

Approfondimento di Statistica

Introduzione alla Data Science

Nicoletta Noceti

Oggi

Impariamo a fare analisi statistiche sui dati che possano permetterci di «capire qualcosa»

Ancora un po' di probabilità

Ricordate il teorema di Bayes?

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

- A e B sono eventi
- $P(A)$ [$P(B)$]: la probabilità che si verifichi evento A [B]
- $P(B|A)$: la probabilità che si verifichi evento B sapendo che si è verificato A
- $P(A|B)$: la probabilità che si verifichi evento A sapendo che si è verificato B

Da dove viene?

Sappiamo che

- $P(A,B) = P(A) P(B|A)$
- $P(B,A) = P(B) P(A|B)$
- $P(A,B) = P(B,A)$

Quindi

$$P(A) P(B|A) = P(B) P(A|B)$$

Da cui deriviamo il teorema

Bayes, ipotesi e dati

- Supponiamo che H sia un'ipotesi relativa a determinati dati e che D siano i dati disponibili
- Possiamo usare il teorema di Bayes per ottenere la probabilità che la nostra ipotesi sia corretta sulla base dei dati disponibili

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

- Qual è la probabilità che la mia ipotesi sia vera sulla base dei dati che ho?

Esempio

- Immaginiamo di incaricare due persone, P_1 e P_2 di scrivere un post
- In base ai post precedenti, vi è piaciuto l'80% dei post scritti da P_1 ed il 50% dei post scritti da P_2 (Il tasso di produzione di P_1 e P_2 è lo stesso)
- Arriva il nuovo post, anonimo, che vi piace molto
- Qual è la probabilità che sia stato scritto da P_2 ?

Esempio: info disponibili

- H = ipotesi = il post è stato scritto da P_2
- D = dati = il post ci è piaciuto
- $P(H|D)$ = la probabilità che il post sia stato scritto da P_2 sapendo che ci è piaciuto
- $P(D|H)$ = la probabilità che il post ci sia piaciuto sapendo che è stato scritto da P_2
- $P(H)$ = la probabilità che il post sia stato scritto da P_2
- $P(D)$ = la probabilità che il post ci sia piaciuto

Esempio: info disponibili

$P(H)$ = la probabilità che il post sia stato scritto da P_2

- Sappiamo che il rate di produzione di P_1 e P_2 è uguale, quindi possiamo supporre che $P(H) = 0.5$

$P(D|H)$ = la probabilità che il post ci sia piaciuto sapendo che è stato scritto da P_2

- Sappiamo essere il 50%, ossia $P(D|H) = 0.5$

$P(D)$ = la probabilità che il post ci sia piaciuto (in generale!!!)

- Usiamo le regole della probabilità:

$$D = (P_2 \text{ AND piaciuto}) \text{ OR } (P_1 \text{ AND piaciuto})$$

$$P(D) = P(P_2 \text{ AND piaciuto}) \text{ OR } P(P_1 \text{ AND piaciuto})$$

$$P(D) = P(P_2) \times P(\text{piaciuto}) + P(P_1) \times P(\text{piaciuto}) = 0.5 \times 0.5 + 0.5 \times 0.8 = 0.65$$

Esempio: torniamo al teorema

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} = \frac{0.5 \cdot 0.5}{0.65} = 0.38$$

Notate

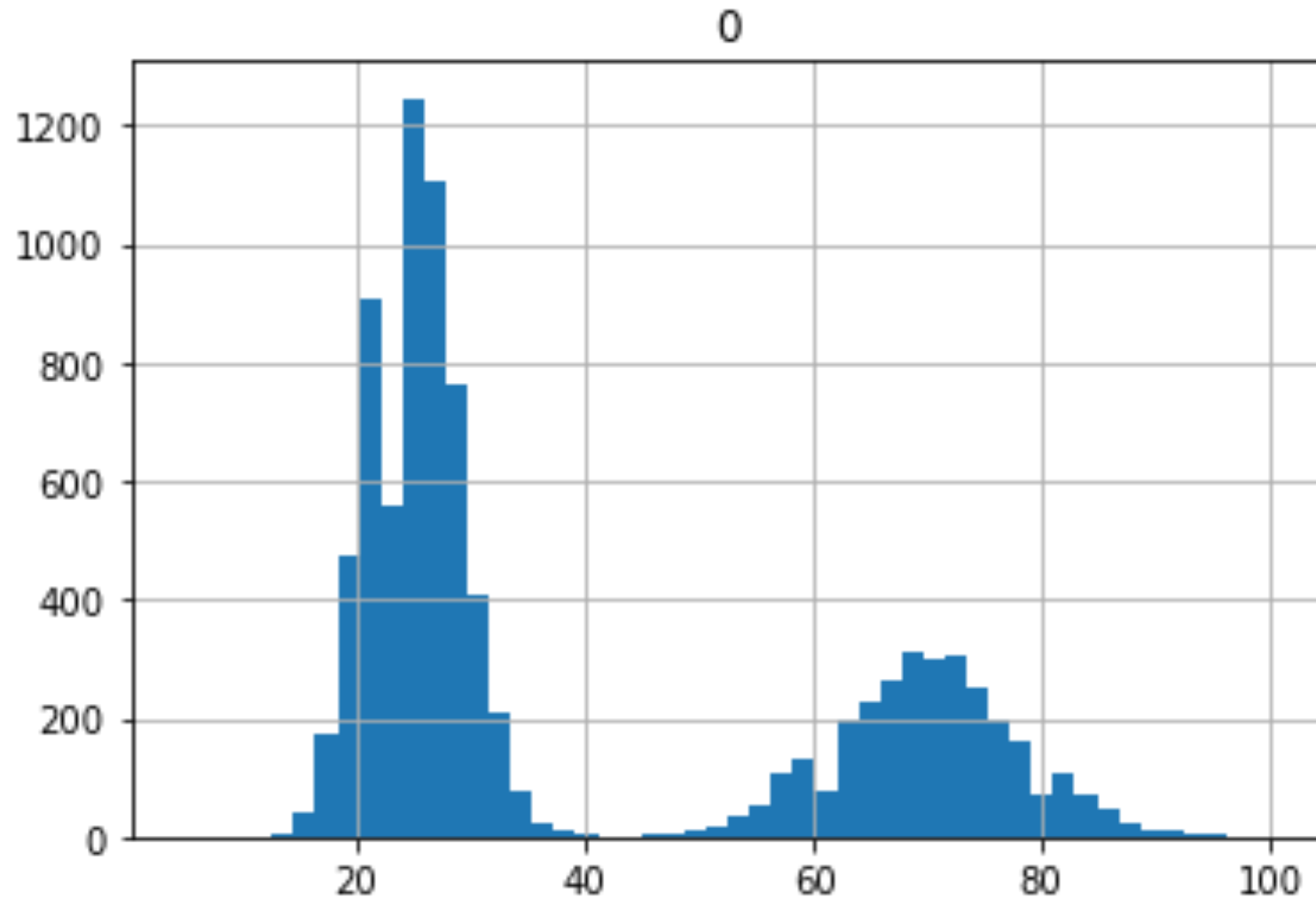
- Se non avessimo avuto dati la probabilità sarebbe stata 0.5
- Abbiamo aumentato le conoscenze a posteriori sulla base di conoscenze a priori, avendo nuovi dati utili e significativi

