

# **Appunti di biostatistica**

Gaia Di Francescantonio

a. a. 2021/2022



# Indice

<b>I</b>	<b>Principi, descrizione ed inferenza</b>	<b>7</b>
<b>1</b>	<b>Introduzione alla statistica (B1)</b>	<b>9</b>
1.1	Statistica . . . . .	9
1.2	Popolazione . . . . .	10
1.3	Campione e stima . . . . .	10
1.4	Stima . . . . .	11
1.5	Dati e variabili . . . . .	12
1.6	Distribuzioni di frequenza e distribuzioni di probabilità . . . . .	14
1.7	Tipi di studi . . . . .	14
<b>2</b>	<b>Archiviazione dei dati (B1)</b>	<b>15</b>
<b>3</b>	<b>Visualizzazione dei dati (B1)</b>	<b>17</b>
3.1	Rappresentare le distribuzioni di frequenza . . . . .	17
3.1.1	Visualizzare variabili categoriche . . . . .	17
3.1.2	Visualizzare variabili numeriche . . . . .	18
3.2	Quantili di una distribuzione di frequenza . . . . .	20
3.3	Distribuzione di frequenza cumulativa . . . . .	20
3.4	Associazioni tra variabili categoriche . . . . .	21
3.5	Confrontare variabili numeriche tra gruppi . . . . .	22
3.6	Visualizzare la relazione tra una coppia di variabili numeriche . . . . .	22
<b>4</b>	<b>Statistiche descrittive (B2)</b>	<b>27</b>
4.1	Misure di tendenza centrale . . . . .	27
4.2	Misure di dispersione . . . . .	28
4.3	Coefficiente di variazione . . . . .	29
4.4	Range ed intervallo . . . . .	29
4.5	Confronto tra misure di posizione e dispersione . . . . .	29
4.6	Mode lasche . . . . .	30
4.7	Distribuzione campionaria . . . . .	30
4.8	Errore standard . . . . .	30
<b>5</b>	<b>Distribuzione normale (B2)</b>	<b>33</b>
5.1	Teorema del limite centrale . . . . .	34
5.2	Intervallo di confidenza . . . . .	34
<b>6</b>	<b>Probabilità (B3)</b>	<b>37</b>
6.1	Istogrammi di frequenza . . . . .	37
6.2	Eventi indipendenti . . . . .	37
6.3	Eventi incompatibili . . . . .	37
6.4	Somma delle probabilità . . . . .	38
6.5	Diagrammi di Venn . . . . .	38
6.6	Prodotto di probabilità . . . . .	39
6.7	Alberi di probabilità . . . . .	41
6.8	Probabilità condizionata e teorema della probabilità totale . . . . .	41
6.9	Teorema della probabilità composta . . . . .	41

6.10	Teorema di Bayes . . . . .	42
<b>7</b>	<b>Test delle ipotesi</b>	<b>45</b>
7.1	Ragionamento logico . . . . .	45
7.2	Metodo scientifico galileiano . . . . .	45
7.3	Formulazione delle ipotesi . . . . .	46
7.4	Verifica delle ipotesi . . . . .	46
7.5	Statistica test . . . . .	47
7.6	Livello di significatività di P . . . . .	48
7.7	Errori nella verifica delle ipotesi . . . . .	49
<b>II</b>	<b>Confronti tra gruppi</b>	<b>53</b>
<b>8</b>	<b>Analisi delle proporzioni (B5)</b>	<b>55</b>
8.1	Distribuzione binomiale . . . . .	55
8.1.1	Distribuzione campionaria di una proporzione . . . . .	56
8.2	Studio di una proporzione: il test binomiale . . . . .	56
8.2.1	Approssimazioni del test binomiale . . . . .	58
8.3	Stima delle proporzioni . . . . .	58
8.3.1	Stima dell'errore standard per una proporzione . . . . .	58
8.3.2	Intervallo di confidenza di una proporzione: metodo Agresti-Coull . . . . .	59
8.3.3	Intervallo di confidenza di una proporzione: metodo Wald . . . . .	59
8.4	Ricavare la distribuzione binomiale . . . . .	60
<b>9</b>	<b>Modelli probabilistici per dati di frequenza (B5)</b>	<b>61</b>
9.1	Modello proporzionale . . . . .	61
9.2	Test del $\chi^2$ di bontà dell'adattamento . . . . .	62
9.2.1	Calcolo del P-value . . . . .	63
9.2.2	Assunzioni del test $\chi^2$ di bontà dell'adattamento . . . . .	64
9.2.3	Test del $\chi^2$ con due sole categorie . . . . .	64
9.3	Adattamento della distribuzione binomiale . . . . .	65
9.3.1	Test binomiale per $n$ grandi . . . . .	67
9.4	Distribuzione di Poisson . . . . .	67
9.4.1	Verifica della casualità con la distribuzione di Poisson . . . . .	69
9.4.2	Indice di dispersione per descrivere il discostamento dalla distribuzione di Poisson . . . . .	71
<b>10</b>	<b>Analisi dell'associazione tra variabili categoriche (B5)</b>	<b>73</b>
10.1	Associazione tra due variabili categoriche . . . . .	73
10.2	Stima dell'associazione in tabelle 2x2: odds ratio . . . . .	74
10.2.1	Errore associato all'odds ratio . . . . .	74
10.3	Test del $\chi^2$ per stabilire la significatività dell'associazione tra due variabili categoriche . . . . .	76
10.3.1	Calcolo del $\chi^2$ . . . . .	77
10.3.2	P-value . . . . .	78
10.3.3	Assunzioni del test $\chi^2$ per l'analisi delle tabelle di contingenza . . . . .	78
10.4	Test esatto di Fisher . . . . .	78
<b>11</b>	<b>Inferenza in una popolazione con distribuzione normale</b>	<b>81</b>
11.1	Riepilogo sulle relazioni tra variabili e statistiche . . . . .	81
11.2	Distribuzione $t$ di Student . . . . .	81
11.3	Intervalli di confidenza e $t$ test per un campione . . . . .	83

<i>INDICE</i>	5
<b>III Associazioni</b>	<b>85</b>
12	87
12.1 . . . . .	87



## **Parte I**

# **Principi, descrizione ed inferenza**





# Capitolo 1

## Introduzione alla statistica (B1)

1

### 1.1 Statistica

I biologi studiano le proprietà degli esseri viventi. Misurare queste proprietà è però un compito assai impegnativo, perché non esistono due individui appartenenti alla stessa popolazione biologica che siano esattamente identici. Inoltre, non siamo quasi mai in grado di misurare ogni individuo presente nella popolazione, e siamo quindi vincolati dal tempo e dai fondi disponibili a limitare le misurazioni a un campione di individui estratto dalla popolazione. Ma il campionamento introduce incertezza. Per effetto del caso, le proprietà del campione non sono uguali a quelle della popolazione dalla quale il campione è stato estratto: le misure ottenute da un campione, infatti, sono influenzate dall'inclusione in quel campione di certi individui piuttosto che altri.

La **statistica** è una disciplina che permette di descrivere e misurare diversi aspetti della natura basandosi su campioni. Soprattutto, la statistica consente di quantificare l'incertezza di queste misure, ovvero di determinare la probabile entità del loro scostamento dal valore vero. La statistica si occupa del processo di **stima**, cioè del processo con cui si inferisce (stima) una grandezza incognita di una popolazione usando i **dati campionari**, ossia i dati provenienti da un campione. Se applicati correttamente, gli strumenti della stima permettono di approssimare pressoché ogni grandezza tipica di una popolazione usando soltanto campioni.

Oltre a stimare grandezze incognite, la statistica permette di valutare le differenze tra gruppi e le relazioni tra variabili.

Tutte le grandezze che descrivono le popolazioni (medie, proporzioni, misure di variazione e misure di relazione) sono dette **parametri**. La statistica dice come si possono stimare nel modo migliore i parametri usando delle misurazioni relative a un campione. Il parametro è quindi il valore vero e la stima (il risultato del processo di stima) è un'approssimazione del valore vero, soggetta a errore. Se fossimo in grado di misurare ogni membro della popolazione, potremmo conoscere il parametro senza errore, ma ciò è possibile solo molto raramente. Invece, si utilizzano stime su dati incompleti per approssimare questo valore vero. Disponendo di strumenti statistici appropriati, siamo in grado di determinare la bontà delle approssimazioni.

La statistica si occupa anche della verifica delle ipotesi (o test delle ipotesi).

La statistica prevede il susseguirsi di alcuni passaggi standard: raccogliere, organizzare, visualizzare, analizzare, interpretare e presentare dati.

---

<sup>1</sup>Capitolo 1, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

## 1.2 Popolazione

La capacità di ottenere misure affidabili delle caratteristiche delle popolazioni e di valutare l'incertezza di queste misure dipende in modo critico da come si campionano le popolazioni stesse. Il primo passo nella raccolta di dati biologici di qualsiasi tipo è decidere quale sia la popolazione da campionare. Una **popolazione statistica** è l'intero insieme di individui o di unità che interessano a un ricercatore e della quale si vogliono conoscere i parametri. Di solito una popolazione è costituita da un grande numero di individui; nella maggior parte dei casi, si suppone in effetti che la popolazione sia infinitamente grande.

Le caratteristiche che si misurano in tale gruppo di elementi devono essere necessariamente variabili (altrimenti la misurazione sarebbe inutile).

Si deve essere in grado di identificare un sistema e di identificarne delle caratteristiche, misurare queste caratteristiche ed operare un'inferenza che permetta di estrapolare le informazioni misurate all'intera **popolazione biologica** (che può coincidere con la popolazione statistica, ma il più delle volte non è così).

## 1.3 Campione e stima

Un **campione statistico** è invece un insieme molto più piccolo di individui selezionati dalla popolazione statistica. Il ricercatore impiega il campione per trarre conclusioni che siano possibilmente valide per l'intera popolazione.

Il campione serve per stimare una caratteristica e la sua variabilità nella popolazione.

Si ricorre alla stima quando non è possibile effettuare la misurazione del parametro su tutti gli elementi della popolazione statistica.

Spesso, l'**unità statistica** di base del campionamento è il singolo individuo. Talvolta, però, l'unità fondamentale di uno studio è un gruppo di individui, nel qual caso un campione è costituito da un sottoinsieme di tali unità.

Un'**unità campionaria** è il soggetto/individuo/evento/caso/osservazione/ecc. su cui si compiono le misurazioni delle variabili. Il concetto di unità statistica ha un valore più ampio, mentre l'unità campionaria è uno degli elementi che compongono il campione, selezionati all'interno della popolazione statistica.

Per designare l'unità di campionamento gli scienziati impiegano diversi termini, quali "unità", "individuo", "soggetto" o "replica".

Il campione deve essere:

- **rappresentativo** della popolazione: quanto maggiore è la variabilità del parametro nella popolazione, tanto più ampio dev'essere il campione al fine di rappresentare adeguatamente la popolazione.
- **casuale**, cioè soddisfa due criteri:
  - ogni unità nella popolazione deve avere un'uguale probabilità di essere inclusa nel campione, quindi il campione è **non distorto**. La **distorsione** (o bias) è una discrepanza sistematica fra le stime e il valore vero della caratteristica della popolazione.
  - la selezione delle unità deve essere **indipendente**, cioè l'inclusione nel campione di un qualsiasi individuo della popolazione non deve influenzare l'inclusione di un altro individuo.

Per ottenere la casualità sono disponibili diversi **disegni di campionamento**, ossia protocolli che definiscono una serie di regole tramite le quali si selezionano le unità statistiche che comporranno il campione; un disegno di campionamento può essere:

- **Casuale**. Si appoggia a strumenti che selezionano elementi casuali (es. con GIS si potrebbe selezionare un poligono e, all'interno di tale poligono, richiedere al software di selezionare delle unità in maniera causale, o anche in modo da assecondare determinate richieste). Quando è possibile, randomizzare è un approccio solido e affidabile, ma a volte non fattibile economicamente.

- **Uniforme.** Si appoggia a strumenti quali una griglia, un intervallo preciso tra le unità, ecc. Un disegno di campionamento **uniforme stratificato** seleziona dei sottoinsiemi su informazioni già note (es. classi di età) in cui campionare in modo differenziale a seconda delle esigenze di ricerca, al fine di cogliere la variabilità nella categoria di interesse.
- **Opportunistico.** Si associa a delle necessità, ha il pericolo di esporsi a un bias di campionamento.

Il campionamento casuale riduce al minimo la distorsione e permette di quantificare l'errore di campionamento.

Un modo per ottenere un campione casuale è quello di generare una lista di tutti gli elementi della popolazione di interesse, assegnare loro un numero ed usare un generatore di numeri casuali per generare  $n$  numeri che corrispondano agli elementi della popolazione che verranno scelti come elementi del campione.

Tale procedura, tuttavia, è difficilmente attuabile nella maggior parte dei casi, quando si ha a che fare con popolazioni molto numerose, come accade solitamente in biologia.

Una possibile soluzione è che l'unità fondamentale di un campionamento non deve essere necessariamente un singolo individuo, ma può anche essere un gruppo.

Ad esempio, è più facile usare una mappa per dividere un tratto di foresta in molte aree di uguali dimensioni, dette anche plot, e poi creare una lista numerata di questi plot piuttosto che produrre una lista numerata di ogni albero. Gli alberi contenuti in un campione casuale di plot non costituiscono un campione casuale poiché gli alberi presenti nello stesso plot non sono campionati in modo indipendente; una semplice tecnica consiste nel calcolare la media delle misure di tutti gli individui entro una certa unità e utilizzare questa media come osservazione indipendente per quell'unità.

I dati reali si basano spesso su campioni non casuali. I biologi affrontano questo problema riconoscendone l'esistenza, indicando dove potrebbero sorgere distorsioni nei loro studi e conducendo ulteriori studi per tentare di risolvere gli eventuali problemi di campionamento.

Un'alternativa indesiderabile al campione casuale è il **campione di convenienza**, basato su individui facilmente disponibili al ricercatore. I ricercatori sono obbligati ad assumere che il campione di convenienza non sia distorto e sia indipendente, come un campione casuale, ma questa assunzione è spesso errata.

Il problema principale riguardo al campione di convenienza è la distorsione.

Un campione di convenienza potrebbe violare anche l'ipotesi dell'indipendenza, se gli individui presenti nel campione fossero più simili l'uno all'altro di quanto siano gli individui scelti casualmente dall'intera popolazione.

Gli studi umani in particolare devono contemplare la possibilità del **bias del volontario**, che è una distorsione derivante da una differenza sistematica tra il pool dei volontari (campione di volontari) e la popolazione a cui appartengono. Il problema sorge quando il comportamento dei soggetti influenza la possibilità che vengano campionati.

## 1.4 Stima

L'**errore di campionamento** è la differenza dovuta al caso tra una stima e il parametro della popolazione che viene stimato.

La dispersione delle stime dovute all'errore di campionamento indica la **precisione** di una stima: più piccolo è l'errore di campionamento, più alta è la precisione.

La stima di un parametro è:

- **corretta** quando il suo valore medio coincide con il valore medio del parametro;
- **distorta** quando il suo valore medio non coincide con il valore medio del parametro;

- **precisa** quando la distribuzione dei suoi valori presenta una bassa variabilità;
- **imprecisa** quando la distribuzione dei suoi valori presenta un'alta variabilità.

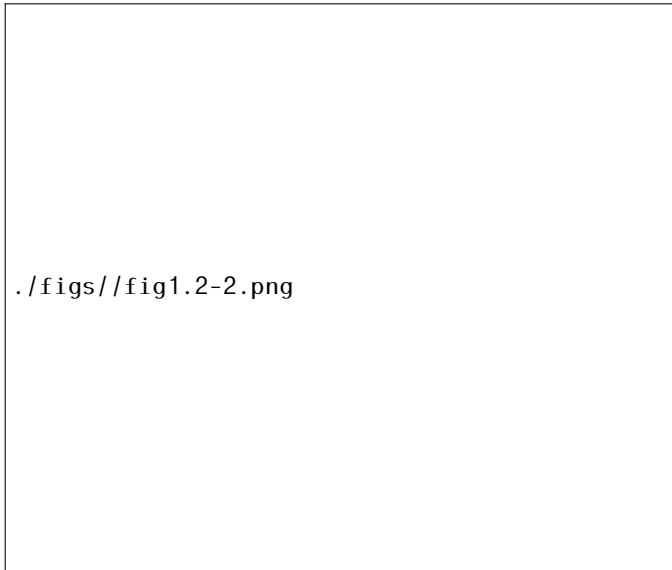


Figura 1.1:

La **variabilità** della misura di una caratteristica può avere molte fonti:

- **naturale**, insita nelle caratteristiche che vogliamo misurare;
- **strumentale**, dovuta all'imprecisione degli strumenti;
- **metodologica**, insita nell'uso improprio degli strumenti o nell'uso degli strumenti non adeguati;
- **dell'operatore**, dovuta all'abilità dell'operatore, che è soggettiva e dipende da esperienza personale e molti altri fattori (stanchezza, stato emotivo, aspettative, autorità, ecc.); data la possibilità di questo tipo di variabilità è utile documentare minuziosamente la misurazione;
- **campionaria**, dipende dal campione selezionato, che solitamente non rispecchia perfettamente la popolazione statistica ed è comunque generalmente differente dal campione che si sceglierebbe randomicamente in un'altra situazione; ciò genera una variabilità ulteriore, oltre a quella propria della popolazione statistica. Maggiori sono le dimensioni del campione, minore è la variabilità campionaria.

La **convergenza della stima al parametro** dipende dalla grandezza dell' $N$  campionario.

## 1.5 Dati e variabili

Una volta ottenuto un campione, si può cominciare a misurare le variabili di interesse.

Una variabile è una qualsiasi caratteristica quantitativa o qualitativa delle unità statistiche, che si presenta in esse con un certo valore (es. sesso, peso, colore dei capelli, velocità di corsa) e che ha una variabilità intrinseca. Sono variabili anche le stime (es. velocità media di corsa di un campione casuale di 10 individui), perché esse differiscono, per effetto del caso, da campione a campione.

Una variabile latente è una variabile che non può essere misurata direttamente, ma dev'essere estrapolata dalla misura di altre variabili caratterizzabili in maniera quantitativa o qualitativa.

Le variabili possono essere qualitative, quantitative, categorie/classi.

La variabile può essere una caratteristica/attributo/campo/colonna misurabile.

I dati sono i risultati grezzi delle misurazioni di una o più variabili effettuate su un campione di individui.

A partire dalle misurazioni si caratterizzano le variabili, le quali possono essere:

- **Qualitative (categoriche).** Descrivono caratteristiche che non possono essere misurate con un numero ma che permettono di inserire gli elementi di un campione in una categoria o gruppo (es. modalità di trasmissione di malattie: acqua, aria, vettore animale, contatto diretto).
- **Quantitative.** Le misure degli individui sono quantitative e ad esse è associato un valore su una scala numerica (es. conteggi, dimensioni, angoli, tassi, percentuali).

Un altro modo (preferibile) per classificare i tipi di variabili è:

- **Nominale.** Sono qualitative, categoriali e le differenti categorie non hanno un ordine intrinseco, ma solo un nome (es. gruppo sanguigno). La variabile nominale risponde solo alla domanda “a che classe l’unità campionaria appartiene”, restituendo, dunque, l’identità degli elementi del campione in base alla variabile misurata.
- **Ordinale.** Sono quantitative nel senso che i valori corrispondenti possono essere ordinati nonostante non siano rappresentabili su una scala numerica (es. fase biologica: uovo, larva, giovane, adulto; stato di salute; colore dei capelli, a meno di non misurarlo con uno strumento quale un esposimetro che potrebbe rendere questa categoria continua). La variabile ordinale categorizza le unità campionarie (restituisce la loro identità) e le pone in un rango.
- **Continue.** Sono quantitative. I valori corrispondenti appartengono a categorie continue (es. temperature, peso, lunghezza). La variabile continua categorizza le unità campionarie, le inserisce in un rango e definisce degli intervalli (si può conoscere la distanza tra due misurazioni). In alcuni casi (es. peso, lunghezza) si aggiunge l’informazione di rapporto, dato che sono scale assolute (una temperatura di 30°C non è il doppio di una temperatura di 15°C, a meno che le misure non si effettuino in scala Kelvin, che è una scala assoluta). Nel caso in cui si tratti di variabili numeriche, queste possono assumere qualsiasi valore numerico reale in un certo intervallo di variazione. Tra due valori qualsiasi di una variabile continua possono esistere infiniti altri valori (es. temperatura corporea, area di un territorio).
- **Numeriche discrete.** Si presentano in unità indivisibili (es. conteggi, numero di accoppiamenti durante la stagione riproduttiva, numero di aminoacidi in una proteina).

L’età è effettivamente una variabile continua, ma a seconda delle possibilità di misurazione può anche essere considerata ordinale.

La misurazione di una variabile qualitativa permette di suddividere l’unità campionaria in gruppi (es. M/F).

Se si passa da una variabile continua ad intervalli si perdono informazioni e la variabile diventa categoriale ordinale. Passare da variabili continue a variabili ordinali può essere utile quando non ci si fida delle misurazioni effettuate (è come concedersi un margine d’errore).

Il fatto che una variabile venga indicizzata con un numero non significa che sia una variabile numerica; si potrebbero usare i numeri anche per denominare categorie. I dati numerici possono essere ridotti a dati categorici attraverso un loro raggruppamento (es. in due gruppi: sopra e sotto la media) ma, in questo modo, parte dell’informazione viene persa.

Nell’analisi statistica si cerca sempre di prevedere una delle variabili, detta **variabile risposta**, in base a una seconda variabile, detta **variabile esplicativa**. Le variabili vengono talvolta classificate come variabili indipendenti o dipendenti; questi termini sono sinonimi di variabili esplicative e risposta, rispettivamente, ma, a rigore, se una di due variabili dipende dall’altra, allora nessuna delle due è indipendente.

## 1.6 Distribuzioni di frequenza e distribuzioni di probabilità

<sup>2</sup> Gli individui in un campione hanno generalmente diversi valori della grandezza analizzata. Si può osservare questa variabilità considerando una distribuzione di frequenza. La **frequenza** di una particolare misura in un campione è il numero di osservazioni di un determinato valore della misura. La **distribuzione di frequenza** rappresenta la frequenza con cui ciascun valore della variabile si presenta nel campione.

Si usa la distribuzione di frequenza di un campione per acquisire informazioni sulla distribuzione della variabile nella popolazione dalla quale il campione è stato estratto.

La distribuzione di una variabile in tutta la popolazione è detta **distribuzione di probabilità**.

Nel caso di una variabile continua (es. spessore del becco nei fringuelli delle Galapagos), la distribuzione nella popolazione viene approssimata spesso con una distribuzione di probabilità teorica, detta distribuzione normale o di Gauss.

## 1.7 Tipi di studi

<sup>3</sup> In biologia, i dati si ottengono da uno studio sperimentale o da uno studio osservazionale.

In uno **studio sperimentale**, il ricercatore assegna casualmente differenti gruppi di trattamento o valore di una variabile esplicativa alle singole unità di studio.

In uno **studio osservazionale**, è la natura ad assegnare gruppi di trattamenti o valori di una variabile esplicativa agli individui, ed il ricercatore non ha alcun controllo su quali unità rientrano in quali gruppi.

Una differenza fondamentale fra studi sperimentali ed osservazionali sta nel fatto che, mentre i primi possono determinare le relazioni di causa ed effetto tra le variabili, i secondi sono in grado soltanto di individuare associazioni. (non spiegato a lezione)

---

<sup>2</sup>non fatto a lezione

<sup>3</sup>non fatto a lezione

## Capitolo 2

# Archiviazione dei dati (B1)

Quando si organizzano i dati che si hanno a disposizione si devono considerare due aspetti: quali sono le **unità campionarie** e quali sono le **qualità** che sono state misurate (o valutate) su queste unità statistiche. Una volta chiari questi aspetti, si può compilare una **tabella di dati**, in cui, solitamente, ad ogni colonna corrisponde una variabile, ad ogni riga corrisponde un'unità statistica, e ad ogni cella un dato.

È sempre necessaria la presenza di un **campo chiave** che identifichi univocamente le unità statistiche.

È importante mantenere una **uniformità di formato** in ciascuna colonna (una stessa colonna deve ospitare dati che abbiano sempre uno stesso formato, es. numerico, alfabetico, alfa-numerico).

