo il test per una proporzione.

Ipotesi statistiche

Test statistici

Il caso di medie e proporzioni

- Test per la media di un carattere non dicotomico
- Test per una proporzione (carattere dicotomico)

Confronto tra due gruppi

Il caso della deviazione standard

Sia ψ la proporzione di unità statistiche in una data popolazione che posseggono una certa caratteristica (es. favorevoli a una data proposta) e sia $\bar{x}$ la corrispondente proporzione in un campione casuale semplice $x_1, \ldots, x_n$; se si vuole verificare l'ipotesi nulla

$$H_0 : \psi = \psi_0$$

contro l'alternativa

- bilatera
$$H_1 : \psi \neq \psi_0$$
- unilatera destra
$$H_1 : \psi > \psi_0$$
- unilatera sinistra
$$H_1 : \psi < \psi_0$$

e $x_1, \ldots, x_n$ è un campione “grande” (diciamo $n\psi_0 \geq 5$ e $n(1-\psi_0) \geq 5$) ci si può basare sulla statistica test osservata

$$z = \frac{\bar{x} - \psi_0}{\sqrt{\frac{\psi_0(1-\psi_0)}{n}}} = \frac{\bar{x} - \psi_0}{se_0}$$

e il p-value sarà

- $p = 2 \times \mathbb{P}\{Z > |z|\}$ se l'alternativa è bilatera
- $p = \mathbb{P}\{Z > z\}$ se l'alternativa è unilatera destra
- $p = \mathbb{P}\{Z < z\}$ se l'alternativa è unilatera sinistra

dove Z ha distribuzione normale standard.

Seguono un paio di applicazioni del test sopra formalizzato.

Supponiamo di avere intervistato un campione casuale semplice di $n = 1200$ cittadini e che $k = 631$ di loro abbiano espresso l'intenzione di rieleggere il Sindaco: abbiamo abbastanza evidenza per concludere che il Sindaco dispone di una maggioranza?

- sistema di ipotesi

$$\begin{array}{llll} H_0: & \psi & \leq & 0.5 \\ H_1: & \psi & > & 0.5 \end{array}$$

- proporzione campionaria osservata

$$\bar{x} = \frac{k}{n} = \frac{631}{1200} = 52.58\%$$

- statistica test osservata (come se $H_0 : \psi = \psi_0$)

$$z = \frac{0.5258 - 0.5000}{\sqrt{\frac{0.5(1-0.5)}{1200}}} = \frac{0.0258}{0.0144} = 1.79$$

- p-value (alternativa unilatera destra)

$$\mathbb{P}\{Z > 1.79\} = 3.7\%$$

Concludiamo rifiutando (al 5%, ma non all'1%) l'ipotesi che i concittadini intenzionati a rieleggere il Sindaco siano una minoranza: il Sindaco dispone di una significativa maggioranza (si noti che, se l'alternativa fosse bilatera, non rifiuteremmo).

Se lanciamo una moneta 50 volte e ottieniamo 16 volte “testa”, potremmo essere interessati al sistema di ipotesi

$$\begin{aligned} H_0 &: \psi = 0.5 \\ H_1 &: \psi \neq 0.5 \end{aligned}$$

dove ψ è la probabilità di ottenere “testa” in ogni singolo lancio (proporzione di “teste” in una successione infinita di lanci).

Vogliamo cioè verificare l’ipotesi che si tratti di una moneta onesta, contro l’alternativa che non lo sia.

A tal fine calcoliamo innanzi tutto la proporzione campionaria osservata: $\bar{x} = 16/50 = 0.32$.

Dopo di che troviamo

$$z = \frac{0.32 - 0.50}{\sqrt{\frac{0.5(1-0.5)}{50}}} = \frac{-0.18}{0.07} = -2.57$$

e il *p*-value varrà

$$\begin{aligned} 2 \times \mathbb{P}\{Z > 2.57\} &= 2 \times (1 - \Phi(2.57)) \\ &= 2 \times (1 - 0.9949) \\ &= 2 \times 0.0051 \qquad = 1.02\%, \end{aligned}$$

onde concludiamo che la moneta è <u>significativamente disonesta</u> (sulla soglia dell'alta significatività).

Cosa facciamo se abbiamo un campione "piccolo"?

