# Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 19 20



## Piano della presentazione





# Associazione e indipendenza



## Associazione e indipendenza

- Il problema che ci poniamo è quello di studiare l'associazione tra variabili qualitative.
- Diciamo che due variabili qualitative sono associate se, nella popolazione, alcune modalità della prima variabile tendono a presentarsi più spesso in relazione ad alcune modalità della seconda variabile.

## Associazione e indipendenza

- Per esempio, negli Stati Uniti, le variabili qualitative "affiliazione religiosa" con modalità (protestante, cattolico, altro) e "gruppo etnico" con modalità (caucasico, afro-americano, latino, altro) sono associate nel senso che:
  - gli appartenenti al gruppo caucasico tendono ad essere protestanti, mentre
  - gli appartenenti al gruppo latino tendono ad essere cattolici.
- La nozione di indipendenza statistica tra variabili qualitative risulta più chiara se le osservazioni sono rappresentate mediante una tavola di contingenza.



## Distribuzioni marginali

L'esempio fornito da Agresti e Finlay per discutere la nozione di associazione tra variabili qualitative è quello che rappresenta l'*identificazione partitica* in funzione del *genere* per un campione di $n = 980$ individui (i dati provengono dal censimento del 1991):

| | identificazione partitica | | | |
|---|---|---|---|---|
| | Democratici | Indipendenti | Repubblicani | |
| Genere | | | | Tot |
| F | 279 | 73 | 225 | 577 |
| M | 165 | 47 | 191 | 403 |
| Tot | 444 | 120 | 416 | 980 |



## Distribuzioni marginali

I totali di riga e di colonna della tavola sono detti distribuzioni marginali.

La distribuzione marginale dell'identificazione partitica, per esempio, è l'insieme delle frequenze marginali $\{444, 120, 416\}$.



## Distribuzioni condizionate

Per studiare la dipendenza dell'identificazione partitica dal genere dobbiamo trasformare le frequenze assolute nelle frequenze di riga:

| | identificazione partitica | | | | |
|---|---|---|---|---|---|
| | Democratici | Indipendenti | Repubblicani | | |
| Genere | | | | Tot % | $n$ |
| F | 48.3 | 12.7 | 39.0 | 100 | 577 |
| M | 40.9 | 11.7 | 47.4 | 100 | 403 |

## Distribuzioni condizionate

Le due distribuzioni di frequenze (relative) all'interno delle classi delle femmine e dei maschi sono dette distribuzioni condizionate dalla variabile risposta, l'identificazione partitica.

- Per le femmine avremo: $\{0.483, 0.127, 0.39\}$;
- e per i maschi: $\{0.409, 0.117, 0.474\}$.



## Distribuzione congiunta

- Un altro modo per rappresentare i dati della tabella precedente è quello di calcolare le proporzioni di osservazioni in ciascuna cella della tavola di contingenza rispetto al totale ($n = 980$):

| | identificazione partitica | | |
|---|---|---|---|
| | Democratici | Indipendenti | Repubblicani |
| Genere | | | |
| F | 0.277 | 0.074 | 0.230 |
| M | 0.168 | 0.048 | 0.194 |

Tot = 980

- Questa distribuzione di frequenze relative è detta distribuzione congiunta



## Indipendenza e associazione

Una volta definite le nozioni di distribuzione marginale, condizionata e congiunta possiamo definire la nozione di indipendenza statistica per variabili qualitative.

Due variabili qualitative si dicono statisticamente indipendenti se, nella popolazione, le distribuzioni condizionate della prima variabile sono identiche per ciascuna modalità della seconda.

Le variabili si dicono associate se le distribuzioni condizionate non sono identiche.



## Indipendenza e associazione

Per esempio, se la distribuzione dell'identificazione partitica nella popolazione fosse diversa all'interno delle classi dei maschi e delle femmine, allora le variabili "genere" e "identificazione partitica" sarebbero associate o dipendenti.

Il problema che ci poniamo è: le differenze nelle distribuzioni condizionate osservate nel campione possono essere attribuite alla variabilità campionaria, oppure sono così grandi da rendere implausibile l'ipotesi di indipendenza nella popolazione?

# Differenza di due proporzioni



## Differenza di due proporzioni

- Un test statistico che consente di rispondere alla domanda precedente, nel caso più semplice di due variabili qualitative, ciascuna con due modalità, è quello sulla differenza di due proporzioni.
- Possiamo infatti pensare ai dati forniti da due proporzioni come ad una tavola di contingenza in cui la variabile esplicativa e la variabile di risposta hanno due modalità ciascuna.



## Differenza di due proporzioni

**Illustrazione.** Consideriamo uno studio relativo all'applicazione della pena di morte negli Stati Uniti nei casi in cui la vittima è di *tipi umano* (t.u.) "bianco".

Le unità sono 214 accusati di omicidio in 20 contee della Florida tra il 1976 e il 1977, classificati secondo il t.u. (modalità: bianco e afro-americano) e il verdetto di condanna a morte (modalità: no e sì).

| | Pena di morte | | |
|---|---|---|---|
| Accusati | Si | No | Totale |
| Bianchi | 19 | 132 | 151 |
| Afro-Americani | 11 | 52 | 63 |
| Totale | 30 | 184 | 214 |



## Illustrazione

- Dato che il numero di accusati bianchi è $151/63 = 2.4$ volte maggiore del numero di accusati afro-americani è difficile fare dei confronti diretti.
- Per stabilire se l'applicazione della pena di morte varia in funzione della razza dell'accusato, esprimiamo in termini percentuali le modalità della variabile risposta (pena di morte), *all'interno* di ciascuna modalità della variabile esplicativa (t.u.).

| | Pena di morte | | |
|---|---|---|---|
| Accusati | Si | No | Totale |
| Bianchi | 12.6 | 87.4 | 100 |
| Afro-Americani | 17.5 | 82.5 | 100 |

## Ipotesi nulla e alternativa

Vogliamo verificare l'ipotesi nulla

$$H_0 : \pi_1 = \pi_2 \quad o \quad H_0 : \pi_1 - \pi_2 = 0$$

la probabilità di un verdetto di condanna a morte è la stessa per gli accusati bianchi e gli accusati afro-americani,

verso l'ipotesi alternativa

$$H_a : \pi_1 < \pi_2 \quad o \quad H_a : \pi_1 - \pi_2 < 0$$

gli accusati bianchi hanno una probabilità minore di ottenere un verdetto di condanna a morte degli accusati afro-americani.



# Inferenze per le tavole di contingenza



## Inferenze per le tavole di contingenza

- La proporzione complessiva di accusati a cui è stata inflitta la pena di morte nel campione è

$$\hat{\pi} = \frac{19 + 11}{151 + 63} = \frac{30}{214} = 0.1402$$

- Il test $z$ per la differenza tra proporzioni è

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}\,(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.1258278 - 0.1746032}{\sqrt{0.1402(1 - 0.1402)\left(\frac{1}{151} + \frac{1}{63}\right)}} = -0.9367$$



## Inferenze per le tavole di contingenza

- Il $p-$valore della statistica $z = -0.9367$ è uguale a

```
pnorm(-0.9367)
## [1] 0.1744564
```

- Quindi questo particolare campione non fornisce evidenze sufficienti per rifiutare l'ipotesi nulla.
- Si noti inoltre che $\chi^2_1 = z^2 = (-0.9367)^2 = 0.8774069$, il cui $p-$valore (diviso due) coincide con il valore precedentemente ottenuto:

```
pchisq(0.8774069,df=1,lower.tail=FALSE) / 2
## [1] 0.1744564
```

## Test $\chi^2$ per due proporzioni

- Supponiamo però di volere confrontare più di due campioni, o che la variabile di risposta abbia più di due modalità. In tali circostanze, un semplice test basato sulla differenza tra due proporzioni non sarà sufficiente.
- Esaminiamo ora un altro metodo eseguire un test sulla differenza tra due proporzioni. Tale metodo può essere esteso al caso più generale in cui la variabile esplicativa e la variabile di risposta sono composte da un numero qualsiasi di modalità.