Stime dei punti e intervalli di confidenza

Definizione

- Una **stima di un punto** è una stima di un parametro della popolazione sulla base dei dati di un campione (es. la media)
- Esempio: supponiamo che esista una società con 9000 dipendenti e che siamo interessati a determinare la durata delle loro pause
- Non è possibile in generale ottenere la risposta da tutti i dipendenti, quindi si considera un campione per calcolare poi, ad esempio, la media
- La media del campione è la nostra stima del punto

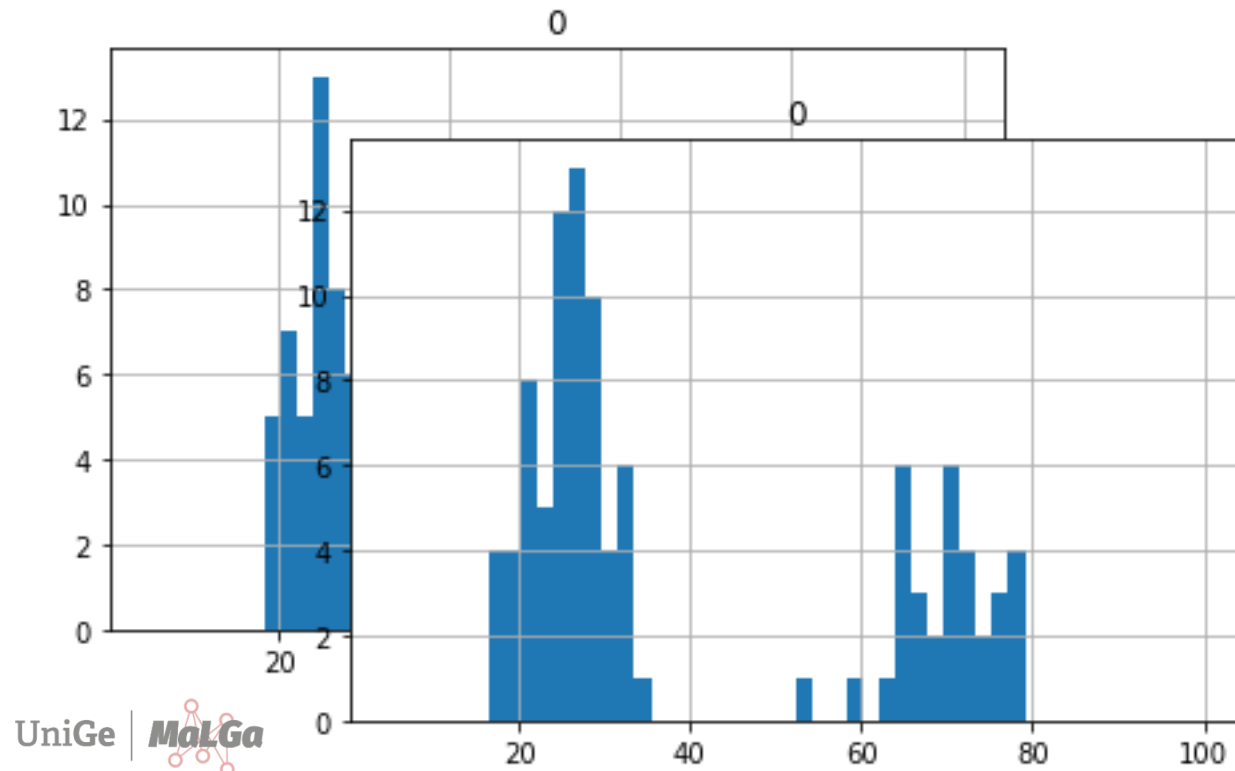
La popolazione (fingiamo di averla)



Media = 39.99 minuti (quindi circa 40 minuti)

Stimiamo il punto

- Vogliamo simulare la situazione in cui chiediamo a solo 100 dipendenti scelti a caso la durata solita delle loro pause
- Campioniamo a caso 100 risposte dalla popolazione



Media = 47 minuti

Media = 40 minuti

Distribuzioni di campionamento

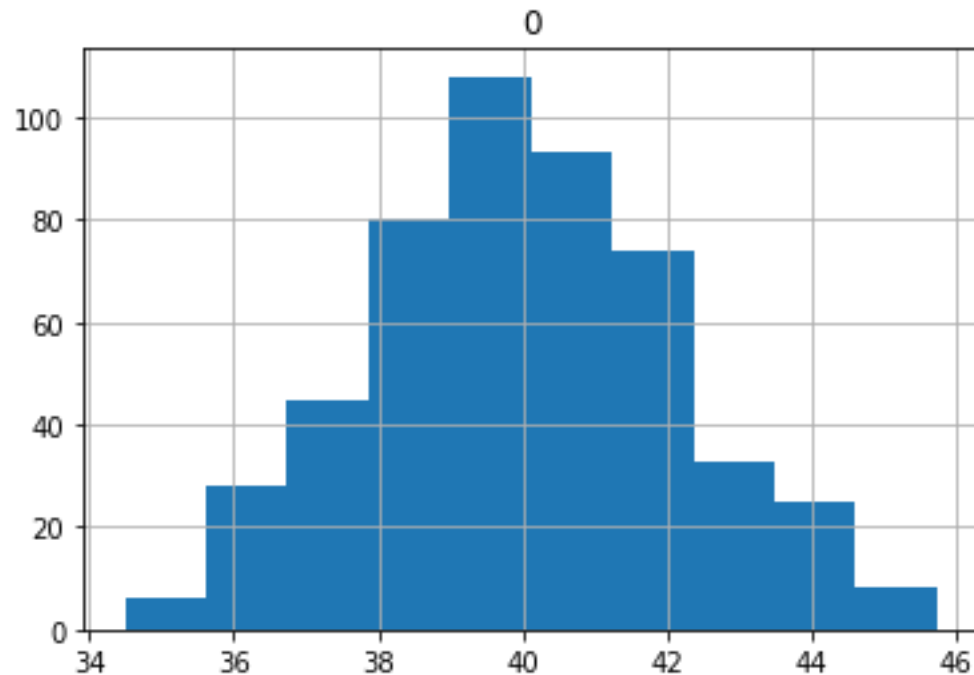
Una distribuzione delle stime di più campioni di uguale dimensione