Se $n\psi_0 < 5$ o $n(1-\psi_0) < 5$ (es. ψ_0 è la proporzione di prodotti che falliscono il controllo di qualità a valle di un processo di produzione molto affidabile) possiamo ricorrere al test esatto basato sulla distribuzione binomiale del totale $X_1 + \cdots + X_n$ (frequenza campionaria assoluta della modalità 1)...

...non approfondiamo questa eventualità (anche perché si tratta di un'eventualità tutto sommato rara).

Un caso di particolare interesse è quello in cui si abbiano due campioni e si vogliano confrontare le medie delle popolazioni da cui questi sono estratti; per esempio si potrebbe volere confrontare:

- il reddito medio di uomini e donne in un certo paese (due popolazioni distinte o meglio due strati di una popolazione);
- la pressione sanguigna di chi assume un certo farmaco e di chi non lo assume (stessa popolazione, ma diverso trattamento);
- il peso di ragazze anoressiche prima e dopo di un certo trattamento (stessa popolazione e stesso campione, ma in momenti diversi).

Nel primo e secondo caso avremo due campioni indipendenti (distinti) mentre nel terzo caso avremo due campioni appaiati (stesse ragazze).

Denotate con μ_Y e μ_Z le medie nelle due popolazioni di interesse (gruppi) spesso interessa verificare l'ipotesi nulla

$$H_0 : \mu_Y = \mu_Z$$

contro l'alternativa bilatera (per fissare le idee)

$$H_1 : \mu_Y \neq \mu_Z$$

avendo osservato i campioni casuali semplici (eventualmente appaiati) $y_1, \ldots, y_n$ e $z_1, \ldots, z_m$ (in generale $m \neq n$) dalle due popolazioni:

> *abbiamo abbastanza evidenza per rifiutare l'ipotesi che i redditi siano uguali? che il farmaco/trattamento sia inefficace?*

Come procederemo?

- Campioni indipendenti: riformuliamo l'ipotesi nulla come

$$\mu_Y - \mu_Z = 0$$

e basiamo la nostra risposta sulla differenza tra le medie campionarie osservate

$$\bar{y} - \bar{z}$$

rapportandola al suo errore standard (stimato); il *p*-value sarà calcolato sulla base della distribuzione campionaria dalla quale il valore ottenuto proviene (normale o *t* di Student).
Non approfondiamo.

- Campioni appaiati: consideriamo il carattere

$$X = Y - Z$$

ovvero la differenza algebrica tra i valori osservati nei due campioni (prima e dopo il trattamento) e riformuliamo l'ipotesi nulla come

$$\mu = 0$$

dove $\mu = \mu_Y - \mu_Z$ è la media di popolazione di X; procederemo come se avessimo osservato $x_i = y_i - z_i$, $i = 1, \ldots, n$, verificando la nullità di μ come se fossimo nel caso di un solo campione (si noti che in questo caso abbiamo necessariamente $m = n$).

Supponiamo per esempio (Agresti & Finlay, 1997, Example 6.4) di avere un campione di ragazze anoressiche il cui peso è stato misurato prima e dopo un trattamento finalizzato all'aumento di peso.

Alla luce di quanto visto, i dati consisteranno delle differenze algebriche di peso x_i, $i = 1, \ldots, n$, dove una differenza positiva indica un aumento di peso, mentre una differenza negativa indica una diminuzione di peso.

Supponiamo di avere trovato

$$\begin{aligned} \bar{x} &= 3.007 \quad \text{Pounds} \\ s &= 7.309 \quad \text{Pounds} \end{aligned}$$

con una numerosità campionaria $n = 40$.

Verificheremo l'ipotesi di irrilevanza del trattamento

$$H_0: \quad \mu \quad = \quad 0$$

nella popolazione di tutte le (potenziali) ragazze anoressiche, contro l'alternativa (prudentemente) bilatera

$$H_1: \quad \mu \quad \neq \quad 0.$$

La statistica test osservata varrà

$$t \quad = \quad \frac{3.007 - 0.000}{\frac{7.309}{\sqrt{40}}} \quad = \quad 2.60$$

di modo che (usando per semplicità la distribuzione normale standard, giustificata perché il campione è "grande"...) troveremo

$$p = 2 \times \mathbb{P}\{Z > 2.60\} = 2 \times 0.47\% = 0.94\%$$

e concluderemo che il trattamento comporta un aumento di peso (altamente) significativo.