## Test $\chi^2$ per due proporzioni

- Se le variabili "tipo umano dell'accusato" e "verdetto di condanna a morte" fossero indipendenti, allora la pena di morte verrebbe inflitta nella stessa proporzione di casi agli accusati bianchi e agli accusati afro-americani.
- Questa proporzione può essere stimata nel modo seguente:

$$\hat{\pi} = \frac{19 + 11}{151 + 63} = \frac{30}{214} = 0.1402$$



## Test $\chi^2$ per due proporzioni

- Nel caso presente, ci sono $n = 151$ accusati bianchi e la proporzione di "successi" è $\hat{\pi} = 0.1402 \approx 0.14$. Sotto l'ipotesi di indipendenza ($\pi_1 = \pi_2$), dunque, ci aspettiamo un verdetto di condanna a morte per il seguente numero di accusati bianchi nel campione:

$$151 \times 0.14 = \frac{151 \times 30}{214} = 21.2$$

- Tale frequenza attesa si calcola allo stesso modo della media di una variabile bernoulliana $X$ con parametri $n = 151$ e probabilità di successo $p = 0.14$, ovvero $E(X) = np$.



## Test $\chi^2$ per due proporzioni

In maniera corrispondente, ci aspettiamo che il numero seguente di accusati bianchi nel campione non ottenga un verdetto di condanna a morte:

$$151 \times 0.86 = \frac{151 \times 184}{214} = 129.8$$

## Test $\chi^2$ per due proporzioni

- Sotto l'ipotesi di indipendenza, ci aspettiamo un verdetto di condanna a morte per il seguente numero di accusati afro-americani nel campione:

$$63 \times 0.14 = \frac{63 \times 30}{214} = 8.8$$

- La frequenza attesa di accusati afro-americani nel campione a cui non dovrebbe essere inflitta la pena di morte è:

$$63 \times 0.86 = \frac{63 \times 184}{214} = 54.2$$



## Test $\chi^2$ per due proporzioni

Queste frequenze attese corrispondono a ciò che ci aspettiamo di osservare *in media* se venissero estratti molti campioni composti da $151$ accusati bianchi e $63$ accusati afro-americani da una popolazione con probabilità di un verdetto di condanna a morte pari a $\pi = 0.14$ per entrambi.

| | Pena di morte | | |
|---|---|---|---|
| Accusati | Si | No | Totale |
| Bianchi | 19 (21.2) | 132 (129.8) | 151 |
| Afro-Americani | 11 (8.8) | 52 (54.2) | 63 |
| Totale | 30 | 184 | 214 |



## Test $\chi^2$ per due proporzioni

Se la razza dell'accusato e il verdetto di condanna a morte sono indipendenti, allora le frequenze attese in ciascuna cella della tavola di contingenza si possono calcolare con la seguente formula:

$$\frac{\text{totale riga} \times \text{totale colonna}}{\text{totale generale}}$$

dove il totale generale è la dimensione $n$ del campione (nel nostro esempio, $n = 214$).



## Test $\chi^2$ per due proporzioni

Per esempio, la frequenza attesa nella prima cella della tavola di contingenza è

$$\frac{151 \times 30}{214} = 21.2$$

## Test $\chi^2$ per due proporzioni

La formula precedente può anche essere concepita come l'applicazione della regola del prodotto per eventi indipendenti.

- La proporzione complessiva di accusati a cui viene inflitta la pena di morte è

$$\hat{\pi}_m = \frac{30}{214} = 0.140$$

- La proporzione di accusati bianchi è

$$\hat{\pi}_b = \frac{151}{214} = 0.706$$



## Test $\chi^2$ per due proporzioni

- Se i due eventi "pena di morte" e "razza dell'accusato" sono indipendenti, allora la proporzione di "accusati bianchi a cui viene inflitta la pena di morte" è

$$P(\text{pena di morte} \cap \text{razza bianca}) = \hat{\pi}_{mb} = \hat{\pi}_m \hat{\pi}_b =$$

$$= 0.140 \times 0.706 = 0.0988$$

- da cui deriva la frequenza attesa di accusati bianchi nel campione a cui viene inflitta la pena di morte:

$$n\hat{\pi}_{mb} = 214 \times 0.0988 = 21.2$$



## Terminologia

Le frequenze attese sotto l'ipotesi di indipendenza dovrebbero essere chiamate frequenze attese stimate in quanto sono basate sulle distribuzioni marginali delle due variabili in un campione (e non sulle frequenze marginali della popolazione).



## Statistica $\chi^2$

La statistica $\chi^2$ confronta le frequenze attese con le frequenze osservate in tutte le celle della tavola di contingenza:

$$\chi^2 = \sum \frac{(\text{frequenze osservate} - \text{frequenze attese})^2}{\text{frequenze attese}}$$

## Statistica $\chi^2$

**Illustrazione. Per l'esempio presente avremo (considerando nei calcoli precedenti almeno tre decimali):**

| f. osservate | f. attese | diff. | $\frac{(\text{oss.} - \text{att.})^2}{\text{att.}}$ |
|---|---|---|---|
| 19 | 21.2 | -2.168 | 0.222 |
| 132 | 129.8 | 2.168 | 0.036 |
| 11 | 8.8 | 2.168 | 0.532 |
| 52 | 54.2 | -2.168 | 0.087 |
| 214 | 214.0 | 0.000 | $\chi^2 = 0.877$ |



## Illustrazione

- Il $p-$valore della statistica $\chi^2 = 0.877$ è $P = 0.349$.
- Il test $\chi^2$ per una tavola di contingenza $2 \times 2$ è soltanto un modo (poco) più complicato di calcolare il test $z$ relativo alla differenza tra due proporzioni. In questo caso, $\chi^2 = z^2$.
- Per l'esempio presente, infatti $(-0.9367)^2 = 0.887$.
- La sola differenza tra i due test è che il test $z$ può essere usato per sottoporre a verifica un'ipotesi nulla unilaterale mentre il test $\chi^2$ è bilaterale. Per questa ragione, nel caso presente il $p-$valore del test $\chi^2$ è pari al doppio (dato l'errore di approssimazione) del $p-$valore del test $z$ unilaterale.