Si procede così:

- Si campiona N volte un insieme di M campioni
- Si costruisce un istogramma delle N stime calcolate (es. media)

Torniamo all'esempio precedente: campioniamo 500 volte un insieme di 100 durate della pausa

Distribuzioni di campionamento



Media della distribuzione: 40.01194!!!

- Siamo passati da una distribuzione bimodale ad una distribuzione normale, su cui possiamo applicare test statistici
- Grazie al teorema centrale del limite, se aumentiamo il numero di campioni, la distribuzione delle stime approssima una distribuzione normale
- Se i dati sono «abbastanza», la media della distribuzione si avvicina alla media della popolazione!

Intervalli di confidenza

- Spesso non è facile (a volte è impossibile) ottenere delle stime abbastanza precise anche da campionamenti della popolazione
- In questi casi è opportuno usare il concetto di intervallo di confidenza, un intervallo di valori basato su una stima che sappiamo contenere il vero parametro della popolazione con un certo grado di confidenza
- Importante: il livello di confidenza rappresenta la frequenza con cui la risposta ottenuta è accurata (probabilità % che l'intervallo contenga il valore esatto)

Esempio: per avere il 95% di probabilità di catturare il vero parametro della popolazione usando la stima, dobbiamo impostare il livello di confidenza a 95%

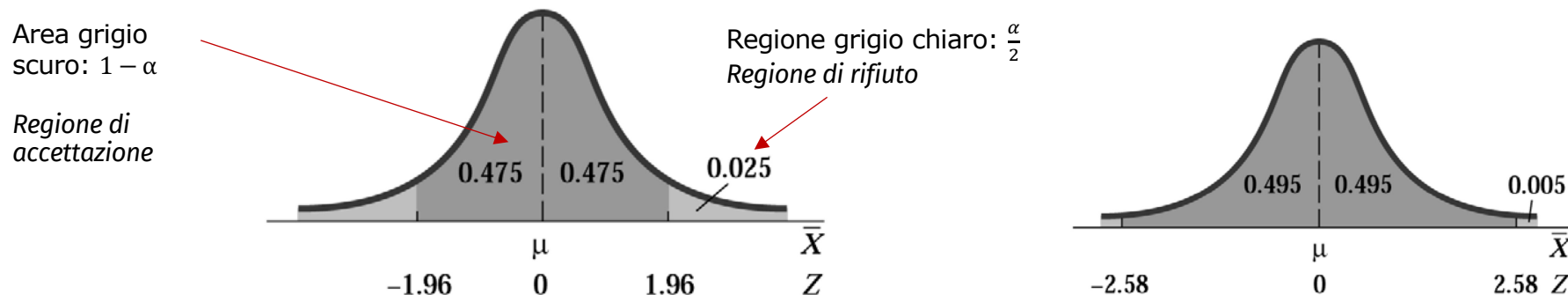
Un po' di teoria

- In generale si parla di livello di confidenza di un intervallo al $(1 - \alpha)\%$ per riferirci alla probabilità che il valore vero della statistica appartenga all'intervallo
- Per calcolare l'intervallo usiamo la formula

$$\bar{X} - Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}}$$

dove $Z_{\frac{\alpha}{2}}$ è il valore critico - valore a cui corrisponde un'area cumulata pari a $\left(1 - \frac{\alpha}{2}\right)$

della distribuzione normale standard



Cosa ci serve [con un esempio]

1. Una stima del punto [media delle pause dal campione]
2. Una stima della deviazione standard della popolazione [deviazione standard del campione / radice quadrata della dimensione del campione]

Per $\alpha = 0.05$ abbiamo valore critico 1.96. Se ad esempio la stima della media sul campione vale 37.3 e la stima della deviazione standard sul campione vale 20.55, otteniamo l'intervallo

$$(37.3 - 1.96 * \frac{20.55}{\sqrt{100}}, 37.3 + 1.96 * \frac{20.55}{\sqrt{100}})$$

Ossia (33.27, 41.33)

Intervallo di confidenza per la durata media della pausa con confidenza 95%

Intervalli di confidenza e test statistici

Verifica delle ipotesi

Usiamo il concetto di livello di confidenza per parlare della verifica delle ipotesi statistiche

Una verifica delle ipotesi è un test statistico per valutare se possiamo presumere che una determinata condizione sia vera per l'intera popolazione dato un campione limitato di dati

Il test ci dice se possiamo accettare l'ipotesi o rigettarla

Ipotesi

- Una verifica delle ipotesi di solito esamina due diverse ipotesi su una popolazione
 - Ipotesi nulla \rightarrow Ipotesi da verificare
 - Ipotesi alternativa
- Usiamo un valore p (p-value) che si basa sul **livello di significatività** per giungere alla conclusione del test (intuitivamente: la fiducia che abbiano nel risultato)
- Esempi di domande che possiamo porci
 - La durata media delle pause è diversa da 40 minuti?
 - Esiste una differenza tra persone che hanno interagito con il sito A e persone che hanno interagito con il sito B (A/B testing)?

