DIAPOSTIVE DI BIOSTATISTICA

DOCENTE: PROF. GIORGIO BERTORELLE

Le diapositive dalla pagina 1 alla pagina 106 sono utilizzate durante il corso di Biostatistica della Laurea Triennale in Scienze Biologiche, Università di Ferrara

Creato il: 7.6.2011 **Pubblicato il:** 8.3.2012 **Autore:** Giorgio Bertorelle

Anno Accademico 2010-2011
Università degli Studi di Ferrara
Corso di Laurea Triennale in Scienze Biologiche

BIOSTATISTICA

(6 crediti)

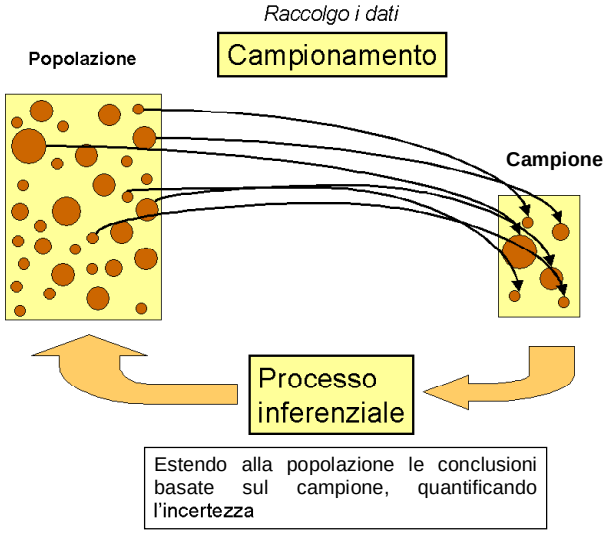
Docente: Prof. Giorgio Bertorelle

DI COSA CI OCCUPEREMO IN QUESTO CORSO?

- Le basi della statistica applicata allo studio dei dati biologici
- Cos'e la statistica?
  - Studio scientifico dei dati, raccolti o ottenuti in un esperimento, al fine di descrivere un fenomeno, interpretarlo, scegliere tra ipotesi alternative
- BIOSTATISTICA: I dati provengono da organismi viventi, che sono altamente variabili
- Fondamentale per capire le proprietà degli organismi viventi, e quindi fondamentale in biologia, medicina, agraria, ecc.
- Parole chiave:
- **Campione**: studio i fenomeni biologici a partire da campioni, non da **popolazioni**
- **Incertezza**: quantifico l'incertezza dovuta al fatto che riesco solo ad analizzare campioni e non popolazioni

DI COSA MI OCCUPO IO?

- Studiare la variabilità genetica per ricostruire il passato dell'uomo e di altri animali
- I dati sulla variabilità genetica devono prima essere "prodotti" in laboratorio (attraverso tecniche di biologia molecolare a partire da materiale organico come sangue, muscolo, peli, ossa, ecc. e anche a partire da campioni scheletrici di individui vissuti migliaia di anni fa) e poi essere analizzati ***statisticamente*** per poter giungere a conclusioni credibili
- Questi studi sono rilevanti per capire l'evoluzione delle specie e per prevenire la perdita di biodiversità



Esempi di popolazioni e campioni

√Tutti i gatti caduti dagli edifici di New York
√Tutti i geni del genoma umano
√Tutti gli individui maggiorenni in Australia
√Tutto i serpenti volanti del paradiso nel Borneo
√Tutti i bambini asmatici di Milano

√I gatti caduti portati in un singolo ambulatorio in un certo intervallo di tempo
√20 geni umani
√Un pub in Australia frequentato da maggiorenni
√Otto serpenti volanti del Borneo
√50 bambini asmatici a Milano

GLI OBIETTIVI DELLA (BIO)STATISTICA

1. Stimare i **parametri** (grandezze incognite importanti) a partire da campioni e capire la precisione delle **stime**

   - Per esempio, altezza media, proporzione di pazienti guariti se trattati con un certo farmaco, differenza tra luce emessa da lucciole in due aree geografiche diverse, concentrazione media di mercurio in un lago

2. Testare (verificare) delle ipotesi: i dati raccolti sono compatibili con una certa **ipotesi nulla**?

   - L'ipotesi nulla (o ipotesi zero, o H_0) è un'affermazione che riguarda in genere un parametro della popolazione

   - Esempi di ipotesi nulla, in genere il punto di vista scettico: il nuovo farmaco non è migliore del precedente (ovvero, la velocità di guarigione non è cambiata); i rospi destrimani sono tanti quanti i rospi mancini (proporzione = 50%); la frazione di maschi nella parte destra dell'aula è la stessa della frazione dei maschi nella parte sinistra (differenza tra proporzioni = 0).

- La statistica serve anche per ragionare su come raccogliere dati in natura o attraverso esperimenti in laboratorio, ovvero per definire le buone pratiche di disegno dell'esperimento e di strategia del campionamento; è una fase cruciale!

- La statistica serve anche per riassumere e rappresentare graficamente i dati raccolti

   - Distinzione tra statistica descrittiva e statistica inferenziale

STRUTTURA DEL CORSO

- Lezioni teoriche in aula con molti esempi di applicazioni in ambito biologico (quasi sempre reali, a volte immaginari)
- Esercizi in aula
- Necessaria sempre la calcolatrice
- Gli esempi permettono di ricordare sia la parte teorica che quella pratica. E' importante ricordare gli esempi.
- NON CONVIENE STUDIARE TEORIA ED ESEMPI DI APPLICAZIONI SEPARATAMENTE
- Ogni argomento è collegato a quelli precedenti, E' QUASI INUTILE SEGUIRE LE LEZIONI SE NON SI STUDIA CON CONTINUITA'

TIPOLOGIA DELL'INSEGNAMENTO E QUALCHE CONSIGLIO

- E' necessario capire e non imparare a memoria
- La teoria serve per capire come analizzare i dati e per svolgere correttamente gli esercizi.
- Gli esercizi sono applicazioni a dati biologici delle tecniche statistiche. Sono una verifica fondamentale della comprensione della parte teorica.

DOMANDE

- Se non capite a lezione, fate domande (utile sempre!)
- Se non capite dopo aver studiato gli appunti e il libro, consultate il docente (prima per email, poi eventualmente per appuntamento). Ricordate che i vostri docenti svolgono anche attività di ricerca
  - ggb@unife.it in generale (orario ricevimento: venerdì dalle 12.30 alle 14.00)
- Non arrivate a fine corso con domande/problemi riscontrati fin dalle prime lezioni!

VALUTAZIONI

- Dello studente
  - Esame finale scritto con domande a scelta multipla e esercizi
    - Revisione degli argomenti in una sessione intermedia
  - Gli appelli successivi per chi non supera l'esame negli appelli a fine corso potranno essere scritti o orali
- Del docente
  - Scheda di valutazione, attenzione a compilarla sulla base delle domande richieste

MATERIALE DIDATTICO

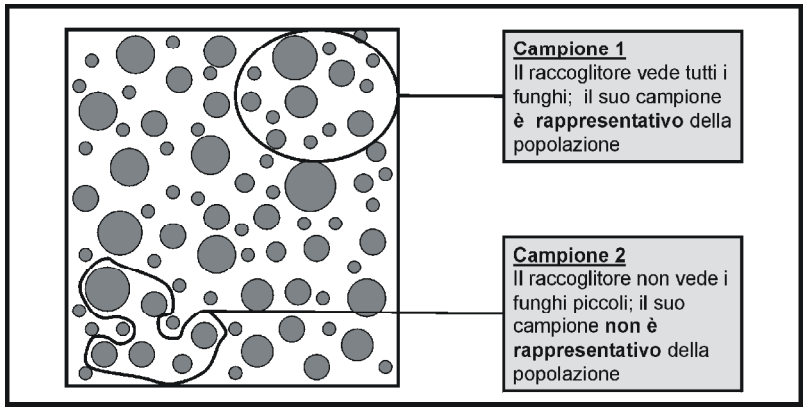
- Vostri appunti (la frequenza è consigliata)
- Libro: MC Whitlock, D Schluter (2010) - ANALISI STATISTICA DEI DATI BIOLOGICI. Edizione italiana a cura di G. Bertorelle - Zanichelli Editore
- Materiale disponibile sito docente riferito alle lezioni dello scorso anno: da consultare in alcuni casi
- [Libri di testo online (in inglese)
  - http://www.statsoft.com/textbook/
  - http://davidmlane.com/hyperstat/]

SITO WEB CORSO
http://docente.unife.it/giorgio.bertorelle/didattica_insegnamenti

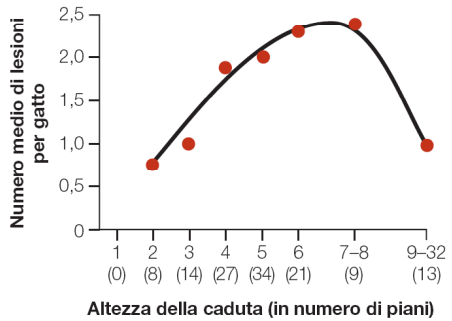
SCHEDA 1 (libro di testo): La biologia e la storia della statistica