Il modo migliore per identificare un **valore mancante** è lasciare il campo vuoto (non si usa lo zero, poiché quello è effettivamente un valore) o utilizzare la sintassi specifica richiesta dal programma che si sta utilizzando (es. NA, NULL, -).

Una volta scelti dei **codici** per operare è importanti che questi siano **uniformi** (es. se per indicare che un individuo è maschio si usa "m" non si può utilizzare anche "M" o "male", etc.)

**Excel** è uno strumento di calcolo su fogli elettronici, non un database.

In un database, ogni variabile è definita precisamente e non è possibile inserire dati che non siano corrispondenti a una formattazione prefissata, mentre in un foglio di calcolo questo inserimento è possibile. Le righe, inoltre, nei database, rappresentano solo le unità statistiche, mentre la prima riga nei fogli di calcolo solitamente rappresenta le etichette.

In un foglio di calcolo è utile creare un foglio di "LUT", ossia una "**look-up table**", un foglio di calcolo in cui vengono descritte le variabili che è possibile trovare nel primo foglio. In ogni LUT (una per ciascuna variabile considerata) vanno inseriti (etichetta colonne):

- codice (i vari dati che è possibile inserire);
- descrizione (la descrizione di ogni possibile dato);
- eventuali note.

Questo non è necessario nel caso sia presente un'interfaccia grafica che restringe la possibilità di immissione nelle celle ai soli valori della look-up table che faccia scegliere da una lista precisa la codifica corretta (es. i vari form su internet in cui, ad esempio, si inserisce la propria provincia di residenza scegliendola da una lista definita e non inserendo i singoli caratteri autonomamente).

La look up table non si usa per le variabili continue, al limite è possibile in alcuni casi definire degli intervalli di valori continui che è possibile inserire, ma generalmente non è necessario.

Le colonne di variabili non devono essere ridondanti, ovvero non devono essere presenti due variabili che indichino una stessa caratteristica.





## Capitolo 3

# Visualizzazione dei dati (B1)

<sup>1</sup>

La visualizzazione dei dati è fondamentale per la divulgazione e la comprensione dei risultati.

Regole basilari:

- identificare il messaggio;
- essere consapevoli che ciò che si sta presentando è importante;
- non spaventare i destinatari con i numeri;
- massimizzare la risoluzione dell'immagine;
- evitare immagini 3D quando non necessarie (es. barre 3D in istogrammi);
- usare il tipo di rappresentazione idoneo;
- non mischiare tipi di grafico inutilmente;
- non usare gli assi per confondere;
- non fare troppo affidamento al colore e usarlo per trasmettere un significato.

Differenti tipi di grafici possono essere utilizzati per rappresentare stessi dati; la scelta deve dipendere dal tipo di obiettivo che si ha, poiché grafici differenti sottolineano aspetti differenti.

### 3.1 Rappresentare le distribuzioni di frequenza

La frequenza relativa è la frazione di osservazione di una data misura, calcolata com'è il rapporto tra la frequenza (assoluta) e il numero totale di osservazioni. La distribuzione di frequenza relativa mostra la proporzione di occorrenze di ciascun valore nell'insieme dei dati.

Sono disponibili diversi metodi per visualizzare graficamente le distribuzioni di frequenza, a seconda che la variabile si è categorica o numerica.

#### 3.1.1 Visualizzare variabili categoriche

Una **tabella di frequenza** è una visualizzazione testuale del numero di occorrenze di ciascuna categoria nel dataset.

Un **diagramma a barre** impiega l'altezza di rettangoli per visualizzare la frequenza (assoluta o relativa) di occorrenza di ciascuna categoria. Sull'asse delle ordinate può essere rappresentata la frequenza assoluta o la frequenza relativa.

Quando si trattano variabili nominali, le categorie non hanno un ordine intrinseco, ma si può

---

<sup>1</sup>Capitolo 2, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

semplificare il confronto della loro importanza relativa disponendole in ordine di occorrenza dalla più alla meno frequente.

Nei diagrammi a barre non si riescono a distinguere le frequenze esatte, ma si ha un quadro chiaro e immediato di come i valori numerici dei crescano rapidamente nelle categorie meno frequenti.

Una variabile nominale si può descrivere con un grafico a barre in cui i valori della variabile nominale sono sulle ascisse mentre sulle ordinate si riporta la percentuale dei casi in cui si sono presentate le singole variabili. Le variabili possono essere elencate in ordine alfabetico o dalla meno alla più frequente, a seconda dell'obiettivo che si ha.

Per essere efficace, un diagramma a barre dovrebbe seguire alcune regole generali:

- nel caso di una variabile categorica ordinale (es. grado di gravità del morso di un serpente), i valori lungo l'asse orizzontale dovrebbero essere ordinati in modo naturale;
- nel caso di variabili categoriche nominali, l'ordinamento delle categorie secondo la frequenza con cui si osservano facilita la rappresentazione visiva delle informazioni;
- i rettangoli dovrebbero essere di uguale larghezza e ciascuno di essi dovrebbe poggiare su una linea base di riferimento corrispondente a  $y = 0$ , affinché l'area e l'altezza di ciascuno siano direttamente proporzionali alla frequenza; una frequenza base diversa da zero o una larghezza dei rettangoli disuguale fornirebbe una visione distorta dell'importanza relativa di ciascuna categoria;
- i rettangoli dovrebbero essere separati, non addossati tra loro;
- è utile indicare il numero totale di osservazioni ( $n$ ) nella didascalia della figura.

Un **grafico a torta** è una valida alternativa attraverso la quale si possono visualizzare sia la frequenza assoluta che la frequenza relativa (ossia percentuale) dei casi.

### 3.1.2 Visualizzare variabili numeriche

Una distribuzione di frequenza per una variabile numerica può essere visualizzata come una tabella di frequenza o con un istogramma.

In un **istogramma** si rappresentano i valori di una variabile che si può assumere continua; un istogramma differisce da un grafico a barre in quanto l'ascissa non contiene le singole categorie nominali, ma categorie che raggruppano le singole categorie nominali in base ad una loro caratteristica (es. invece di inserire le varie specie sull'asse delle  $x$ , si inseriscono categorie di abbondanza, in modo da rappresentare con l'istogramma la frequenza delle specie che hanno una determinata abbondanza). La somma delle superfici di un grafico rappresentato mediante isto-

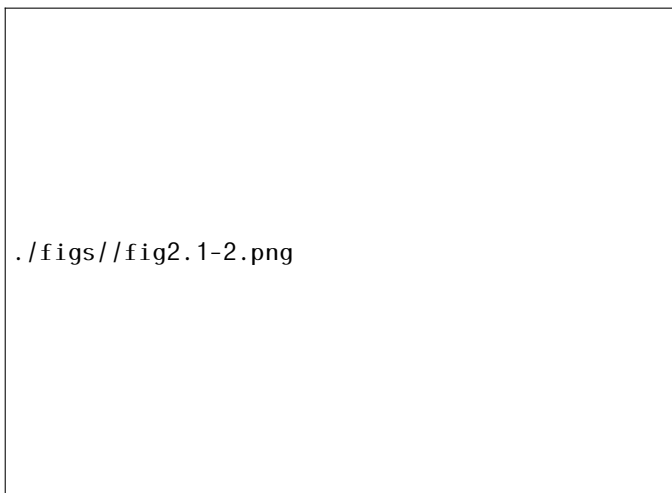


Figura 3.1:

gramma è uguale al numero totale di specie. Se invece di rappresentare la frequenza assoluta se ne rappresentasse la frequenza relativa, la somma delle aree dei rettangoli corrisponderebbe ad 1.

Un istogramma utilizza l'area di bar rettangolari per visualizzare la frequenza. I valori dei dati vengono suddivisi in intervalli consecutivi (**bin**) generalmente di uguale ampiezza, e viene rappresentata la frequenza delle osservazioni che cadono in ciascun bin.

Anche se l'asse delle ordinate riporta la frequenza (assoluta o relativa) in un certo intervallo, l'istogramma deve essere inteso come rappresentazione areale: sono le aree dei vari rettangoli che devono essere proporzionali alle frequenze corrispondenti, non le altezze. Quando tutti i bin hanno la stessa ampiezza, comunque, ragionare in termini di altezze o di aree delle barre è indifferente. Se però si decidesse di raggruppare due bin in un solo intervallo, lasciando gli altri inalterati, l'altezza della nuova barra con la base più larga non dovrebbe essere pari alla somma delle frequenze nei due bin, ma alla loro media. Solo in questo modo la rappresentazione della distribuzione di frequenza basata sulle aree delle barre risulterebbe non distorta. Alternativamente si potrebbe rappresentare sull'asse delle ordinate, per ogni intervallo, non la frequenza delle osservazioni ma la densità, cioè la frequenza divisa per l'ampiezza dell'intervallo. In questo modo l'area dei rettangoli sarebbe uguale, non solo proporzionale, alla frequenza, essendo il prodotto della base (l'ampiezza dell'intervallo) per l'altezza (la densità).

L'istogramma di solito si usa per descrivere variabili tipicamente continue. Nel caso delle variabili nominali, l'istogramma dice solo quante sono le specie che hanno diverse classi di abbondanza (le unità statistiche non sono più i singoli uccelli osservati, ma le specie e la proprietà osservata è il numero di specie).

Un intervallo nella distribuzione di frequenza notevolmente più frequenti di quelli circostanti è detto **picco**.

La **moda** è l'intervallo corrispondente al picco più alto.

Un istogramma può assumere diverse forme, che riflettono diversi tipi di distribuzioni:

- **uniforme**: la frequenza delle varie categorie è molto simile;
- **a campana (normale)**: il valore centrale è quello più frequente e a partire da esso si ha un simmetrico decremento della frequenza (spesso, le biometrie presentano questa distribuzione);
- **a campana asimmetrica**: il decremento rispetto al picco non è uniforme nelle due direzioni (es. livello di infezione da parassiti negli animali: pochi soggetti che portano molti parassiti);
- **bimodale**: ci sono due modalità di distribuzione che indicano la potenziale presenza di due popolazioni statistiche.

Una distribuzione di frequenza è detta distribuzione **simmetrica** se il pattern di frequenze nella metà sinistra del istogramma è l'immagine speculare del pattern nella metà destra. La distribuzione uniforme e la distribuzione a campana sono simmetriche. Una distribuzione di frequenza che non è simmetrica è detta **asimmetrica**. Una distribuzione di frequenza per una variabile numerica, se non è simmetrica, può essere asimmetrica **positiva** (a destra) o **negativa** (a sinistra) a seconda della posizione della coda più lunga.

Valori estremi molto distanti dal resto delle osservazioni sono detti **outlier** (valori anomali). Gli outlier sono comuni nei dati biologici: possono derivare da errori commessi nella registrazione dei dati o possono rappresentare caratteristiche reali della natura. Gli outlier dovrebbero essere sempre analizzati attentamente e si dovrebbe eliminarli dall'insieme dei dati soltanto se si riesce a dimostrare che sono frutto di errori.

L'ampiezza degli intervalli di un istogramma può influenzare le conclusioni tratte dal grafico.

L'ampiezza delle classi (bin) ed il loro numero sono legati, e definirli è importante poiché si tratta di fattori che influenzano la risoluzione dei casi (es. una distribuzione bimodale potrebbe sembrare una asimmetrica se i bin venissero allargati).

Per definire ampiezza e numero delle classi si può ricorrere a diversi approcci, ad esempio:

- **H. Sturges** (1926):

$$C = 1 + \frac{10}{3} \log_{10}(N) \quad (3.1)$$

In cui  $C$  è pari al numero di classi.

- **D. Scott** (1979):

$$h = \frac{3.5 \cdot s}{\sqrt{N}} \quad (3.2)$$

In cui  $h$  è l'ampiezza delle classi, ed  $s$  è la deviazione standard.

In entrambi i casi,  $N$  è la dimensione campionaria.

Non esistono regole rigorose riguardo al numero di intervalli che si devono usare nelle tabelle di frequenza e negli istogrammi; alcuni programmi impiegano la regola empirica di Sturges, in cui il numero di intervalli è  $1 + \ln(n)/\ln(2)$ , in cui  $n$  è il numero di osservazioni. Il numero di intervalli che si ottiene con questa formula viene poi arrotondato al numero intero più alto.

In generale, si dovrebbe scegliere il numero di intervalli in modo da far risaltare nella maniera migliore i pattern e le eccezioni nei dati, il che richiede buon senso piuttosto che regole rigorose.

Per costruire un istogramma si dovrebbero seguire alcune semplici regole:

- il lato inferiore di ogni rettangolo deve giacere sulla linea di base corrispondente a  $y = 0$ , affinché le aree siano proporzionali alle frequenze;
- diversamente da quanto accade nei diagrammi a barre, i rettangoli vicini di un istogramma non sono separati, il che rafforza la percezione di una scala numerica, grazie al passaggio graduale da ciascun rettangolo a quello successivo;
- quando si suddividono i dati in intervalli si devono usare numeri facilmente leggibili per i punti di suddivisione tra intervalli (es. 0.5 invece che 0.482);
- è utile indicare il numero totale di individui nella didascalia che accompagna l'istogramma.

## 3.2 Quantili di una distribuzione di frequenza

Tutte le informazioni relative alla forma della distribuzione di frequenza di una variabile numerica sono contenute nei suoi percentili e nei quantili strettamente correlati.

L' $x$ -esimo **percentile** di una distribuzione è il valore al di sotto del quale è situato l' $x\%$  degli individui; lo stesso valore anche rappresentabile attraverso il **quantile**  $x/100$ .

## 3.3 Distribuzione di frequenza cumulativa

La **distribuzione di frequenza cumulativa** è il diagramma di tutti i quantili di una variabile numerica. Per costruire una distribuzione di frequenza cumulativa dell'abbondanza di tutte le specie di uccelli del deserto dell'Esempio 2.1B, prima sono state ordinate tutte le misure dalla più piccola alla più grande, poi è stata calcolata la frazione di osservazioni minori o uguali a ciascun valore. Questa frazione, detta frequenza relativa cumulativa, è rappresentata graficamente dall'altezza della curva sopra il corrispondente valore osservato nei dati. Ogni gradino è piatto, ma la curva si innalza di  $1/n$  in corrispondenza di ogni misura osservata, dove  $n$  è il numero di osservazioni, fino a un massimo di 1. Possono esservi salti più grandi di  $1/n$  in corrispondenza di una misura se la stessa misura si ripete in più punti. La curva fornisce molti particolari perché illustra ogni singola osservazione.

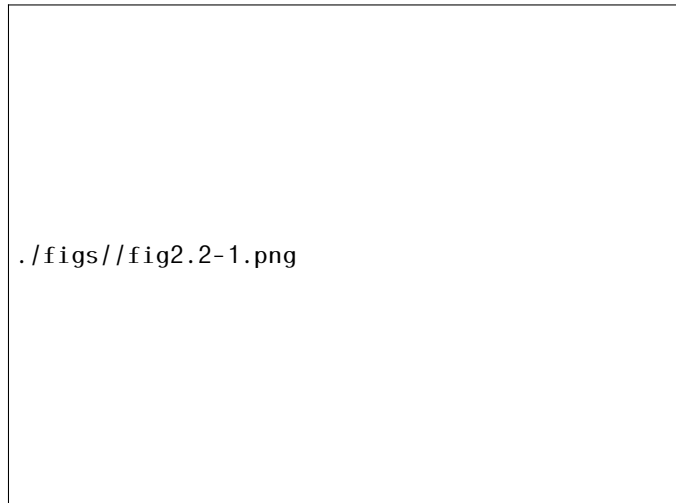


Figura 3.2:

### 3.4 Associazioni tra variabili categoriche

La visualizzazione dei dati è molto utile quando si studiano l'associazione tra variabili e le differenze tra gruppi. I metodi che si usano per presentare graficamente le relazioni dipendono dai tipi di variabili di interesse, cioè dal fatto che esse siano categoriche oppure numeriche.

Le **tabelle di contingenza** sono tabelle di frequenza per due o più variabili categoriche che mostrano come i valori di una variabile siano associati e valori di una seconda variabile.

Una cella è una combinazione di categorie delle variabili riga e colonna. La variabile esplicativa è presentata nelle colonne, mentre la variabile risposta, quella che viene prevista, è presentata nelle righe.

Il **diagramma a barre raggruppate** si rettangoli per rappresentare graficamente la frequenza di tutte le combinazioni di due (o più) variabili categoriche.

I diagrammi a barre raggruppate seguono tutte le convenzioni valide per i diagrammi a barre, con la differenza che i gruppi (per esempio, malaria e assenza di malaria) sono indicati con differenti colori o sfumature di colore. Lo spazio tra i rettangoli dei diversi gruppi della variabile esplicativa (per esempio, controllo e sottrazione di uova) è bene che sia più ampio dello spazio tra i rettangoli della variabile risposta.

Un **grafico a mosaico** (mosaic plot, diagramma di Marimekko) è simile a un diagramma a barre raggruppate, con la differenza che i rettangoli relativi allo stesso gruppo sono impilati l'uno sull'altro invece di essere disposti fianco a fianco.

L'area dei rettangoli in un grafico a mosaico rappresenta la frequenza relativa (ossia la proporzione) entro ciascun gruppo.

La larghezza di ciascuna di la verticale è direttamente proporzionale al numero di osservazioni per quel gruppo.

L'area di ciascun rettangolo, quindi, è direttamente proporzionale alla frequenza relativa di quella combinazione di variabili nell'intero insieme di dati.

L'ordine dei gruppi lungo l'asse orizzontale è predeterminato nel caso dei dati ordinali ed arbitrario nel caso dei dati nominali.

Un grafico a mosaico fornisce una panoramica dei dati e consente di riconoscere le relazioni tra variabili considerate. Ad esempio, l'indipendenza viene mostrata quando le caselle nelle categorie hanno tutte le stesse aree. L'area delle tessere, nota anche come dimensione del bin, è proporzionale al numero di osservazioni all'interno di quella categoria.

Il principale vantaggio del diagramma a barre raggruppate, rispetto alla tabella di contingen-

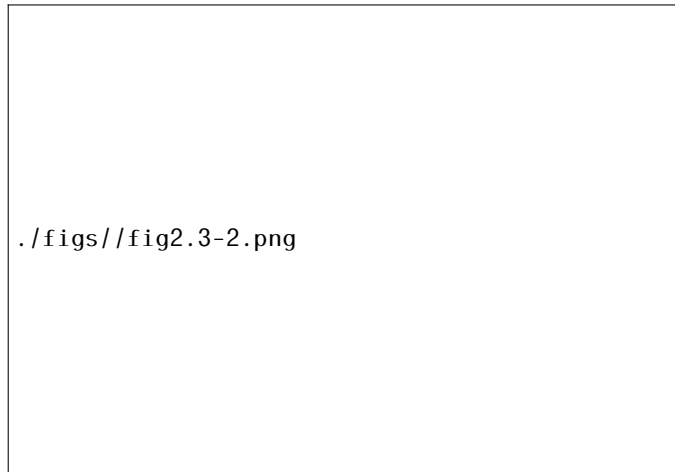


Figura 3.3:

za, è il fatto che risulta più facile confrontare visivamente l'altezza o l'area dei rettangoli che i numeri. Questo vantaggio si riduce se le variabili esplicative e le variabili hanno molte categorie, poiché in questo caso aumenta la complessità del diagramma. Il grafico a mosaico, d'altra parte, è più semplice da leggere rispetto al diagramma a barre raggruppate. Per decidere quale tipo di visualizzazione sia più efficace per una data circostanza è meglio sperimentare più metodi e scegliere tra essi sulla base delle informazioni che esprimono, della chiarezza e della semplicità.

### 3.5 Confrontare variabili numeriche tra gruppi

Gli istogrammi di frequenza relativa prescindono dalle dimensioni del campione, quindi ci si concentra su posizione e forma dell'istogramma, e non sull'altezza assoluta.

Dunque, un metodo utile per confrontare più distribuzioni di frequenza consiste semplicemente nel costruire un istogramma per ciascun gruppo. Se poi gli istogrammi vengono presentati in serie, uno sotto l'altro, facendo attenzione che la scala dell'asse delle x sia la stessa, il confronto risulta più semplice e immediato.

Esempio di applicazione degli istogrammi: livelli di emoglobina in maschi umani a diverse altitudini. Sono tutte distribuzioni discretamente simmetriche. L'ipotesi (aumento della concentrazione di emoglobina nel sangue all'aumentare dell'altitudine) non sembra essere verificata, poiché i valori di tendenza centrale dei primi tre grafici sono molto simili; solo sulle Ande si riscontra una palese differenza.

Prendendo tutti i dati assieme si potrebbe ottenere una distribuzione bimodale.

Lo stesso confronto si può fare usando delle curve di frequenze cumulative.

Si può esaminare anche l'associazione tra una variabile numerica e una variabile categorica confrontando le distribuzioni di frequenze cumulative. Questo approccio presenta le distribuzioni per più gruppi in una singola visualizzazione compatta.

### 3.6 Visualizzare la relazione tra una coppia di variabili numeriche

Ci sono tre metodi per visualizzare la relazione tra due variabili numeriche: diagrammi di dispersione (scatter plot), diagrammi a linee, mappe.

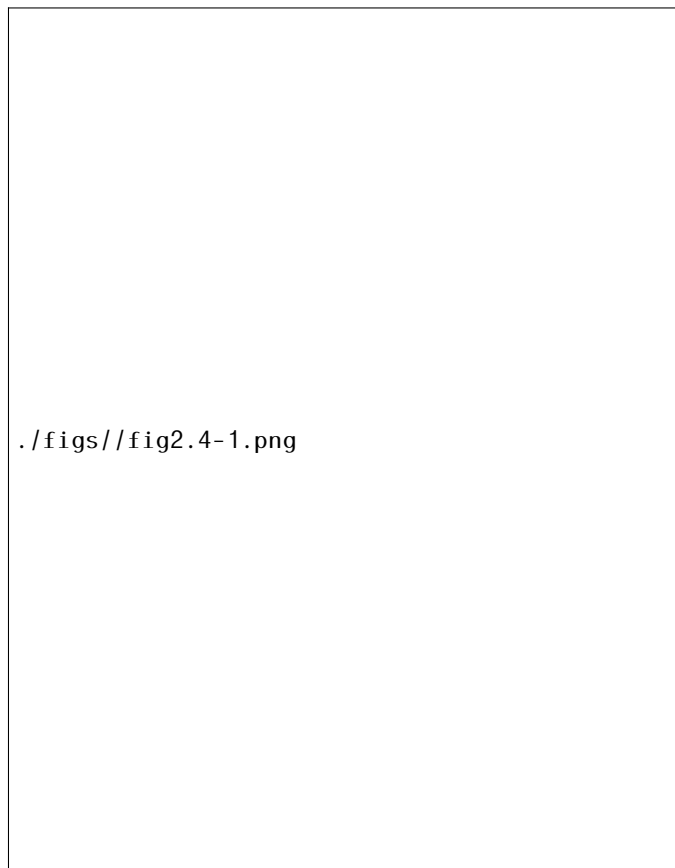


Figura 3.4:

Uno **scatter plot** (diagramma di dispersione, grafico a punti) visualizza il pattern di associazione tra due variabili numeriche. Ogni osservazione è rappresentata da un punto su un diagramma formato da due assi (diagramma cartesiano). La posizione lungo l'asse orizzontale (l'asse delle ascisse o delle  $x$ ) indica la misura della variabile esplicativa; la posizione lungo l'asse verticale (l'asse delle ordinate o delle  $y$ ) indica la misura della variabile risposta. Il pattern, nella risultante nube di punti, indica se un'associazione tra le due variabili sia positiva (nel qual caso i punti tendono a distribuirsi nel diagramma dal basso a sinistra verso l'alto a destra), negativa (i punti tendono a decorrere dall'alto a sinistra verso il basso a destra) o assente (non c'è un pattern distinguibile). Un diagramma di dispersione riesce anche a rivelare se la relazione tra due variabili possa essere rappresentata da una retta oppure da una curva più articolata.

Es. in una specie di pesci, la proporzionalità diretta tra ornamentazione del padre con l'attrattività del figlio, le unità statistiche sono le coppie padre-figlio. L'ornamentazione del padre è un proxy per l'attrattività del figlio. Quando si ha una sequenza temporale dei dati che si stanno rappresentando si possono collegare i vari punti con una linea, ottenendo un diagramma a linee. Un **diagramma a linee** è uno strumento efficace per visualizzare tendenze nel tempo o rispetto a qualche altra variabile ordinata. Un diagramma a linee è simile a un diagramma di dispersione, eccetto che viene riportato soltanto una misura  $y$  per ogni osservazione  $x$ . I punti adiacenti lungo l'asse  $x$  sono uniti da un segmento.