Condurre una verifica delle ipotesi

1. Specificare l'ipotesi

- Formuliamo le due ipotesi: quella nulla e quella alternativa
- Usiamo la notazione H_0 e H_a

2. Determinare le dimensioni del campione per il test

3. Scegliere un livello di significatività

- Di solito fissato a 0.05

4. Raccogliere i dati

5. Decidere se rigettare o accettare l'ipotesi nulla (dipende dal tipo di test...)

Tre tipi di verifica delle ipotesi

- T-test su un campione
- Correttezza chi-quadrato
- Test del chi-quadrato dell'associazione o indipendenza

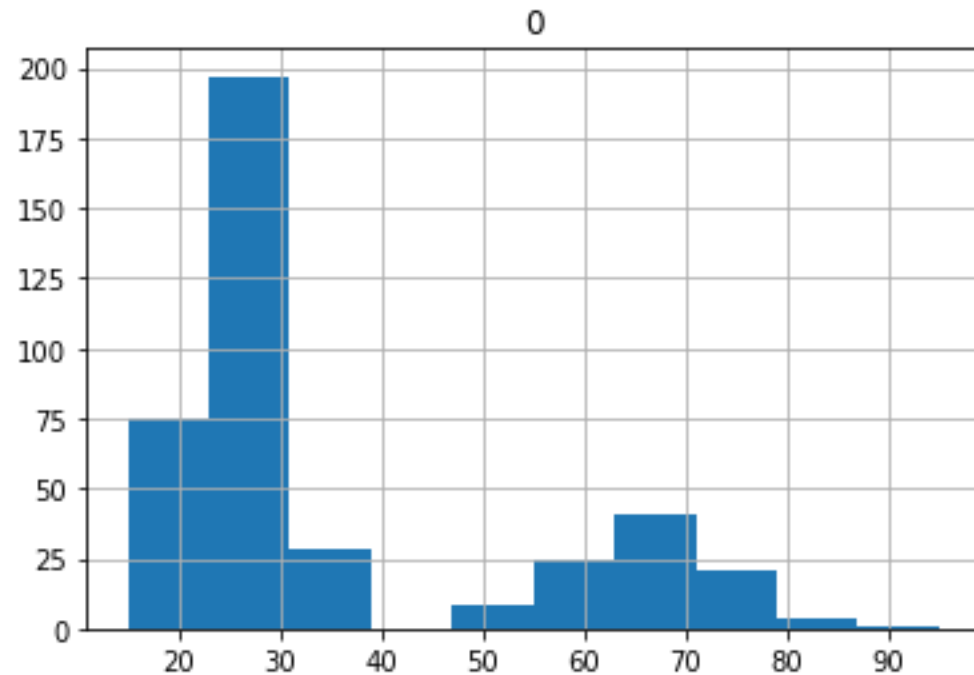


T-test

Definizione

- Il t-test per un campione è un test statistico per determinare se un campione di dati numerici (quantitativi) differisce in modo significativo da un altro dataset (come la popolazione o un altro campione)
- Continuiamo con il nostro esempio: immaginiamo di volere conoscere la durata della pause di un certo reparto. Possiamo generare dei dati simulando il fatto di avere selezionato solo i dipendenti del reparto stesso (400)

Esempio



La media è circa 34.82, che differisce di circa 5 minuti dalla media delle popolazione

NOTA BENE: di solito non abbiamo l'informazione qui sopra!!! Dato che siamo noi ad aver generato i dati, in questo caso sappiamo tutto...

Esempio

Scopo: valutare se esista una differenza statisticamente significativa tra la distribuzione della durata delle pause dell'intera popolazione e quella del reparto selezionato

Conduciamo un t-test con livello di confidenza al 95% (significatività 0.05) per trovare (o no) una differenza

Due condizioni importanti:

- La distribuzione della popolazione deve essere normale e il campione deve essere ampio ($n \geq 30$)
- La dimensione della popolazione deve essere almeno 10 volte superiore a quella del campione ($10n < N$) → Questo garantisce che il campione sia tratto in modo indipendente

I 5 passi

1. Specificare l'ipotesi

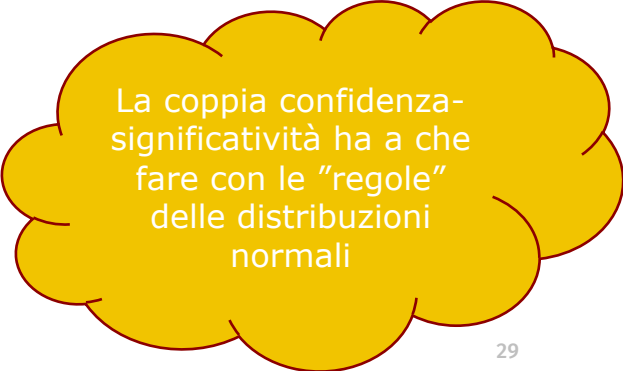
- H_0 = il reparto ha le stesse abitudini dell'azienda (stessa durata delle pause)
- H_a = la media del campione è differente dalla media dell'azienda (t-test a due code) ←
OPPURE
- H_a = la media del campione è inferiore [superiore] a quella dell'azienda (t-test a una coda)

2. Determinare le dimensioni del campione per il test

- Il campione ha almeno 30 punti? Sì, ne ha 400
- $10 \cdot 400 < 9000$ Sì

3. Scegliere un livello di significatività

- Fissando un livello di confidenza del 95% il livello di significatività è 0.05



La coppia confidenza-significatività ha a che fare con le "regole" delle distribuzioni normali

I 5 passi

4. Raccogliere i dati
5. Decidere se rigettare o accettare l'ipotesi nulla: dobbiamo calcolare la statistica del test e il valore di p , o p -value. Intuizione: il p -value ci fa capire se la differenza tra il risultato osservato e quello ipotizzato è dovuta alla casualità del campionamento dei dati o se è statisticamente significativa.