## Il test $\chi^2$ con R

Il test $\chi 2$ può essere eseguito con la funzione chisq.test di R:

```
deathPenalty <-
matrix(c(19, 132, 11, 52),
byrow=T, ncol=2,
dimnames = list("Razza" = c("Bianca", "Nera"),
"Pena di morte" = c("Si", "No")))

deathPenalty
```

```
       Pena di morte
Razza    Si  No
  Bianca 19 132
  Nera   11  52
```



## Il test $\chi^2$ con R

```
chisq.test(deathPenalty, correct=FALSE)
```

```
Pearson's Chi-squared test

data:  deathPenalty
X-squared = 0.8774, df = 1, p-value = 0.3489
```

## Il test $\chi^2$ in R

Come in precedenza, l'argomento `correct=FALSE` fa in modo che non venga applicata la correzione per la continuità.

```
  chisq.test(deathPenalty, correct=TRUE)

Pearson's Chi-squared test with Yates' continuity correction

data:  deathPenalty
X-squared = 0.5194, df = 1, p-value = 0.4711
```



# Inferenze per le tavole di contingenza



## Tavole di contingenza $r \times c$

- I dati della tavola seguente provengono da un'indagine condotta nel 1989 sulla popolazione canadese e sono riportati da Sniderman, Northrup, Fletcher, Russell e Tetlock (1993).
- L'indagine riguardava gli atteggiamenti dei canadesi nei confronti dei diritti civili.
- Tra gli altri temi trattati vi era l'antisemitismo e, a questo proposito, il questionario conteneva la seguente domanda.



## Tavole di contingenza $r \times c$

*We realize, NO STATEMENT is true of all people in a group, but GENERALLY SPEAKING, please tell me whether you agree or disagree with the following statements. The first one is: Most Jews don't care what happens to people who aren't Jewish. Would you say that you agree strongly, agree somewhat, disagree somewhat, or disagree strongly with that statement?*

## Tavole di contingenza $r \times c$

I dati possono essere presentati in una tavola di contingenza $4 \times 2$ avente come righe le modalità della risposta e come colonne le modalità della variabile esplicativa (gruppo linguistico).

| | Lingua | | |
|---|---|---|---|
| *Jews don't care* | Inglese | Francese | Totale |
| *Agree* | 261 | 171 | 432 |
| *Disagree Somewhat* | 576 | 159 | 735 |
| *Disagree Strongly* | 561 | 73 | 634 |
| *Don't Know or Refused* | 134 | 117 | 251 |
| Totale | 1532 | 520 | 2052 |



## Tavole di contingenza $r \times c$

Per interpretare più facilmente la tavola di contingenza calcoliamo le percentuali di colonna – dato che la variabile esplicativa è riportata in colonna.

| | Lingua | |
|---|---|---|
| *Jews don't care* | Inglese | Francese |
| *Agree* | 17.1 | 33.0 |
| *Disagree Somewhat* | 37.6 | 30.5 |
| *Disagree Strongly* | 36.6 | 14.1 |
| *Don't Know or Refused* | 8.7 | 22.5 |
| Totale | 100.0 | 100.1 |
| Numero di casi | 1532 | 520 |



## Tavole di contingenza $r \times c$

- Non è possibile riassumere la distribuzione di risposte per gli intervistati di lingua inglese e francese con una singola percentuale dato che la variabile di risposta (*Jews don't care*) ha più di due modalità.
- Di conseguenza, il test $z$ per la differenza di due proporzioni non può essere applicato.
- Possiamo però calcolare le frequenze attese assumendo che vi sia indipendenza tra la variabile "gruppo linguistico" e la variabile "antisemitismo", e poi usare il test $\chi^2$ per confrontare le frequenze attese e le frequenze osservate.



## Tavole di contingenza $r \times c$

La frequenza attesa per gli intervistati di lingua inglese che concordano con l'affermazione *Jews don't care*, per esempio, è

$$\text{freq. attesa} = \frac{\text{totale riga} \times \text{totale colonna}}{\text{totale generale}}$$

$$= \frac{432 \times 1532}{2052} = 322.5$$

## Tavole di contingenza $r \times c$

Le frequenze attese in ciascuna cella della tavola di contingenza sono le seguenti:

| | Lingua | | |
|---|---|---|---|
| *Jews don't care* | Inglese | Francese | Totale |
| *Agree* | 322.5 | 109.5 | 432 |
| *Disagree Somewhat* | 548.7 | 186.3 | 735 |
| *Disagree Strongly* | 473.3 | 160.7 | 634 |
| *Don't Know or Refused* | 187.4 | 63.6 | 251 |
| Totale | 1531.9 | 520.1 | 2052 |

Si noti che, tenendo conto dell'errore di arrotondamento, la somma delle frequenze attese è uguale ai totali di riga e di colonna.



## Tavole di contingenza $r \times c$

- Il test $\chi^2$ viene calcolato nel modo seguente:

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{frequenze osservate} - \text{frequenze attese})^2}{\text{frequenze attese}} \\ &= \frac{(261 - 322.5)^2}{322.5} + \frac{(171 - 109.5)^2}{109.5} + \ldots + \frac{(117 - 63.6)^2}{63.6} \\ &= 175.8\end{aligned}$$

- Prima di trovare il $p-$valore della statistica test dobbiamo esaminare le proprietà delle distribuzioni $\chi^2$.



## Distribuzioni $\chi^2$

Come nel caso delle distribuzioni $t$ di Student, anche le distribuzioni $\chi^2$ formano una famiglia di funzioni di densità parametrizzate dai gradi di libertà.

- C'è una diversa distribuzione $\chi^2$ a seconda dei numeri di grado di libertà $\nu = 1, 2, \ldots$.
- Diversamente dalle distribuzioni $t$ di Student e dalla distribuzione normale, la distribuzione $\chi^2$ possiede un'asimmetria positiva.
- L'asimmetria della distribuzione diminuisce al crescere dei gradi di libertà.



## Distribuzioni $\chi^2$

Le proprietà matematiche della distribuzione consentono di risolvere molti più problemi di quelli che illustreremo in questo corso, dove ci limiteremo ad esporre solo i seguenti argomenti:

(a) confronto tra frequenze osservate e frequenze teoriche per lo studio dell'indipendenza di una tabella di contingenza;

(b) calcolo dell'intervallo di confidenza per la varianza.

## Distribuzioni $\chi^2$

Si chiama $\chi^2$ la sommatoria dei quadrati di $n$ variabili (casuali e indipendenti) normali standardizzate, che è espressa dalla seguente equazione:

$$\chi^2_n = \sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2$$

- Dato che la variabile casuale $\chi^2$ è generata dalla somma dei quadrati di $n$ *valori indipendenti* di una variabile normale standardizzata, i gradi di libertà coincideranno con: $\nu = n$.
- Quando tali valori *non sono indipendenti*, è necessario stabilire le condizioni che li vincolano. Sottraendo tali vincoli si ottiene il numero di gradi di libertà.
- I gradi di libertà sono dunque il parametro che caratterizza ogni distribuzione di $\chi^2$ e vengono molto spesso indicati con la lettera greca $\nu$.



## Approfondimento: i momenti caratteristici di una distribuzione

**1.** Si definisce momento di ordine $k$ la quantità:

$$E\left(x_i - \mu\right)^k.$$

Il momento di ordine $k = 0$ è 1, il momento di ordine $k = 1$ è $E\left(x_i - \mu\right) = 0$, infine il momento di ordine $k = 2$ è la varianza $\sigma^2 = E\left(x_i - \mu\right)^2$.