Campionare le popolazioni

- I dati che analizziamo (il nostro campione), siano osservazioni (Es: alberi sui quali misuro tasso fotosintetico) o frutto di esperimenti (Es: lo stato di salute dei pazienti trattati o meno con un farmaco) devono essere *rappresentativi* della popolazione. Se non lo sono, si dice che il campione è distorto (o affetto da distorsione o *bias)*
- Vediamo un esempio di un campione distorto



- Vediamo un altro esempio di campione distorto



Proprietà dei buoni campioni

1. Basso errore di campionamento
2. Bassa distorsione delle stime

L'errore di campionamento è la differenza dovuta al caso tra stima e parametro
A parità di tutti gli altri fattori, l'errore di campionamento si riduce (e la precisione della stima aumenta) quando la dimensione (detta anche numerosità) del campione è elevata

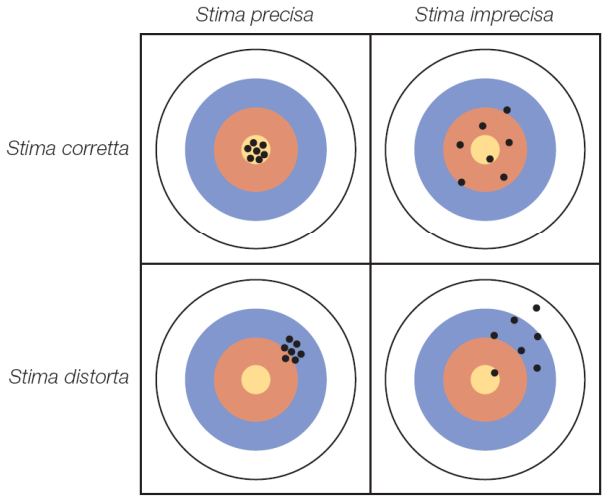
La distorsione di una stima è una discrepanza sistematica tra stima e parametro. Un campione **casuale** riduce la distorsione (e permette anche di quantificare correttamente l'errore di campionamento). La stima della dimensione dei funghi, o delle ferite dei gatti a diversi piani, sono distorte.

Altri esempio di stime distorte

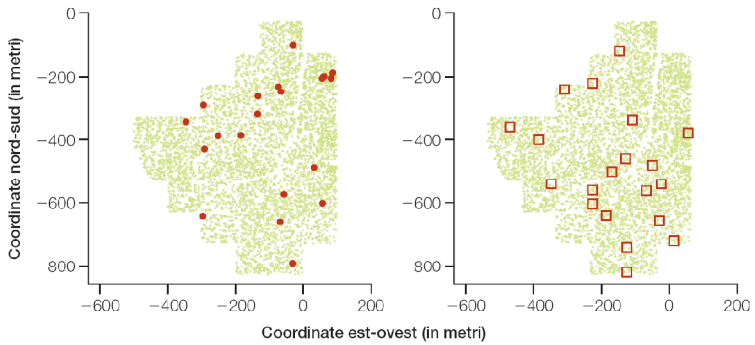
a) Piante raccolte solo sul bordo della strada
b) Stime basate su sondaggi telefonici
c) Dimensioni medie di una specie di pesci catturati con reti da pesca con maglie troppo grandi
d) Campionamenti in aree non rappresentative

Nel **campionamento casuale**, ogni individuo di una popolazione ha una probabilità uguale e indipendente di essere selezionato

I concetti di distorsione e precisione: 2 cose diverse!



Come si ottiene un campione casuale?
E' sempre possibile ottenerlo?
Vediamo un esempio con i 5699 alberi nella foresta di Harvard



Il **campione di convenienza** e il **campione di volontari** sono spesso distorti (non rappresentativi)

Esempi di **campioni di convenienza**
- Lesioni dei gatti che cadono dai cornicioni stimati sulla base dei gatti "ospedalizzati"
- Merluzzi stimati sulla base della pesca
- Inchieste telefoniche

Esempi di **campione di volontari (uomo)**
- Campioni provenienti da individui pagati
- Campioni di individui che si offrono di rispondere a domande "imbarazzanti"

Dal campione alle variabili

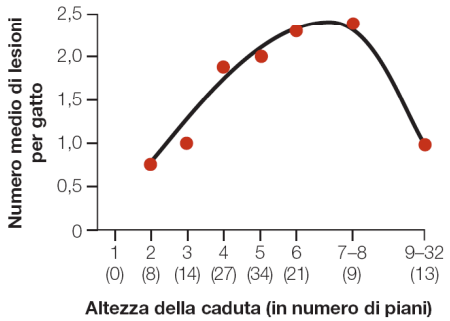
Cos'è una *variabile*?

- una qualsiasi caratteristica misurata o registrata in un'unità campionaria. Generalmente le variabili sono indicate con lettere maiuscole e i valori che possono assumere con lettere minuscole, spesso indicizzati per indicare il valore assunto dalla variabile in una specifica osservazioni

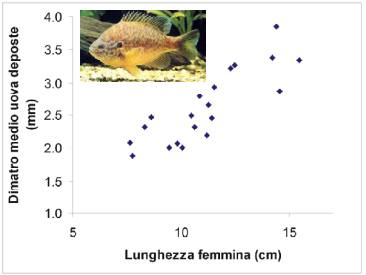
- *V*ariabili **categoriche** (o qualitative), **nominali** o **ordinali**
  - Es: Lingua parlata; gravità del morso di serpente; gruppo sanguigno; classe di dimensione; mutazione presente; sesso
- Variabili **numeriche** (quantitative), **continue** o **discrete**
  - Es: temperatura; peso; numero di accoppiamenti; concentrazione; numero medio di sigarette al giorno;numero di amminoacidi in una proteina
  - Dati numerici possono essere ridotti a dati categorici

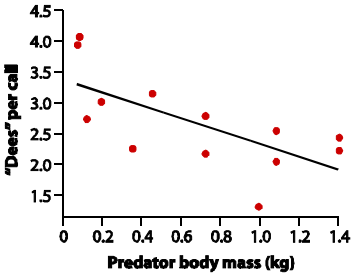
Variabili risposta e variabili esplicative (dipendenti e indipendenti)

- Ipertensione arteriosa e rischio di ictus
- Piani edificio e lesione gatti



- Dimensione femmina e dimensione uova deposte
- Peso del predatore in avvicinamento e numero di richiami d'allarme





Studi sperimentali e studi osservazionali

- Negli studi sperimentali, il ricercatore assegna casualmente diversi trattamenti agli individui
  - Es: alcuni topi, scelti a caso, riceveranno un trattamento farmacologico, gli altri non lo ricevono
- Negli studi osservazionali, non è il ricercatore che assegna i trattamenti
  - Es: analizzo la relazione tra colorazione e predazione: non scelgo io il colore (trattamento) da assegnare a ciascun individuo (può diventare sperimentale? Come?)
  - Es: studio la relazione tra fumo e tumore: non scelgo io i soggetti a cui somministrare il "trattamento fumo"

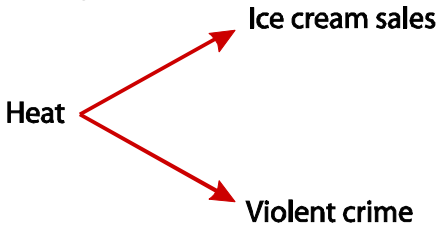
Negli studi osservazionali, una associazione può essere dovuta ad una causa comune, non ad una relazione di causa ed effetto tra le due variabili analizzate.

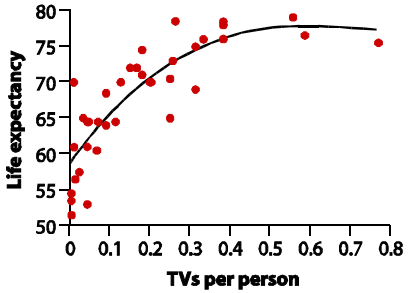
Per esempio, i pesci rossi sono meno predati di quelli rosa, ma in realtà potrebbe esserci una terza variabile (salute media) che determina colore e livello di predazione.

Oppure, potrebbero essere gli individui più depressi che fumano, e il rischio di tumore potrebbe dipendere dalla depressione e non dal fumo.