I segmenti di retta che congiungono i punti rendono più visibili le tendenze della popolazione. L'inclinazione più o meno ripida dei segmenti rispecchia la velocità di variazione della popolazione.

Si tratta di una serie temporale. Questi grafici possono mettere in evidenza meccanismi naturali (es. relazione preda-predatore: effetto rimozione, effetto disponibilità risorse per preda, effetto stress da predazione, etc).

Una **mappa** è l'equivalente spaziale del diagramma a linee: visualizza una misura numerica di

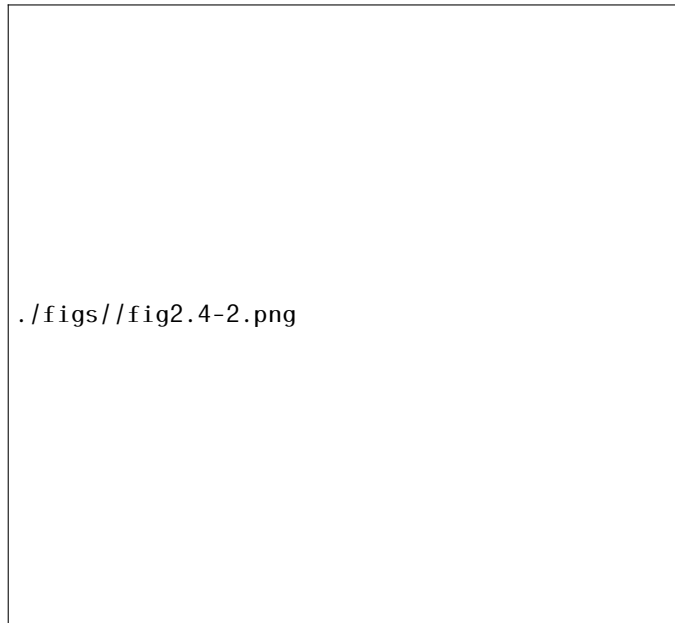


Figura 3.5:

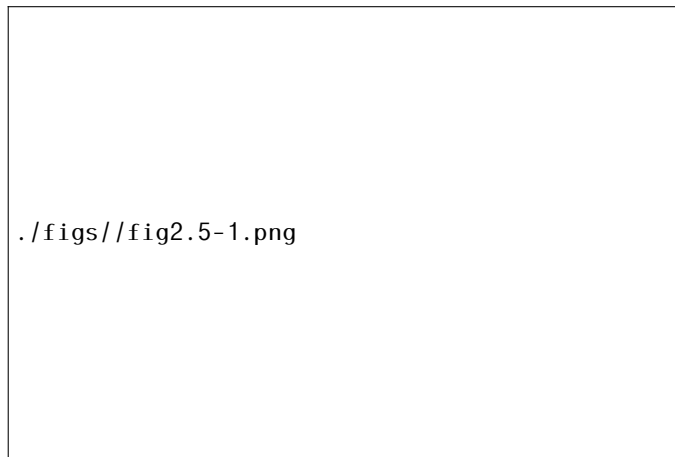


Figura 3.6:

risposta in più posizioni su una superficie. La variabile esplicativa è sempre una serie ordinata, costituita da punti nello spazio. Viene visualizzata una misura  $y$  per ogni punto o regione sulla superficie.

Le mappe possono indicare misure in punti su qualsiasi superficie. La superficie può essere un reticolo spaziale o una regione della superficie terrestre delimitata da confini politici o geologici. Le mappe possono essere usate anche per indicare misure su rappresentazioni di qualsiasi oggetto bidimensionale o tridimensionale.

Le rappresentazioni grafiche e le tabelle hanno un duplice scopo: essere uno strumento per l'analisi dei dati e comunicare e sintetizzare i pattern nei dati a un pubblico più vasto. Questi obiettivi sono ampiamente convergenti, perché le visualizzazioni più informative sono anche quelle più adatte per identificare pattern nei dati e per comunicarli. In tutti i casi, la visualizzazione dei dati deve essere chiara, onesta ed efficiente.

Alcuni consigli generali per aumentare l'efficacia delle rappresentazioni grafiche sono:

- Mostrare i dati.



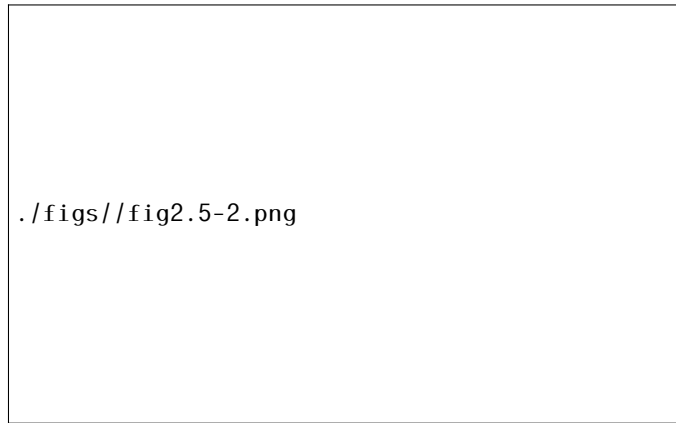


Figura 3.7:

- Rappresentare accuratamente le grandezze. Una linea base corrispondente allo zero è essenziale nei diagrammi a barre e negli istogrammi, perché l'occhio interpreta istintivamente l'area dei rettangoli come direttamente proporzionale al valore relativo. Altri diagrammi, come il diagramma a linee, non sempre hanno questo scopo e quindi non sono vincolati a possedere questa caratteristica. L'obiettivo principale di un diagramma di dispersione è visualizzare un'associazione, non un valore numerico o una frequenza, e la scelta della linea base dovrebbe essere fatta tenendo presente questo criterio.
- Cose superflue sono gli effetti tridimensionali, che non soltanto non sono necessari, ma distorcono anche l'altezza e l'area dei rettangoli, ingannando di nuovo l'occhio. Questi abbellimenti inutili e ingannevoli sono noti come "chartjunk" ("spazzatura del diagramma"). Gli elementi grafici inessenziali dovrebbero essere eliminati per permettere che i pattern nei dati possano emergere con maggior chiarezza.
- Utilizzare elementi grafici in maniera chiara riducendo la confusione e le ridondanze.
- Fare in modo che la rappresentazione grafica sia di facile interpretazione. Questo avviene attraverso una scelta mirata dei caratteri tipografici e dei colori e attraverso l'indicazione delle unità di misura e l'uso di simboli grafici chiaramente distinguibili per rappresentare differenti gruppi.
- Identificare chiaramente gli assi.
- Inserire didascalie che siano self-explanatory (non deve essere necessario leggere il testo, ad esempio, dell'articolo che contiene il grafico, per interpretarlo) e dunque ridondanti rispetto alle didascalie di altri grafici presenti nello stesso lavoro o al lavoro stesso.

Le tabelle possono avere scopi diversi:

- immagazzinamento dei dati, ovvero la conservazione dei dati grezzi a scopo di riferimento; tabelle di questo tipo sono spesso grandi e non adatte alla comunicazione dei risultati finali; quando vengono pubblicate, vengono presentate di solito come appendici o allegati.
- visualizzazione di pattern ed eccezioni nei dati (es. tabelle di frequenza). Per produrre tabelle chiare, oneste e deficienti si devono seguire in generale gli stessi criteri già elencati per le rappresentazioni grafiche. Queste tabelle devono essere compatte e presentare il numero minimo di cifre significative necessarie ad evidenziare il pattern. L'organizzazione dei dati in tabella deve facilitare la lettura e l'identificazione delle caratteristiche principali dei dati.

I diagrammi sono il mezzo migliore per mostrare pattern ed eccezioni, ma forniscono pochi particolari quantitativi; le tabelle, quindi, sono più appropriate quando è importante comunicare anche aspetti quantitativi dei dati.

## CONSIDERAZIONE PERSONALE

La rappresentatività di un campione non dipende solo dal numero di elementi costituenti il campione, ma anche dalla variabilità del parametro al quale si è interessati (caso estremo: 0 variabilità, anche solo un elemento è rappresentativo; variabilità=2, anche quattro su 100 è più probabile che sia rappresentativo che nel caso in cui ci fosse varietà=10).

(ma in effetti vale solo se si è interessati ad avere un'idea della variabilità del parametro e non del rapporto tra le  $n$  varietà del parametro).

## Capitolo 4

# Statistiche descrittive (B2)

1

I dati si possono descrivere graficamente o attraverso indici (statistiche) sintetiche.

I parametri delle popolazioni si indicano con lettere greche, mentre per le stime si usa il normale alfabeto.

Utile prima rappresentare i dati graficamente e poi in base ai risultati scegliere le statistiche descrittive adatte.

Una distribuzione può essere:

- **uniforme:** tutto il range di variazione della  $x$  è rappresentato in maniera uniforme nella popolazione;
- **a campana:** i valori centrali del range di distribuzione della  $x$  sono i più frequenti nella popolazione (molte biometrie hanno tale distribuzione);
- **asimmetrica:** valori di  $x$  più frequenti lontani dai valori centrali;
- **bimodale:** i valori di  $x$  più frequenti corrispondono a due diversi punti del range di distribuzione di  $x$ ; porta a concludere che nella popolazione ci sono in realtà due popolazioni statistiche.

Per descrivere numericamente una distribuzione si deve tener conto di due sue caratteristiche:

- **tendenza centrale:** la misura che meglio descrive la maggior parte dei dati o che comunque esprime una loro centralità. Indici della tendenza centrale sono: quartili (misure di posizione) e media.
- **dispersione:** indica come le misure si collocano attorno a quella di tendenza centrale, rappresenta la variabilità della popolazione statistica in relazione alla variabile misurata. Misure di dispersione sono: distanza interquartile (misura di posizione), scarto medio assoluto, varianza, deviazione standard.

Le misure di posizione misurano la posizione dei valori, non le differenze tra essi in termini numerici.

### 4.1 Misure di tendenza centrale

Su una serie di dati ordinati in ordine crescente, i **quartili** sono i tre valori che dividono la distribuzione in quattro blocchi, ognuno dei quali rappresenta un quarto complessivo dei record (25%). Il valore centrale è la **mediana** (o secondo quartile, il valore centrale che divide la sequenza di dati ordinati in due gruppi di eguale numerosità, 50%), gli altri due sono il primo (il 25-esimo percentile) ed il terzo quartile (il 75-esimo percentile). Per calcolare la mediana:

---

<sup>1</sup>Capitolo 3, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

- se il numero di osservazioni ( $n$ ) è **dispari**, allora la mediana è l'osservazione centrale:

$$mediana = Y_{\frac{n+1}{2}} \quad (4.1)$$

- se il numero di osservazioni ( $n$ ) è **pari**, allora la mediana è la media della coppia centrale di valori:

$$mediana = \frac{Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1}}{2} \quad (4.2)$$

Primo e terzo quartile non sono necessariamente simmetrici rispetto alla mediana, non si può sapere, solo conoscendo i loro valori, quanto essi siano distanti dalla mediana.

In generale, si parla di **quantili**, statistiche descrittive che indicano la posizione del valore nella distribuzione.

I **decili** dividono la serie ordinata di numeri in blocchi di 10%. I **percentili** possono corrispondere ad una percentuale arbitraria.

I **boxplot** (box & whiskers) sono grafici che rappresentano la variabilità dei valori rispetto ai quartili.

I punti che giacciono al di fuori della distribuzione sono gli **outliers**. Si usa come criterio l'ampiezza del box: un outlier è considerato tale quando dista dal box (solitamente) più di una volta e mezzo l'ampiezza del box (si possono scegliere anche altre distanze, più o meno permissive, in base ad un ragionamento sensato).

I **baffi** sono i valori estremi misurati (tra inizio distribuzione e primo quartile e tra terzo quartile ed ultimo valore), di solito il minimo ed il massimo, una volta esclusi gli outliers.

Il boxplot è, dunque, una rappresentazione grafica che NON assume la simmetria della distribuzione. La **media** è la somma di tutti i valori osservati diviso il loro numero.

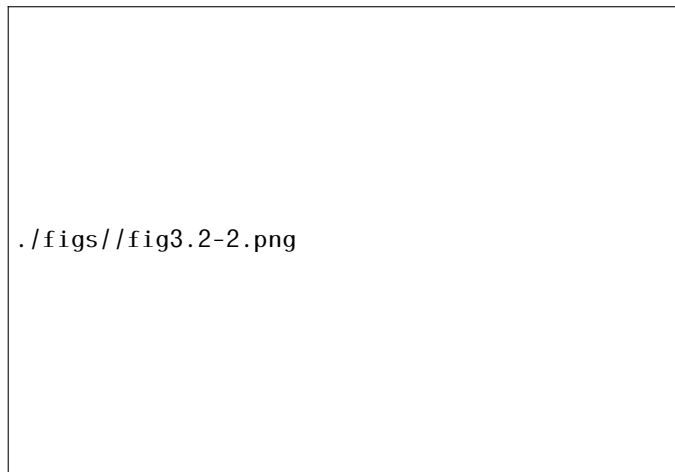


Figura 4.1:

La **media ponderata** tiene conto del “peso” di ciascun valore, si ha quando la variabile continua è riclassificata in una variabile ordinale.

Si attribuisce a ciascuna classe un valore, spesso il valore medio dell'intervallo, che si moltiplica per il numero di soggetti che hanno tale valore medio. La media ponderata è meno precisa di quella aritmetica classica.

La media non discrimina tra distribuzioni con la stessa tendenza centrale ma diversa dispersione.

## 4.2 Misure di dispersione

La **distanza interquartile** è la misura di variabilità quando si usano statistiche descrittive basate sui quartili.

Per calcolare la dispersione:

- **scarto medio assoluto (Sm)**: media di quanto ciascun caso dista dal valore medio. Data la simmetria attorno alla media, i valori positivi e quelli negativi si annullerebbero, per questo è necessario considerare i valori assoluti, per cui si deve calcolare il quadrato e poi fare la radice.

$$S_m = \frac{\sum |x_i - \bar{x}|}{n} \quad (4.3)$$

- **devianza (SQ)**: si sommano le differenze ottenute da ciascun valore, elevate alla seconda per evitare che si ottenga un valore nullo; la devianza non è pesata sul numero di casi, cosa necessaria quando si vogliono confrontare le dispersioni di popolazioni con diverso N.

$$SQ = \sum (x_i - \bar{x})^2 \quad (4.4)$$

- **varianza (QM)**: permette di pesare la devianza sul numero di casi considerati. Si usa spesso  $n-1$  al denominatore perché si fa una correzione: più la numerosità è elevata più la varianza campionaria è affidabile e la correzione non ha peso; quando l' $N$  campionario è molto basso, invece, la correzione permette di non rischiare di avere una sottostima della variabilità. Tale correzione ha senso quando si fanno stime della varianza, non quando questa viene calcolata esattamente (ovvero quando si hanno i dati riguardanti tutta la popolazione statistica).

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad (4.5)$$

- **deviazione standard (o scarto quadratico medio, DS o SD)**: misura più utilizzata nelle statistiche descrittive.

$$s = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}} \quad (4.6)$$

### 4.3 Coefficiente di variazione

Più la dimensione campionaria è bassa, più la variabilità misurata è elevata, quindi a volte non è immediato confrontare variabilità di distribuzioni diverse (di stessa natura o di natura diversa); esistono delle statistiche di variabilità che consentono di misurare la variabilità rispetto al valore della media; una di queste è il **coefficiente di variazione**:

$$CV = \frac{\sigma}{\mu} 100 \quad (4.7)$$

In cui  $\sigma$  è la deviazione standard,  $\mu$  è la media, ed il loro rapporto viene moltiplicato per 100 perché è comodo usare le percentuali.

### 4.4 Range ed intervallo

Altre misure di dispersione sono i range e l'intervallo.

Il **range** è la differenza tra il valore massimo ed il minimo della serie di dati, mentre l'**intervallo** definisce quali sono il valore massimo e quello minimo.

### 4.5 Confronto tra misure di posizione e dispersione

Queste misure alternative di posizione e di dispersione danno informazioni simili quando la frequenza di distribuzione è simmetrica e unimodale, ma la media e la deviazione standard diventano meno informative rispetto alla mediana e alla differenza interquartile quando i dati includono osservazioni estreme.

Alcune condizioni che si verificano quando la distribuzione è normale:

- somma degli scarti è pari a zero (quindi la media è un punto di equilibrio, mentre la mediana solo di centralità);
- devianza è un minimo (la distanza di tutti i punti da qualunque altro valore che non sia la media è maggiore della devianza);
- media delle medie ponderate è pari alla media (associatività).

## 4.6 Mode lasche

La **moda** è il valore più frequente della distribuzione.

Moda e mediana vengono dette “**mode lasche**”, perché danno relativamente poche informazioni.

Nelle distribuzioni normali, media, moda e mediana corrispondono.

## 4.7 Distribuzione campionaria

La **stima** è il processo con cui si inferisce un parametro della popolazione a partire dai dati ottenuti da un campione. Si utilizza il termine stima, però, anche per indicare il risultato del processo di stima.

La stima non è quasi mai esattamente uguale al valore del parametro della popolazione che viene valutato, perché il campionamento è influenzato dal caso.

Per conoscere la precisione della stima si deve conoscere a grandi linee come il processo di campionamento potrebbe influenzare la stima che otteniamo. Si usa a questo scopo la distribuzione campionaria della stima, cioè la distribuzione di probabilità di tutti i valori di una stima che si potrebbero ottenere campionando molte volte la popolazione.

La distribuzione campionaria rappresenta la “popolazione” di valori per una stima.

Non si tratta di una popolazione reale, ma una popolazione immaginaria di valori per una stima.

La dispersione della distribuzione campionaria di una stima dipende dalla dimensione del campione.

Più grande è la dimensione del campione, più stretta è la distribuzione campionaria.

E, più stretta è la distribuzione campionaria, più precisa sarà la stima, visto che una singola stima sarà mediamente più vicina al valore vero.

La distribuzione della media campionaria è normale anche se la distribuzione iniziale non lo è.

## 4.8 Errore standard

La distribuzione campionaria si usa per misurare la precisione di una stima.

La deviazione standard della distribuzione campionaria di una stima è detta **errore standard (ES)**. Dato che riflette le differenze tra una stima e il parametro preso in considerazione, l'errore standard misura la precisione della stima. Le stime affette da errori standard più piccoli sono più precise di quelle affette da errori standard più grandi: minore è l'errore standard, minore è l'incertezza riguardo al parametro della popolazione.

$$ES = \frac{s}{\sqrt{n}} \quad (4.8)$$

In cui  $s$  è la deviazione standard della stima.

Quando si usa la deviazione standard, dato che questa si basa sulla media, si sta assumendo che la distribuzione almeno tenda alla normalità.

Tutte le statistiche parametriche si basano sulla media; nel momento in cui la media non è un buon descrittore, non si possono usare più statistiche parametriche.

Quando si riporta una media, si deve almeno indicare l'errore standard.





## Capitolo 5

# Distribuzione normale (B2)

<sup>1</sup>

Con  $n$  molto alti ci si può permettere di avere bin di ampiezza molto ridotta che, una volta rappresentati, possono essere approssimati da una curva.

La misura di tendenza centrale più informativa nei casi di distribuzione normale è la media aritmetica, che in tali casi corrisponde con moda e mediana.

La distribuzione normale è molto comune in natura (molte biometrie seguono una distribuzione normale), probabilmente perché i processi che si osservano in natura sono il risultato di un processo di selezione.

I parametri che determinano la forma delle curve normali sono media e varianza ( $\sigma^2$ ), mentre gli altri sono costanti.

Le curve **platicurtiche** sono schiacciate (es. viola), mentre le **leptocurtiche** sono appuntite (es. rossa). Le curve normali sono asintotiche rispetto all'asse delle  $x$ .

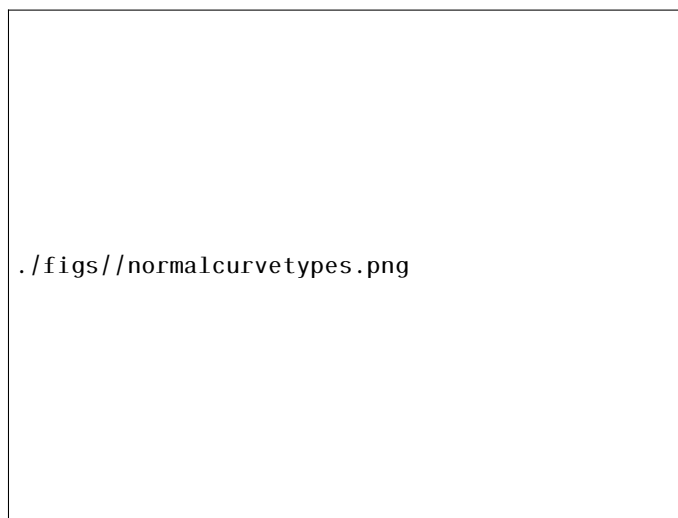


Figura 5.1:

La **curva normale standardizzata**, che ha media 0 e varianza 1, si può ottenere a partire da una distribuzione normale qualsiasi di una qualsiasi variabile continua. Considerando una curva

---

<sup>1</sup>Capitolo 10, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

normale, si trasforma ogni valore della distribuzione ( $y_i$ ) in unità di deviazione standard ( $Z$ ):

$$Z = \frac{y_i - \bar{y}}{s} \quad (5.1)$$

In cui  $s$  è la deviazione standard osservata per il campione.

Il valore ottenuto fa capire se il valore corrispondente sta sopra o sotto la media rispetto ad una misura di variabilità. Con riferimento alla distribuzione standardizzata, il 95% dei valori ricade

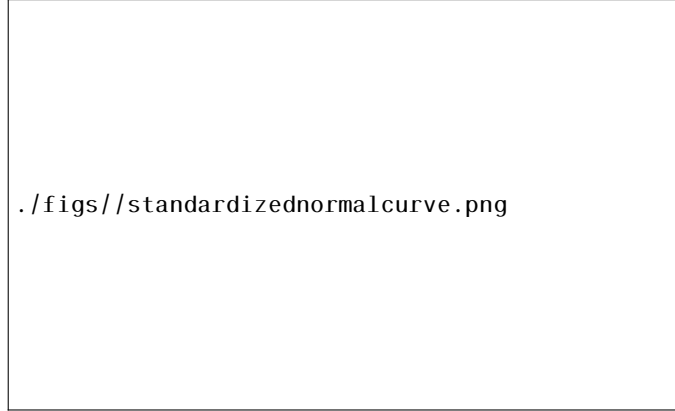


Figura 5.2: no 1.98 ma 1.96 volte  $\sigma$

all'interno dell'intervallo che si estende tra -1.96 volte la deviazione standard e +1.96 volte la deviazione standard, mentre tra -1 e +1 volta la deviazione standard si ha il 66.27% dei casi. La curva normale standardizzata ha proprietà definite.

In un boxplot di una distribuzione normale, il 50% dei casi è racchiuso nel box.

Queste considerazioni valgono anche per la distribuzione campionaria delle medie.

## 5.1 Teorema del limite centrale

La sommatoria di molte distribuzioni distribuite in maniera non normale genera una distribuzione normale.

Secondo il **teorema del limite centrale**, la somma o la media di  $n$  misure da una distribuzione quasi o non normale genera una distribuzione normale.

## 5.2 Intervallo di confidenza

L'errore standard è la misura della ds nella media campionaria. Siccome, per la media campionaria, l'errore standard è uguale alla ds per una variabile qualsiasi, anche in questo modo si può indicare la precisione della media.

Per la curva normale, se si campiona  $n$  volte, nel 95% dei casi, la media avrà un intervallo di confidenza che ricadrà tra  $media - 1.96 * ES$  e  $media + 1.96 * ES$ .

In questo modo gli IC possono essere usati come metodo inferenziale.