Quando i dati presentano prove molto forti contro l'ipotesi nulla, la statistica del test tende a crescere (in positivo o negativo) ed il valore di p diventa molto piccolo
→ significa che il test mostra risultati netti e quello che dimostra non è dovuto al caso

Il valore del test nel nostro esempio

t-test = -5.74 → la deviazione della media del campione rispetto all'ipotesi nulla

p-value = 0.00000018 → la frequenza con cui il risultato ottenuto si otterrebbe per caso

Se p-value < livello di significatività → rigettiamo ipotesi nulla

[Se p-value << livello di significatività → rigettiamo ipotesi nulla in favore di quella alternativa]

Se p-value > livello di significatività → non possiamo rigettare ipotesi nulla

Noi rigettiamo l'ipotesi nulla: il reparto NON ha le stesse abitudini dell'azienda

Test chi-quadrato

Test chi-quadrato dell'idoneità (o correttezza)

- Lavora su dati QUALITATIVI ragionando in termini di conteggi
- Si usa quando
 - Vogliamo analizzare una variabile categorica da una popolazione
 - Vogliamo determinare se una variabile segue una certa distribuzione, specificata oppure attesa

→ Confrontiamo quanto osservato con quanto previsto

Requisiti:

- Tutti i conteggi previsti devono essere almeno 5
- Le singole osservazioni devono essere indipendenti e le dimensioni della popolazione devono essere almeno 10 volte quelle del campione

Esempio

- Immaginiamo di volere verificare se la distribuzione dei gusti di caramelle nei sacchetti commercializzati è quella attesa
- Ogni sacchetto contiene 100 caramelle che dovrebbero essere equamente distribuite tra i 5 gruppi. Usiamo un campione di 10 sacchetti e contiamo i gusti presenti all'interno di ciascun sacchetto. Scopriamo che...

Gusto	Numero contato	Numero atteso
Mela	180	200
Lime	250	200
Ciliegia	120	200
Arancio	225	200
Uva	225	200

- Tutti i conteggi previsti sono almeno 5? Sì
- Le singole osservazioni sono indipendenti? Sì
- Le dimensioni della popolazione sono almeno 10 volte quelle del campione? Sì

Da https://www.jmp.com/it_it/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.html

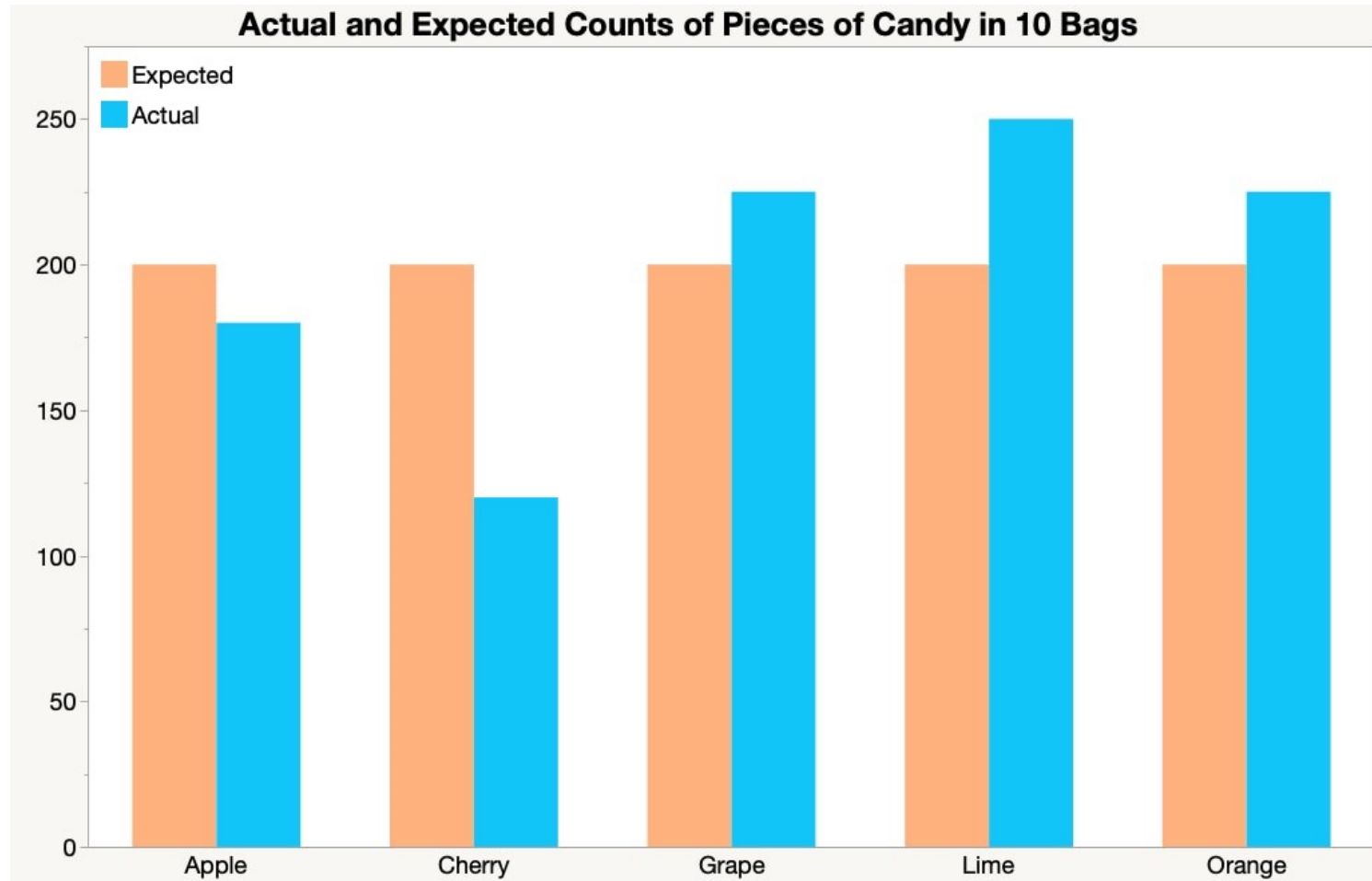
Esempio

- H_0 : la proporzione tra i gusti in ogni sacchetto di caramelle è la stessa
- H_a : la proporzione tra i gusti in ogni sacchetto di caramelle NON è la stessa
- Eseguiamo un test con livello di significatività pari a 0.05
- I gradi di libertà sono $5-1=4$ (numero di classi – 1)

$$\chi = \sum \frac{(Osservato - Atteso)^2}{Atteso}$$

- Il valore del test risulta essere 52,75
- Come facciamo a capire se il risultato è statisticamente significativo?

Si può vedere «a occhio»?



Ci chiediamo: i dati rilevati dal nostro campione sono “abbastanza vicini” al risultato atteso da poter concludere che la proporzione tra i gusti nei sacchetti della popolazione in esame sia uguale oppure no?