Se da uno studio osservazionale passo ad uno studio sperimentale (per esempio, in un campione pesci, metà scelti a caso li coloro di rosso e metà di rosa; oppure, scelgo a caso un certo numero di topi e li metto in gabbie con fumo, un altro numero in gabbie senza fumo), posso capire molto di più riguardo le relazioni di causa ed effetto (la **variabile di confondimento** di distribuisce a caso nei gruppi confrontati)

Esempi in Scheda 4, Pg. 105





PROBLEMI 10, 11, 12, 13, 17 , pagine 12 e 13

Visualizzare, descrivere e sintetizzare i dati nel campione (statistica descrittiva)

- Dai dati alle frequenze (di occorrenza) alla tabella di frequenza alla distribuzione di frequenza

- Es: 22 nidi di merlo: contiamo il numero di uova
  - unità campionaria = nido
  - variabile numerica discreta: numero di uova

- $x_1 = 0$; $x_2 = 2$; $x_3 = 2$; $x_4 = 0$; $x_5 = 1$; $x_6 = 3$; $x_7 = 3$; $x_8 = 2$; $x_9 = 2$; $x_{10} = 4$; $x_{11} = 1$; $x_{12} = 4$; $x_{13} = 2$; $x_{14} = 1$; $x_{15} = 2$; $x_{16} = 3$; $x_{17} = 3$; $x_{18} = 6$; $x_{19} = 4$; $x_{20} = 2$; $x_{21} = 3$; $x_{22} = 3$,

- dove x_i, indica il valore assunto dalla variabile X nella i-esima osservazione, con l'indice i che varia da 1 a n (n = 22 = dimensione del campione)

- ***La tabella di frequenza***:

Numero di uova	*Frequenza (numero di nidi)*
0	2
1	3
2	7
3	6
4	3
5	0
6	1
Totale	22

- *Ovviamente la somma di tutte le frequenze è pari a n*

- ***frequenze assolute*** *(quelle nella tabella precedente, indicate con* n_i, dette anche *numerosità*)
  - Ovviamente $\sum n_i = n$

- ***frequenze relative*** (f_i, o, a volte, p_i,)
  - Ovviamente varia tra 0 e 1

$$f_i = p_i = \frac{n_i}{n}$$

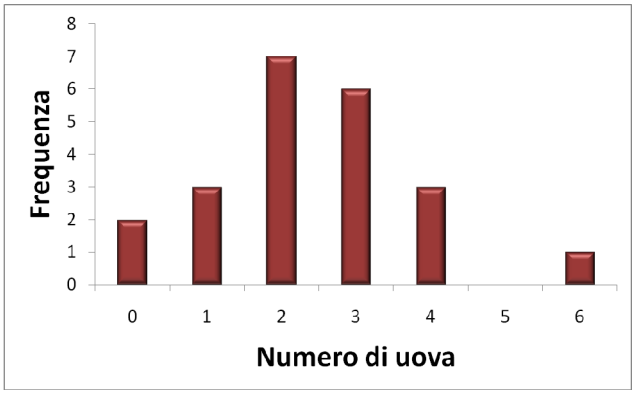
- ***frequenza percentuale***

$$f_i(\%) = f_i \times 100$$

Il termine generico *frequenza* è spesso utilizzato per indicare cose diverse

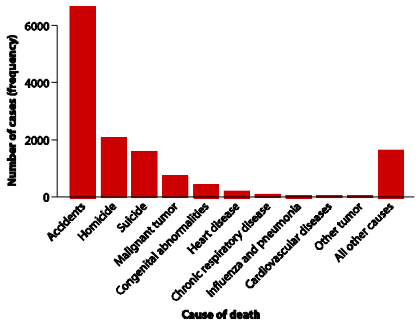
- ***La distribuzione di frequenza (diagramma a barre)***

o



- ***Tabella e distribuzione di frequenza per una variabile categorica nominale (diagramma a barre)***

Cause	No.deaths
Accidents	6688
Homicide	2093
Suicide	1615
Malignant tumor	745
Heart disease	463
Congenital abnormalities	222
Chronic respiratory disease	107
Influenza and pneumonia	73
Cerebrovascular diseases	67
Other tumor	52
All other causes	1653



- ***I dati sull'abbondanza (variabile numerica discreta) in 43 specie***

Tabella 2.1-2

Abbondanza di ciascuna specie di uccello osservata durante quattro rilevamenti nell'Organ Pipe Cactus National Monument.

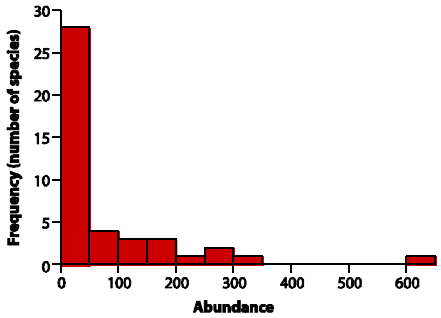
Specie	Abbondanza	Specie	Abbondanza
Geococcyx californianus (corridore della strada, roadrunner)	1	*Cathartes aura* (avvoltoio tacchino)	23
Archilocus alexandri	1	*Tachycineta thalassina* (rondine verdeviola)	23
Tyrannus verticalis (re dei tiranni occidentale)	1	*Chordeles acutipennis* (succiacapre minore)	25
Quiscalus mexicanus (gracula codalunga)	1	*Icterus parisorum* (oriolo di Scott)	28
Molothrus aeneus (vaccaro bronzeo)	1	*Progne subis* (rondine viola)	33
Bubo virginianus (gufo reale della Virginia)	2	*Amphispiza bilineata* (passero golanera)	33
Calypte costae (colibrì di Costa)	2	*Molothrus ater* (vaccaro testabruna)	59
Catherpes mexicanus (scricciolo dei canyon)	2	*Caragyps atratus* (avvoltoio nero	64
Pipilo fuscus (toui dei canyon)	2	*Vermivora luciae* (parula di Lucy)	67
Parabuteo unicinctus (poiana di Harris)	3	*Colaptes chrysoides* (picchio dorato)	77
Lanius ludovicianus (averla stolida)	3	*Myiarchus tyrannulus* (pigliamosche crestabruna)	128
Icterus cucullatus (oriolo dal bavaglino)	4	*Zenaida macroura* (tortora piangente americana)	135
Mimus polyglottos (mimo settentrionale)	5	*Callipepla gambelii* (quaglia di Gambel)	148
Falco sparverius (gheppio americano)	7	*Polioptila melanura* (pigliamoschini codanera)	152
Columba livia (piccione selvatico)	7	*Myiarchus cinerascens* (pigliamosche golacenere)	173
Vireo bellii (vireo di Bell)	10	*Toxostoma curvirostre* (mimo beccocurvo)	173
Corvus corax (corvo imperiale comune)	12	*Campylorhynchus brunneicapillus* (scricciolo dei cactus)	230
Cardinalis cardinalis (cardinale settentrionale)	13	*Auriparus flaviceps* (verdino)	282
Passer domesticus (passero domestico, passera europea, passera oltremontana)	14	*Carpodacus mexicanus* (fringuello delle case)	297
Picoides scalaris (picchio scalare)	15	*Melanerpes uropygialis* (picchio di Gila)	300
Buteo jamaicensis (poiana codarossa)	16	*Zenaida asiatica* (tortora alibianche)	625
Phainopepla nitens (fainopepla)	18		

➢ *Tabella e distribuzione di frequenza per una variabile numerica discreta (istogramma)*

Tabella 2.1-3

Distribuzione di frequenza dell'abbondanza di specie di uccelli nell'Organ Pipe Cactus Monument.[a]

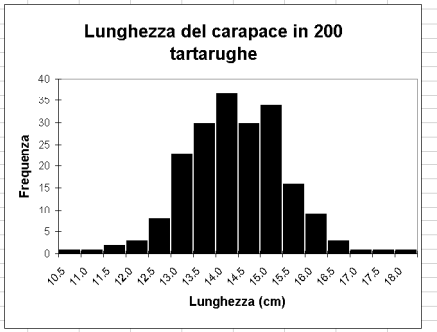
Abbondanza	Frequenza (numero di specie)
0 ÷ 50	28
50 ÷ 100	4
100 ÷ 150	3
150 ÷ 200	3
200 ÷ 250	1
250 ÷ 300	2
300 ÷ 350	1
350 ÷ 400	0
400 ÷ 450	0
450 ÷ 500	0
500 ÷ 550	0
550 ÷ 600	0
600 ÷ 650	1
Totale	43



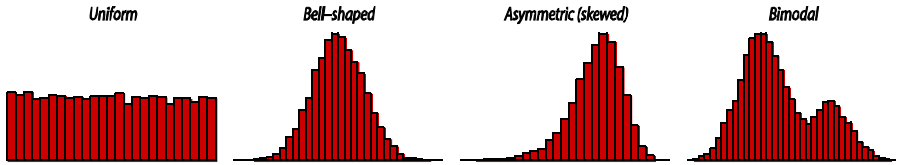
Presenza di un oulier

➢ *Tabella e distribuzione di frequenza per una variabile numerica continua (istogramma)*

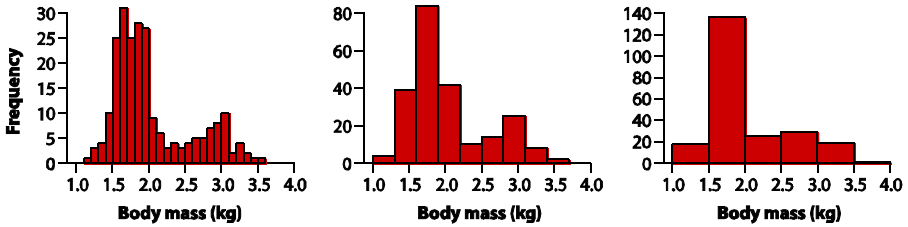
	Classe (estremo sup.)	Frequenza
3	10.5	1
4	11.0	1
5	11.5	2
6	12.0	3
7	12.5	8
8	13.0	23
9	13.5	30
10	14.0	37
11	14.5	30
12	15.0	34
13	15.5	16
14	16.0	9
15	16.5	3
16	17.0	1
17	17.5	1
18	18.0	1



➢ *La forma di una distribuzione di frequenza*



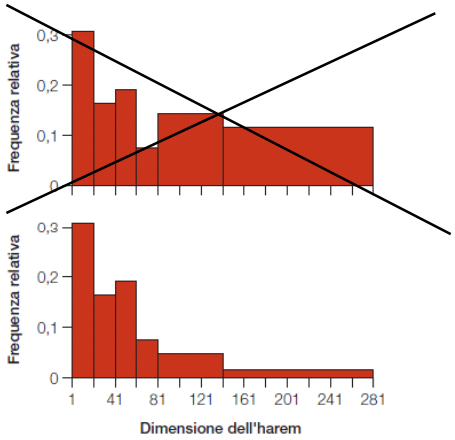
➢ Come costruire un buon istogramma?