In conclusione: se siamo in grado di verificare la nullità di una media, siamo in grado di confrontare due gruppi (in modo immediato nel caso in cui abbiamo campioni appaiati, come visto, con un po' più di fatica nel caso in cui abbiamo campioni indipendenti, perché dobbiamo calcolare un opportuno errore standard...).

Ipotesi statistiche

Test statistici

Il caso di medie e proporzioni

Confronto tra due gruppi

Il caso della deviazione standard

Sia $x_1, \ldots, x_n$ un campione casuale semplice per il carattere X con deviazione standard di popolazione σ; se X ha distribuzione di popolazione normale o con code della stessa "pesantezza" della normale e il campione è "grande" (diciamo $n \geq 30$) per verificare l'ipotesi nulla

$$H_0 : \sigma = \sigma_0$$

contro l'alternativa

- bilatera
$$H_1 : \sigma \neq \sigma_0$$
- unilatera destra
$$H_1 : \sigma > \sigma_0$$
- unilatera sinistra
$$H_1 : \sigma < \sigma_0$$

ci si può basare sulla statistica test osservata

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

dove s è la deviazione standard "corretta" di $x_1, \dots x_n$.

Al livello di signficatività 2α rifiuteremo se

- $\chi^2 < \chi^2_{1-\alpha}$ o $\chi^2 > \chi^2_\alpha$ se l'alternativa è bilatera
- $\chi^2 > \chi^2_{2\alpha}$ se l'alternativa è unilatera destra
- $\chi^2 < \chi^2_{1-2\alpha}$ se l'alternativa è unilatera sinistra

dove χ^2_α è il valore critico della distribuzione chi-quadrato con $n-1$ gradi di libertà che individua una coda destra di probabilità α.

Se per esempio si trova $s = 8$ in un campione casuale semplice di $n = 20$ altezze e si vuole verificare l'ipotesi

$$H_0 : \sigma = 7$$

al livello di significatività del 10% contro l'alternativa

$$H_1 : \sigma \neq 7$$

il valore osservato della statistica test sarà

$$\chi^2 = \frac{19 \times 8^2}{7^2} \simeq 24.82$$

da confrontare con...

. . . i valori critici

$$\begin{aligned} \chi^2_{95\%} &= 10.12 \\ \chi^2_{5\%} &= 30.14 \end{aligned}$$

della distribuzione chi-quadrato con $n - 1 = 19$ gradi di libertà (desunti per esempio dalla tabella di Borra & Di Ciaccio, 2008, p. 494).

I dati non forniscono dunque abbastanza evidenza per concludere che la deviazione standard di popolazione è diversa da 7; se questo è il valore relativo alla generazione precedente, accetteremo (fino a prova contraria) l'ipotesi nulla che la variabilità (come misurata dalla deviazione standard) sia rimasta la stessa.

Se per esempio, provando $n = 30$ batterie, si trova $s = 242$ ore e si vuole verificare l'ipotesi

$$H_0 : \sigma \leq 200$$

al livello di significatività del 10% contro l'alternativa

$$H_1 : \sigma > 200$$

la statistica test osservata varrà

$$\chi^2 = \frac{29 \times 242^2}{200^2} \simeq 42.46$$

da confrontare con...

. . . il valore critico

$$\chi^2_{10\%} \quad = \quad 39.09$$

della distribuzione chi-quadrato con $n - 1 = 29$ gradi di libertà (desunto per esempio dalla tabella di Borra & Di Ciaccio, 2008, p. 494).

I dati forniscono dunque abbastanza evidenza per concludere che la deviazione standard di popolazione è maggiore di 200; se questo valore corrisponde alla variabilità massima dichiarata dal costruttore, rifiuteremo l'ipotesi che questi stia affermando il vero (naturalmente le nostre conclusioni varrano al livello di significatività del 10%).

- AGRESTI, A. & FINLAY, B. (1997).
  Statistical Methods for the Social Sciences.
  Prentice-Hall, Upple Saddle River.
- BORRA, S. & DI CIACCIO, A. (2008).
  Statistica: Metodologie per le Scienze Economiche e Sociali (Seconda Edizione).
  McGraw-Hill, Milano.