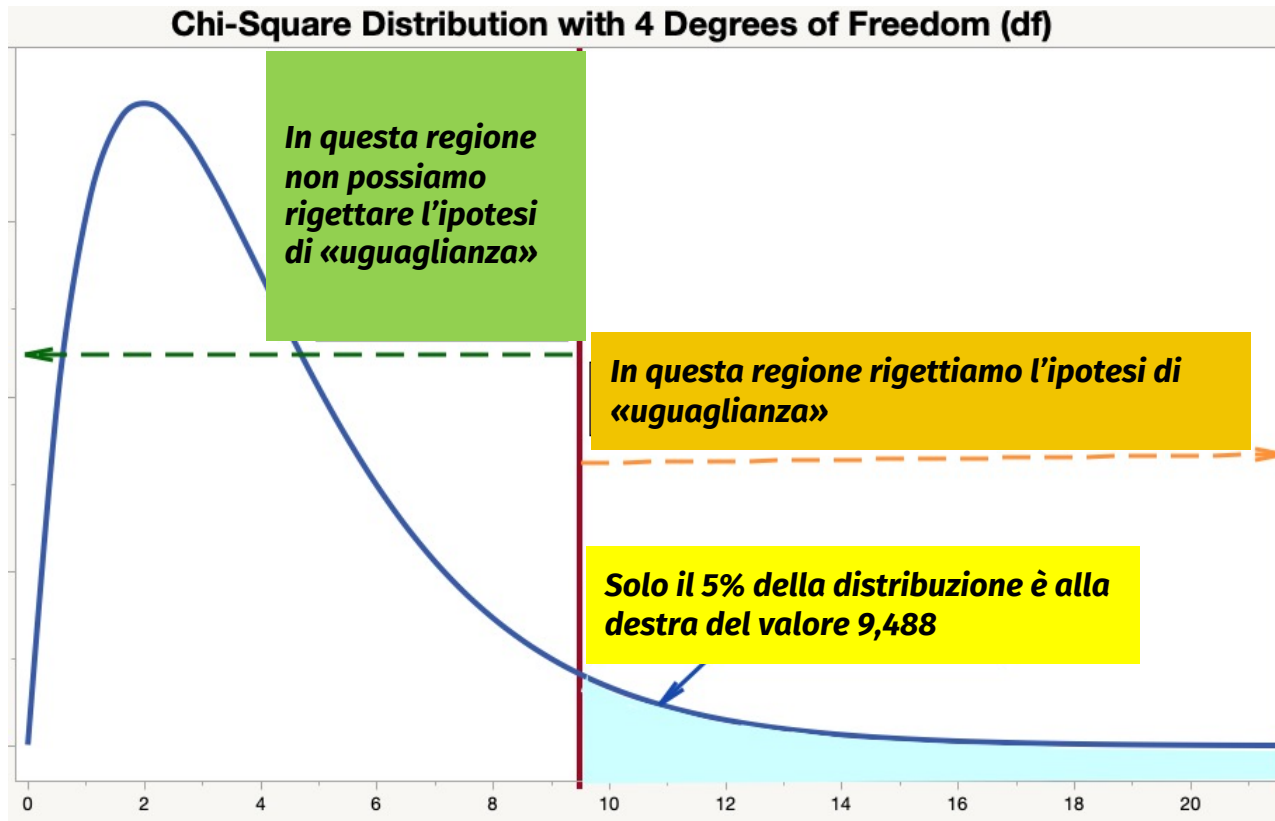
https://www.jmp.com/it_it/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.html

Esempio

- Per trarre una conclusione avremo bisogno di confrontare il valore ottenuto dal test ed il valore della distribuzione de chi-quadrato corrispondente a livello di confidenza e gradi di libertà del nostro problema
- Cosa può succedere:
 - La statistica di test è inferiore al valore del chi-quadrato → non possiamo rifiutare ipotesi nulla
 - La statistica di test è superiore al valore del chi-quadrato → rifiutiamo ipotesi nulla
- Controllando la tabelle della distribuzione dei chi-quadrato (ad es qui: https://it.wikipedia.org/wiki/Distribuzione_chi_quadrato) scopriamo che il valore del chi-quadrato nel nostro caso -- ossia significatività 0.05 e 4 gradi di libertà -- è 9,488
- Poiché $52.75 > 9.488$, possiamo rifiutare l'ipotesi nulla secondo cui la proporzione tra i gusti di caramelle sarebbe la stessa.