**2.** Per la *distribuzione normale*, possiamo esprimere i momenti di ordine $k > 1$ in funzione di $\sigma^2$ attraverso la seguente formula generale:

$$(|k-1|)\sigma^2 E\left(x - \mu\right)^{|k-2|}$$

da cui facilmente ricaviamo che:

$$\begin{array}{lll} k=1; & (|1-1|)\sigma^2 E\left(x-\mu\right)^{|1-2|} & = 0 \times \sigma^2 \times 0 \quad = 0 \\ k=2; & (|2-1|)\sigma^2 E\left(x-\mu\right)^{|2-2|} & = 1 \times \sigma^2 \times 1 \quad = \sigma^2 \\ k=3; & (|3-1|)\sigma^2 E\left(x-\mu\right)^{|3-2|} & = 2\sigma^2 \times 0 \quad = 0 \\ k=4; & (|4-1|)\sigma^2 E\left(x-\mu\right)^{|4-2|} & = 3\sigma^2\sigma^2 \quad = 3(\sigma^2)^2 \end{array}$$



## Approfondimento: i momenti caratteristici di una distribuzione

**3.** Si definisce l'indice di asimmetria come il rapporto tra il momento di ordine $k = 3$ e il cubo della deviazione standard:

$$\text{asimmetria} = \frac{E\left(x-\mu\right)^3}{\sigma^3} = E\left[\frac{(x-\mu)}{\sigma}\right]^3,$$

Detto anche *momento standardizzato di ordine terzo*, assume valore 0 nella distribuzione normale standard (vedi punto 2; per $k = 3$)

**4.** L'indice di curtosi, o *momento standardizzato di ordine quarto*:

$$\text{kurtosi} = \frac{E\left(x-\mu\right)^4}{(\sigma^2)^2} = E\left[\frac{(x-\mu)}{\sigma}\right]^4,$$

assume valore 3 nella distribuzione normale e quantifica l'appiattimento di una distribuzione. Le distribuzioni piatte con code ampie sono chiamate *platicurtiche* (es. $t - Student$), quelle appuntite con code piccole sono chiamate *leptocurtiche*. Una distribuzione con la stessa kurtosi della distribuzione normale è chiamata *mesocurtica*. (vedi punto 2; per $k = 4$)



## Premesse

Si dimostrano le seguenti equivalenze asintotiche (per $n \to \infty$):

5 Il valore atteso di un valore $z^2 \sim \chi^2_{\nu=1}$ è uguale al valore 1:

$$E\left[\chi^2_1\right] = E\left[\frac{\left(x_i - \mu_i\right)^2}{\sigma_i^2}\right] = \left[\frac{1}{\sigma_i^2} E\left(x_i - \mu_i\right)^2\right] = \left[\frac{1}{\sigma_i^2}\ \sigma_i^2\right] = 1$$

6 Il valore atteso di un valore $z^4 \sim (\chi^2_{\nu=1})^2$ è uguale al valore 3:

$$E\left[\left(\chi^2_1\right)^2\right] = E\left[\frac{\left(x_i - \mu_i\right)^4}{(\sigma_i^2)^2}\right] = \left[\frac{E\left(x_i - \mu_i\right)^4}{(\sigma_i^2)^2}\right] = \text{indice di curtosi} = 3$$

7 La somma di $n$ valori $\chi^2_{\nu=1}$ è un $\chi^2_{\nu=n}$:

$$\sum_{i=1}^{n} \left(\chi^2_1\right)_i = \sum_{i=1}^{n} \left(\frac{\left(x_i - \mu_i\right)^2}{\sigma_i^2}\right) = \sum_{i=1}^{n} z_i^2 = \chi^2_n$$

## Distribuzioni $\chi^2$ - valore atteso

Si dimostra che:

*Il valore atteso di una variabile $\chi^2_n$ è uguale al numero dei gradi di libertà della variabile stessa:*

$$E\left(\chi^2_n\right) = n = \nu$$

Applicazione dei punti 5. e 7.

$$E\left[\chi^2_n\right] = E\left[\sum_{i=1}^{n}\left(\chi^2_1\right)_i\right] = \sum_{i=1}^{n}\left[E\left(\chi^2_1\right)_i\right] = n[1] = n$$



## Distribuzioni $\chi^2$ - varianza

*la varianza è pari a due volte i gradi di libertà:*

$$Var\left(\chi^2_n\right) = 2n = 2\nu$$

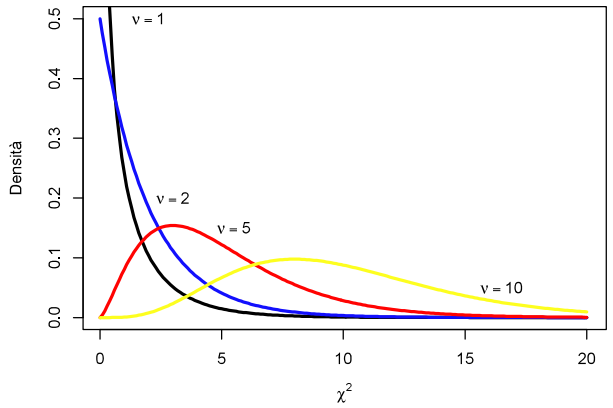
Applicazione dei punti 5. 6. e 7.

$$\begin{aligned} Var\left[\chi^2_n\right] = & \quad Var\left[\sum_{i=1}^{n}\left(\chi^2_1\right)_i\right] = \sum_{i=1}^{n}\left[Var\left(\chi^2_1\right)_i\right] = \\ = & \quad \sum_{i=1}^{n}\left[E\left[\left(\chi^2_1\right)^2\right] - E\left[\chi^2_1\right]^2\right] = \\ = & \quad n\left[3 - 1^2\right] = 2n \end{aligned}$$



## Distribuzioni $\chi^2$

- La distribuzione $\chi^2$ è di tipo continuo e non può mai essere negativa; perciò si trova sempre compresa nel primo quadrante degli assi cartesiani ed ha forme diverse a seconda del valore $\nu$.






## Distribuzioni $\chi^2$

- La variabile $\chi^2$ è definita nell'intervallo $[0; \infty]$.
- La statistica $\chi^2$ dipende dal quadrato delle differenze tra frequenze osservate e attese e, quindi, non può assumere valori negativi.
- Se le frequenze osservate e le frequenze attese sono uguali, la statistica $\chi^2$ ha valore 0.
- Tanto maggiore è la differenza tra frequenze osservate e attese, tanto più grande sarà il valore della statistica $\chi^2$.

## Test $\chi^2$

- Il test $\chi^2$ è unilaterale:
- l'ipotesi nulla di indipendenza tra le due variabili rappresentate in una tavola di contingenza viene rifiutata quando la statistica $\chi^2$ è sufficientemente grande.
- L'ipotesi alternativa, invece, è non-direzionale: *una qualunque forma di associazione* tra le due variabili.



## Gradi di libertà

- I gradi di libertà per il test $\chi^2$ in tavole di contingenza $r \times c$ sono

$$\text{gdl} = (r-1)(c-1)$$

  dove $r$ è il numero di righe e $c$ è il numero di colonne.
- Nel primo esempio, $r = 2$, $c = 2$, $\text{gdl} = (2-1)(2-1) = 1$.
- Nel secondo esempio, $r = 4$, $c = 2$, $\text{gdl} = (4-1)(2-1) = 3$.