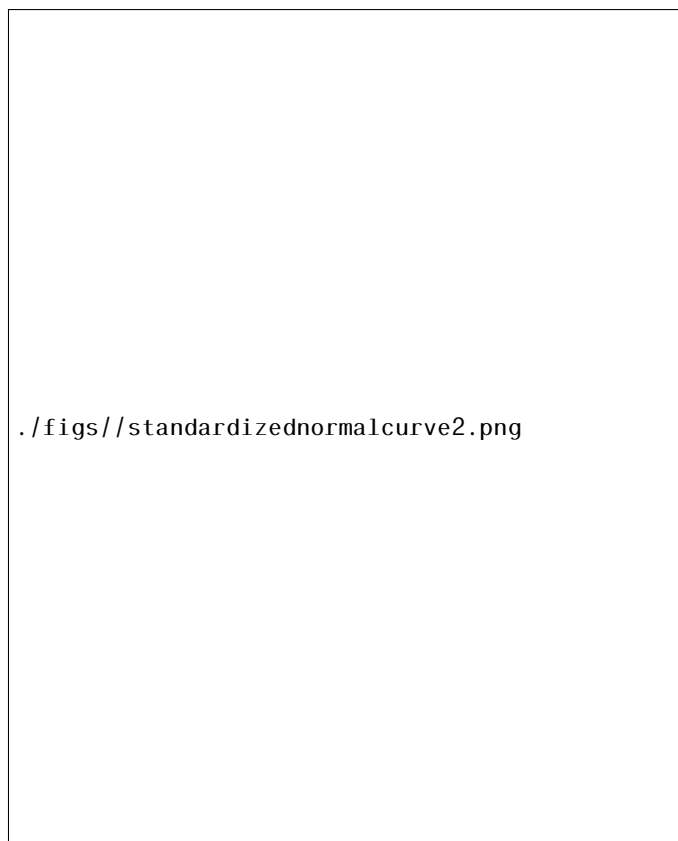


Figura 5.3:



# Capitolo 6

## Probabilità (B3)

1

Lo **spazio campionario** è la lista di tutti i possibili risultati di una prova casuale.

Una **prova** è **casuale** quando è aleatoria.

Un **evento** è un qualsiasi sottoinsieme potenziale dello spazio campionario. Per esempio, se lanciamo un dado a 6 facce, sono possibili 6 risultati: i numeri da 1 a 6. Questi 6 differenti numeri rappresentano nel loro insieme lo spazio campionario. Possiamo definire molti eventi che potrebbero interessare, quali "il risultato è un numero pari", "il risultato è un numero maggiore di 3" o anche il semplice evento "il risultato è 4".

La **probabilità** di un evento è la proporzione di tutte le prove casuali in cui si verifica l'evento specificato rispetto a tutti gli eventi possibili, quando la stessa prova casuale viene ripetuta un tot di volte in condizioni definite.

La proporzione di casi di uno degli eventi dello spazio campionario sul totale degli eventi possibili è la migliore stima della probabilità di tale evento.

Le probabilità, essendo proporzioni, devono essere sempre comprese tra 0 e 1, estremi inclusi: un evento ha probabilità 0 se non si verifica mai e ha probabilità 1 se si verifica sempre.

La sommatoria di tutte le probabilità di tutti gli eventi possibili è pari a 1.

Es. lanciando un dado, ogni lancio (prova casuale) può risultare in una qualsiasi delle sei facce del dado (spazio campionario), e c'è una probabilità di  $\frac{1}{6}$  che una determinata faccia esca.

Un evento può essere anche una combinazione di più eventi (es. lanciando il dado due volte, probabilità che esca un 2 ed un 5?), ed è definito da una probabilità (quante volte si osserva un dato evento sul totale delle volte che si compie l'osservazione).

### 6.1 Istogrammi di frequenza

Si potrebbe creare un istogramma di frequenza relativa per descrivere la probabilità di un evento.

### 6.2 Eventi indipendenti

Due eventi si dicono **indipendenti** se il verificarsi di uno dei due non influenza in alcun modo la probabilità che si verifichi anche l'altro.

### 6.3 Eventi incompatibili

Quando due eventi sono indipendenti e mutualmente esclusivi, ovvero non possono verificarsi simultaneamente, si parla di **eventi incompatibili**. Per esempio, un singolo dado lanciato una

---

<sup>1</sup>Capitolo 5, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

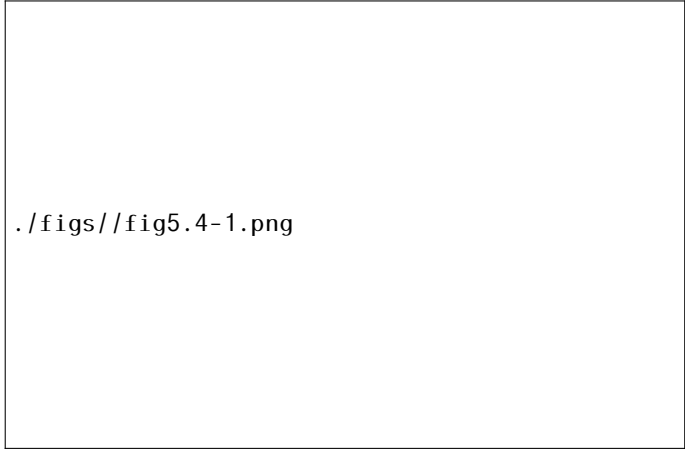


Figura 6.1:

sola volta non può presentare simultaneamente sia 1 che 6: gli eventi "1" e "6" sono incompatibili. In termini matematici, due eventi sono incompatibili se  $Pr[A \text{ e } B] = 0$ , in cui  $Pr[A \text{ e } B]$  significa la probabilità che si verifichi sia A che B.

## 6.4 Somma delle probabilità

Se gli eventi che si tenta di combinare sono incompatibili, allora il calcolo della probabilità che si verifichi l'uno o l'altro è semplicemente la somma delle probabilità per ciascuno di quegli eventi considerati separatamente. Ottenere 7 e ottenere 11 lanciando due dadi sono eventi incompatibili, perciò, la probabilità di ottenere 7 o 11 lanciando due dadi è la probabilità che esca 7 più la probabilità che esca 11:  $Pr[\text{ottenere 7 o ottenere 11}] = Pr[\text{ottenere 7}] + Pr[\text{ottenere 11}]$ .

Questa proprietà additiva delle probabilità di eventi incompatibili è detta **regola della somma** (un caso particolare del teorema della somma delle probabilità).

Dunque, secondo la regola della somma, se due eventi A e B sono incompatibili, allora:

$$Pr[A \text{ o } B] = Pr[A] + Pr[B] \quad (6.1)$$

La regola della somma può essere estesa a più di due eventi, purché essi siano tutti incompatibili.

La somma delle probabilità di tutti i possibili risultati incompatibili di una prova casuale deve essere uguale ad 1; ciò significa che la probabilità che un evento non si verifichi è semplicemente 1 meno la probabilità che si verifichi.

Se due eventi non sono incompatibili (non mutualmente esclusivi ma indipendenti), la probabilità che l'uno oppure l'altro evento possa verificarsi viene data dalla **regola della somma generalizzata**:

$$Pr[A \text{ o } B] = Pr[A] + Pr[B] - Pr[A \text{ e } B] \quad (6.2)$$

Se non si sottraesse la probabilità che si verifichino sia A che B, si conterebbero due volte i risultati in cui gli eventi si verificano contemporaneamente. In realtà, la regola potrebbe essere applicata anche a casi in cui si hanno eventi mutualmente esclusivi, perché

## 6.5 Diagrammi di Venn

Un modo utile per ragionare sulle probabilità degli eventi è usare uno strumento grafico, detto **diagramma di Venn**. L'area del diagramma di Venn rappresenta tutti i risultati possibili di una prova casuale, e si possono rappresentare vari eventi come aree del diagramma.

Se il box per ciascun risultato ha la stessa area, i risultati sono equiprobabili. Es. probabilità che, lanciando due dadi a 6 facce, si ottengano due numeri che, sommati, restituiscono un valore di 7

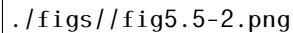


Figura 6.2:

o di 11.

La probabilità di ottenere 7 o 11 è additiva; in questo caso, la funzione di probabilità non è più uniforme, ma segue una curva a campana.

Dunque, le probabilità di eventi alternativi ma non incompatibili sono additive.

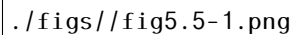


Figura 6.3:

## 6.6 Prodotto di probabilità

Quando due eventi sono indipendenti, la probabilità che si verifichino entrambi è data dal prodotto della probabilità del primo evento per la probabilità del secondo. Questa proprietà degli eventi indipendenti, detta regola del prodotto (un caso particolare del teorema della probabilità composta), è essenziale per l'analisi dei dati perché permette di determinare se due o più variabili siano associate.

Secondo la regola del prodotto, se due eventi A e B sono indipendenti, allora:

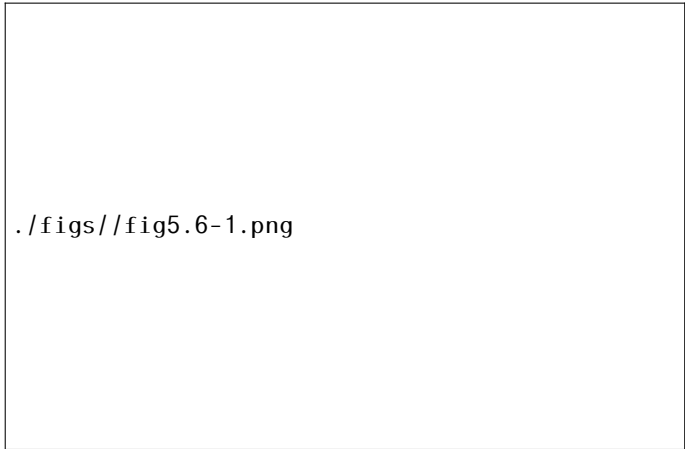
$$Pr[A \text{ e } B] = Pr[A] \times Pr[B] \quad (6.3)$$

La regola del prodotto riguarda combinazioni che implicano "e", cioè la probabilità che si verifichino entrambi gli eventi. Volendo conoscere la probabilità che si verifichino questo e quello, e se i due eventi sono indipendenti, possiamo moltiplicare le probabilità di ciascuno di essi per



./figs//istogrammaperdiagramma.png

Figura 6.4:



./figs//fig5.6-1.png

Figura 6.5:

ottenere la probabilità che si verifichino entrambi.

OPPURE

La probabilità di ottenere due eventi indipendenti mutualmente esclusivi si ottiene invece calcolando il prodotto delle probabilità dei singoli eventi.

Es. probabilità che su due lanci esca 3 e 3.

Nel caso di eventi indipendenti non mutualmente esclusivi si calcola il prodotto dei due o più eventi per ottenere la probabilità della combinazione degli eventi considerati.

#### Esempio 6.6.a: slot machines

Con una slot machine si vince il 9.8% delle volte ( $Pr[vincita] = 0.098$ ), indipendentemente dalla giocata precedente. Qual è la probabilità che un giocatore perda 8 volte di seguito?

$$Pr[perdita] = 1 - Pr[vincita] = 0.902 \quad (6.4)$$

Vincita e perdita sono incompatibili, quindi:

$$Pr[8\ perdita] = Pr[perdita]^8 = 0.438 \quad (6.5)$$



## 6.7 Alberi di probabilità

Un **albero di probabilità**, o albero probabilistico, è un diagramma che può essere utilizzato per calcolare le probabilità di combinazioni di eventi che sono i risultati di più prove casuali.

## 6.8 Probabilità condizionata e teorema della probabilità totale

La **probabilità condizionata** (o subordinata) è la probabilità di un evento condizionata dal verificarsi di un altro evento.

La probabilità condizionata si indica come:  $\Pr[\text{evento}|\text{condizione}] = \text{probabilità}$ . Il che indica la probabilità che si verifichi l'evento quando la condizione è soddisfatta.

La barra verticale al centro di questa espressione sta per "dato che" o "quando la condizione seguente è soddisfatta".

Volendo conoscere la probabilità complessiva di un particolare evento, si devono sommare tutte le sue probabilità per ogni possibile condizione, pesate sulla probabilità di quella condizione. Questa regola è nota come **teorema della probabilità totale (o globale)**, secondo cui la probabilità di un evento  $X$  è:

$$\Pr[X] = \sum \Pr[Y_i] \Pr[X|Y_i] \quad (6.6)$$

## 6.9 Teorema della probabilità composta

Le probabilità condizionate ci permettono di calcolare la probabilità di una combinazione di due eventi anche se essi non sono indipendenti. In questa situazione, possiamo valutare la probabilità che si verifichino entrambi moltiplicando la probabilità di un evento per la probabilità condizionata dell'altro evento, ovvero per la probabilità dell'altro evento condizionata al verificarsi del primo evento. Si tratta della regola del prodotto generalizzata, detta anche **teorema della probabilità composta**.

La regola del prodotto generalizzata permette di trovare la probabilità che si verifichino due eventi anche quando sono dipendenti:

$$\Pr[A \text{ e } B] = \Pr[A]\Pr[B|A] \quad (6.7)$$

Riflettendo attentamente, possiamo renderci conto di come questa regola sia ragionevole. Affinché due eventi ( $A$  e  $B$ ) si verifichino, deve verificarsi l'evento  $A$ . Per definizione, ciò avviene con probabilità  $\Pr[A]$ . Sapendo che  $A$  si è verificato, la probabilità che si verifichi anche  $B$  è  $\Pr[B|A]$ . Moltiplicando tra loro queste probabilità otteniamo la probabilità che si verifichino sia  $A$  sia  $B$ .

Vale anche l'inverso:

$$\Pr[A \text{ e } B] = \Pr[B]\Pr[A|B] \quad (6.8)$$

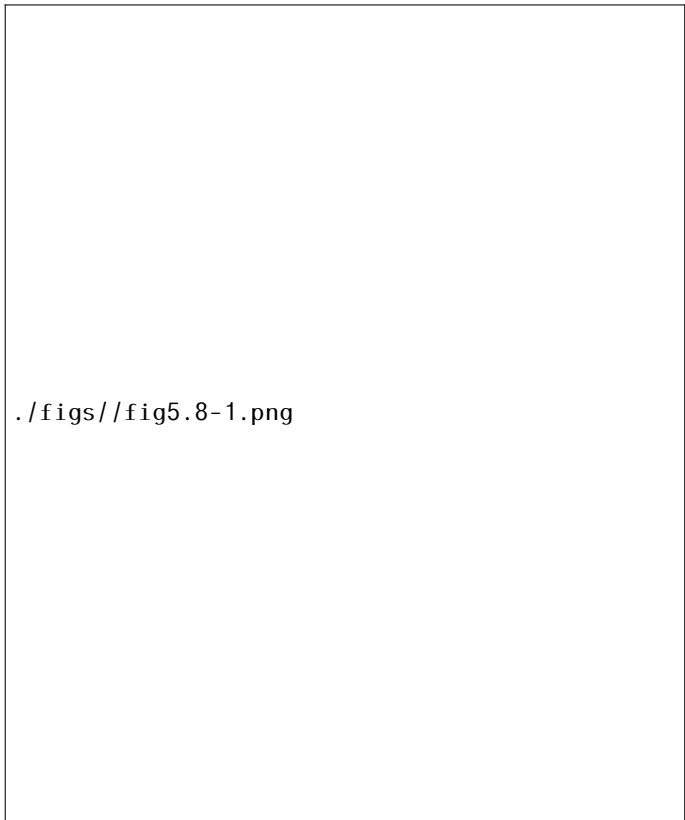
Dunque, il teorema della probabilità composta è applicabile anche ad eventi non indipendenti ed è simmetrico.

### Esempio 6.9.a: *Nasonia vitripennis*

La vespa parassita *Nasonia vitripennis* depone uova in larve di mosca decidendone il sesso in base a se la larva è già parassitata (90% maschi perché questi riusciranno a fecondare tutte le uova delle altre femmine competendo bene con i maschi dell'altra parassitazione) o no (soprattutto femmine e pochi maschi perché questi pochi potrebbero fecondare tutte le femmine). In questi casi la probabilità di parassitosi è dello 0.2 (assumendo che questa sia la stessa per larve già parassitate e non). Il problema può essere rappresentato mediante diagramma di Venn (in questo caso molto simile ad un mosaic plot) o albero delle probabilità.

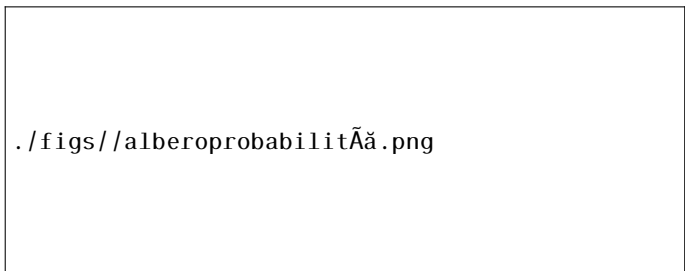
Qual è la probabilità che un uovo sia maschio? Sommando le due probabilità relative:

$$\begin{aligned} \Pr[M] &= \Pr[\text{ospite parassitato}] \Pr[M|\text{ospite parassitato}] + \\ &+ \Pr[\text{ospite non parassitato}] \Pr[M|\text{ospite non parassitato}] \end{aligned} \quad (6.9)$$



./figs//fig5.8-1.png

Figura 6.6:



./figs//alberoprobabilit .png

Figura 6.7:

## 6.10 Teorema di Bayes

Partendo dalla regola del prodotto generalizzata, dato che:

$$Pr[A \text{ e } B] = Pr[A] Pr[B|A] \quad (6.10)$$

E che:

$$Pr[B \text{ e } A] = Pr[B] Pr[A|B] \quad (6.11)$$

Allora è anche vero che:

$$Pr[B] * Pr[A|B] = Pr[A] Pr[B|A] \quad (6.12)$$

Dividendo entrambi i membri per  $Pr[B]$  si ottiene il **teorema di Bayes** (sostituire E con B):

$$Pr[A|B] = \frac{Pr[B|A] Pr[A]}{Pr[B]} \quad (6.13)$$

Ovvero: la probabilità di un evento, dato un altro evento che la condiziona, è uguale alla probabilità dell'evento per la probabilità dell'evento che lo condiziona, dato l'evento iniziale, diviso la probabilità dell'evento che lo condiziona.

**Esempio 6.10.a: sindrome di Down**

La prevalenza (frequenza di casi che si ha su un certo numero di eventi) della sindrome di Down è di  $\frac{1}{1000}$  ( $Pr[DS] = \frac{1}{1000} = 0.001$ ) feti.

L'amniocentesi porta ad aborto in  $\frac{1}{200}$ .

Il Triplo Test (TT) è un'alternativa all'amniocentesi che ha una sensibilità (probabilità di rilevare un vero positivo) di 0.6 (60% dei casi rileva un feto con DS,  $Pr[TTpos | DS] = 0.6$ ), e 0.05 della probabilità di avere un falso positivo (5% di avere un falso positivo).

Qual è la probabilità che si abbia un vero positivo quando si ottiene un TT positivo ( $Pr[DS | TTpos]$ )?

Applicando il teorema di Bayes:

$$Pr[TTpos] = Pr[TTpos|DS] Pr[DS] + Pr[TTpos|no DS] Pr[no DS] = 0.05055 \quad (6.14)$$

$$Pr[DS|TTpos] = \frac{Pr[DS] Pr[TTpos|DS]}{Pr[TTpos]} = 0.012 \quad (6.15)$$



# Capitolo 7

## Test delle ipotesi

1

Una teoria non confutabile non è necessariamente falsa, ma non rientra nell'ambito scientifico, bensì in quello metafisico.

Il neopositivismo impiega il criterio di verificabilità: per essere corretta, una teoria deve essere verificata, il che può essere fatto mediante la logica.

Non potendo fare esperienza dell'universale si deve adottare l'**approccio falsificazionista**: bisogna provare che un'ipotesi sia sbagliata per smentirla.

Un'ipotesi non si verifica e non si può accettare, si possono portare prove a suo sostegno e si può falsificare.

### Esempio 7.0.a: sesso e altezza

Data un'altezza media dei maschi pari ad  $\bar{y}_m = 175.6 \text{ cm}$ , con deviazione standard  $s_m = 7.1$ , ed un'altezza media delle femmine pari ad  $\bar{y}_f = 162.6 \text{ cm}$ , con deviazione standard  $s_f = 6.4$ , i maschi sono davvero più alti delle femmine?

Un valore di 193 cm quanto è anomalo nella popolazione maschile?

$$193 - 175.6/7.1 = 2.\text{qualcosa} \quad (7.1)$$

Dunque, dato che soglia oltre la quale si ha solo il 5% dei casi, avere 193 è solo in meno del 5% dei casi.

(stessa cosa per femmine e 155 cm?)

## 7.1 Ragionamento logico

Il ragionamento scientifico può seguire percorsi alternativi:

## 7.2 Metodo scientifico galileiano

Il metodo scientifico si basa su una combinazione di processi induttivi e deduttivi; può essere infatti scomposto in due fasi:

- **induttiva** (dal particolare all'universale):
  - osservazione;
  - formulazione di ipotesi;
- **deduttiva** (dall'universale al particolare):
  - formulazione della teoria e dimostrazione matematica;
  - verifica sperimentale.

---

<sup>1</sup>Capitolo 6, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

<b>Deduzione</b>	Dall'universale al particolare attraverso postulati e concatenazioni logiche (conseguente).	<b>Regola:</b> tutti i fagioli di questo sacchetto sono bianchi <b>Caso:</b> questi fagioli vengono da questo sacchetto <b>Risultato:</b> questi fagioli sono bianchi
<b>Induzione</b>	Dal particolare all'universale attraverso osservazioni ripetute (non sempre conseguente).	<b>Caso:</b> questi fagioli vengono da questo sacchetto <b>Risultato:</b> questi fagioli sono bianchi <b>Regola:</b> tutti i fagioli di questo sacchetto sono bianchi
<b>Abduzione</b>	Dal particolare all'universale, si tratta più di intuito che di reale logica (non sempre conseguente).	<b>Risultato:</b> questi fagioli sono bianchi <b>Regola:</b> tutti i fagioli di questo sacchetto sono bianchi <b>Caso:</b> questi fagioli vengono da questo sacchetto (non è assolutamente detto)

Tabella 7.1:

### 7.3 Formulazione delle ipotesi

Per formulare un'ipotesi biologica bisogna definire:

- **enunciato** chiaro e specifico;
- **razionale**, ciò che giustifichi l'ipotesi nella sua formulazione (alla fine dell'introduzione si elencano le ipotesi, il razionale si evince dal resto dell'introduzione);
- **previsione**, che deve essere precisa e misurabile.

### 7.4 Verifica delle ipotesi

La verifica formale delle ipotesi parte dall'enunciazione chiara di due ipotesi, l'ipotesi nulla e l'ipotesi alternativa, riguardo a un parametro della popolazione. L'ipotesi nulla si può considerare l'ipotesi di base, di partenza, mentre l'ipotesi alternativa comprende in genere ogni altra possibilità rispetto a quelle enunciate nell'ipotesi nulla.

L'**ipotesi nulla** è un enunciato specifico riguardo al valore di un parametro della popolazione. Viene formulata ai fini del ragionamento e include spesso il punto di vista scettico. L'ipotesi nulla spesso afferma che un parametro nella popolazione di interesse è pari a zero (cioè, nessun effetto, nessuna preferenza, nessuna correlazione o nessuna differenza). In generale, questa ipotesi è un enunciato che sarebbe interessante rifiutare. L'ipotesi nulla, chiamata anche ipotesi zero, o semplicemente H-zero, e indicata con il simbolo  $H_0$ , è sempre specifica: identifica un particolare valore del parametro studiato.

Ogni ipotesi nulla è associata a un'**ipotesi alternativa** ( $H_A$ ) che generalmente rappresenta tutti gli altri valori possibili di un parametro rispetto a quelli previsti nell'ipotesi nulla.

Per questo motivo spesso coincide, ma non sempre, con l'ipotesi che il ricercatore spera essere vera.

A differenza dell'ipotesi nulla, l'ipotesi alternativa non è specifica. Essa include ogni valore possibile per una caratteristica della popolazione o per una differenza tra popolazioni, eccettuato il valore specificato dall'ipotesi nulla.

L'ipotesi alternativa può essere a una o due code: a una coda se è accettabile solo il cambiamento in una delle due direzioni, a due code se vanno bene entrambe le direzioni.

Se un'ipotesi alternativa sia ad **una** o **due code** si deve decidere a priori e la decisione deve essere supportata da un forte razionale.