ATTENZIONE: la tabella ha su ogni colonna la confidenza!!! 0.95 nel nostro caso

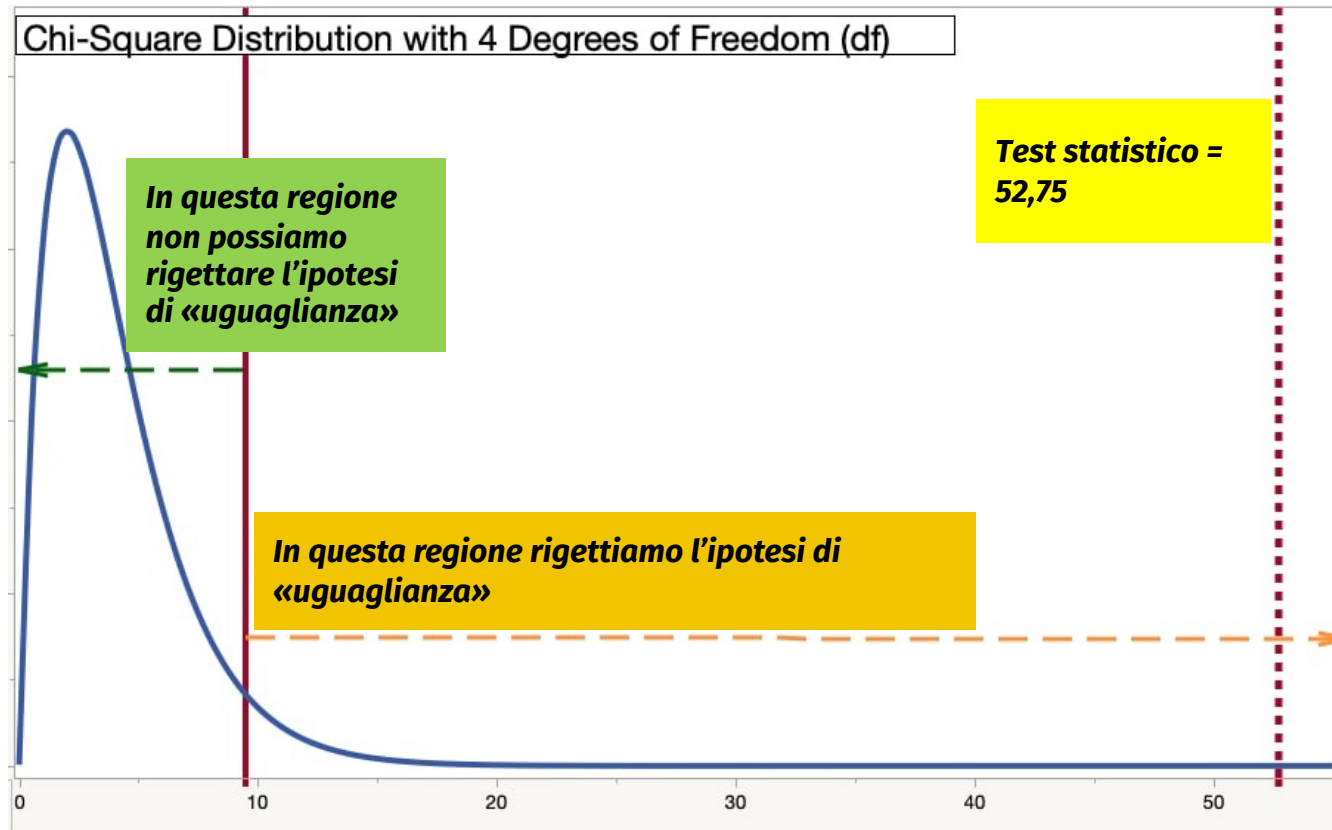
Cerchiamo di capire meglio



- Stiamo verificando se la statistica di test è un valore più estremo del valore critico nella distribuzione.
- Ma dove è la nostra statistica? Ricordiamo che valeva circa 53

https://www.jmp.com/it_it/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.html

Cerchiamo di capire meglio



Il p-value di un test è la probabilità che un altro campione delle stesse dimensioni possa produrre una statistica di test più estrema rispetto a quella del campione studiato, dietro l'assunzione che l'ipotesi nulla sia vera

Il calcolo è difficile da eseguire manualmente, usiamo software appositi

Test Chi-quadrato dell'associazione/indipendenza

- Ci aiuta a determinare se due variabili categoriche sono indipendenti fra loro
- Le condizioni necessarie sono le stesse del test chi-quadrato
- **Esempio** (preso da https://www.jmp.com/it_it/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html):
abbiamo raccolto informazioni circa il consumo di snack al cinema, vogliamo scoprire se ci sia una relazione tra genere di film e snack

	Snack	No snack
Azione	50	75
Commedia	125	175
Family	90	30
Horror	45	10

Esempio

- H_0 : genere di film e consumo di snack sono variabili indipendenti
- H_a : genere di film e consumo di snack NON sono variabili indipendenti
- Nel trarre una conclusione avremo di nuovo bisogno di confrontare il valore ottenuto dal test ed il valore della distribuzione de chi-quadrato corrispondente a livello di confidenza e gradi di libertà del nostro problema
- Cosa può succedere:
 - La statistica di test è inferiore al valore del chi-quadrato, per cui non è possibile rifiutare l'ipotesi di indipendenza.
 - La statistica di test è superiore al valore del chi-quadrato, per cui l'ipotesi di indipendenza viene rifiutata.

Esempio

Questa volta dobbiamo calcolare i conteggi attesi, usiamo la tabella di contingenza

	Snack	No snack	Totale riga
Azione	50	75	125
Commedia	125	175	300
Family	90	30	120
Horror	45	10	55
Totale colonna	310	290	600

Per calcolare i conteggi attesi usiamo la formula $\text{totale riga} \times \text{totale colonna} : \text{totale}$

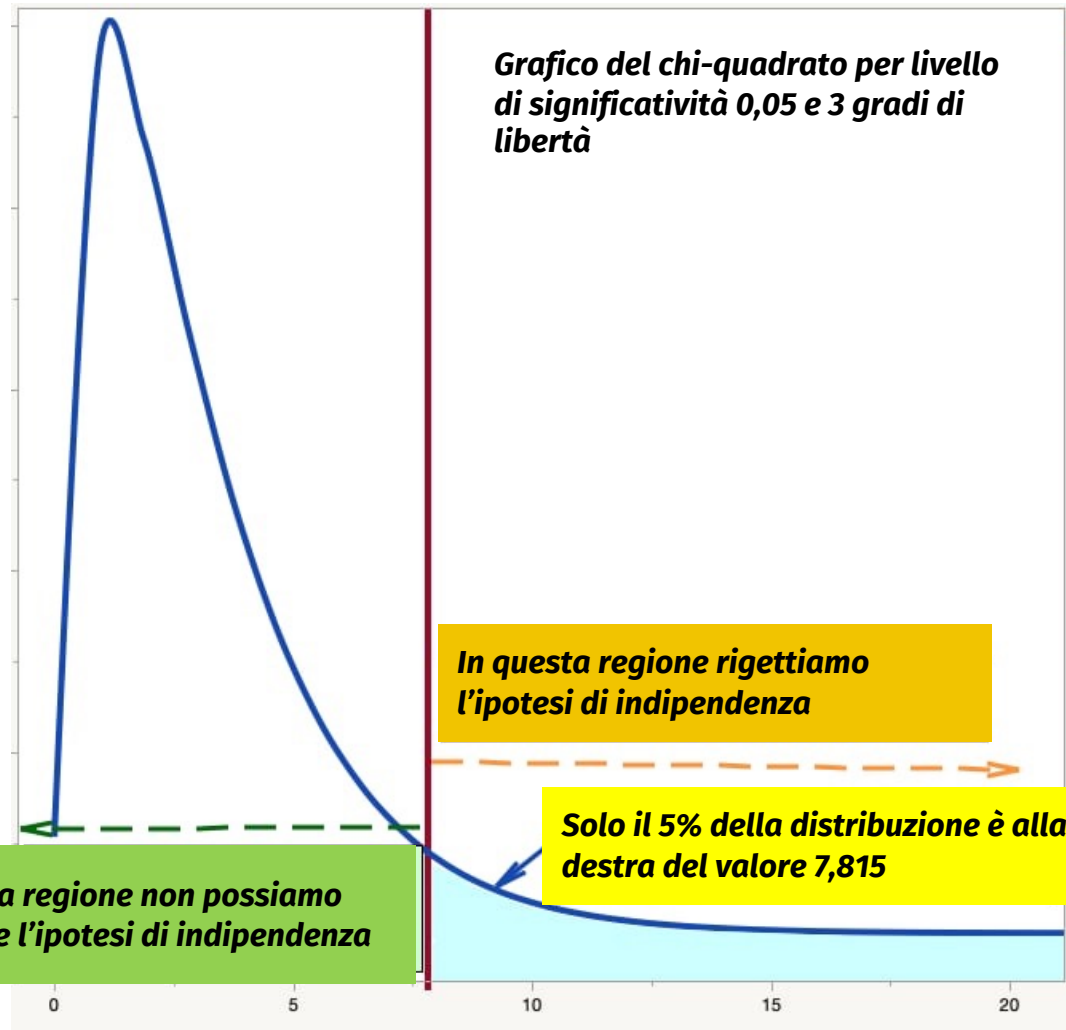
	Snack	No snack
Azione	64,58	60,42
Commedia	155	145
Family	62	58
Horror	28,42	26,58