## Gradi di libertà

- I gradi di libertà per il test $\chi^2$ in una tavola di contingenza obbediscono al seguente vincolo:

  *i totali marginali attesi devono essere uguali ai totali marginali osservati.*

- Quindi, se calcoliamo le frequenze attese per tutte le celle della tavola in un riga o colonna *tranne l'ultima*, il valore rimanente può essere calcolato per sottrazione dai valori già considerati.



## Gradi di libertà

**Illustrazione.** Nel caso dell'esempio presente, all'interno di ciascuna colonna è sufficiente specificare soltanto 3 frequenze attese (l'ultima frequenza attesa è determinata dalle altre, dato che il totale marginale è fisso).

| | Lingua | | |
|---|---|---|---|
| *Jews don't care* | Inglese | Francese | Totale |
| *Agree* | 322.5 | ? | 432 |
| *Disagree Somewhat* | 548.7 | ? | 735 |
| *Disagree Strongly* | 473.3 | ? | 634 |
| *Don't Know or Refused* | ? | ? | 251 |
| Totale | 1531.9 | 520.1 | 2052 |

## Test $\chi^2$

- Il $p-$valore della statistica $\chi^2$ può essere trovato con R oppure consultando una tavola sinottica che riporta i valori critici della distribuzione $\chi^2$.
- La tavola C nel testo di Agresti e Finlay è costruita come la tavola della distribuzione $t$ di Student e riporta i gradi di libertà ($df$) sulle righe e la probabilità nella coda destra della distribuzione $\chi^2$ sulle colonne.

Primo esempio $\chi^2_1 = 0.877 \rightarrow P > 0.25$.

Secondo esempio $\chi^2_3 = 175.8 \rightarrow P < 0.001$.

```
 1 - pchisq(0.877, 1)
## [1] 0.3490247
 1 - pchisq(175.8, 3)
## [1] 0
```



## Test $\chi^2$ in R

Il test $\chi^2$ per l'esempio presente può essere eseguito con R nel modo seguente.

```
antiSemitism <-
matrix(c(261, 171, 576, 159, 561, 73, 134, 117),
byrow=T, ncol=2,
dimnames = list("Jews Don't Care" =
c("Agree", "Disagree Somewhat", "Disagree Strongly",
"Don't Know or Refused"),
"Language" = c("English", "French")))

antiSemitism
```

```
                        Language
Jews Don't Care       English French
  Agree                   261    171
  Disagree Somewhat       576    159
  Disagree Strongly       561     73
  Don't Know or Refused   134    117
```



## Test $\chi^2$ in R

```
 chisq.test(antiSemitism, correct=FALSE)

Pearson's Chi-squared test

data:  antiSemitism
X-squared = 175.7601, df = 3, p-value < 2.2e-16
```



# Condizioni di validità del test $\chi^2$

## Condizioni di validità del test $\chi^2$

Il test $\chi^2$ può essere usato per verificare l'esistenza di un'associazione in una tavola di contingenza.

- È necessario avere dei c.c.i. di osservazioni per ciascuna modalità della variabile esplicativa.
- Il $p-$valore della statistica $\chi^2$ non è accurato se il campione rappresenta una frazione considerevole della popolazione ($> 10\%$).
- Il test $\chi^2$ può essere utilizzato anche quando i soggetti vengono assegnati in maniera casuale a diversi trattamenti sperimentali e la variabile di risposta è categoriale.



## Condizioni di validità del test $\chi^2$

- La distribuzione $\chi^2$ fornisce soltanto un'approssimazione alla distribuzione esatta della statistica $\chi^2$.
- L'approssimazione è adeguata se
  - il campione è abbastanza grande ($n > 100$ oppure più restrittivamente $n > 200$), e
  - entro ogni casella solo poche frequenze attese (non oltre il $20\%$) sono inferiori a 5 e nessuna frequenza attesa è minore di 1.



## Residui di Pearson

- Il test $\chi^2$, così come gli altri test di significatività, fornisce poche informazioni. Se il $p-$valore del test è molto piccolo, i dati consentono di rifiutare l'ipotesi nulla di indipendenza tra i due criteri di classificazione. Questo significa che le variabili sono associate.
- Il test $\chi^2$, però, non dice nulla a proposito della natura di tale associazione.
- L'analisi dei residui dovrebbe essere fatta seguire al rifiuto dell'ipotesi nulla in quanto il confronto tra le frequenze osservate e le frequenze attese può contribuire a rivelare la fonte dell'associazione tra le variabili.



## Residui di Pearson

La differenza

$$\text{frequenza osservata} - \text{frequenza attesa}$$

è chiamata residuo.

- Gli scarti tra le frequenze osservate e le frequenze attese non risentono solamente delle differenze realmente esistenti tra le due distribuzioni a confronto, ma anche delle variazioni casuali.
- Quanto grandi devono essere i residui affinché si possa escludere che siano dovuti agli errori di campionamento?

## Residui di Pearson

Pearson ha dimostrato che, sotto $H_0$, il rapporto

$$z = \frac{\text{frequenza osservata} - \text{frequenza attesa}}{\sqrt{\text{frequenza attesa}}}$$

segue la distribuzione normale standardizzata, ovvero $z \sim N(0, 1)$.

Un grande residuo standardizzato (diciamo $> 3$) può quindi essere preso come evidenza che, nella corrispondente cella della tavola di contingenza, la frequenza osservata si discosta in maniera significativa dalla frequenze attesa in base all'ipotesi di indipendenza.



## Residui di Pearson e R

- Utilizzando R, le frequenze attese si ottengono nel modo seguente:

```
out <- chisq.test(antiSemitism, correct=FALSE)
out$expected
```

```
                    Language
Jews Don't Care      English    French
 Agree               322.5263 109.47368
 Disagree Somewhat   548.7427 186.25731
 Disagree Strongly   473.3372 160.66277
 Don't Know or Refused 187.3938  63.60624
```

- Calcoliamo il residuo standardizzato per la prima cella della tavola di contingenza:

```
(261-322.5263)/sqrt(322.5263)
## [1] -3.425928
```



## Residui di Pearson e R

- I residui standardizzati in tutte le celle sono:

```
out$residuals
```

```
                    Language
Jews Don't Care      English    French
 Agree              -3.425929  5.880389
 Disagree Somewhat   1.163586 -1.997222
 Disagree Strongly   4.029302 -6.916041
 Don't Know or Refused -3.900434  6.694847
```



## Residui di Pearson e R

- Nel gruppo di lingua inglese, troviamo una frequenza significativamente minore di quella attesa per la categoria "Agree" mentre, per la stessa categoria, il gruppo francese esibisce una frequenza significativamente maggiore a quella attesa.
- Nella cella "Disagree Strongly" troviamo una frequenza significativamente maggiore di quella attesa per il gruppo inglese e significativamente minore di quella attesa per il gruppo francese.

## Residui di Pearson e R

- Nella cella "Don't Know or Refused", infine, troviamo una frequenza minore di quella attesa in base all'ipotesi di indipendenza per il gruppo inglese e maggiore di quella attesa per il gruppo francese.
- In conclusione, l'esame dei residui indica che l'associazione tra antisemitismo e gruppo linguistico è dovuta al fatto che, in questo campione, atteggiamenti anti-semiti sono presenti in misura maggiore all'interno della comunità francese.