➢ Regolette di base

- Numero di classi: inizialmente provare la radice quadrati di n oppure (1 + ln(n)/ln(2))
- Evitare classi vuote
- Usare limiti chiari e leggibili
- Inserire valori uguali al limite tra due classi nella classe superiore

- **Come costruire un buon istogramma?**



Attenzione ad ampiezze diverse!

- **Attenzione:**

o ***distribuzione di frequenza*** : si riferisce al campione

o ***distribuzione di probabilità****: si riferisce alla popolazione*

o ***distribuzione di probabilità teorica****: è una funzione matematica*

PROBLEMI 10, 11, 12, 13, 17 , pagine 12 e 13

10) D,C,C,D,C

11) Osservazionale: il ricercatore non ha il controllo su quale donna avrà un aborto e quale perderà il feto per altre cause

12) ND, NC, CO, NC, CO, NC, CN, ND, CN, NC

13) Osservazionale; sottospecie del pesce e lunghezza d'onda di massima sensibilità; esplicativa:sottospecie; risposta: d'onda di massima sensibilità

17) Le 60 misure non sono un campione casuale. I 6 tuffi artificiali effettuati per ogni animale non sono indipendenti. Le 6 misure ottenute nei 6 tuffi di un animale sono probabilmente più simili tra loro di quanto sarebbero 6 misure prese da 6 animali diversi.

- **Confrontare gruppi e visualizzare eventuali associazioni: tabelle e distribuzioni di frequenza per due variabili categoriche**

Tabella 2.3-1

Tabella di contingenza che mostra l'incidenza della malaria aviaria nelle femmine di cinciallegra in relazione al trattamento sperimentale.

	Gruppo di trattamento sperimentale		
	Gruppo di controllo	Gruppo a cui sono state sottratte due uova	Totale delle righe
Malaria	7	15	22
Assenza di malaria	28	15	43
Totale delle colonne	35	30	65

Tabella di contingenza (2 variabili categoriche, una esplicativa e una risposta, ciascuna con due possibili valori)

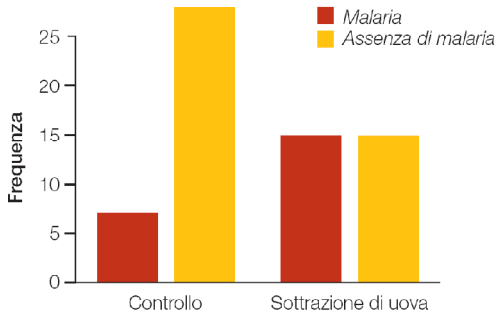
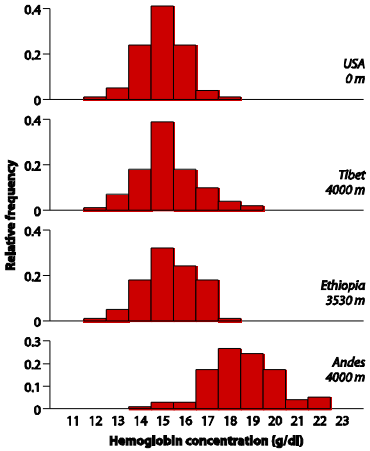


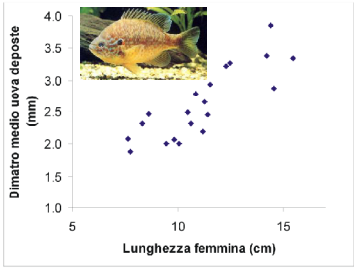
Figura 2.3-1
Diagramma a barre raggruppate sulla relazione fra lo sforzo riproduttivo e la malaria aviaria nelle cinciallegre. I dati sono tratti dalla Tabella 2.3-1, dove $n = 65$ è il numero di uccelli.

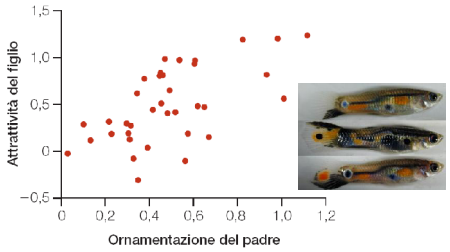
Diagramma a barre raggruppate

- **Confrontare gruppi e visualizzare eventuali associazioni: confronto tra istogrammi di variabili numeriche in più gruppi (una variabile numerica e una variabile categorica)**



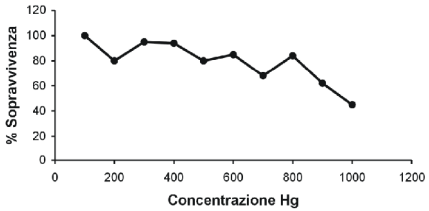
- **Visualizzare eventuali associazioni tra due variabili numeriche: lo scatter plot e il diagramma a linee**
  - ***Scatterplot***, o **nube di punti** (due esempi)





- Associazione positiva (come nelle figure), negativa, o nulla
- Relazione lineare o non lineare

- ***Diagramma a linee*** (due esempi)



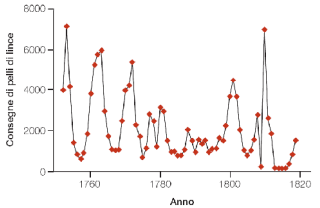
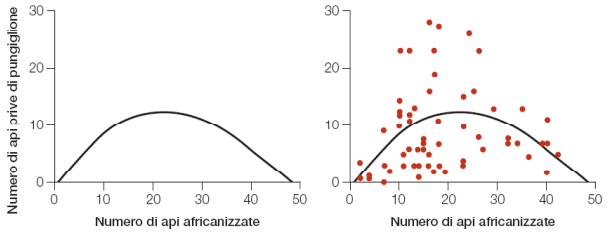


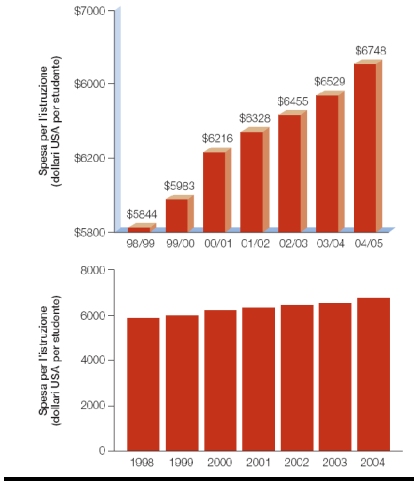
Figura 2.5-2
Diagramma a linee che mostra le consegne di pelli di lince della Hudson's Bay Company dal 1752 al 1819.

Considerazioni generali sulla visualizzazione grafica

- Chiarezza
- Completezza
- Onestà

Figura 2.6-1
Curva che rappresenta la relazione tra il numero di api prive di pungiglione (meliponini) tropicali indigene autoctone e il numero di api africanizzato (*Apis mellifera scutellata*)su arbusti in fiore nella Guyana Francese. I punti relativi alle osservazioni dati sono stati cancellati nella parte sinistra della figura. Da Roubik (1978); ridisegnato.

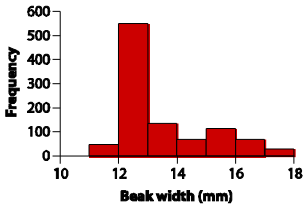




PROBLEMI 14,16,18,19,23,24. pagine 34 e 35

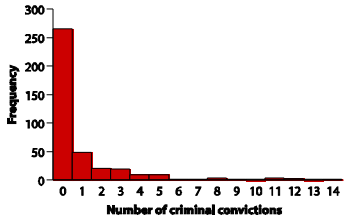
PROBLEMI 14,16,18,19,23,24 (escluso c). Pagine 34 e 35

14) Tra 12 e 13 cm; circa 50%; Provare altre ampiezze degli intervalli; bimodale

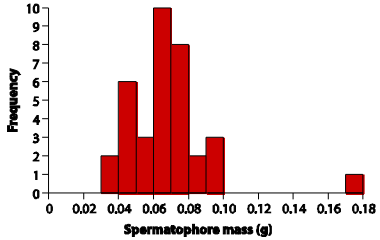


16) Tabella di frequenza; Una (numero di condanne); 21; 265 su 395 (0.67); istogramma (meglio sarebbe stato diagramma a barre); asimmetrico a destra, unimodale, no outlier; no, non è un campione casuale di ragazzi britannici

39



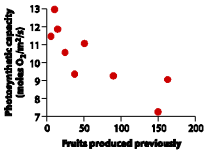
18) istogramma; la maggior parte della distribuzione è simmetrica, con moda intorno ai valori 0.06-0.07; c'e' una misura estrema; outlier