L'ipotesi nulla è l'unica ipotesi con la quale vengono effettivamente messi alla prova i dati. Se

i dati sono compatibili con l'ipotesi nulla, allora diciamo che non possiamo rifiutarla (non "accettiamo" mai l'ipotesi nulla). Se i dati sono invece incompatibili, la scartiamo e accettiamo l'ipotesi alternativa. Rifiutare  $H_0$ , significa escludere il valore che  $H_0$  specifica. Il rifiuto dice anche in quale direzione si trova verosimilmente il valore vero rispetto a quello specificato dall'ipotesi nulla, senza però suggerire nulla sul valore numerico del parametro della popolazione. Per stimare i valori dei parametri e il loro errore usiamo il processo di stima.

#### Esempio 7.4.a: lateralità nei rospi

Gli esseri umani sono in prevalenza destrimani. Questa lateralità è presente anche in altri animali? Bisazza et al. (1996) hanno studiato a tal proposito il rospo comune (*Bufo bufo*), campionando 18 rospi selvatici e sottoponendoli a un esperimento. Assumiamo prima di tutto che il campione fosse casuale. I rospi furono trasferiti in laboratorio e furono sottoposti, uno alla volta, alla stessa umiliazione: venne avvolto un palloncino attorno alla testa di ogni individuo. Poi i ricercatori registrarono quale dei due arti anteriori ogni rospo usava per togliersi il palloncino dalla testa, trovando che ogni rospo tendeva a usare uno più dell'altro. A questo punto ricercatori si chiesero: i rospi destrimani e quelli mancini sono presenti con uguale frequenza nella popolazione dei rospi, oppure uno dei due tipi di lateralità è più frequente dell'altro, come nell'uomo? Dei 18 rospi sottoposti al test, 14 erano destrimani e 4 erano mancini. Questi risultati si possono considerare una chiara evidenza che nei rospi uno dei due tipi di lateralità prevale sull'altro?

Il valore numerico di interesse è la frazione di rospi destrimani nella popolazione. Chiamiamo  $p$  questa proporzione. Secondo l'ipotesi di partenza ( $H_0$ ), i rospi mancini e quelli destrimani sono ugualmente frequenti nella popolazione ( $p = 0.5$ ). Se l'ipotesi nulla è falsa, allora nei rospi prevale l'utilizzo di una zampa, dunque, secondo  $H_A$ , i rospi mancini e quelli destrimani non sono ugualmente frequenti nella popolazione ( $p \neq 0.5$ ).

L'ipotesi alternativa è bilaterale (o a due code). Ciò significa semplicemente che essa contempla due possibilità: che  $p$  sia maggiore di 0,5 (i rospi destrimani sono più frequenti di quelli mancini nella popolazione) oppure che  $p$  sia minore di 0,5 (i rospi mancini predominano). Nessuna delle due possibilità può essere esclusa prima dell'esperimento, quindi entrambe devono essere incluse nell'ipotesi alternativa. In un test bilaterale (o test a due code), l'ipotesi alternativa include valori da entrambi i lati (maggiori e minori) del valore specificato dall'ipotesi nulla. La definizione "bilaterale" deriva dal fatto che l'ipotesi alternativa contiene valori da entrambe le parti rispetto al valore specificato dall'ipotesi nulla, mentre la definizione "a due code" si riferisce alle code della distribuzione campionaria, dove una "coda" è la regione all'estremità superiore o all'estremità inferiore della distribuzione.

## 7.5 Statistica test

La **statistica test** è una grandezza, calcolata sulla base dei dati osservati, che viene utilizzata per valutare il grado di compatibilità dei risultati con il risultato che ci aspettiamo se fosse vera l'ipotesi nulla.

Serve una statistica test che rappresenti la previsione per verificare le ipotesi.

Nello studio sui rospi usiamo come statistica test il numero osservato di rospi destrimani.

Purtroppo non sempre i dati rispecchiano perfettamente la realtà. A causa degli effetti del caso durante il campionamento, non ci attendiamo di osservare esattamente 9 rospi destrimani quando campioniamo 18 rospi dalla popolazione, anche quando l'ipotesi nulla è vera. Anche in questo caso, infatti, esiste generalmente una discrepanza, dovuta al caso, tra il risultato osservato e quello atteso. Questa discrepanza può essere piuttosto grande, pur essendo  $H_0$  vera, in particolare se i dati non sono molto numerosi. Per decidere se i dati siano compatibili con l'ipotesi nulla, dobbiamo calcolare la probabilità di un disaccordo tanto ampio quanto quello osservato, assumendo che l'ipotesi nulla sia vera.

Per ottenere questa probabilità dobbiamo determinare la distribuzione campionaria della statistica test assumendo che l'ipotesi nulla sia vera. Dobbiamo cioè determinare i valori della statistica

test che sono possibili sotto l'ipotesi nulla e le probabilità a essi associate. La distribuzione di probabilità dei valori della statistica test quando si assume che l'ipotesi nulla sia vera è detta "distribuzione campionaria sotto  $H_0$ " o, più semplicemente, distribuzione nulla.

La distribuzione nulla è la distribuzione campionaria dei possibili valori che può assumere una statistica test quando si ipotizza che sia vera l'ipotesi nulla.

Analizzando questa distribuzione, possiamo concludere che ogni numero di rospi destrimani tra 0 e 18 può essere osservato in un campione casuale di 18 individui quando è vera  $H_0$ , ma alcuni valori sono molto più probabili di altri.

## 7.6 Livello di significatività di P

Il metodo comunemente utilizzato per descrivere il disaccordo tra i dati e un'ipotesi nulla è calcolare la probabilità di ottenere quei dati, oppure di ottenere i dati che sono ancor più diversi da quelli attesi (cioè meno probabili), assumendo vera l'ipotesi nulla. In altre parole, vogliamo conoscere la probabilità complessiva, se fosse vera  $H_0$ , di tutti i risultati ugualmente insoliti o più insoliti rispetto a quelli osservati. Se questa probabilità risulterà piccola, allora l'ipotesi nulla verrà considerata incompatibile con i dati e verrà rifiutata in favore dell'ipotesi alternativa. Se questa probabilità non risulterà piccola, allora sarebbe ingiustificato mettere in dubbio l'ipotesi nulla, che quindi non verrà rifiutata. La probabilità di ottenere i dati osservati, o i dati che sono ancor meno in accordo con l'ipotesi nulla, assumendo che sia vera l'ipotesi nulla, è detta **P-value** (o valore P). Se il P-value è piccolo, allora l'ipotesi nulla è incompatibile con i dati e la rifiutiamo. Altrimenti essa non può essere rifiutata.

Il P-value (o valore P) è la probabilità di ottenere i dati osservati (oppure i dati che hanno una differenza altrettanto grande o ancora maggiore rispetto al valore dell'ipotesi nulla) se l'ipotesi nulla fosse vera (ovvero per caso). In pratica, calcoliamo il P-value in base alla distribuzione nulla della statistica test.

P, riferito ad una curva di distribuzione di frequenza, rappresenta la probabilità di osservare per caso un valore uguale o più o meno estremo di quello che si osserva (quindi quanto è probabile il valore della statistica test se l'ipotesi nulla è vera). Se  $P > 0.05$  (test non significativo, la differenza osservata potrebbe essere semplicemente dovuta alla variabilità naturale o campionaria, non si può escludere che siano dovute al caso) non si può rifiutare l'ipotesi nulla e supportare l'ipotesi alternativa. Se  $P \leq 0.05$  (test significativo \*) allora si rifiuta l'ipotesi nulla e si supporta quella alternativa, a maggior ragione se  $P \leq 0.01$  (test molto significativo \*\*) o 0.001 (test altamente significativo \*\*\*); questi tre livelli sono standard internazionali. (la differenza osservata è difficilmente osservabile per caso).

I valori soglia con cui si confronta il p-value sono indicati con  $\alpha$  ( $0.05$  = meno di 5 volte su 100 si potrebbero trovare valori della statistica test uguali o più estremi,  $0.01$ ,  $0.001$ ).

La nostra ipotesi alternativa  $H_A$  è bilaterale, quindi dobbiamo considerare nel calcolo del P-value i risultati tanto insoliti quanto i dati osservati, ho ancora più insoliti, localizzati su entrambe le code della distribuzione.

La probabilità di avere 14 o più rospi destrimani, sotto  $H_0$ , è pari a:

$$Pr[\geq 14 \text{ destrimani}] = Pr[14] + Pr[15] + Pr[16] + Pr[17] + Pr[18] = 0,0155 \quad (7.2)$$

In cui  $Pr[14]$  è la probabilità di osservare esattamente 14 rospi destrimani. Le probabilità relative a 14, 15, 16, 17 e 18 rospi destrimani possono essere sommate perché i singoli risultati sono tra di loro incompatibili (cioè mutuamente esclusivi). Questa somma non è però il P-value, perché non include ancora i risultati altrettanto estremi giacenti sulla coda sinistra della distribuzione nulla (quelli cioè che implicano una predominanza di rospi mancini). Il metodo più rapido per includere le probabilità dei risultati ugualmente estremi localizzati sull'altra coda della distribuzione è moltiplicare per 2 la somma precedente, ottenendo il P-value desiderato. In altre parole, se assumiamo che l'ipotesi nulla sia vera, la probabilità di un risultato tanto estremo quanto "14 rospi destrimani osservati su 18 campionati", o ancora più estremo, è pari a  $P = 0,031$ .



Se il P-value è "piccolo" rifiutiamo l'ipotesi nulla; altrimenti, non la rifiutiamo. Per convenzione, nella maggior parte dei settori della ricerca biologica il confine tra P-value piccoli e P-value non piccoli è 0,05. In altre parole, se  $P < 0,05$  rifiutiamo l'ipotesi nulla, mentre se  $P > 0,05$  non la rifiutiamo.

Il P-value per i dati sui rospi,  $P = 0,031$ , è in effetti minore di 0,05: rifiutiamo quindi l'ipotesi nulla per la quale i rospi mancini e quelli destrimani sono ugualmente frequenti nella popolazione di rospi. In base a questi dati possiamo concludere che la maggioranza di rospi nella popolazione è destrimane.

Questa soglia di decisione per  $P$  (cioè,  $P = 0,05$ ) è detta livello di significatività, e si indica con  $\alpha$ . In biologia, il livello di significatività più usato è  $\alpha = 0,05$ , ma potrete incontrare alcuni studi che usano un differente valore di  $\alpha$ . Dopo 0,05, il successivo livello di significatività più utilizzato è 0,01.

Quando si devono comunicare i risultati di uno studio con un articolo scientifico o una relazione, è necessario includere sempre il valore della statistica test, la dimensione campionaria ed il P-value. Inoltre, è sempre utile fornire gli intervalli di confidenza, o almeno gli errori standard, per i parametri di interesse.

#### Esempio 7.6.a: voti in biostatistica

Le ipotesi sono:

- $H_0$ : medie di ragazzi e ragazze in biostatistica sono uguali;
- $H_{A1}$ : ragazze hanno voti più alti in biostatistica dei ragazzi;
- $H_{A2}$ : ragazzi hanno voti più alti in biostatistica delle ragazze.

Bisogna confrontare le medie di ragazze e ragazzi; la statistica test in questo caso è il calcolo della differenza tra le medie: se  $H_0$  è verificata, la differenza è pari a zero o quasi (si deve considerare la dispersione, generata dalla variabilità insita nella popolazione e dalla dimensione del campione).

$$\frac{\bar{Y}_F - \bar{Y}_M}{ES_{\bar{Y}_F - \bar{Y}_M}} \quad (7.3)$$

## 7.7 Errori nella verifica delle ipotesi

Rifiutare  $H_0$  non significa necessariamente che l'ipotesi nulla sia falsa.

Analogamente, il mancato rifiuto di  $H_0$ , non significa necessariamente che l'ipotesi nulla sia vera. Le ipotesi riguardano infatti le popolazioni, ma avendo a disposizione solo l'informazione contenuta nei campioni, che sono inevitabilmente influenzati dal caso, è possibile che una conclusione sulle ipotesi sia errata.

Ovviamente non saremo in grado di dire, dopo aver quantificato il P-value e deciso se rifiutare o no l'ipotesi nulla, se abbiamo commesso un errore. È possibile, però, ragionare sugli errori e in parte quantificarli se i dati provengono da un campione casuale, valutando quindi con maggiore consapevolezza la decisione finale.

Quando si rifiuta un'ipotesi nulla vera, si commette un **errore di tipo I**. Se invece non si rifiuta un'ipotesi nulla falsa, allora si commette un **errore di tipo II**.

Il livello di significatività  $\alpha$  corrisponde alla probabilità di commettere un errore di tipo I. Se seguiamo la convenzione e usiamo un livello di significatività 0,05, allora dobbiamo rifiutare  $H_0$  quando  $P$  è inferiore o uguale a 0,05. Ciò significa che, se l'ipotesi nulla fosse vera, la rifiuteremmo erroneamente 1 volta su 20. I biologi considerano accettabile il rischio di commettere un errore di primo tipo.

Si potrebbe ridurre il tasso di errore di tipo I usando un livello di significatività inferiore a 0,05. Tuttavia, questa scelta avrebbe anche l'effetto collaterale di aumentare la probabilità di commettere un errore di tipo II. Ridurre  $\alpha$  rende più difficile rifiutare l'ipotesi nulla quando è vera, ma rende anche più difficile rifiutarla quando è falsa. Questo è un motivo per cui si usa per conven-

zione 0.05.

Il mancato rifiuto di un ipotesi nulla falsa è un errore di tipo II.

Se in un certo studio, applicando una certa statistica, la probabilità di commettere un errore di tipo II è bassa, si dice che il test ha un'elevata potenza. La **potenza** è la probabilità che un campione casuale estratto da una popolazione, quando viene analizzato, determini il rifiuto di un ipotesi nulla falsa.

$$\text{potenza del test} = 1 - \beta \quad (7.4)$$

A parità di tutti gli altri fattori, un test è migliore se ha una potenza più elevata.

Poiché non si conosce mai il valore vero, non si è in grado di prevedere con precisione la potenza del test utilizzato in uno studio.

Quando però si è in grado di fare qualche previsione, anche approssimativa, sull'entità della possibile deviazione dalle ipotesi nulla, allora si può stimare la potenza. In generale, un test ha una potenza maggiore se la dimensione campionaria è grande, se la discrepanza vera dalle ipotesi nulla è grande e se la variabilità della popolazione è bassa.

Si vuole solitamente una potenza di almeno 0.8 e quindi un  $\beta$  di 0.2.



Figura 7.1:

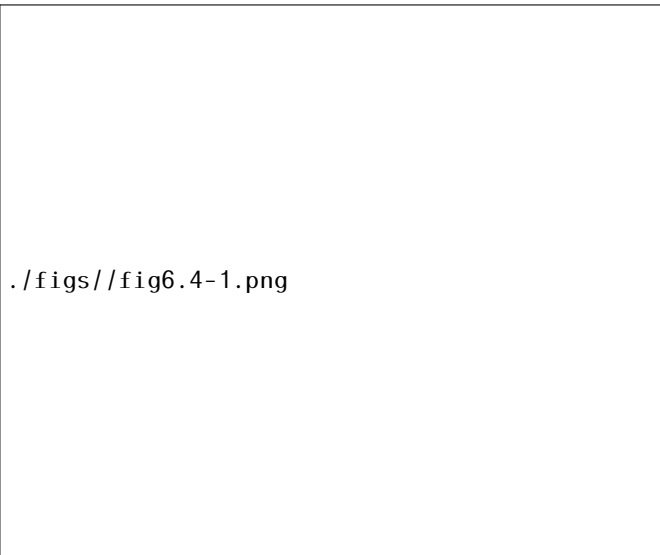
#### Esempio 7.7.a: fiori sinistrorsi

Essere sinistrorsi o destrorsi per quanto riguarda la posizione dei pistilli di alcuni fiori è determinato da un fenomeno di dominanza genetica.

Si è visto che:

- da un incrocio di puri destrorsi e sinistrorsi si ottengono solo destrorsi (ds);
- dall'incrocio dell' $F_1$  si ottengono 6 sinistrorsi (ss) e 21 ds.

Questo è coerente con quanto ci si aspetterebbe ( $H_0$ : 1 ss ogni 3 ds, come ci si aspetterebbe dalla teoria classica (0.25 della probabilità di ottenere ss)?



./figs//fig6.4-1.png

Figura 7.2:

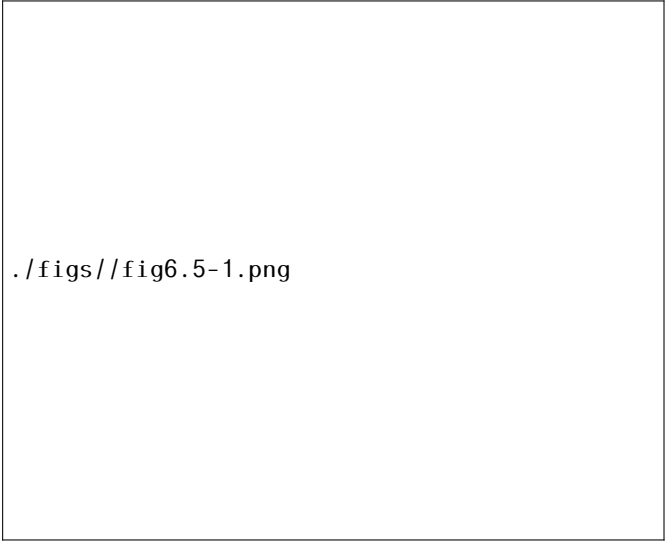
In questo caso, la statistica test è la frazione di fiori ss trovata, che va confrontata con la distribuzione attesa, per vedere quanto sia frequente ottenere i valori ottenuti. Il valore atteso di fiori ss sulla base dell  $H_0$  è 6.75, mentre il valore osservato è 6. Il p-value è rappresentato, nel grafico, da tutte le barre rosse, che rappresentano i valori uguali a 6 o più estremi, dunque il p-value è abbastanza coerente con la  $H_0$ . Dunque non si rifiuta l'ipotesi nulla.

#### **Esempio 7.7.b: riconoscimento della paternità**

18 prove indipendenti: soggetti diversi con foto diversi. Le ipotesi sono:

- $H_0$ : è impossibile attribuire ad un uomo la paternità di una figlia guardando foto dei due;
- $H_A$ : è possibile.

Potrebbe essere un test ad una coda nel caso in cui la relazione padre-figlia potesse rendere sia più somiglianti che meno somiglianti, mentre nella realtà si tratta di un test ad una coda, poiché la relazione padre-figlia può rendere i due soggetti solo più somiglianti, e dunque potrebbe essere plausibile che l'osservatore riesca a riconoscere la relazione guardando le due foto, e dunque che il numero di risposte corrette sia più alto di quello che ci si aspetterebbe se l' $H_0$  fosse vera, mentre non è possibile che l'osservatore sbagli nell'indovinare la parentela significativamente di più di quanto ci si aspetti per caso.



./figs//fig6.5-1.png

Figura 7.3:

Con un P-value pari a 0.048 si rifiuta l' $H_0$ , poiché il risultato ottenuto è significativamente diverso da ciò che ci si aspetterebbe se  $H_0$  fosse vera, almeno prendendo in considerazione un valore di  $\alpha$  pari a 0.05.

**Parte II**

**Confronti tra gruppi**



## Capitolo 8

# Analisi delle proporzioni (B5)

1

### 8.1 Distribuzione binomiale

Una misurazione effettuata su degli individui suddivide questi ultimi in due gruppi mutualmente esclusivi in base ad una certa caratteristica come, ad esempio, "successo" ed "insuccesso".

Nella popolazione, una proporzione fissa "p" di individui rientra in uno dei due gruppi, mentre i restanti rientrano nell'altro gruppo.

Se si estrae un campione casuale di n individui a partire dalla popolazione, la distribuzione campionaria del numero di individui che rientrano nella categoria "successo" è descritta dalla **distribuzione binomiale** (binomiale perché esistono soltanto due risultati possibili ed entrambi sono categorie con un certo nome).

La distribuzione binomiale fornisce la distribuzione di probabilità di X successi in (1) un numero fisso (n) di prove (2) indipendenti (l'una dalle altre) quando (3) la probabilità di successo (p) è la stessa in ogni prova ed (4) i possibili risultati della prova sono solamente due:

$$P[X \text{ successi}] = \binom{n}{X} p^X (1-p)^{n-X} \quad (8.1)$$

Il termine  $\binom{n}{X}$  è detto **coefficiente binomiale**, e corrisponde al numero di combinazioni semplici di n elementi presi X alla volta; tale termine rappresenta il numero di sequenze ordinate uniche (diverse tra loro) di successi ed insuccessi che portano esattamente ad X successi in n prove, e sta per:

$$\binom{n}{X} = \frac{n!}{X! (n-X)!} \quad (8.2)$$

In cui n! (n fattoriale) sta per:

$$n! = n \times (n-1) \times (n-2) \times \dots \times 2 \times 1 \quad (8.3)$$

#### Esempio 8.1.a: Fiori sinistrorsi

In una popolazione di una specie floristica, p = 0.25 ha fiori sinistrorsi e 1-p = 0.75 ha fiori destrorsi. Con questa situazione si può usare la distribuzione binomiale per determinare la probabilità di ogni dato numero di successi (es. X = 6, con n = 27, ovvero con 27 fiori campionati):

$$P[6 \text{ fiori sinistrorsi}] = \binom{27}{6} 0.25^6 (0.75)^{27-6} = 0.1719 \quad (8.4)$$

Dunque, esiste una probabilità di circa il 17% che 6 fiori su 27 siano sinistrorsi, se la proporzione di fiori sinistrorsi nella popolazione è pari a 0.25.

<sup>1</sup>Capitolo 7, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

Si possono calcolare le probabilità associate ad ogni possibile numero di successi (in questo caso, da 0 a 27), e la distribuzione di probabilità così ottenuta si può riportare graficamente:

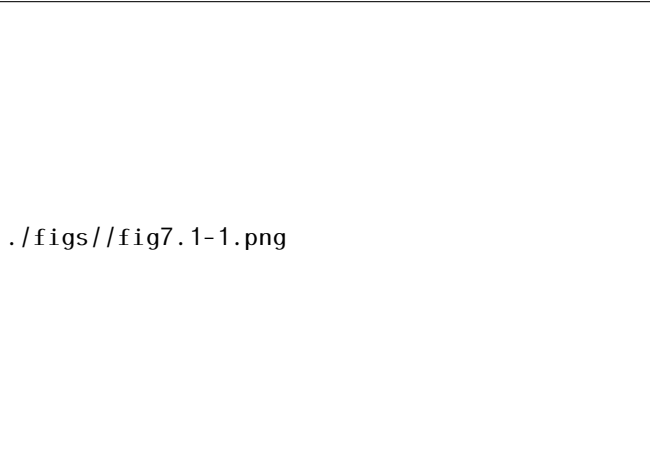


Figura 8.1:

Una distribuzione estremamente simile si può ottenere estraendo un gran numero di campioni casuali di  $n$  fisso (27, in questo caso), anche se la distribuzione binomiale fornisce probabilità più esatte, mentre le simulazioni ne forniscono di approssimate, soprattutto per  $n$  insufficienti; per verificare le ipotesi nulle si possono quindi utilizzare le distribuzioni binomiali.

### 8.1.1 Distribuzione campionaria di una proporzione

La proporzione di successi in una popolazione ( $X/n$ ,  $X$  su  $n$ ) si indica con  $p$ , mentre la proporzione di successi in un campione, ovvero la **proporzione campionaria**, si indica con  $\hat{p}$  ( $p$  hat). La proporzione di successi in campioni casuali è uguale *in media* alla proporzione di successi nella popolazione:  $\hat{p}$  è quindi una stima non distorta della proporzione nella popolazione perché fornisce, in media, la risposta corretta.

Come si nota nella Figura sottostante, la dimensione campionaria influisce sulla larghezza della distribuzione campionaria di  $\hat{p}$ : Quando  $n$  è grande, la distribuzione campionaria è stretta. Questo effetto è quantificato dall'**errore standard di  $\hat{p}$** :

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \quad (8.5)$$

Campioni di grandi dimensioni ottenuti campionando più volte una stessa popolazione avranno proporzioni stimate più vicine tra loro e più vicine alla proporzione nella popolazione, rispetto ai campioni piccoli: i campioni più grandi forniscono dunque stime più precise (**legge dei grandi numeri**).

## 8.2 Studio di una proporzione: il test binomiale

Il **test binomiale** viene usato nella verifica delle ipotesi quando una variabile in una popolazione ha due stati possibili (successo, insuccesso) e si vuole verificare se la frequenza relativa di successi nella popolazione ( $p$ ) sia pari al valore atteso assumendo vera l'ipotesi nulla ( $p_0$ ).