Questi sono i valori che avremmo se non ci fosse dipendenza tra le due variabili categoriche

Applichiamo di nuovo il chi-quadrato

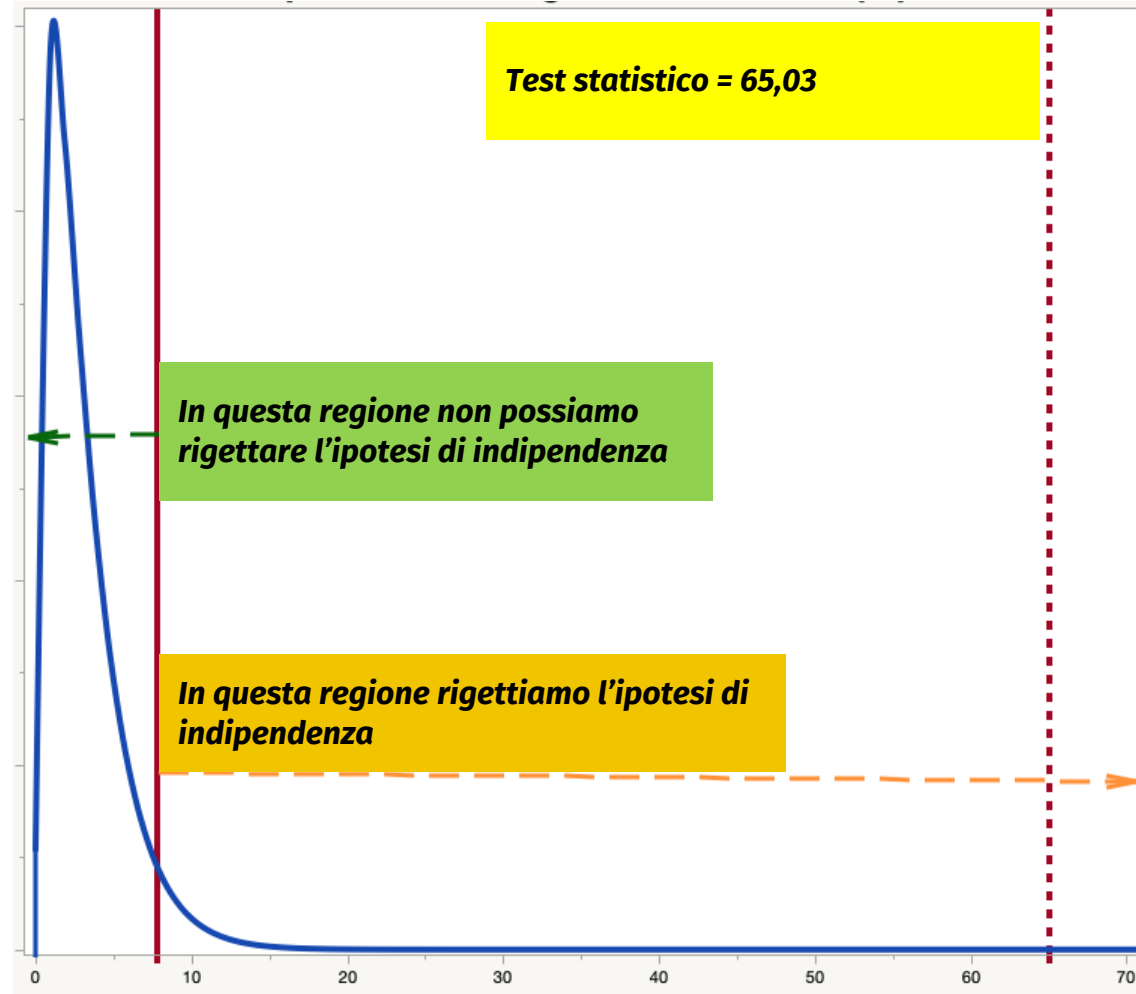
- Applicando la formula del test del chi-quadrato otteniamo 65,03
- Per capire se la statistica è significativa dobbiamo confrontare il valore preso dalla distribuzione del chi-quadrato per (livello di confidenza, gradi di libertà)
- Come al solito usiamo livello di significatività pari a 0,05 → Livello di confidenza 95% (0,95)
- I gradi di libertà in questo caso si calcolano come $(\text{righe}-1) \times (\text{colonne}-1)$, quindi abbiamo 3 gradi di libertà
- Controllando la tabelle della distribuzione dei chi-quadrato (ad es qui: https://it.wikipedia.org/wiki/Distribuzione_chi_quadrato) scopriamo che il valore del chi-quadrato nel nostro caso è 7,815
- Dato che $65,03 > 7,815$ possiamo rifiutare l'ipotesi secondo cui genere e film siano indipendenti

Cerchiamo di capire meglio



- Stiamo verificando se la statistica di test è un valore più estremo del valore critico nella distribuzione.
- Ma dove è la nostra statistica? Ricordiamo che valeva circa 65

Cerchiamo di capire meglio



UniGe