## Residui di Pearson e R

- I residui nella categoria "Don't Know or Refused" per i due gruppi potrebbero significare, però, che i membri della comunità francese hanno opinioni meno forti su tale questione degli intervistati di lingua inglese.



# Misure di associazione



## Misure di associazione

- Una misura di associazione è una statistica che rivela la forza della dipendenza statistica tra due variabili.
- Nella tabella, l'opinione nei confronti della legalizzazione dell'aborto (favorevole verso contraria) è indipendente dal *tipo umano* (t.u.) degli intervistati (bianco verso nero).

| | Opinione | | |
|---|---|---|---|
| | Favorevole | Contraria | |
| T.u. | | | Totale |
| Bianco | 360 | 240 | 600 |
| Nero | 240 | 160 | 400 |
| Totale | 600 | 400 | 1000 |

## Misure di associazione

- Nella tavola precedente, le opinioni a favore sono il $60\%$ del totale, in entrambi i gruppi. La variabile opinione non è associata alla variabile *t.u.*.
- Nella tavola seguente, invece, l'associazione è perfetta: tutti i bianchi esprimono un opinione favorevole, tutti i neri un'opinione contraria. Per questi dati, la variabile *opinione* è completamente determinata dalla variabile *t.u.*

| | Opinione | | |
|---|---|---|---|
| | Favorevole | Contraria | |
| T.u. | | | Totale |
| Bianco | 600 | 0 | 600 |
| Nero | 0 | 400 | 400 |
| Totale | 600 | 400 | 1000 |



## Misure di associazione

- Le misure di associazione tra variabili categoriali rivelano quanto una tavola di contingenza sia simile a uno o l'altro degli estremi che rappresentano l'assenza di associazione o la perfetta associazione.
- Il test $\chi^2$ verifica l'ipotesi nulla di indipendenza. Se l'ipotesi di indipendenza viene rifiutata, però, il test $\chi^2$ non consente di valutare l'intensità dell'associazione tra due variabili.



# Odds



## Odds

- L'odds ratio (*ratio of odds*) è la più importante misura della forza di una associazione tra due variabili categoriali.
- Per comprendere questa misura, occorre introdurre il concetto di "odds" (*quote di scommessa*).
- Supponiamo di avere una variabile risposta binaria $B$ con modalità 0 (*insuccesso*) e 1 (*successo*).
- Oltre alle probabilità di successo spesso si considerano anche le quote di scommessa (gli odds).

## Odds

- Gli odds di successo $\omega$ sono per definizione il rapporto tra la probabilità di successo ($\pi$) e la probabilità di insuccesso ($1-\pi$):

$$\omega = \frac{\pi}{(1-\pi)}$$

- La quota di scommessa è un indice non negativo che misura quanti successi ci si attendono per ogni insuccesso.



## Odds

Si noti che

- $\omega = 1$: la probabilità di successo è uguale alla probabilità di insuccesso;
- $\omega < 1$: la probabilità di successo è minore della probabilità di insuccesso;
- $\omega > 1$: la probabilità di successo è maggiore della probabilità di insuccesso.



## Odds

- Se $\pi = 0.75$ la quota è $\omega = 0.75/0.25 = 3$ . Scommettendo su $B = 0$ (l'allibratore perde e il vostro "pollo" vince) si riceve 3 volte la posta.
- Se l'odds è minore di 1, possiamo calcolare il reciproco $(1/\omega)$ e interpretare il risultato come riferito all'evento complementare.

  Per esempio, se $\omega$ è $0.3333$ , ci aspettiamo $0.3333$ successi per ogni insuccesso ossia $1/0.3333 = 3$ *insuccessi per ogni successo*.



## Odds

La relazione inversa tra probabilità di successo e odds è

$$\pi = \frac{\omega}{1+\omega}$$

- Per esempio, se $\omega = 3$, la probabilità di successo sarà

$$\pi = \frac{3}{1+3} = 0.75$$

quindi, $\pi = 0.75$ come nell'esempio precedente,

## Odds

- Consideriamo ora il rapporto degli odds in una tavola di contingenza.
- In una tavola di contingenza, poniamo la variabile esplicativa sulle righe e la variabile di risposta sulle colonne.
- Per la ciascuna modalità della variabile esplicativa,
  - la probabilità di successo è uguale al rapporto tra il numero di successi e il totale di riga;
  - la probabilità di insuccesso è uguale al rapporto tra il numero di insuccessi e il totale di riga.



## Odds

Consideriamo nuovamente i dati relativi alla pena di morte.

| | Opinione | | |
|---|---|---|---|
| | Favorevole | Contraria | |
| T.u. | | | Totale |
| Bianco | 19 | 132 | 151 |
| Nero | 11 | 52 | 63 |
| Totale | 30 | 184 | 214 |



## Odds

- Per gli accusati bianchi, la probabilità di un verdetto di condanna a morte è $19/151 = 0.1258$ e la probabilità dell'evento complementare è $132/151 = 0.8742$.
- Per gli accusati afro-americani, la probabilità di un verdetto di condanna a morte è $11/63 = 0.1746$ e la probabilità dell'evento complementare è $52/63 = 0.8254$ .



## Odds

- Per gli accusati bianchi l'odds della pena di morte è

$$\frac{0.1258}{0.8742} = 0.1439$$

- Per gli accusati afro-americani, l'odds della pena di morte è

$$\frac{0.1746}{0.8254} = 0.2115$$

# Odds ratio



## Odds ratio

L'odds ratio (*ratio of odds*) è il rapporto tra gli odds di due eventi.

Per una tavola di contingenza, il rapporto delle quote (*odds-ratio*) detto anche *rapporto incrociato* (*cross-product ratio*) è il rapporto tra gli odds di due righe $i$ e $j$:

$$\theta = \frac{\omega_i}{\omega_j}$$



## Odds ratio

**Illustrazione.** Per l'esempio presente, l'odds ratio tra gli accusati bianchi e gli accusati afro-americani è

$$\theta = \frac{0.1439}{0.2115} = 0.6804$$

il suo reciproco, ovvero l'odds ratio tra accusati afro-americani e bianchi è

$$\theta = \frac{0.2115}{0.1439} = 1.469771$$



## Interpretazione

- Come si interpreta il rapporto degli odds?
- Nel caso presente diremo che, per gli accusati bianchi, l'odds della pena di morte è $0.6804$ volte più piccolo che per gli accusati afro-americani.
- In maniera equivalente, possiamo calcolare il reciproco di $\theta$ e dire che, per gli accusati afro-americani, l'odds della pena di morte è $1/0.6804 = 1.4698$ più grande che per gli accusati bianchi (circa il $46.98 \approx 50\%$ in più)

## Interpretazione

- Questo NON significa che la probabilità di un verdetto di condanna a morte sia $1.47$ volte maggiore per gli accusati afro-americani che per gli accusati bianchi.
- L'odds ratio è una rapporto tra odds, non un rapporto di probabilità.
- ponendo che negli accusati bianchi ci sia $1$ verdetto di condanna a morte ogni $1$ verdetto di non condanna a morte, $\omega_b = \frac{1}{1} = 1$, allora negli accusati afro-americani, per ogni non condanna a morte, ci sono $1.47$ condanne a morte, ossia $\omega_a = \frac{1.47}{1} = 1.47$.

$$\theta = \frac{1.47}{1} = 1.47$$



## Odds ratio

- Quando $\theta = 1$, le probabilità di successo sono uguali in entrambe le righe della tavola di contingenza. Le variabili riga e colonna sono dunque indipendenti
- Quando $\theta > 1$, le probabilità di successo sono maggiori nella prima riga della tavola: $\pi_1 > \pi_2$.
- Quando $\theta < 1$, le probabilità di successo sono minori nella prima riga della tavola: $\pi_1 < \pi_2$.
- Valori di $\theta$ diversi da 0, nelle due direzioni, rappresentano dunque la forza dell'associazione tra le variabili riga e colonna.