Le ipotesi nulla ed alternativa assumono particolari forme:

- $H_0$ : la frequenza relativa di successi nella popolazione è  $p_0$ ;
- $H_A$ : la frequenza relativa di successi nella popolazione è diversa da  $p_0$ .



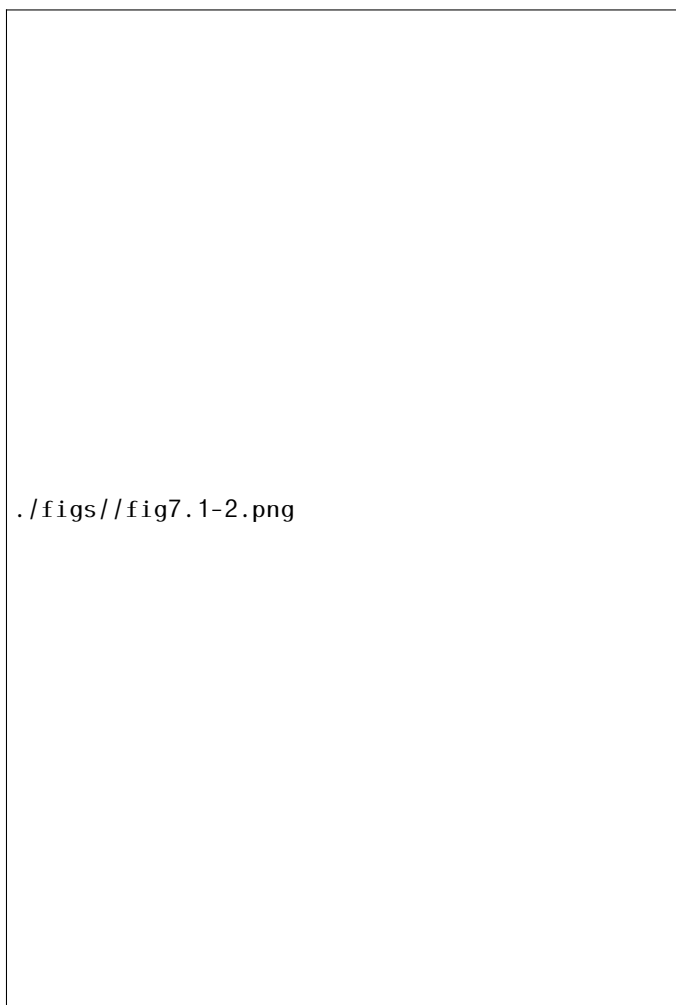


Figura 8.2:

Il valore atteso sotto l'ipotesi nulla può essere dato da una qualsiasi proporzione compresa tra 0 ed 1, estremi inclusi.

**Esempio 8.2.a: sesso e cromosoma X**

Si vuole provare che i geni legati alla spermatogenesi si trovino preferenzialmente sul cromosoma X.

Su 25 geni campionati, 10 (40%) risiedono effettivamente sul cromosoma X.

Se i geni per la spermatogenesi fossero distribuiti casualmente nel genoma, solo il 6.1% di essi risiederebbe sul cromosoma X, poiché su tale cromosoma è presente solo il 6.1% del genoma.

Definiamo le ipotesi:

- $H_0$ : la probabilità che un gene per la spermatogenesi risieda sul cromosoma X è pari a 0.061 ( $p_0 = 0.061$ );
- $H_A$ : la probabilità che un gene per la spermatogenesi risieda sul cromosoma X è diversa da 0.061 ( $p \neq 0.061$ ).

In questo caso, dato che è teoricamente possibile che i geni per la spermatogenesi risiedano meno sul cromosoma X che su altre parti del genoma, l'ipotesi alternativa prevede tutti i casi in cui la probabilità è diversa da 0.061, sia quelli in cui la probabilità eccede tale valore, sia quelli in cui essa ne è minore; pertanto, in questo caso, il test è a due code.

Nel caso del test binomiale, la statistica test utile per confrontare il risultato osservato con quello atteso sotto l'ipotesi nulla è il numero di successi osservato; in questo caso, dunque, la

statistica test assume il valore di 10 (10 geni per la spermatogenesi osservati sul cromosoma X). Il valore atteso sotto l'ipotesi nulla è invece  $0.061 \times 25 = 1.525$ . Dunque, nel campione, il numero di geni per la spermatogenesi osservati sul cromosoma X è maggiore di quello atteso sotto l'ipotesi nulla.

Ciò che è necessario stabilire è quanto sarebbe improbabile osservare un eccesso di geni per la spermatogenesi sul cromosoma X come quello osservato se la distribuzione di tali geni fosse casuale. Per stabilire ciò, si deve ottenere la distribuzione nulla, ovvero la distribuzione campionaria della statistica test assumendo che l'ipotesi nulla sia vera; tale distribuzione è descritta dalla distribuzione binomiale:

$$Pr[X \text{ successi}] = \binom{25}{X} 0.061^X (1 - 0.061)^{25-X} \quad (8.6)$$

La distribuzione nulla permette di calcolare il P-value, ovvero la probabilità di ottenere un risultato che sia altrettanto o più estremo di  $X = 10$  quando il valore atteso sotto l'ipotesi nulla è 1.525.

Dal momento che il test è a due code, P è la probabilità di ottenere 10 o più geni per la spermatogenesi sul cromosoma X, più la probabilità di risultati altrettanto o più estremi nell'altra coda della distribuzione nulla, corrispondenti a troppo pochi geni d'interesse sul cromosoma X. Il modo più semplice per includere tutti i risultati estremi è moltiplicare per due la probabilità di ottenere 10 o più geni.

La probabilità di ottenere 10 o più geni per la spermatogenesi sul cromosoma X, assumendo che l'ipotesi nulla sia vera, è pari alla somma di tutte le singole probabilità corrispondenti ad un numero esatto di successi, da 10 a 25, calcolate con la formula della distribuzione binomiale e mutualmente esclusive. Moltiplicando tale valore per 2 (per includere l'altra coda) si ottiene un P-value pari a  $1.98 \times 10^{-6}$ , che è inferiore al livello convenzionale di significatività  $\alpha = 0.05$ . Dunque, si rifiuta l'ipotesi nulla e si conclude che sul cromosoma X è presente un numero sproporzionatamente elevato di geni per la spermatogenesi.

La migliore stima della proporzione di geni per la spermatogenesi localizzati sul cromosoma X è, quindi:

$$\hat{p} = \frac{10}{25} = 0.40 \quad (8.7)$$

### 8.2.1 Approssimazioni del test binomiale

Il test binomiale fornisce un P-value esatto e può essere applicato a dati di qualsiasi tipo classificabili in due categorie; tuttavia, per  $n$  elevati, esistono alternative più rapide che permettono di giungere a valori di P non esatti ma, se sono soddisfatte determinate condizioni, approssimati molto bene; tali alternative sono il test del  $\chi^2$  di bontà dell'adattamento (sezione 9.2) ed il test dell'approssimazione normale alla distribuzione binomiale (SEZIONE?).

## 8.3 Stima delle proporzioni

### Esempio 8.3.a: figli dei radiologi

Si vuole calcolare la proporzione di maschi tra i discendenti dei radiologi maschi.

In un campione di 87 figli (da assumere casuale), 30 sono maschi.

La migliore stima della proporzione di figli maschi nella popolazione è:

$$\hat{p} = \frac{X}{n} = \frac{30}{87} = 0.345 \quad (8.8)$$

### 8.3.1 Stima dell'errore standard per una proporzione

Come visto nella sezione 4.8, la deviazione standard della distribuzione campionaria di una stima costituisce il suo errore standard. Nella sottosezione 8.1.1 si è visto che la deviazione

standard di  $\hat{p}$  (e quindi il suo errore standard) è:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \quad (8.9)$$

Tale valore solitamente non si può calcolare, perché non si conosce  $p$ ; tuttavia, si può ottenere l'errore standard, usando la stima della proporzione:

$$ES_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \quad (8.10)$$

Nel caso dell'esempio corrente:

$$ES_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\frac{0.345(1-0.345)}{87-1}} = 0.051 \quad (8.11)$$

Tale valore indica quando la stima campionaria  $\hat{p}$  sia verosimilmente vicina, in media, alla proporzione della popolazione  $p$ .

### 8.3.2 Intervallo di confidenza di una proporzione: metodo Agresti-Coull

Come visto (RIFERIMENTO! libro 4.3), l'intervallo di confidenza comprende i valori più plausibili del parametro da stimare in base ai dati. L'intervallo di confidenza al 95% di una proporzione comprende il valore vero della proporzione il 95% delle volte che lo si calcola sulla base dei nuovi dati.

Con il **metodo Agresti-Coull** per calcolare un intervallo di confidenza approssimato per una proporzione si deve inizialmente calcolare una quantità detta  $p'$ :

$$p' = \frac{X+2}{n+4} \quad (8.12)$$

L'intervallo di confidenza per una proporzione è quindi dato da:

$$p' - Z\sqrt{\frac{p'(1-p')}{n+4}} < p < p' + Z\sqrt{\frac{p'(1-p')}{n+4}} \quad (8.13)$$

In cui  $Z$  è una variabile normale standardizzata che dipende dall'intervallo di confidenza da calcolare (i valori sono tabulati, es. per un intervallo al 95%,  $Z = 1.96$ ). Tornando all'esempio, l'intervallo di confidenza al 95% di  $\hat{p}$  usando  $X = 30$  ed  $n = 87$  con il metodo Agresti-Coull si ottiene calcolando:

$$p' = \frac{30+2}{87+4} = 0.352 \quad (8.14)$$

$$0.352 - 1.96\sqrt{\frac{0.352(1-0.352)}{87+4}} < p < 0.352 + 1.96\sqrt{\frac{0.352(1-0.352)}{87+4}}$$

$$\Rightarrow 0.254 < p < 0.450$$

Tale intervallo non include il valore 0.512, che è la proporzione di figli maschi che si osserva tipicamente nella popolazione umana, dunque si può supportare l'idea che la proporzione di figli maschi dei radiologi sia molto più bassa della media della popolazione, se i dati appartengono effettivamente ad un campione casuale.

### 8.3.3 Intervallo di confidenza di una proporzione: metodo Wald

Un metodo molto usato per il calcolo degli intervalli di confidenza delle proporzioni è il **metodo di Wald**, che tuttavia presenta dei limiti: il metodo, infatti, è preciso solo quando  $n$  è grande e la proporzione della popolazione  $p$  non è vicina a 0 o 1; un intervallo di confidenza al 95%,

però, dovrebbe comprendere il parametro della popolazione nel 95% dei campioni e, purtroppo, quando  $n$  è piccolo o quando  $p$  è vicina a 0 o 1, l'intervallo di confidenza determinato dal metodo di Wald contiene il valore vero meno del 95% delle volte.

In ogni caso, il metodo di Wald delimita un intervallo intorno alla stima della popolazione  $\hat{p}$  attraverso un multiplo del suo errore standard:

$$\hat{p} - Z ES_{\hat{p}} < p < \hat{p} + Z ES_{\hat{p}} \quad (8.15)$$

L'intervallo di confidenza al 95% per i dati relativi all'esempio corrente calcolato con il metodo di Wald corrisponde a  $0.244 < p < 0.445$ .

## 8.4 Ricavare la distribuzione binomiale

### Esempio 8.4.a: sesso di tot individui

Si estrae un campione casuale di  $n$  individui da una popolazione e si considerano in ordine, uno alla volta, da 1 ad  $n$ . Si vuole calcolare la probabilità di ottenere  $X$  successi. Per prima cosa, si deve determinare il numero di sequenze di successi ed insuccessi che conducono ad  $X$  successi in totale. Si campionano 5 persone e si vuole conoscere la probabilità di ottenere 3 maschi e 2 femmine; in questo caso, tutte le possibili sequenze con  $n = 5$  prove che contengono 3 maschi sono 10:

MMMFF MMFMF MMFFM FMMMF MFMFM  
MFFMM FMMMF FMMFM FMFMM FFMMM

In generale, il numero di sequenze che danno esattamente  $X$  successi è dato dal coefficiente binomiale  $\binom{n}{X}$ . Assumendo che le prove siano indipendenti si può procedere.

In questo caso, ogni successo si verifica con stessa probabilità  $p$ , ed ogni insuccesso si verifica con la stessa probabilità  $1 - p$ ; dunque, la probabilità di ogni sequenza di successi ed insuccessi è il prodotto di queste probabilità per ciascun evento (regola del prodotto RIFERIMENTO! libro cap5). Quindi, una singola sequenza che ha  $X$  successi ed  $n - X$  insuccessi ha probabilità pari a:

$$p^X (1 - p)^{n-X} \quad (8.16)$$

In ultimo, si somma la probabilità di tutte le sequenze che danno esattamente  $X$  successi (regola della somma RIFERIMENTO! libro cap5). Dato che ciascuna delle sequenze è mutualmente esclusiva e ciascuna ha la stessa probabilità, per trovare la probabilità complessiva di  $X$  successi in  $n$  prove si moltiplica la probabilità di ogni sequenza per il numero delle sequenze possibili:

$$Pr[X \text{ successi}] = \binom{5}{3} p^X (1 - p)^{n-X} \quad (8.17)$$

Questa è la formula della distribuzione binomiale introdotta ad inizio capitolo.

## Capitolo 9

# Modelli probabilistici per dati di frequenza (B5)

1

### 9.1 Modello proporzionale

Un modello proporzionale è un modello probabilistico in cui la frequenza con cui si verificano i diversi eventi è proporzionale al numero di volte che essi hanno la possibilità di verificarsi.

#### Esempio 9.1.a: borse di studio

LEZIONE 15/11

Definiamo le ipotesi:

- $H_0$ :
- $H_A$ :

#### Esempio 9.1.b: nascite nei weekend

Sotto il modello proporzionale, ci si aspetta che i bambini nascano con la stessa frequenza tutti i giorni della settimana.

La tabella sottostante elenca il numero di nascite in ciascun giorno in un campione casuale di 350 nati nel 1999.

Giorno	Numero di nati
domenica	33
lunedì	41
martedì	63
mercoledì	63
giovedì	47
venerdì	56
sabato	47
<b>Totale</b>	<b>350</b>

Tabella 9.1:

I dati mostrano una grande variazione di nascite rispetto ai giorni della settimana.

Sotto il modello proporzionale, il numero di nascite in un giorno dovrebbe essere direttamente proporzionale al numero di volte in cui tale giorno c'è stato nel 1999, eccettuate le differenze

<sup>1</sup>Capitolo 8, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

dovute al caso.

L'adattamento del modello proporzionale ai dati può essere verificato con il test  $\chi^2$ .

## 9.2 Test del $\chi^2$ di bontà dell'adattamento

Il test del  $\chi^2$  di bontà dell'adattamento impiega un indice chiamato  $\chi^2$  per misurare la discrepanza tra una distribuzione di frequenza osservata e le frequenze previste da un modello casuale che rappresenta l'ipotesi nulla. Il modello casuale viene rifiutato se la discrepanza è troppo grande. Secondo il modello proporzionale, la probabilità di una nascita dovrebbe essere la stessa per ogni giorno della settimana; dato che quello proporzionale è il modello più semplice possibile, esso coincide con l'ipotesi nulla:

- $H_0$ : la probabilità di una nascita è la stessa per ogni giorno della settimana;
- $H_A$ : la probabilità di una nascita non è la stessa per ogni giorno della settimana.

Dato che il modello proporzionale è l'ipotesi nulla, lo si usa per calcolare le frequenze attese sotto l'ipotesi nulla: nel 1999, ogni giorno della settimana è occorso 52 volte, eccetto il venerdì, che è occorso 53 volte; dividendo questi numeri per 365, il numero di giorni nel 1999, si ottengono le proporzioni corrispondenti ad ogni giorno della settimana nell'anno di interesse, che si possono usare per calcolare le frequenze attese delle nascite per ogni giorno della settimana, assumendo vero il modello proporzionale:

$$\text{frequenza attesa per giorno } X = n_{\text{nati}} \frac{n_X}{365} \quad (9.1)$$

Le frequenze attese possono avere componenti frazionarie, anche se nei dati il numero di individui per categoria è ovviamente un intero; ciò avviene perché le frequenze attese sono i valori attesi medi, assumendo il modello nullo. La somma dei valori attesi deve essere uguale alla somma dei valori osservati (350 nell'esempio corrente), fatta eccezione dell'errore per arrotondamento.

### Esempio 9.2.a: covid-19

LEZIONE 15/11 Definiamo le ipotesi:

- $H_0$ :
- $H_A$ :

La statistica  $\chi^2$  misura la discrepanza tra frequenza osservata ed attesa:

$$\chi^2 = \sum_i \frac{(\text{osservato}_i - \text{atteso}_i)^2}{\text{atteso}_i} \quad (9.2)$$

In cui  $\text{osservato}_i$  è la frequenza di individui osservata nella  $i$ -esima categoria ed  $\text{atteso}_i$  è la frequenza attesa in quella categoria sotto l'ipotesi nulla.

Il numeratore dell'espressione è una differenza elevata al quadrato affinché le deviazioni positive e negative siano trattate allo stesso modo.

La statistica  $\chi^2$  impiega le frequenze assolute osservate ed attese, e non le proporzioni, che sono frequenze relative.

Riprendendo l'esempio 9.1.b,  $i$  può assumere solo valori da 1 a 7. Se le frequenze osservate in tutte le categorie corrispondessero a quelle attese,  $\chi^2$  sarebbe pari a 0. Maggiore è il valore del  $\chi^2$ , maggiore è la discrepanza tra osservato ed atteso.

Nell'esempio:

$$\chi_{\text{domenica}}^2 = \frac{(33 - 49.863)^2}{49.863} = 5.70 \quad (9.3)$$

Ripetendo il calcolo per i giorni restanti e sommando i risultati si ottiene un  $\chi^2$  pari a 15.05.

Per stabilire se il valore di  $\chi^2$  ottenuto è sufficiente per rifiutare l'ipotesi nulla è necessario ottenere la distribuzione campionaria della statistica test  $\chi^2$  sotto l'ipotesi nulla.

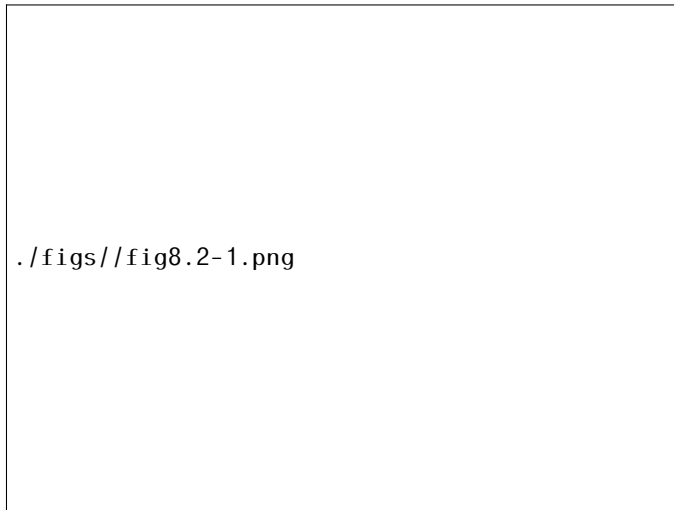


Figura 9.1:

Tale distribuzione si può ottenere mediante simulazioni al computer, oppure attraverso la distribuzione teorica del  $\chi^2$ , che ha una forma matematica nota e che approssima bene la distribuzione nulla. Le caratteristiche essenziali della distribuzione teorica del  $\chi^2$  (d'ora in avanti semplicemente "distribuzione  $\chi^2$ ") sono state tabulate in tavole statistiche.

La distribuzione  $\chi^2$  è in realtà una famiglia di distribuzioni, e quella particolare che serve per analizzare i dati sulle nascite è specificata dal numero di **gradi di libertà** (df, degrees of freedom):

$$df = (\text{numero di categorie}) - 1 - (\text{numero di parametri stimato in base ai dati}) \quad (9.4)$$

Nell'esempio corrente, l'ultimo termine dell'equazione è pari a zero (successivamente si capirà a cosa si riferisce). I dati sulle nascite hanno 7 categorie, una per giorno della settimana, quindi i gradi di libertà sono  $7 - 1 = 6$ , quindi il valore di  $\chi^2$  calcolato in base ai dati osservati ( $\chi^2 = 15.05$ ) deve essere confrontato con la distribuzione  $\chi^2_6$ .

Tale distribuzione è rappresentata dalla curva nera nella figura 9.2

### 9.2.1 Calcolo del P-value

Per il test  $\chi^2$  di bontà dell'adattamento, il P-value è la probabilità di ottenere un valore del  $\chi^2$  maggiore del valore osservato, ovvero calcolato in base ai dati.

Se i dati rispecchiassero esattamente quanto previsto dall'ipotesi nulla,  $\chi^2$  sarebbe 0, mentre più è marcata la differenza dai valori attesi, maggiore è il valore di  $\chi^2$ , per cui per calcolare il P-value si usa solo la coda destra della distribuzione  $\chi^2$ .

La distribuzione  $\chi^2$  è una distribuzione di probabilità continua, quindi la probabilità è data dall'area della regione sottesa dalla curva. La probabilità di ottenere un valore del  $\chi^2$  maggiore di un singolo valore specificato, necessario per calcolare il P-value, è uguale all'area sotto la curva a destra di tale valore fino all'infinito.

Nell'esempio, la probabilità di ottenere un valore uguale o superiore a 15.05 è uguale all'area sottesa dalla curva  $\chi^2_6$  oltre 15.05: Per calcolare l'area a destra del valore di  $\chi^2$  calcolato a partire dai dati si hanno due opzioni:

- Utilizzare un programma statistico al computer.  
Il P-value che si ottiene in tale modo è  $P = 0.0199$  che, al livello di significatività standard  $\alpha = 0.05$  induce a rifiutare l'ipotesi nulla. I dati forniscono quindi evidenze del fatto che le nascite non sono distribuite casualmente nei giorni della settimana.
- Utilizzare dei valori critici.  
Un **valore critico** è il valore di una statistica test che identifica il confine di un'area specificata nella coda (o nelle code) della distribuzione campionaria sotto l'ipotesi nulla; ad

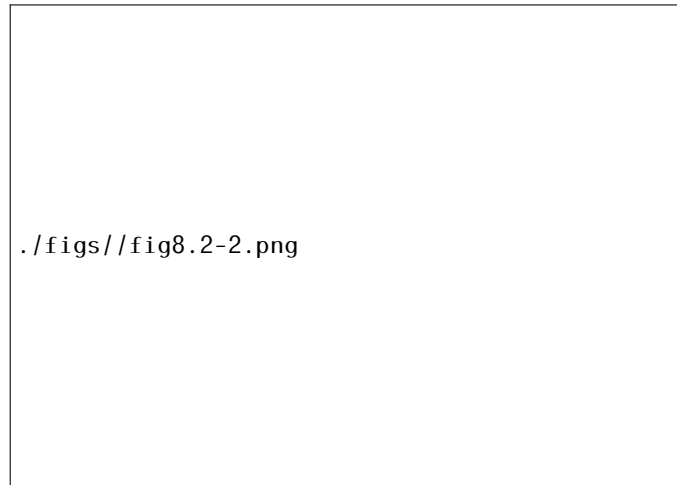


Figura 9.2:

esempio, con un  $\alpha = 0.05$ , il valore di  $\chi^2$  per il quale l'area sottesa dalla curva alla sua destra è pari a 0.05 è detto "valore critico corrispondente ad  $\alpha = 0.05$ ".

I valori critici della distribuzione  $\chi^2$  sono tabulati nell'appendice

Nell'esempio, il valore critico è pari a 12.59, quindi, sotto l'ipotesi nulla, la probabilità di ottenere un valore di  $\chi^2$  altrettanto o più estremo di 12.59 è pari a 0.05:

$$Pr[\chi_6^2 \geq 15.05] = 0.05 \quad (9.5)$$

e si rifiuta quindi l'ipotesi nulla.

### 9.2.2 Assunzioni del test $\chi^2$ di bontà dell'adattamento

Il test  $\chi^2$  di bontà dell'adattamento assume che gli individui nell'insieme dei dati siano un campione casuale estratto dall'intera popolazione, ovvero che ogni individuo sia stato scelto indipendentemente da tutti gli altri e che ogni individuo della popolazione ha avuto la stessa probabilità di essere incluso.