40

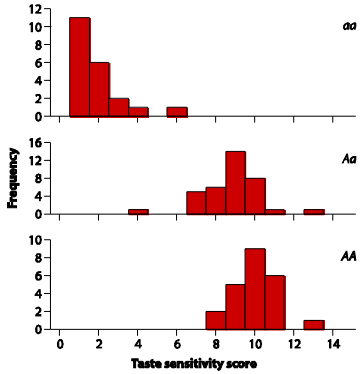
19) Due variabili numeriche continue; scatter plot; relazione positiva non lineare; No, il campione no è casuale perché non tutte le misure sono indipendenti (ogni pesce è stato misurato più volte)

23) Scatter plot; numero di frutti prodotti in precedenza (vogliamo capire come questo spieghi eventualmente la capacità fotosintetica); negativa



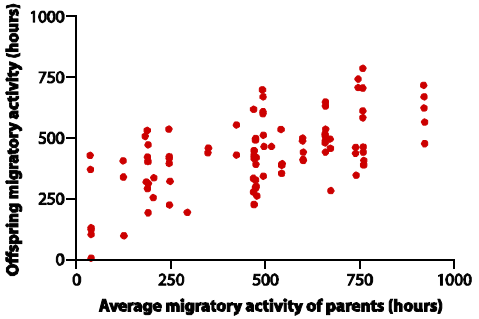
41

24)



Istogrammi raggruppati; V. esplicativa: genotipo al gene PCT; V. risposta: sensibilità gustativa; prima è categorica, la seconda numerica

42



Scatter plot; V.e: attività migratoria dei genitori; V.r.: attività migratoria della prole; Numeriche;

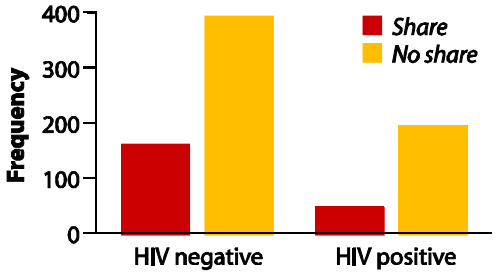
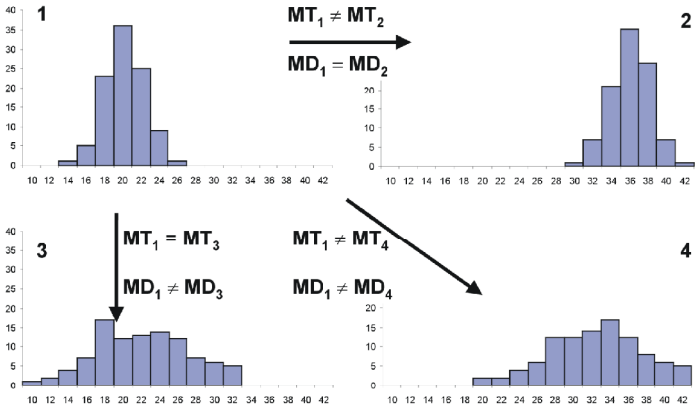


Diagramma a barre raggruppate; V.e.: stato di HIV (positivo o negativo); V.r.: condivisione ago; Categoriche.

Indici sintetici di una distribuzione: le statistiche descrittive

Per variabili numeriche

- Misure di posizione (o di tendenza centrale)
  - cercano di identificare il valore "tipico" di una distribuzione, ovvero la posizione, nella scala della variabile analizzata, intorno alla quale si concentrano le osservazioni
- Misure di dispersione
  - sintetizzano il grado di variabilità dei dati
- Le misure di posizione ci informano su una situazione media, quelle di dispersione su quanto questa situazione media abbia significato rispetto alla variazione tra individui



MT : misura di tendenza centrale
MD : misura di dispersione

- Conoscere la dispersione dei dati equivale a conoscere qualcosa sul valore di ogni singolo valore per la comprensione di un fenomeno.
- Se la dispersione è molto elevata, le singole osservazioni possono essere anche molto diverse, e quindi singolarmente di scarso valore.
- Si può dire quindi che all'aumentare della dispersione il numero di osservazioni necessarie per trarre delle conclusioni generali a partire da un campione deve aumentare.
- Quando la variabilità è molto bassa può anche non essere necessario effettuare molte osservazioni, e forse nemmeno ricorrere alla statistica inferenziale.

Per variabili categoriche

- Alcuni indici di posizione
- Proporzione (o frequenza relativa)

Misure di posizione

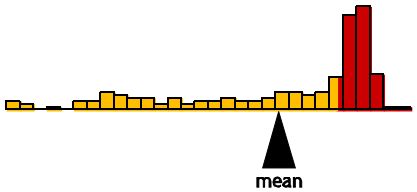
La media aritmetica (la media)

- M*edia campionaria*, della variabile *X*, indicata con $\bar{x}$.

$$\bar{x} = \frac{\sum_{1}^{n} x_i}{n}$$

- La somma delle differenze dei singolo valori dalla media (detti *scarti dalla media*) è uguale a 0 e quindi la media si può considerare il baricentro del campione dove si bilanciano gli scarti.

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$$



- Media a partire da una tabella di frequenza :

$$\bar{x} = \frac{\sum x_i n_i}{n}$$

Esempio

Numero di uova	*Frequenza (numero di nidi)*
0	2
1	3
2	7
3	6
4	3
5	0
6	1
Totale	22

$$\overline{x} = \frac{\sum x_i n_i}{n} = \frac{0\times 2 + 1\times 3 + 2\times 7 + 3\times 6 + 4\times 3 + 5\times 0 + 6\times 1}{22} = \frac{53}{22} = 2{,}41$$

La mediana

- La *mediana* è il valore centrale in una serie di dati ordinati. Per esempio

Dati: 30, 49, 74, 40, 63, 295, 60

Dati ordinati: 30, 40, 49, 60, 63, 74, 295

- La mediana è quindi il valore che divide un campione di dati ordinati in due parti ugualmente numerose. In altre parole, metà dei valori nel campione sono più piccoli della mediana, e metà sono più grandi. E' evidente quindi che la mediana è una misura della tendenza centrale

- Attenzione quando n è pari

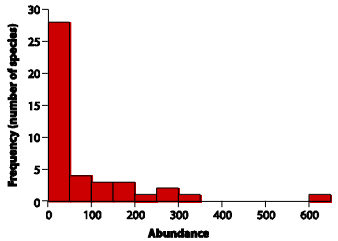
La moda

- La *moda* è semplicemente il valore osservato più spesso nel campione (il valore corrispondente al picco più alto nella distribuzione di frequenza)

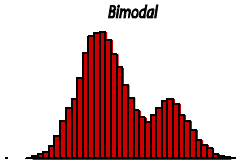
Dati: 0, 1, 5, 2, 2, 2, 3, 3, 3, 2, 4, 4, 1,2

valori	*frequenza*
0	1
1	2
2	5
3	3
4	2
5	1

La moda è quindi pari a 2



- Classe modale (se i dati sono raggruppati in classi contenenti valori diversi, come nell'istogramma)



- Distribuzione detta bimodale anche se esiste una sola classe modale

Proprietà della media

- la media implica la somma di valori numerici e quindi

  ⇒ ha un significato solo per le variabili quantitative;

  ⇒ risente molto degli outliers; se un singolo valore nel campione è per esempio molto più grande di tutti gli altri, la media non identifica un valore tipico del campione

  ⇒ non è calcolabile se alcune osservazioni sono “fuori scala”

- nel caso di distribuzioni multimodali, la media raramente identifica un valore tipico

Proprietà della mediana

- Il calcolo della mediana non implica l’elaborazione dei dati numerici osservati
  - L’informazione sul peso relativo dei singoli valori viene perduta

- E’ però spesso un buon indicatore della tendenza centrale di un set di dati
  - è calcolabile anche se la variabile è qualitativa (ma deve essere ordinabile!)
  - non risente degli outliers
  - è calcolabile anche se alcune osservazioni sono “fuori scala”

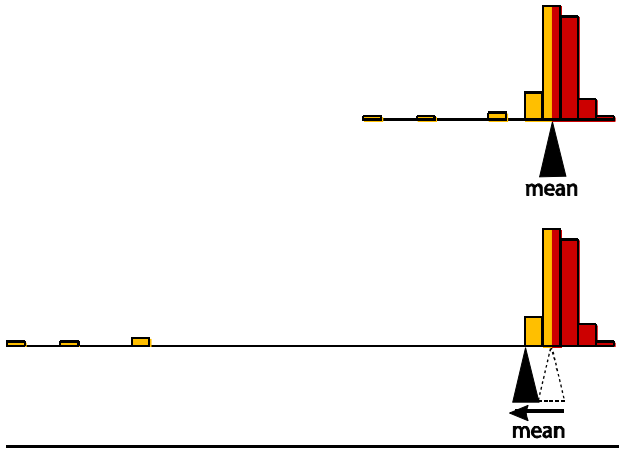
- La mediana, però, soffre dello stesso inconveniente della media, ovvero può portare ad un valore assolutamente non rappresentativo quando la distribuzione non è unimodale.

Proprietà della moda

- La moda è una statistica molto semplice e intuitiva per riassumere una distribuzione di frequenza attraverso il suo “picco” più elevato. Anche se, come la mediana, non considera il peso delle singole osservazioni, ha alcune proprietà importanti:

- è possibile identificare la moda in qualsiasi tipo di variabile, quindi anche nelle variabili qualitative non ordinabili
- indica sempre un valore realmente osservato nel campione
- non è influenzata dai valori estremi
- nel caso di distribuzioni di frequenza molto asimmetriche, la moda è forse il miglior indice per descrivere la tendenza centrale di un campione
- è collegata direttamente al concetto di probabilità (che vedremo meglio nei prossimi capitoli): la moda di una popolazione è il valore della variabile con la maggior probabilità di essere osservata

Effetto degli ouliers su media, mediana e moda



Esempio con oulier: Supponiamo di sacrificare 12 trote campionate in natura per contare in ciascuna di esse il numero di parassiti intestinali di una certa specie.