## Tavola di contingenza $2 \times 2$

- Nel caso di una tavola di contingenza $2 \times 2$, l'odds ratio si può calcolare direttamente dalle frequenze nella tavola, senza calcolare gli odds.
- L'odds ratio è il rapporto tra il prodotto delle frequenze sulla diagonale principale e il prodotto delle frequenze sulla diagonale opposta. Per questa ragione, l'odds ratio è anche detto rapporto incrociato.

| | |
|---|---|
| a | b |
| c | d |

$$\theta = \frac{ad}{bc}$$



## Tavola di contingenza $2 \times 2$

**Illustrazione.** Per l'esempio presente avremo

$$\theta = \frac{19 * 52}{132 * 11} = 0.6804$$

Dalle stesse quantità si ricava la formula abbreviata del $\chi^2$:

$$\begin{aligned} \chi^2 &= \frac{N\,(ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)} \\ &= \frac{214\,((19 \times 52) - (132 \times 11))^2}{(19+132)(19+11)(11+52)(132+52)} = 0.8774 \end{aligned}$$

## Tavola di contingenza $2 \times 2$

**Illustrazione.**
che può essere opportunamente *corretta per la continuità* (Yates):

$$\chi^2 = \frac{N\left(|ad - bc| - N/2\right)^2}{(a+b)(a+c)(c+d)(b+d)}$$

$$= \frac{N\left(|(19 \times 52) - (132 \times 11)| - 214/2\right)^2}{(19+132)(19+11)(11+52)(132+52)} = 0.5194$$

e che possiamo ottenere in R con il comando `chisq.test`



## Tavola di contingenza $2 \times 2$

```
  chisq.test(deathPenalty,correct=FALSE)

Pearson's Chi-squared test

data:  deathPenalty
X-squared = 0.8774, df = 1, p-value = 0.3489

  chisq.test(deathPenalty,correct=TRUE)

Pearson's Chi-squared test with Yates' continuity correction

data:  deathPenalty
X-squared = 0.5194, df = 1, p-value = 0.4711
```



# Rischio relativo



## Rischio relativo

Il rischio relativo è il rapporto tra la probabilità che un evento si verifichi dato uno stato del mondo e la probabilità che tale evento si verifichi dato un altro stato del mondo.

**Illustrazione.** Per l'esempio presente avremo

$$RR = \frac{P\left(\text{pena di morte}|\text{afro-americano}\right)}{P\left(\text{pena di morte}|\text{bianco}\right)} = \frac{0.1746}{0.1258} = 1.3879$$

che non è uguale all'odds ratio ($\theta = 1.4698$).

## Relazione tra $RR$ e $\theta$

- Anche se $\theta$ non è uguale al rischio relativo, questi due indici sono simili se la probabilità di successo $\pi \approx 0$. In tali circostanze, il rapporto degli odds si può interpretare come un rischio relativo.
- Che cosa ci dice $\theta$ del rischio relativo quando le probabilità $\pi_1$ e $\pi_2$ non sono prossime allo zero?
  - $\theta = 1$ implica che i criteri di classificazione $R$ e $C$ sono indipendenti;
  - $\theta > 1$ implica $RR > 1$;
  - $\theta < 1$ implica $RR < 1$;
  - $RR$ è sempre più simile a 1 di $\theta$



## Calcolo dell'odds ratio con R

- E' semplice calcolare l'odds ratio. Tuttavia, se vogliamo usare R procediamo nel modo seguente.

```
  library(vcd)  # scaricare dal CRAN
  oddsratio(deadPenalty, log=FALSE)
## [1] 0.6718622

 (19.5*52.5)/(132.5*11.5)
## [1] 0.6718622
```

- Si noti che ai dati è stata applicata la correzione per la continuità (0.5 è stato sommato a ciascuna cella della tavola).



# Conclusioni



## Conclusioni

- Il test $\chi^2$ viene usato per verificare l'ipotesi nulla di indipendenza tra una variabile esplicativa posta sulle righe e una variabile di risposta posta sulle colonne di una tavola di contingenza.
- Sotto l'ipotesi di indipendenza, le frequenze attese in ciascuna cella della tavola di contingenza $r \times c$ si calcolano come

$$\text{frequenza attesa} = \frac{\text{totale riga} \times \text{totale colonna}}{\text{totale generale}}$$

## Conclusioni

- La statistica test

$$\chi^2 = \sum \frac{(\text{frequenze osservate} - \text{frequenze attese})^2}{\text{frequenze attese}}$$

è approssimativamente distribuita come $\chi^2$ con $(r-1)(c-1)$ gradi di libertà.



## Conclusioni

Affinché il test $\chi^2$ sia accurato,

- entro ogni casella solo poche frequenze attese (non oltre il $20\%$) possono essere inferiori a 5 e nessuna frequenza attesa deve essere minore di 1;
- le dimensioni del campione devono essere grandi;
- le frequenze osservate in ciascuna cella della tavola di contingenza devono essere un c.c.i. estratto dalla popolazione.



## Conclusioni

- Una volta stabilito che due variabili categoriali sono tra loro associate, dobbiamo stabilire
  1- da che cosa dipende tale associazione, e
  2- quale è la forza dell'associazione.
- L'esame dei residui standardizzati consente di individuare le celle di una tavola di contingenza che si discostano in maniera significativa dalle frequenze attese in base all'ipotesi di indipendenza.



## Conclusioni

- L'*odds ratio* consente di confrontare gli *odds* di successo per le diverse modalità della variabile esplicativa.
- Se le probabilità di successo per le modalità della variabile esplicativa sono prossime allo zero (successo = evento raro), allora l'odds ratio si può interpretare come un rischio relativo.

# Applicazioni derivate del $\chi^2$



## Distribuzione campionaria di $s^2$

- Possiamo utilizzare la distribuzione $\chi^2$ per ricavare la distribuzione campionaria della statistica

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- utilizzeremo tale informazione per fare inferenze sulla varianza incognita della popolazione.
- In particolare, svilupperemo la stima per intervalli di confidenza del parametro $\sigma$, a partire dal valore campionario $s^2$.