La distribuzione campionaria della statistica  $\chi^2$  segue una distribuzione  $\chi^2$  solo approssimativamente, anche se l'approssimazione è molto buona quando si seguono due regole:

- nessuna delle categorie deve avere una frequenza attesa minore di 1;
- non più del 20% delle categorie deve avere frequenze attese minori di 5.

Queste restrizioni si riferiscono alle frequenze attese e non a quelle osservate; se tali condizioni non sono soddisfatte, il test diventa inaffidabile.

Se una delle condizioni non è soddisfatta, si hanno due opzioni:

- raggruppare, se possibile, alcune delle categorie con frequenze attese piccole, per ottenere meno categorie, che abbiano però frequenze attese maggiori (si devono cambiare di conseguenza i gradi di libertà);
- trovare un'alternativa al test del  $\chi^2$  (es. simulazioni al computer).

### 9.2.3 Test del $\chi^2$ con due sole categorie

Il test del  $\chi^2$  funziona anche quando vi sono solo due categorie, per cui si può usare anche in sostituzione del test binomiale, purché siano soddisfatte le relative assunzioni. I calcoli sono più rapidi ma meno esatti di quelli di un test binomiale.



**Esempio 9.2.b: contenuto genico del cromosoma X umano**

Ad ora sappiamo che il cromosoma X dell'uomo contiene 781 dei 20290 geni trovati finora, e che il cromosoma X contiene circa il 5.2% del DNA totale.

Assumendo il modello proporzionale, ci si aspetterebbe che il 5.2% dei geni fosse localizzato su tale cromosoma. È così?

Definiamo le ipotesi:

- $H_0$ : la percentuale dei geni umani sul cromosoma X è 5.2%;
- $H_A$ : la percentuale dei geni umani sul cromosoma X non è 5.2%

Il numero atteso di geni sul cromosoma X umano è  $20290 \times 0.052 = 1055$ , quindi un numero molto maggiore rispetto a quello dei geni osservati (781).

Si vuole stabilire qual è la probabilità di un risultato altrettanto o più estremo di quello osservato se fosse vera l'ipotesi nulla.

Calcolare il P-value con il test binomiale sarebbe impegnativo, poiché il numero di prove (geni) è  $n = 20290$  e la probabilità che un gene sia presente sul cromosoma X è  $p = 0.052$ . Quindi, il valore di P si calcolerebbe come:

$$P = 2 \times \Pr[X \leq 781] = 2 \times (\Pr[X = 0] + \Pr[X = 1] + \Pr[X = 2] + \dots + \Pr[X = 781]) \quad (9.6)$$

Usando il test  $\chi^2$ , invece:

$$\chi^2 = \frac{(781 - 1055)^2}{1055} + \frac{(19509 - 19235)^2}{19235} = 75.1 \quad (9.7)$$

Questa statistica ha due categorie, e quindi un solo grado di libertà ( $df = 2 - 1 = 1$ ). Dall'appendice si vede che il valore critico per  $\chi^2_1$  per un  $\alpha = 0.05$  è pari a 3.84 e, poiché il  $\chi^2$  osservato, 75.1, è maggiore di 3.84, si può dire che  $P < 0.05$  e rifiutare l'ipotesi nulla.

In realtà, dal momento che il  $\chi^2$  calcolato è maggiore anche del più grande valore critico tabulato per un  $df[\chi^2_{1,(0.001)} = 10.83]$ , si può dire che  $P < 0.001$ , e dunque che, nella specie umana, il numero di geni sul cromosoma X è significativamente minore di quello che ci si aspetterebbe in base alle dimensioni di tale cromosoma.

Dunque, nel caso di due sole categorie, il test binomiale è l'opzione migliore per  $n$  piccoli e quando le frequenze attese sono troppo basse per soddisfare le assunzioni del test  $\chi^2$ ; inoltre, tale test è preferibile anche per  $n$  grandi quando si dispone di un computer, poiché fornisce un P-value esatto.

## 9.3 Adattamento della distribuzione binomiale

La distribuzione binomiale descrive la probabilità di ottenere  $X$  successi in  $n$  prove indipendenti.

**Esempio 9.3.a: famiglie pianificate**

Ci si aspetta che il numero di maschi e di femmine nati in famiglie con 2 figli siano conformi ad una distribuzione binomiale, con  $n = 2$  e  $p$  uguale alla probabilità di avere un maschio in ogni singola prova.

È questo ciò che si osserva?

Per le famiglie con 2 figli sono possibili 3 risultati: 0, 1 o 2 maschi; la tabella 9.2 riporta i dati raccolti per 244 famiglie:

Numero maschi	Numero osservato di famiglie
0	530
1	1332
2	582
<b>Totale</b>	<b>2444</b>

Tabella 9.2:

L'adattamento della distribuzione binomiale ai dati si può verificare con il test del  $\chi^2$ . Definiamo le ipotesi:

- $H_0$ : il numero di maschi nella famiglie con 2 figli ha una distribuzione binomiale;
- $H_A$ : il numero di maschi nella famiglie con 2 figli non ha una distribuzione binomiale.

Si sta quindi verificando l'adattamento di una distribuzione ai dati osservati in più famiglie, non un'ipotesi sulla proporzione media di maschi.

Quando si testa l'adattamento ad una distribuzione binomiale, si studiano i risultati ottenuti in più set di prove, confrontando un set di frequenze con i corrispondenti valori previsti dalla distribuzione. Tale procedura è diversa dall'impiego della distribuzione binomiale per verificare un'ipotesi nulla su una singola proporzione: in un test binomiale si ha un solo set di prove.

In questo caso, l'ipotesi nulla non specifica  $p$ , la probabilità che un singolo figlio sia maschio, quindi si deve stimare  $p$  in base ai dati prima di poter calcolare le frequenze attese.

Nello studio ci sono 4888 figli; il numero totale di maschi è  $(2 \times 582) + 1332 = 2496$ , perciò, la probabilità che un figlio sia maschio è:

$$\hat{p} = \frac{2496}{4888} = 0.5106 \quad (9.8)$$

Tale stima, assieme alla distribuzione binomiale con  $n = 2$ , si usano per calcolare le frequenze attese sott l'ipotesi nulla; ad esempio, la frazione attesa di famiglie con 2 figli che hanno 1 solo maschio è:

$$Pr[1 \text{ maschio}] = \binom{2}{1} (0.5106)^1 (1 - 0.5106)^1 = 0.49977 \quad (9.9)$$

Quindi, la frequenza attesa di 2444 famiglie con 2 figli ed 1 solo maschio è:

$$Atteso (1 \text{ maschio}) = 2444 \times 0.49977 = 1221.4 \quad (9.10)$$

Le frequenze attese per tutti i risultati possibili sono riportate in tabella 9.3:

Numero maschi	Numero osservato di famiglie	Numero atteso di famiglie
0	530	585.3
1	1332	1221.4
2	582	637.3
<b>Totale</b>	<b>2444</b>	<b>2444.0</b>

Tabella 9.3:

E i valori attesi assieme ai dati osservati sono rappresentati nella figura 9.3.a:

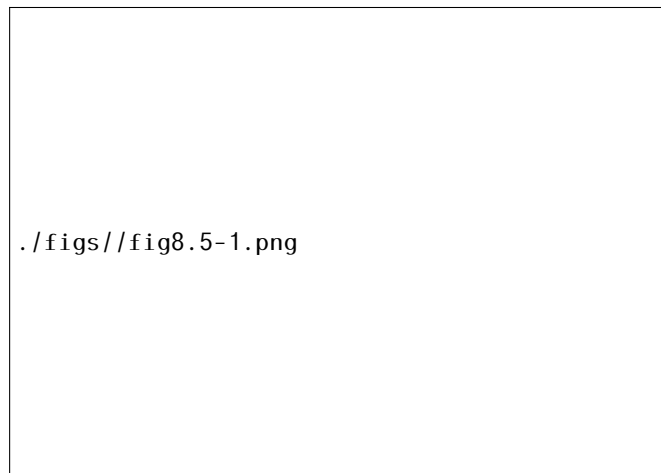


Figura 9.3:

Nell'esempio corrente, le frequenze osservate non sono uguali a quelle attese; esiste infatti un eccesso di famiglie con 2 figli ed 1 solo figlio maschio, rispetto alle altre due tipologie. Si vuole stabilire se tale differenza sia attribuibile al caso o se sia motivo di rifiuto dell'ipotesi nulla.

Si calcola il  $\chi^2$ :

$$\chi^2 = \frac{(530 - 585.3)^2}{585.3} + \frac{(1332 - 1221.4)^2}{1221.4} + \frac{(582 - 637.3)^2}{637.3} \quad (9.11)$$

Per calcolare il numero di gradi di libertà si considera che si hanno 3 categorie; di norma, dunque, i gradi sarebbero 2; tuttavia, in questo caso si è dovuto stimare un parametro a partire dai dati per calcolare le frequenze attese ( $p$ ), quindi:

$$df = 3 - 1 - 1 = 1 \quad (9.12)$$

Il valore critico per la distribuzione  $\chi^2_1$  con un grado di libertà ed  $\alpha = 0.05$  è pari a 3.84. Poiché 20.04 è localizzato verso destra nella coda della distribuzione rispetto a 3.84,  $P < 0.05$ , quindi si rifiuta l'ipotesi nulla.

Dunque, la distribuzione di frequenza del numero di maschi e femmine nelle famiglie con 2 figli non è governata dalla distribuzione binomiale, quindi in questi dati una delle assunzioni della distribuzione binomiale deve essere rifiutata: la probabilità di avere un maschio varia da famiglia a famiglia e/o gli individui nella stessa famiglia non sono indipendenti l'uno dall'altro.

Una probabile spiegazione è che le famiglie con 2 senza maschi o senza femmine siano insoddisfatte e decidano di avere un terzo figlio, rimuovendosi dall'insieme di famiglie con 2 soli figli.

### 9.3.1 Test binomiale per $n$ grandi

2

Una distribuzione di casi con  $p$  vicini allo 0.5 ed  $n$  vicini al 40 tende ad assumere la forma di una curva normale; quindi quando si hanno distribuzioni più o meno binomiali, invece di usare la distribuzione binomiale si può, usando i parametri:

- $\mu = np$ ;
- $\sigma = \sqrt{np(1-p)}$ .

stimare il valore  $z$ :

$$z = \frac{x_i - \mu}{\sigma} \quad (9.13)$$

Se il valore di  $z$  che si ottiene è superiore ad 1.96, si può rifiutare l'ipotesi nulla.

## 9.4 Distribuzione di Poisson

La **distribuzione di Poisson** indica la probabilità di ottenere  $X$  successi in un certo intervallo di tempo o in una certa porzione di spazio, quando ogni successo si verifica indipendentemente e con uguale probabilità in ogni punto del tempo e dello spazio.

### Esempio 9.4.a: Esempio: ragni dopo l'eruzione

Dopo un'eruzione, i ragni sono i primi animali a ricolonizzare il territorio, trasportati dai venti. Come sarebbe la distribuzione dei ragni, se l'atterraggio fosse casuale nello spazio?

Si fanno due assunzioni:

- la probabilità che un ragno atterri in un dato punto è la stessa in ogni punto del territorio considerato;
- il fatto che un ragno atterri in un dato punto è indipendente dagli atterraggi in ogni altro

<sup>2</sup>da spiegazione del prof. Massolo, fine della lezione 20211115

punto (quindi i ragni non si raggruppano né si respingono tra loro).

Per contare i ragni, si suddivide il territorio d'interesse in tante porzioni di uguale dimensione, grandi a sufficienza da contenere molti potenziali siti di atterraggio.

Se entrambi gli assunti sono soddisfatti, la distribuzione di frequenza del numero di ragni che atterrano nelle porzioni di suolo seguirà una distribuzione di Poisson.

È lecito aspettarsi una distribuzione di Poisson per certi conteggi in biologia (es. numero di mutazioni in ogni individuo di una popolazione, numero di salmoni catturati in un giorno da un predatore, numero di semi germinanti per ogni pianta). Un fenomeno diventa interessante quando non segue il modello semplice di partenza, perché in tal caso si scopre che alcune delle assunzioni fatte sono false, il che suggerisce l'esistenza di processi particolari (es. alcuni individui possono essere effettivamente più inclini alle mutazioni di altri, alcuni predatori più bravi nella cattura di altri, alcune piante produrre semi di maggior successo).

Dunque, mentre la distribuzione di Poisson prevede una distribuzione **casuale** dei successi, le alternative ad essa prevedono distribuzioni non casuali nel tempo e nello spazio; i successi, ad esempio, possono essere:

- **raggruppati** (o agglomerati), ovvero possono avvenire più vicini l'uno a l'altro di quanto ci si aspetterebbe per effetto del caso, come potrebbe succedere, ad esempio, quando la presenza di un successo aumenta la probabilità che si verifichino altri successi in prossimità (es. epidemie di malattie contagiose);
- **dispersi**, ovvero possono avere una distribuzione più uniforme di quella attesa per effetto del caso, come può accadere quando la presenza di un successo diminuisce la probabilità che ci siano altri successi in prossimità (es. distribuzione degli animali territoriali).

./figs//fig8.6-1.png

Figura 9.4:

Le deviazioni da pattern casuali possono quindi evidenziare processi biologici interessanti.

La formula matematica della distribuzione di Poisson descrive la probabilità che si verifichino  $X$  successi in qualsiasi intervallo di tempo o porzioni di spazio:

$$Pr[X \text{ successi}] = \frac{e^{-\mu} \mu^X}{X!} \quad (9.14)$$

In cui:

- $\mu$  è il numero medio di successi indipendenti nello spazio e nel tempo, espressi come conteggio riferito all'unità di tempo o di spazio;
- $e$  è la base dei logaritmi naturali (2.718...).

Se si immagina di dividere un intervallo di tempo in moltissimi intervalli più brevi (stesso ragionamento si potrebbe applicare alla dimensione spaziale), la probabilità che ha un evento di verificarsi in ognuno di essi sarà prossima a zero, quindi per ciascuno di tali intervalli potranno esserci due possibili situazioni:

- l'evento di verifica una volta, con probabilità molto piccola;
- l'evento non si verifica, con probabilità molto grande.

Non potranno quindi esistere intervalli in cui un evento si verifica più di una volta. In questo modo, si riconduce la distribuzione di Poisson del numero di eventi che possono avvenire in un certo intervallo di tempo (o in una certa area) ad una distribuzione binomiale; ci sono infatti due soli eventi possibili nella singola prova, successo o insuccesso, ma in questo caso il numero di prove (numero di potenziali successi, es. potenziali siti di atterraggio dei ragni in ogni plot) è enorme, e la probabilità di successo della singola prova (un ragno che atterra in un punto specifico) tende a zero. Si può in effetti dimostrare che la distribuzione di Poisson si può ottenere come limite della distribuzione binomiale per  $p$  che tende a zero ed  $n$  che tende ad infinito. Il prodotto  $np$  corrisponde al numero medio di eventi per intervallo (o area), ovvero al parametro  $\mu$  della distribuzione di Poisson.

#### 9.4.1 Verifica della casualità con la distribuzione di Poisson

In biologia, la distribuzione di Poisson si impiega principalmente per formulare ipotesi nulle e verificare se i successi siano distribuiti casualmente nel tempo e nello spazio.

Di solito non si conosce il tasso esatto con cui si verificano i successi, quindi per poter fare previsioni sulla probabilità di diversi risultati in base ad una distribuzione di Poisson si deve prima stimare il tasso di successi in base ai dati.

##### Esempio 9.4.b: estinzioni di massa

La tabella 9.4 riporta il numero di estinzioni di famiglie di invertebrati marini registrate in 76 intervalli di tempo di durata simile, stimate sulla base di testimonianze fossili.

Numero estinzioni ( $X$ )	Frequenza
0	0
1	13
2	15
3	16
4	7
5	10
6	4
7	2
8	1
9	2
10	1
11	1
12	0
13	0
14	1
15	0
16	2
17	0
18	0
19	0
20	1
>20	0
<b>Totale</b>	<b>76</b>

Tabella 9.4:

Le estinzioni si verificano casualmente, o ci sono periodi in cui i tassi di estinzione sono insolitamente elevati (estinzioni di massa)?

Se l'estinzione delle famiglie fosse casuale nel tempo, il numero di estinzioni in ogni intervallo dovrebbe seguire una distribuzione di Poisson; le alternative sono (1) che le estinzioni siano raggruppate nel tempo e che avvengano dunque delle estinzioni di massa, oppure (2) che le estinzioni avvengano più uniformemente nel tempo di quanto ci si aspetterebbe casualmente. Per capire quale dei tre sia il caso, si confronta la distribuzione di frequenza delle estinzioni delle famiglie con quella prevista da una distribuzione di Poisson usando il test del  $\chi^2$ .

Le ipotesi sono:

- $H_0$ : il numero di estinzioni per ogni intervallo di tempo ha una distribuzione di Poisson;
- $H_A$ : il numero di estinzioni per ogni intervallo di tempo non ha una distribuzione di Poisson.

Prima di effettuare il test, si calcola  $\mu$ , che in questo caso corrisponde al numero medio di di estinzioni per ogni intervallo di tempo, attraverso la media campionaria:

$$\bar{X} = \frac{(0 \times 0) + (13 \times 1) + (15 \times 2) + \dots}{76} = 4.21 \quad (9.15)$$

Tale media campionaria è usata al posto di  $\mu$  nella formula per la distribuzione di Poisson per generare le frequenze attese.

L'istogramma in figura 9.4.b rappresenta la distribuzione di frequenza osservata del numero di estinzioni per intervallo di tempo, mentre il diagramma a segmenti congiunge le frequenze previste dell'ipotesi nulla (distribuzione di Poisson).

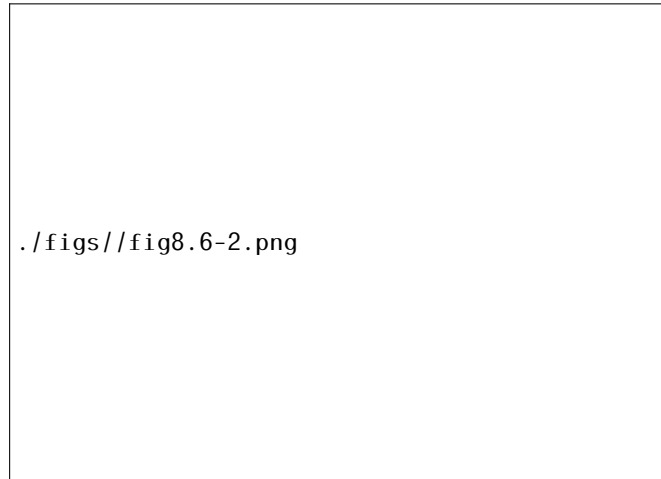


Figura 9.5:

Si nota una discrepanza: rispetto alla distribuzione di Poisson, le testimonianze fossili presentano troppi intervalli di tempo con moltissime estinzioni e troppi intervalli di tempo con pochissime estinzioni.

Il divario che si osserva è maggiore di quello che ci si aspetterebbe per effetto del caso? Per stabilirlo si usa il test del  $\chi^2$ .

Per calcolare le frequenze attese nelle diverse categorie di estinzioni, tranne l'ultima, si applica la formula della distribuzione di Poisson per il calcolo delle probabilità attese, che poi vanno moltiplicate per il numero totale di intervalli di tempo (76); ad esempio, la probabilità attesa e poi la frequenza attesa degli intervalli di tempo con tre estinzioni sono:

$$Pr[3 \text{ estinzioni}] = \frac{e^{-\mu} \mu^3}{3!} = \frac{e^{-4.21} 4.21^3}{3!} = 0.1846 \quad (9.16)$$

$$Frequenza \text{ attesa}[3 \text{ estinzioni}] = 76 \times 0.1846 = 14.03 \quad (9.17)$$

Nell'ultima categoria si raggruppano tutte le  $X \geq 10$  estinzioni perché la frequenza attesa dei numeri più grandi diventa molto piccola; per calcolarla, si sottrae la somma di tutti i valori

attesi precedenti da 76, ovvero il numero totale di intervalli di tempo.

Le frequenze attese non soddisfano gli assunti del test del  $\chi^2$ , dato che una di esse è minore di 1 e più del 20% sono minori di 5; in questi casi, si possono raggruppare le categorie e tentare nuovamente, ad esempio, si possono riunire in un'unica categoria  $X = 0$  ed  $X = 1$ , ed in un'altra tutte le categorie con  $X \geq 8$ , dato che le categorie raggruppate sono simili. I dati che si ottengono hanno 8 categorie, riportate nella tabella 9.5:

Numero estinzioni ( $X$ )	Frequenza osservata di intervalli di tempo	Frequenza attesa di intervalli di tempo
0 o 1	13	5.88
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.72
7	2	5.24
>8	9	4.91
<b>Totale</b>	<b>76</b>	<b>76</b>

Tabella 9.5:

Usando la formula standard per la statistica  $\chi^2$  si ottiene:

$$\chi^2 = \frac{(13 - 5.88)^2}{5.88} + \frac{(15 - 10)^2}{10} + \frac{(16 - 14.03)^2}{14.03} + \dots = 23.93 \quad (9.18)$$

Per il test si hanno 6 gradi di libertà, dato che si deve tener conto del parametro  $\mu$  che è stato calcolato a partire dai dati, e quindi  $df = 8 - 1 - 1 = 6$ .

Il valore critico per la  $\chi^2_{6,(0.05)}$  è 12.59. Dato che la statistica  $\chi^2$  è interna alla coda della distribuzione rispetto al valore critico, il P-value è minore di 0.05 (ed anche di 0.001, dato che 23.93 è anche maggiore di 22.46, che è il valore critico per  $\alpha = 0.001$ ). Si rifiuta quindi l'ipotesi nulla e si conclude che le estinzioni d'interesse non si adattano ad una distribuzione di Poisson.

### 9.4.2 Indice di dispersione per descrivere il discostamento dalla distribuzione di Poisson

3

L'indice di dispersione ( $D$ , o variance to mean ratio,  $VMR$ ) esprime il rapporto tra varianza e media:

$$VMR = \frac{\sigma^2}{\mu} \quad (9.19)$$

Nella distribuzione di Poisson, varianza e media sono uguali, quindi il  $VMR$  è pari ad 1.

Se i due valori differiscono si possono avere due casi:

- $VMR > 1$  ( $\sigma^2 > \mu$ ): la distribuzione è dispersa (come nella distribuzione binomiale negativa);
- $VMR < 1$  ( $\sigma^2 < \mu$ ): la distribuzione è sotto-dispersa, quindi raggruppata (nella distribuzione è binomiale  $0 < VMR < 1$ ).

<sup>3</sup>da spiegazione del prof. Massolo, lezione 20211117 (1:17:00)





## Capitolo 10

# Analisi dell'associazione tra variabili categoriche (B5)

1

### 10.1 Associazione tra due variabili categoriche

L'esistenza di un'associazione tra due variabili categoriche implica che esse non siano indipendenti.

Se le due variabili fossero indipendenti, la probabilità attesa di osservare una combinazione qualsiasi delle categorie appartenenti alle due variabili (es. categoria A per la variabile 1 e categoria C per la variabile 2) sarebbe:

$$Pr[A \text{ e } C] = Pr[A]Pr[C] \quad (10.1)$$

Se invece non sono indipendenti, la probabilità di ottenere la combinazione è diversa dal semplice prodotto.

#### Esempio 10.1.a: naufragio del Titanic

Durante il naufragio del Titanic, sono morte meno donne che uomini, anche tenendo conto delle abbondanze relative delle due categorie.

Per stabilire se la differenza osservata può essere attribuibile al caso o se ci sia stata un'associazione tra sesso e mortalità si può ricorrere al test del  $\chi^2$ , dal momento che si vogliono confrontare frequenze assolute.

Nel caso dell'esempio corrente, sesso e mortalità non sono indipendenti.

I dati sono riportati nella tabella 10.1, mentre i mosaic plots 10.1.a mostrano la relazione tra sesso e mortalità osservata (a sinistra) ed attesa nel caso di variabili indipendenti (destra).

	Deceduti (D)	Sopravvissuti (S)	
M	1364	367	1731
F	126	344	470
	1490	711	2201

Tabella 10.1:

<sup>1</sup>Capitolo 9, Whitlock and Schluter, Analisi statistica dei dati biologici (2010)

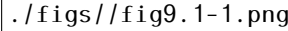


Figura 10.1:

## 10.2 Stima dell'associazione in tabelle 2x2: odds ratio

L'**odds ratio** misura l'entità dell'associazione tra due variabili categoriche quando ciascuna di esse ha solo due categorie; una delle due variabili è la **variabile risposta**, che ha le due categorie "successo" ed "insuccesso", mentre l'altra è la **variabile esplicativa**, le cui due categorie identificano i due gruppi all'interno dei quali le probabilità di successo vengono calcolate e confrontate.