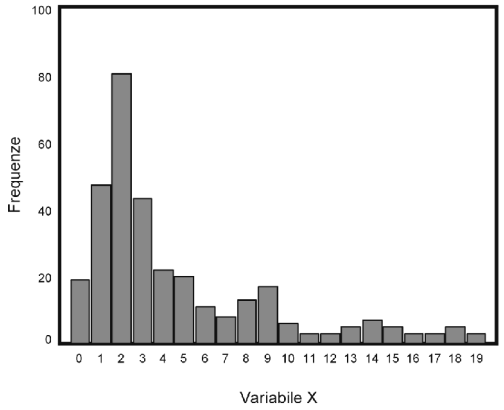
Dati: 3, 2, 3, 4, 6, 2, 44, 8, 5, 3, 4, 2.

- Media = 7,16; Mediana = 3,5; Moda = 3

Esempio con valori non misurabili: Nove cavie sono sottoposte ad un test cognitivo all'interno di un labirinto, e per ogni animale si misura il tempo impiegato a percorrere un certo tracciato. I risultati ottenuti, in minuti, sono i seguenti:

Dati: 23 ,25, 25, 22, 15, >120, 32, 20, >120

- Media = ??; Mediana = 25; Moda = 25



Questa distribuzione presenta una forte asimmetria a destra. Moda = 2 < Mediana = 3 < Media = 5,24

Misure di dispersione

- Basate sugli scarti dalla media
  - Varianza
  - Deviazione standard
  - Coefficiente di variazione
- Semplici indici basati sui dati ordinati
  - Range interquartile

La varianza

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

- E' una media delle distanze (dette *scarti*) al quadrato dalla media
- Il numeratore si chiama devianza, e si può calcolare anche con un'altra formula

$$\sum (x_i - \bar{x})^2 = Devianza = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

- Il denominatore è *n*-1 e non *n*. Serve per correggere la distorsione che avrebbe la statistica (=stima) s^2 se usassi *n*.

La deviazione standard

$$s = DS = \sqrt{s^2}$$

Molto usata per quantificare la dispersione perché ha la stessa unità di misura della variabile

[Quante cifre dopo la virgola nelle statistiche riassuntive? Attenzione, gli arrotondamenti solo alle fine!]

Il coefficiente di variazione

- E' una sorta di deviazione standard rielaborata per fare confronti tra dispersioni in gruppi con medie molto diverse o per confrontare dispersioni in diverse variabili

$$CV = \frac{s}{\bar{x}} \times 100$$

La differenza interquartile

- Cosa sono i quartili?
  - Imparentati con la mediana, solo che invece di separare l'insieme dei dati ordinati in due gruppi lo separano il quattro
  - Ogni gruppo contiene il 25% delle osservazioni: il primo quartile, Q_1, è il valore che separa il primo 25% delle osservazioni ordinate dal restante 75%, il secondo è la mediana, e il terzo quartile, Q_3, è il valore che separa il primo 75% delle osservazioni dal restante 25%.
- La differenza interquartile è data dalla differenza Q_3-Q_1, e identifica quindi l'intervallo centrale della distribuzione di frequenza all'interno del quale cade il 50% delle osservazioni.

Esempio con n = 16

Tabella 3.2-1

Velocità di corsa (cm/s) dei maschi dei ragni del genere Tidarren prima e dopo l'autoamputazione di un pedipalpo.

Ragno	Velocità prima	Velocità dopo	Ragno	Velocità prima	Velocità dopo
1	1,25	2,40	9	2,98	3,70
2	2,94	3,50	10	3,55	4,70
3	2,38	4,19	11	2,84	4,94
4	3,09	3,17	12	1,64	5,06
5	3,41	5,26	13	3,22	3,22
6	3,00	3,22	14	2,87	3,52
7	2,31	2,32	15	2,37	5,45
8	2,93	3,31	16	1,91	3,40

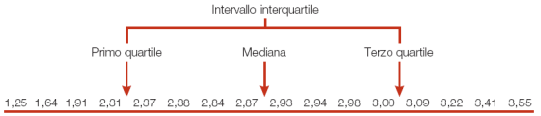
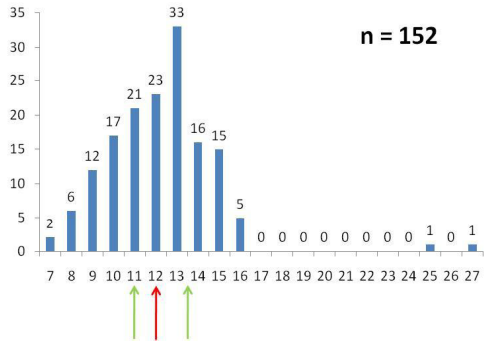


Figura 3.2-1
Il primo quartile, la mediana, e il terzo quartile suddividono l'insieme dei dati in quattro parti uguali. La mediana è il valore centrale, e il primo e il terzo quartile sono i valori centrali della prima e della seconda metà dei dati. La distanza interquartile (o range interquartile) è l'estensione della metà centrale dei dati.

Differenza interquartile = 0.705

Esempio con n = 152

7	10	11	12	12	13	14	15
7	10	11	12	13	13	14	15
8	10	11	12	13	13	14	15
8	10	11	12	13	13	14	15
8	10	11	12	13	13	14	15
8	10	11	12	13	13	14	16
8	10	11	12	13	13	14	16
8	10	11	12	13	13	14	16
9	10	11	12	13	13	14	16
9	10	11	12	13	13	14	16
9	10	11	12	13	13	15	25
9	10	11	12	13	13	15	27
9	10	11	12	13	13	15	
9	10	11	12	13	13	15	
9	10	11	12	13	14	15	
9	10	11	12	13	14	15	
9	10	11	12	13	14	15	
9	11	11	12	13	14	15	
9	11	12	12	13	14	15	
9	11	12	12	13	14	15	



Differenza interquartile = 2.5
La differenza interquartile non risente di misure estreme

- Diagramma *box plot*

  - Mediana, primo e terzo quartile, valore minimo e massimo escludendo outliers (definiti a volte come valori a una distanza dalla scatola superiore a 1.5 volte la differenza interquartile), outliers

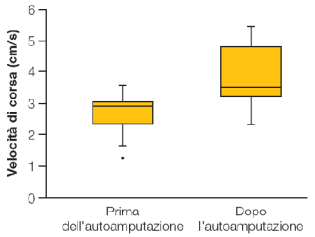


Figura 3.2-2
Box plot della velocità di corsa di 16 maschi di ragno prima e dopo l'autoamputazione di un pedipalpo.

- La proporzione (frequenza relativa)

= numero di osservazioni nella categoria di interesse diviso per il numero totale di osservazioni

$$\hat{p} = \frac{numero\ osservazioni\ nella\ categoria}{n}$$

PROBLEMI 10,11,13,16,17

Le basi della statistica inferenziale

- Il processo inferenziale consente di generalizzare, con un certo grado di sicurezza, i risultati ottenuti osservando uno o più campioni

- E' necessario però anche aggiungere con quale grado di precisione riteniamo che la nostra stima (per esempio di una media o di una proporzione) o generalizzazione (per esempio riguardo ad un'ipotesi) sia corretta

- E' molto importante che il campione sia casuale

Due ambiti principali della statistica inferenziale

Stima dei parametri

1. Si stima un parametro di una popolazione attraverso una statistica (per esempio, stima di p con $\hat{p}$)

2. Fondamentale però quantificare la precisione della stima (solitamente attraverso l' *errore standard* e l' *intervallo di confidenza*)

Test delle ipotesi

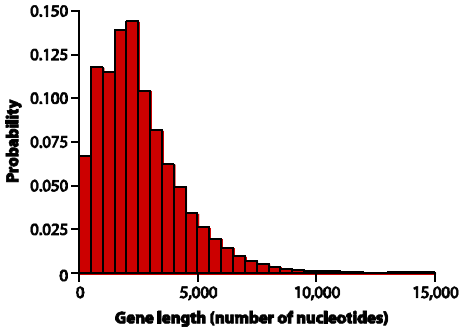
1. Si definisce un'ipotesi nulla e un'ipotesi alternativa

2. Si confrontano i dati osservati con quanto previsto dall'ipotesi nulla

3. Si calcola una misura di probabilità (basata sulla distanza tra quanto previsto dall'ipotesi nulla e quanto osservato) utile per definire quanto si può essere confidenti di una conclusione basata su un campione ma che riguarda la popolazione