## Distribuzione campionaria di $s^2$

- Concentriamoci sul numeratore

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- che possiamo esprimere come sottrazione di $\mu$:

$$\sum_{i=1}^{n}\left[(x_i - \mu) - (\bar{x} - \mu)\right]^2;$$

- e da cui sviluppando il quadrato, otteniamo:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2\right].$$

## Distribuzione campionaria di $s^2$

$$\begin{aligned}\sum_{i=1}^{n}\left[(x_i-\mu)-(\bar{x}-\mu)\right]^2 &= \sum_{i=1}^{n}(x_i-\mu)^2+n(\bar{x}-\mu)^2-2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)\\ &= \sum_{i=1}^{n}(x_i-\mu)^2+n(\bar{x}-\mu)^2-2(\bar{x}-\mu)(n\bar{x}-n\mu)\\ &= \sum_{i=1}^{n}(x_i-\mu)^2+n(\bar{x}-\mu)^2-2n(\bar{x}-\mu)^2\\ &= \sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu)^2\end{aligned}$$

Quindi

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i-\bar{x})\right]^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu)^2\right]$$



## Distribuzione campionaria di $s^2$

- Moltiplicando entrambi i termini dell'equazione per la costante $\frac{n-1}{\sigma^2}$ si ottiene

$$\begin{aligned}\left(\frac{n-1}{\sigma^2}\right)s^2 &= \left(\frac{n-1}{\sigma^2}\right)\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu)^2\right]\\ &= \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}-\frac{(\bar{x}-\mu)^2}{\sigma^2/n}\\ &= \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2-\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2\\ &= \sum_{i=1}^{n}z_i^2-z_{\bar{x}}^2\sim\chi^2_{\nu=n-1}\end{aligned}$$



## Distribuzione campionaria di $s^2$

- dunque la quantità $\left(\frac{n-1}{\sigma^2}\right)s^2$ corrisponde ad una sommatoria di variabili normali standardizzate, che approssima una distribuzione $\chi^2$ con $\nu=n-1$ gradi di libertà.
- Si noti infatti che $\frac{x_i-\mu}{\sigma}$ è una variabile (*normale*) standardizzata, con $\nu_1=n$ gradi di libertà ($\mu$ non è stimata), mentre $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ è la media campionaria standardizzata, con $\nu_2=1$ grado di libertà (ancora, $\mu$ non è stimata);
- allora, la loro sottrazione si distribuisce secondo la legge $\chi^2_{\nu_1-\nu_2=n-1}$:

$$\left(\frac{n-1}{\sigma^2}\right)s^2\sim\chi^2_{n-1}.$$



## Valore atteso e varianza di $s^2$

- Possiamo ricavare facilmente il valore atteso e la varianza di $s^2$ sfruttando l'approssimazione $\left(\frac{n-1}{\sigma^2}\right)s^2\sim\chi^2_{n-1}$:

$$E(s^2)=E\left(\frac{n-1}{\sigma^2}s^2\frac{\sigma^2}{n-1}\right)=E\left(\chi^2_{n-1}\frac{\sigma^2}{n-1}\right)=E\left(\chi^2_{n-1}\right)\frac{\sigma^2}{n-1}=\sigma^2$$

$$Var\left(s^2\right)=Var\left(\chi^2_{n-1}\frac{\sigma^2}{n-1}\right)=Var\left(\chi^2_{n-1}\right)\frac{\left(\sigma^2\right)^2}{(n-1)^2}=$$

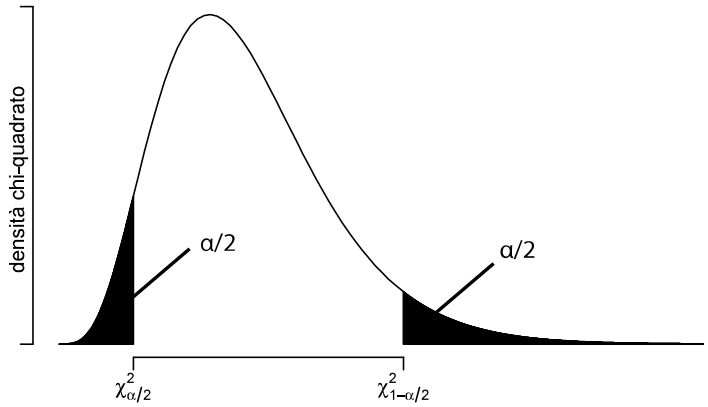
$$=2\left(n-1\right)\frac{\left(\sigma^2\right)^2}{(n-1)^2}=\frac{2\sigma^4}{(n-1)}$$

## Stima per intervalli di $\sigma^2$

- In certi casi può essere necessario fare un'inferenza sulla varianza della popolazione $\sigma^2$.
- Abbiamo detto che $s^2 = \sum (x - \bar{x})^2/(n-1)$ è uno stimatore corretto di $\sigma$, e quindi serve naturalmente a tale scopo.
- Sappiamo inoltre che
  - avendo a disposizione un campione di $n$ valori indipendenti $x_1, x_2, \ldots, x_n$;
  - calcolando la statistica $(n-1)\, s^2$;
  - dalle tavole della distribuzione $\chi^2_{n-1}$, per un coefficiente di fiducia $\alpha$, si possono ottenere i valori $\chi^2_{(\alpha/2)}$ e $\chi^2_{(1-\alpha/2)}$ atti a delimitare un intervallo centrale contenente una probabilità di $1-\alpha$.



## Stima per intervalli di $\sigma^2$






## Stima per intervalli di $\sigma^2$

Dato che la distribuzione $\chi^2$ è asimmetrica, almeno per $n$ non troppo elevati, la centralità dell'intervallo va intesa in senso probabilistico, avendo posto $\alpha/2$ di probabilità su ogni coda:

$$P\left(\chi^2_{(\alpha/2)} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}\right) = 1-\alpha$$

prendendo i reciproci e quindi invertendo la diseguaglianza

$$P\left(\frac{1}{\chi^2_{(\alpha/2)}} \geq \frac{\sigma^2}{(n-1)s^2} \geq \frac{1}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

e moltiplicando gli estremi per $(n-1)s^2$

$$P\left(\frac{(n-1)s^2}{\chi^2_{(\alpha/2)}} \geq \sigma^2 \geq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$



## Stima per intervalli di $\sigma^2$

**Illustrazione.** Si supponga di aver scelto a caso 16 scolaresche dalla popolazione italiana, omogenee come numero, sesso ed età dei componenti. Per queste scolaresche si rileva il tempo dedicato ad attività ricreative. Le statistiche riassuntive del campione danno una media pari a 3.3375, una varianza $s^2 = 2.897$ e $s = 1.702$.

Si calcoli un intervallo di confidenza al $95\%$ per $\sigma^2$.

Si userà la varianza campionaria $s^2 = 2.897$ come stimatore non distorto di $\sigma^2$ e quindi ci si servirà dei valori

$$\frac{(n-1)s^2}{\chi^2_{0.975}} \quad \text{e} \quad \frac{(n-1)s^2}{\chi^2_{0.025}}$$

rispettivamente, come estremo inferiore e superiore dell'intervallo di fiducia.

## Stima per intervalli di $\sigma^2$

**Illustrazione.**
Dalle tavole del $\chi^2$ per $\nu = n - 1 = 15$ gradi di libertà, si trova $\chi^2_{0.975;15} = 27.49$, mentre $\chi^2_{0.025;15} = 6.26$ lo si può ricavare con R

```
 qchisq(0.975,15)
## [1] 27.48839
 qchisq(0.025,15)
## [1] 6.262138
```

quindi l'intervallo di fiducia per $\sigma^2$ al $95\%$ sarà:

$$P\left(\frac{(15)2.897}{6.26} \geq \sigma^2 \geq \frac{(15)2.897}{27.49}\right) = 1 - \alpha$$

$$P\left(6.94 \geq \sigma^2 \geq 1.58\right) = 1 - \alpha$$