Nell'esempio 10.1.a, la categoria variabile è "morte", con categorie "successo" ed "insuccesso", mentre la variabile esplicativa è "sex", che può assumere i valori "maschio" e "femmina".

L'odds ratio confronta la proporzione di successi ed insuccessi tra i due gruppi.

L'**odds** ( $O$ ) di un successo è dato da:

$$O = \frac{p}{1-p} \quad (10.2)$$

In cui  $p$  è la probabilità di successo, e  $1-p$  è la probabilità di insuccesso.

Se  $O = 1$ , si verifica un successo per ogni insuccesso.

Nell'esempio 10.1.a, la probabilità di morire se si è maschi o femmine è data da:

$$O_M = \frac{p_{D|M}}{1-p_{D|M}} \quad (10.3)$$

$$O_F = \frac{p_{D|F}}{1-p_{D|F}} \quad (10.4)$$

In cui  $p_{D|M}$  e  $p_{D|F}$  corrispondono al rapporto tra deceduti maschi o femmine ed il totale di maschi o femmine.

L'odds ratio è il rapporto tra i due odds (conviene mettere al denominatore la categoria della variabile esplicativa che si pensa possa avere più successo):

$$OR = \frac{\frac{p_{D|M}}{1-p_{D|M}}}{\frac{p_{D|F}}{1-p_{D|F}}} \quad (10.5)$$

Le situazioni possono essere due:

- odds ratio = 1 (quindi odds dei maschi ed odds delle femmine sono uguali), quindi tra le due variabili considerate non c'è nessuna associazione;
- odds ratio  $\neq 1$ , quindi tra le due variabili considerate c'è un'associazione.

Ovviamente, la differenza da 1 dev'essere statisticamente significativa.

### 10.2.1 Errore associato all'odds ratio

Quando l'odds ratio si calcola a partire da proporzioni osservate di successi in base ad un campione, sia le proporzioni che l'odds ratio sono stime (e dunque si indicano come  $\hat{p}$  e  $\hat{OR}$ ). In tali casi, alla stima dell'odds ratio va accompagnata anche una stima dell'errore ad esso associato, e questo si fa calcolando l'**intervallo di confidenza dell'odds ratio**.

Per ottenere l'intervallo di confidenza dell'odds ratio è necessario calcolare il logaritmo naturale

dell'odds ratio ( $\ln(\hat{OR})$ ), poiché l'odds ratio ha una distribuzione non normale. L'errore standard del logaritmo naturale dell'odds ratio si calcola come:

$$ES[\ln(\hat{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad (10.6)$$

In cui  $a$ ,  $b$ ,  $c$  e  $d$  si riferiscono alle frequenze osservate nelle celle della tabella di contingenza (nell'esempio 10.1.a, i quattro valori principali nella tabella 10.1).

L'intervallo di confidenza viene quindi calcolato come:

$$\ln(\hat{OR}) - Z ES[\ln(\hat{OR})] < \ln(\hat{OR}) < \ln(\hat{OR}) + Z ES[\ln(\hat{OR})]$$

$$\Rightarrow e^{\ln(\hat{OR}) - Z ES[\ln(\hat{OR})]} < \hat{OR} < e^{\ln(\hat{OR}) + Z ES[\ln(\hat{OR})]}$$

In cui il valore di  $Z$  dipende dal tipo di intervallo di confidenza che si vuole calcolare (es. per un intervallo di confidenza al 95%  $Z = 1.96$ ).

### Esempio 10.2.a: Esempio: aspirina

29876 donne sono state sottoposte casualmente a due trattamenti diversi: 19934 hanno ricevuto 100 mg di aspirina a giorni alterni, 19942 hanno ricevuto un placebo.

Tali donne, ignare riguardo il tipo di trattamento ricevuto, sono state monitorate per 10 anni, durante i quali 1438 che ricevevano aspirina e 1427 che ricevevano il placebo hanno avuto una diagnosi di cancro invasivo.

La tabella 10.2 riporta i dati, che sono rappresentati nel mosaic plot nella figura 10.2.a.

	Aspirina (A)	Placebo (P)
Cancro (C)	a = 1438	b = 1427
No cancro (N)	c = 18496	d = 18515

Tabella 10.2:

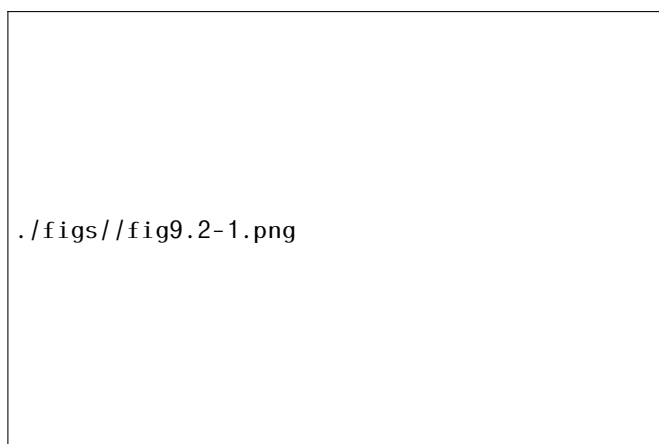


Figura 10.2:

Il rischio di sviluppare cancro invasivo nelle donne è più elevato nelle pazienti che ricevono aspirina?

RISOLVI QUESITO!

### Esempio 10.2.b: covid-19

LEZIONE 20211124

### 10.3 Test del $\chi^2$ per stabilire la significatività dell'associazione tra due variabili categoriche

Il test del  $\chi^2$  è il test statistico più usato per verificare l'associazione tra due variabili categoriche; tale test verifica la bontà di adattamento ai dati osservati di un modello nullo che assume l'indipendenza delle variabili.

#### Esempio 10.3.a: parassitosi e predazione

Un trematode parassitizza un pesce, presumibilmente modificandone il comportamento ed inducendolo a nuotare a profondità minori, il che rende il pesce più vulnerabile alla predazione da parte di un uccello, che è un altro ospite del parassita.

In una vasca all'aperto vengono posti tre diversi gruppi di pesci: non infestati, lievemente infestati e fortemente infestati; nella tabella 10.3 è riportato il numero di pesci predati dagli uccelli a seconda del loro livello di infestazione:

	Non infestati	Lievemente infestati	Fortemente infestati	Totali righe
Predati	1	10	37	48
Non predati	49	35	9	93
Totali colonne	50	45	46	141

Tabella 10.3:

I dati sono rappresentati nel mosaic plot 10.3.a:

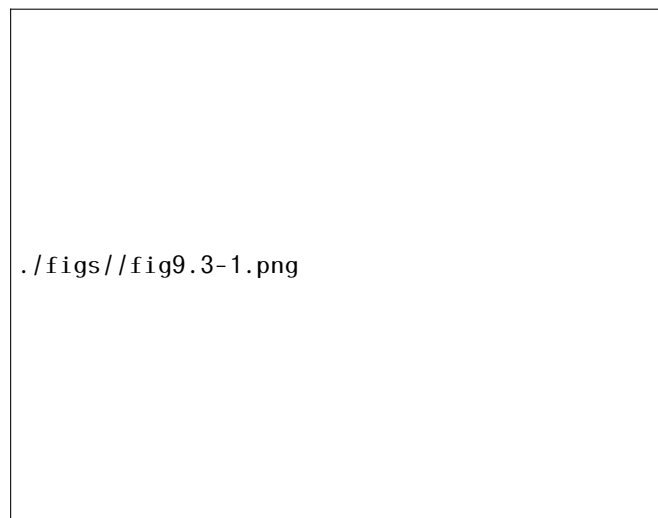


Figura 10.3:

Dai dati ottenuti risulta che solo il 2% dei pesci non infestati viene predato, mentre il 22% e l'80% dei pesci lievemente e fortemente infestati viene predato.

Si vuole verificare se la probabilità che i pesci vengano predati dipenda dal loro stato di infestazione parassitaria.

Le ipotesi sono:

- $H_0$ : parassitosi ed essere/non essere predati sono indipendenti;
- $H_A$ : parassitosi ed essere/non essere predati non sono indipendenti.

Per effettuare il test del  $\chi^2$  sulla tabella di contingenza si devono calcolare le frequenze attese per ciascuna cella nella tabella 10.3, assumendo che sia vera l'ipotesi nulla di indipendenza.

Secondo la regola del prodotto (6.6), se due eventi sono indipendenti, la probabilità che si verifichino entrambi è pari alla probabilità che si verifichi un evento per la probabilità che si verifichi l'altro; tale regola può essere impiegata per calcolare le proporzioni attese di pesci in ciascuna categoria, e poi le frequenze attese sotto l'ipotesi nulla.

Dunque, se l'infestazione e la predazione sono indipendenti, la probabilità che un pesce non infestato venga predato è:

$$Pr[\text{non infestato e predato}] = Pr[\text{non infestato}] Pr[\text{predato}] \quad (10.7)$$

Per calcolare la frequenza attesa di pesci non infestati e predati si deve prima stimare la probabilità che un pesce sia non infestato e la probabilità che un pesce venga predato, valori che si possono stimare in base ai dati della tabella 10.3; dunque, la probabilità stimata che un pesce non sia infestato è data dal numero totale di pesci non infestati nel campione (50), diviso il numero totale di pesci (141), e quindi è pari a 0.3546.

Stesso ragionamento vale per la probabilità stimata di essere predati, che quindi risulta essere 0.3404.

Perciò, assumendo vera l'ipotesi nulla di indipendenza, la probabilità stimata che un pesce sia non infestato e predato è pari a 0.1207, calcolata con la formula 10.7. Dunque, la frequenza attesa di pesci non infestati e predati sotto l'ipotesi nulla è data da tale probabilità (0.1207), moltiplicata per il numero totale di individui dell'insieme di dati (141), ed è quindi pari a 17. Ripetendo tale procedura per tutte le altre celle della tabella 10.3 si ottengono le frequenze attese per tutte le possibili combinazioni di casi, riportate nella tabella 10.4:

	<b>Non infestati</b>	<b>Lievemente infestati</b>	<b>Fortemente infestati</b>	<b>Totali righe</b>
<b>Predati</b>	17.0	15.3	15.7	48
<b>Non predati</b>	33.0	29.7	30.3	93
<b>Totali colonne</b>	50	45	46	141

Tabella 10.4: Frequenze attese di pesci predati e non, in base allo stato di infestazione.

I totali di righe e colonne ottenuti sono uguali ai corrispondenti osservati, dal momento che sono state usate le proporzioni relative ai dati per calcolare le frequenze attese.

### 10.3.1 Calcolo del $\chi^2$

Per verificare se le discrepanze tra i valori osservati e quelli attesi siano maggiori di quelle che ci si aspetterebbe per effetto del caso, si calcola il  $\chi^2$ .

Indicando con "c" il numero di colonne e con "r" il numero di righe:

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{[\text{osservato}(\text{colonna}, \text{riga}) - \text{atteso}(\text{colonna}, \text{riga})]^2}{\text{atteso}(\text{colonna}, \text{riga})} \quad (10.8)$$

Inserendo i dati:

$$\chi^2 = \frac{(1-17)^2}{17} + \frac{(49-33)^2}{33} + \frac{(10-15.3)^2}{15.3} + \frac{(35-29.7)^2}{29.7} + \frac{(37-15.7)^2}{15.7} + \frac{(9-30.3)^2}{30.3} = 69.5 \quad (10.9)$$

La distribuzione del  $\chi^2$  sotto l'ipotesi nulla è approssimata dalla distribuzione teorica del  $\chi^2$ . Per calcolare i gradi di libertà per la distribuzione del  $\chi^2$  di interesse, si contano il numero di righe e di colonne nella tabella; il numero di gradi di libertà è dato da:

$$df = (r-1)(c-1) \quad (10.10)$$

Nell'esempio corrente, dunque, i gradi di libertà sono  $(2-1)(3-1) = 2$ .

### 10.3.2 P-value

Il valore critico per la distribuzione  $\chi^2$  con  $df = 2$  ed  $\alpha = 0.05$  è pari a 5.99. Il valore osservato (69.5) si colloca verso l'esterno nella coda della distribuzione, essendo molto maggiore del valore critico; si rifiuta quindi l'ipotesi nulla: lo stato di infezione dei pesci influenza la probabilità che questi vengano predati.

Esiste anche una formula rapida per il calcolo delle frequenze attese (valore atteso nella cella di una tabella, corrispondente ad una certa riga e ad una certa colonna):

$$atteso[riga i, colonna j] = \frac{(totale\ riga\ i)(totale\ colonna\ j)}{totale\ generale} \quad (10.11)$$

Si può calcolare l'ultima cella in una riga o in una colonna anche per sottrazione, perché la somma delle frequenze attese per una data riga o colonna è uguale alla somma dei valori osservati. In una tabella di contingenza, il numero di celle in cui le frequenze non possono essere calcolate per differenza a partire dai totali di riga o colonna corrisponde anche ai gradi di libertà della distribuzione  $\chi^2$  che dovrà essere utilizzata nel test; infatti, i gradi di libertà in un test del  $\chi^2$  possono essere anche interpretati come il numero di categorie indipendenti, ovvero il numero di categorie sufficiente a calcolare, noti i totali di riga e colonna, tutti i valori nelle altre celle.

Il test del  $\chi^2$  che si usa per l'analisi delle tabelle di contingenza è un'applicazione particolare del più generale test di bontà dell'adattamento; in questo caso, il modello probabilistico il cui possibile adattamento ai dati osservati viene verificato con il test è il modello che assume l'indipendenza delle variabili.

### 10.3.3 Assunzioni del test $\chi^2$ per l'analisi delle tabelle di contingenza

Le assunzioni del test del  $\chi^2$  per l'analisi delle tabelle di contingenza sono le stesse del test del  $\chi^2$  di bontà dell'adattamento.

Se tali assunzioni non sono rispettate, sono disponibili almeno tre soluzioni:

- se la tabella è più grande di una  $2 \times 2$ , allora due o più categorie nelle righe (o nelle colonne) possono essere combinate per ottenere frequenze attese maggiori; in tal caso si deve comunque evitare di ottenere categorie prive di significato;
- se la tabella è una  $2 \times 2$ , si può impiegare il **test esatto di Fisher** (10.4);
- si utilizza un test di randomizzazione.

## 10.4 Test esatto di Fisher

<sup>2</sup> Il **test esatto di Fisher** si applica solo alle tabelle  $2 \times 2$  (teoricamente anche di dimensioni maggiori, ma molto complesso).

$$p = \quad (10.12)$$

#### Esempio 10.4.a: cancro

Si hanno 15 casi

	Trattamento	No trattamento		
Cancro	a = 2	b = 4	<b>6</b>	720
No cancro	c = 8	d = 1	<b>9</b>	362880
	<b>10</b>	<b>5</b>	<b>15</b>	
	362880	120		

Tabella 10.5:

<sup>2</sup>non integrato con il libro, che ne parla a pagina 132.

Le frequenze attese corrispondenti sono riportate nella tabella 10.6(es. calcolo per cella "a":  $(6 \cdot 10)/15$ ):

	Trattamento	No trattamento
Cancro	4	2
No cancro	6	3

Tabella 10.6: Frequenze attese

Se non ci fossero violazioni delle assunzioni del test del  $\chi^2$  si potrebbe proseguire con tale statistica; tuttavia, nell'esempio corrente, tre delle quattro celle hanno dei valori attesi inferiori a 5. Si usa dunque il test di Fisher.

La formulazione del test di Fisher con quattro celle è:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!} \quad (10.13)$$

Il test calcola direttamente la probabilità (QUALE?) .

Applicando tale formula al caso d'interesse, si ottiene una probabilità di ottenere per caso la distribuzione osservata pari a  $1.032 \cdot 10^{-8}$ , un valore estremamente basso, per cui si rifiuta l'ipotesi nulla; le due variabili (trattamento e cancro), quindi, non sono indipendenti.

Il test di Fisher è uno strumento molto potente, dal momento che non assume nessuna distribuzione (invece con il  $\chi^2$  si confronta la distribuzione con quella del  $\chi^2$ ). La probabilità calcolata con Fisher è la reale probabilità di ottenere la combinazione osservata date tutte le combinazioni possibili.

L'unica assunzione del test di Fisher è che i campioni siano casuali.





# Capitolo 11

## Inferenza in una popolazione con distribuzione normale

### 11.1 Riepilogo sulle relazioni tra variabili e statistiche

Una variabile può essere:

- **nominale**: può assumere valori che sono categorie;
- **ordinale**: può assumere valori che sono categorie che non possono essere in un ordine diverso da quello definito;
- **numerica**.

I **conteggi** sono tipo di misura che si può fare sulle variabili di tipo categoriale (nominale ed ordinale); potrebbero essere definiti variabili numeriche discrete.

Le variabili si possono studiare in diversi modi: si possono confrontare i valori che una variabile assume in un gruppo, con quelli che assume in un altro, oppure si può valutare l'associazione tra variabili diverse. Nella tabella sono riportati i test utilizzabili per fare ciò, in base al tipo di variabili d'interesse:

	Confronto	Associazione
<b>Nominali</b>	binomiale, $\chi^2$ , Z	$\chi^2$
<b>Ordinali</b>	$\chi^2$ , ?	?
<b>Conteggi</b>	binomiale e poisson e $\chi^2$ , ?	?
<b>Numeriche</b>	IC, ?	?

Tabella 11.1:

### 11.2 Distribuzione $t$ di Student

Uno dei problemi dell'approccio che prevede l'utilizzo del valore Z per il confronto tra valori osservati ed attesi è che la deviazione standard non è nota, perché non si conoscono i parametri della popolazione, ma solo le stime legate ai campioni, quindi i valori teorici non descrivono bene la realtà.

Man mano che  $n$  aumenta, la deviazione standard della media campionaria diminuisce, poiché essa è l'errore standard, ovvero l'errore dovuto al campionamento, e più  $n$  è elevato, più tale errore si riduce.

Il fatto che la stima dipenda da  $n$  è problematico: l'approccio del valore Z è utile solo quando i parametri sono noti:

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} \quad (11.1)$$

Se non si conoscono i parametri, c'è bisogno di includere l'incertezza campionaria nei confronti; si introduce quindi l'errore standard della media campionaria, che corrisponde alla deviazione standard della distribuzione campionaria, nella statistica; in questo modo si ottiene la statistica  $t$ , ovvero lo scarto tra media osservata (campionaria) e media teorica, rapportato all'errore standard della media campionaria:

$$t_{df} = \frac{\bar{Y} - \mu}{ES_{\bar{Y}}} \quad (11.2)$$

In cui:

$$ES_{\bar{Y}} = \frac{s}{\sqrt{n}} \quad (11.3)$$

In cui  $s$  è la deviazione standard osservata nel campione.

Utilizzando  $t$ , lo scarto tra medie non viene più messo in relazione con una deviazione standard che non è nota, ma ad un errore standard noto, calcolabile.

La statistica  $t$  ha dei gradi di libertà che  $Z$  non ha e che sono pari a:

$$df = n - 1 \quad (11.4)$$

Un esempio di una serie di curve di  $t$  (più la curva normale, quella tratteggiata) è riportato nella figura 11.2:



Figura 11.1:

Il  $t$  è quindi una famiglia di curve che variano al variare dei gradi di libertà, ovvero alla numerosità del campione.

Quando, con variabili normali, la numerosità del campione aumenta a sufficienza (attorno alle 40-50 unità), il  $t$  e la  $Z$  restituiscono gli stessi risultati, ovvero il  $t$  soglia e la  $Z$  soglia tendono a convergere per  $n$  elevati (cioè all'aumentare di  $n$ ,  $t$  tenderà ad 1.96, per tot gradi di libertà, con un  $\alpha$  di 0.05, dato che con  $Z$  1.96 è il valore soglia di  $Z$  che taglia a destra e sinistra della curva il 2.5% della curva sottesa, per un totale del 5%).

Al diminuire di  $n$ , il  $t$  soglia è più elevato della  $Z$  soglia.

Il  $t$  si basa sulla stima di una media e di una deviazione standard ed assume che vi sia una distribuzione normale (e quindi che la media sia un valore rappresentativo).

Dunque, quello del  $t$  di Student è un approccio migliore di quello che utilizza  $Z$  quando i parametri sono ignoti ed  $n$  è relativamente basso (<40-50).

### 11.3 Intervalli di confidenza e $t$ test per un campione

Lo stesso problema che si riscontra nel confronto tra valori attesi ed osservati si ha nella definizione degli intervalli di confidenza: utilizzare  $Z$  non permette di tenere conto del fatto che si sta utilizzando un campione e che quindi non si conosce la variabilità teorica.

Per calcolare gli intervalli di confidenza si utilizza dunque la  $t$ , e non la  $Z$ , che comunque corrispondono per  $n$  elevati ( $>40-50$ , quindi usare  $t$  per il calcolo degli IC solo per  $n$  inferiori). Dunque, quando opportuno, l'**intervallo di confidenza della media campionaria** si calcola come:

$$\bar{Y} - t_{\alpha(2),df} ES_{\bar{Y}} < \mu < \bar{Y} + t_{\alpha(2),df} ES_{\bar{Y}} \quad (11.5)$$

In questo caso, il parametro di riferimento è la media ( $\mu$ ), ed i gradi di libertà sono  $n - 1$ . Le ipotesi sono:

- $H_0$ : la media campionaria non differisce da una media ipotetica;
- $H_A$ : la media campionaria differisce da una media ipotetica.

Il test è a due code, poiché l'ipotesi alternativa prevede sia i casi in cui la media campionaria è più grande, sia quelli in cui essa è più piccola della media ipotetica.

#### Esempio 11.3.a: temperatura corporea

La temperatura corporea dell'uomo è pari a 98.6°F. Le ipotesi sono:

- $H_0$ : la temperatura corporea dell'uomo è di 98.6°F;
- $H_A$ : la temperatura corporea dell'uomo non è di 98.6°F.

Si analizzano 25 soggetti sani e si ottengono  $\bar{Y} = 98.524$  ed  $s = 0.678$ .

Per prima cosa si dovrebbe plottare la distribuzione di frequenza della temperatura dei 25 soggetti e vedere se questa è grossomodo normale:

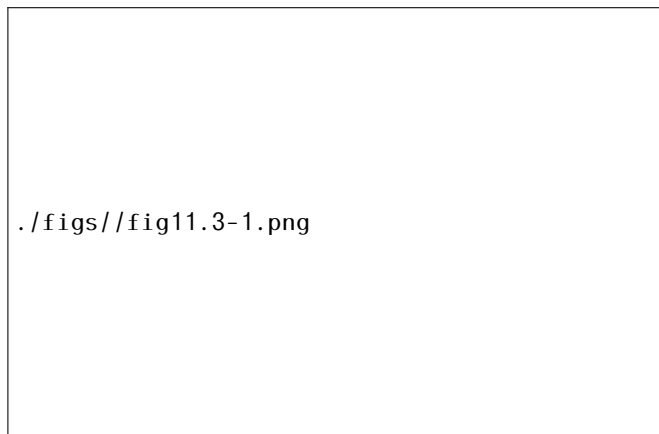


Figura 11.2: Frequenza assoluta, dato che sull'asse delle y è riportato il conteggio

La media della temperatura osservata differisce significativamente dalla media attesa?

**FAI CON IC95% E T TEST**

#### Esempio 11.3.b: guida distratta

Dieci persone vengono osservate alla guida in due casi, con il cellulare e dopo aver bevuto due birre. Nella tabella 11.2 sono riportati l'identificativo di ogni soggetto (il soggetto uno per il cellulare e quello per la birra sono persone diverse, e così via) ed il numero di birilli fatto cadere in entrambi i casi:

ID	Cellulare	Birra
1	14	28
2	44	36
3	17	38
4	69	42
5	77	57
6	80	60
7	82	80
8	86	81
9	99	83
10	99	91

Tabella 11.2:

Da fare:

- calcolare IC95% per i due risultati dei test;
- confrontare le medie ottenute con quelle attese (usando i valori attesi di  $t$ );
- se  $n = 20$  invece di 10, cambierebbe qualcosa?

## **Parte III**

# **Associazioni**



# Capitolo 12

## 12.1