La distribuzione campionaria di una stima

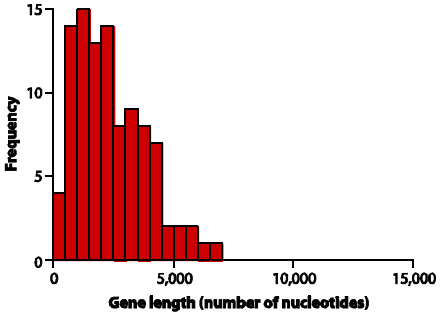
- E' necessario capire cosa potremmo ottenere campionando molte volte la popolazione (teoria del campionamento)
- Ovviamente non conosciamo praticamente mai la popolazione, ma per capire la logica della statistica inferenziale assumiamo che la popolazione sia disponibile
- La popolazione in questo ragionamento è quella dei 20290 geni noti nell'uomo; di ogni gene si conosce la lunghezza in nucleotidi (la nostra variabile numerica discreta)



Distribuzione di **probabilità** (nella popolazione)

$$\mu = 2622.0$$
$$\sigma = 2036.9$$

- Estraiamo un campione con *n* = 100



Distribuzione di **frequenza** (in **<u>un</u>** campione casuale con *n* = 100)

$$\bar{y} = 2411.8$$
$$s = 1463.5$$

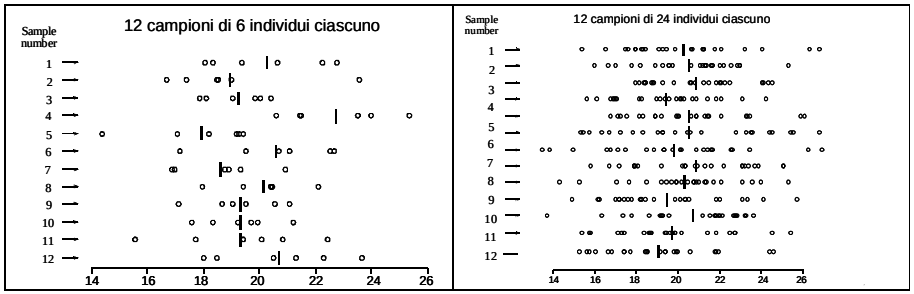
La distribuzione campionaria di $\bar{y}$

- E' la distribuzione di probabilità della stima che otterremmo se potessimo campionare infinite volte la popolazione
- Nel nostro caso, se estraessimo infiniti campioni, ciascuno di 100 geni, dalla popolazione "geni nel genoma", in ogni campione calcolassimo $\bar{y}$, e con tutti questi infiniti valori di $\bar{y}$ costruissimo una distribuzione di frequenza, quella distribuzione sarebbe la distribuzione campionaria di $\bar{y}$ per campioni con n=100 ($\bar{y}$ è anche una variabile!)
- Una singola media, ottenuta per esempio in un campionamento "ordinario", deve essere vista come un valore estratto dalla distribuzione campionaria della stima

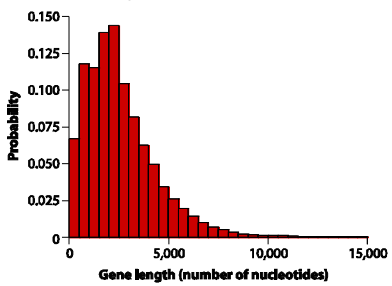
Esempio con un'altra situazione

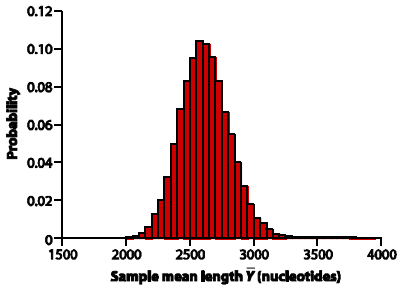
a sinistra: 12 campioni, ciascuno costituito da 6 lupi che vengono pesati
a destra: 12 campioni, ciascuno costituito da 24 lupi che vengono pesati

Per ogni campione, i cerchietti indicano i singoli valori, e la barra verticale la media campionaria. Se invece di 12 campioni immagino di averne infiniti, potrei costruire le distribuzioni della media campionaria per n=6 e per n=24



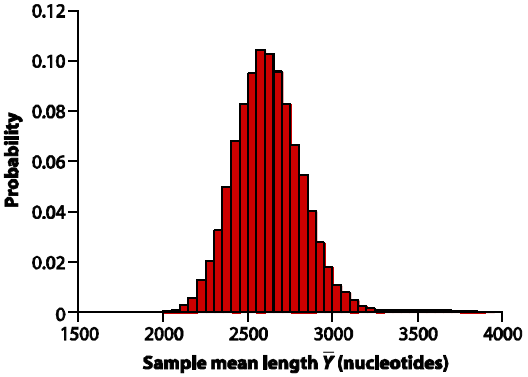
- Due cosa da notare nella distribuzione di $\bar{y}$
  1. Ha una forma a campana (diversa dalla distribuzione di y)
  2. E centrata sulla media della popolazione μ (quindi $\bar{y}$ è una stima corretta, ossia non distorta)
  3. Ha un'ampiezza inferiore all'ampiezza della distribuzione di y





- Torniamo alla variabile $\bar{y}$, ossia la variabile "media campionaria della lunghezza genica in campioni con *n* = 100"

- Usiamo un computer per capire qualcosa sulla distribuzione di $\bar{y}$



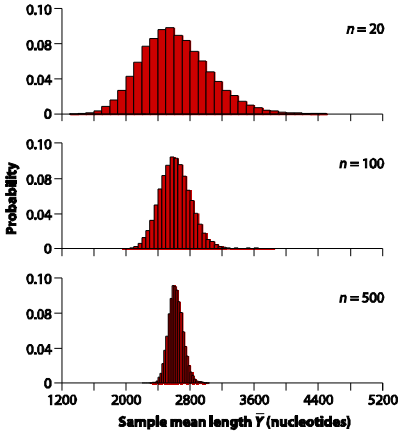
- La distribuzione campionaria di una stima è fondamentale in statistica inferenziale: ci informa sulla precisione di una singola stima

- Molta parte della teoria statistica serve per capire come sono fatte le distribuzioni campionarie delle stime e delle statistiche in generali

- Infatti, queste distribuzioni non sono praticamente mai ricavabili dai dati

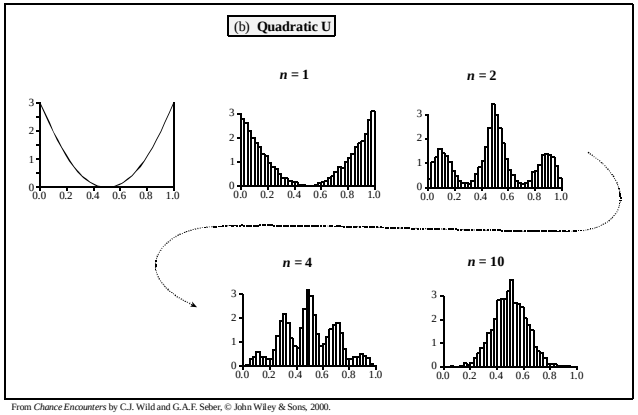
- [il caso dei geni è un caso particolare usato solo per sviluppare il ragionamento]

- Vediamo un'altra caratteristica importante della distribuzione campionaria della media



Aumentando la dimensione del campione (*n*) si riduce la dispersione della distribuzione campionaria: l'errore di campionamento si riduce e aumenta la precisione della stima

- E ancora una caratteristica della distribuzione di $\overline{y}$



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Pensiamo anche per esempio alla media dei punteggi lanciando 2, oppure 5, oppure 10 dadi

http://actuary.my/blog/applet/the-dice-experiment/#postTabs_ul_762

Misurare l'incertezza di una stima: l'errore standard

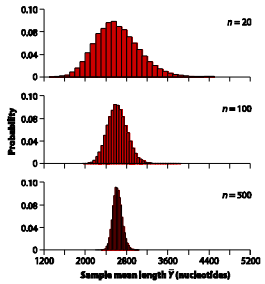
- Errore standard di una stima: deviazione standard della distribuzione campionaria della stima

- E' un indice della dispersione della stima rispetto al valore da stimare (il parametro): misura quindi la precisione della stima

- Se estraessimo moltissimi campioni, ciascuno di 100 geni, dalla popolazione "geni nel genoma", in ogni campione calcolassimo $\overline{y}$, e con tutti questi valori di $\overline{y}$ calcolassimo la deviazione standard, avremmo ottenuto l'errore standard

- Teoricamente si può dimostrare che questo valore, anche senza campionare molte volte la popolazione, ma conoscendo però la deviazione standard della variabile Y *nella popolazione* (σ), posso (*potrei*...) calcolarlo con

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$



- Per esempio, visto che la deviazione standard nella popolazione di geni è pari a σ = 2036.9, l'errore standard nelle tre distribuzioni qui sopra è pari a 455.5, 203.7, e 91.1

- A questi stessi valori sarei giunto se avessi estratto moltissimi campioni 20 geni e avessi calcolato la deviazione standard delle moltissime $\bar{y}$, e poi avessi fatto lo stesso per *n* = 100 e *n* = 500

- Ma, nella pratica usuale, non posso estrarre tanti campioni ma ne posseggo uno solo, e non conosco σ ma solo *s*

- Se il campione è casuale, una approssimazione dell'errore standard $\sigma_{\bar{y}}$ è data da

$$ES_{\bar{y}} = s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

- Da ogni set di dati, frutto di un'osservazione o di un esperimento, possiamo quindi calcolare una approssimazione dell'errore standard della media, ossia di un indice della distanza tra media nel campione e media nella popolazione

- Ogni media calcolata su un campione dovrebbe essere accompagnata dal suo errore standard

- Esempio: abbiamo calcolato la media delle altezze in un campione di 10 individui, la media $\bar{y}$ è risultata pari a 168,2 centimetri e la varianza s^2 = 92,3 cm^2

$$ES_{\bar{y}} = \frac{s}{\sqrt{n}} = 3.0$$

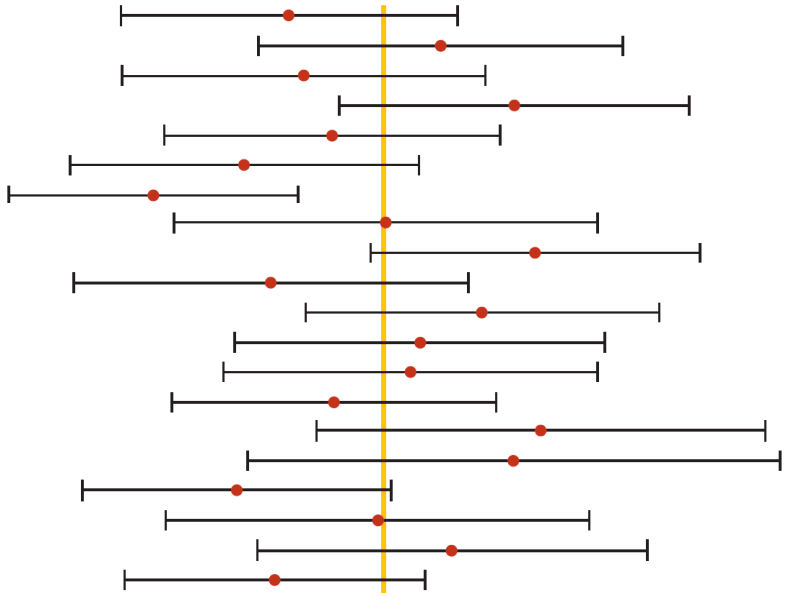
$$\bar{y} = 168.2 \pm 3.0\ (ES)$$

[è un modo usuale di riportare una media]

- Ogni stima (non solo la media) ha una sua distribuzione campionaria e un suo errore standard

Misurare l'incertezza di una stima: l'intervallo di confidenza

- E' un intervallo, definito intorno alla stima di un parametro, che ha alta probabilità di contenere il parametro (ossia, il valore vero)
- Si può definire per molti parametri, come medie, proporzioni, differenze tra medie, ecc.
- Esempio: intervallo di confidenza della media al 95%: siamo fiduciosi al 95% che l'intervallo calcolato conterrà la media della popolazione
- La probabilità si applica all'intervallo, non al parametro (che è fisso)



<u>Regoletta pratica (ma approssimata) per il calcolo dell'IC di una media</u>

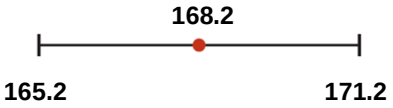
$$IC_{95\%} = \bar{y} \pm 2ES_{\bar{y}}$$

- Nell'esempio di un campione di 100 geni estratti dal genoma umano (quello con $\bar{y}$ della lunghezza pari a 2411.8 e s = 1463.5), l'IC al 95% diventa

$$IC_{95\%} = 2411.8 \pm 2 \times \frac{1463.5}{\sqrt{100}}, \text{ ovvero}$$

- Nell'esempio del campione di 10 individui ai quali era stata misurata l'altezza estratti dal genoma umano (quello con $\bar{y}$ pari a 168.2 e s^2 = 92.3), l'IC al 95% diventa

$$IC_{95\%} = 168.2 \pm 2 \times 3.0, \text{ ovvero}$$

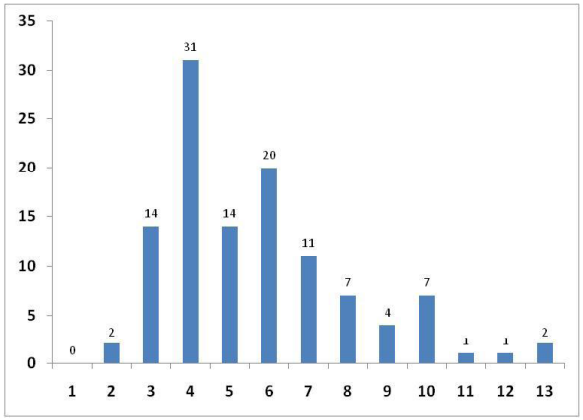


Esercizio 9 pg.47, continua in esercizio 1 pg. 57

I dati

n = 114

x_i	n_i
1	0
2	2
3	14
4	31
5	14
6	20
7	11
8	7
9	4
10	7
11	1
12	1
13	2



a) Calcolo la media

	x_i	n_i	$x_i n_i$
	1	0	0
	2	2	4
	3	14	42
	4	31	124
	5	14	70
	6	20	120
	7	11	77
	8	7	56
	9	4	36
	10	7	70
	11	1	11
	12	1	12
	13	2	26

Somme 114 648

Media 5,68 ore

b) Calcolo varianza e deviazione standard

x_i	n_i	$x_i n_i$	$(x_i - x_{medio})^2$	$n_i(x_i - x_{medio})^2$
1	0	0	21,942	0,000
2	2	4	13,573	27,147
3	14	42	7,205	100,870
4	31	124	2,837	87,934
5	14	70	0,468	6,554
6	20	120	0,100	1,994
7	11	77	1,731	19,044
8	7	56	5,363	37,540
9	4	36	10,994	43,978
10	7	70	18,626	130,382
11	1	11	28,258	28,258
12	1	12	39,889	39,889
13	2	26	53,521	107,042
Somme	114	648		630,6315789

Media 5,68 ore

s^2 5,58 ore^2

s 2,36 ore

c) Calcolo frazione di **osservazioni** in intervallo richiesto

Numero di **osservazioni** entro una DS dalla media =	83	[numero **osservazioni** tra (5,7-2,4) e (5,7+2,4)]
Frazione di **osservazioni** entro una DS dalla media =	83/114=0.73	[frazione **osservazioni** tra (5,7-2,4) e (5,7+2,4)]

d) Determino la mediana e commento

[media tra i valori in posizione 57 e 58 nella lista ordinata]

x_i	n_i
1	0
2	2
3	14
4	31
5	14
6	20
7	11
8	7
9	4
10	7
11	1
12	1
13	2

a) Determino ES

$$ES_{\bar{y}} = \frac{s}{\sqrt{n}} = \frac{2.36}{\sqrt{114}} = 0.22\ ore$$

b) e c)

Problemi Cap 3. (10-11-13-16-17)

10.
a) Media: 5.5 (in unità logaritmiche)
b) Deviazione standard: 0.26 (in unità logaritmiche)
c) 39/39 = 1.0 (100%)

13.
a) Mediana: 8 (valore in 64esima posizione nella lista ordinata)
b) Primo quartile: 3; Secondo quartile 17; Range interquartile: 14
c) Non si può, non abbiamo i valori nell'ultima classe della tabella di frequenza

16.
a) Mediana (ma ci sono comunque pochissime osservazioni)
b) Distanza iterquartile

Cap. 4 Ulteriori problemi 8, 11, 13

La pseudoreplicazione

- Ogni individuo deve avere una probabilità uguale e indipendente di essere campionato (campione casuale)
- Se le singole misure non sono indipendenti si compie l'errore di pseudoreplicazione
- Tre esempi: Esercizio 17, pg.13. Studio delle preferenze del canto a pg. 59. Studio sulla glicemia a pg. 59
- E' un problema perché si crede di avere molta informazione sulla popolazione ma se ne ha poca; la precisione viene sovra-stimata (per esempio, si ottengono CI troppo stretti)
- A volte basta fare medie di valori non indipendenti prima di usare i dati per fare statistica inferenziale

La probabilità: concetti di base

- La teoria della probabilità è molto complessa, ma il concetto di probabilità è molto intuitivo
- Abbiamo una scatola (urna) con 3 palline rosse e 7 palline nere
- Estraggo una pallina e ne osservo il colore; ho svolto una **prova casuale**
- La prova casuale è un processo o un esperimento che ha due o più risultati possibili che non possono essere predetti
- Altri esempi di prova casuale: lancio moneta e verifico se viene testa (o croce); campiono un individuo e misuro il peso; campiono 10 individui e determino la proporzione di femmine
- Torniamo all'urna e alle palline rosse e nere
- Qual è la probabilità P di estrarre una pallina rossa?

  P = 0.3 (30%)
- Ma cosa significa esattamente che la probabilità è uguale a 0.3?

  Se ripetessi questa estrazione (prova casuale) un numero elevatissimo di volte, ogni volta reinserendo la pallina nell'urna dopo l'estrazione (oppure immaginando che l'urna contenga infinite palline nella proporzione 30:70), ….
- La probabilità di un evento (pallina rossa nel nostro caso) è la frazione o proporzione di tutte le prove casuali con cui si verificherebbe l'evento specificato se si ripetesse la prova casuale moltissime volte

- E quindi la rappresentazione teorica della frequenza, ovvero il valore a cui tende la frequenza quando il numero di ripetizioni dell'evento è molto grande

- Questa definizione implica anche che una tabella di frequenza tende ad una tabella di probabilità se il campione è molto grande (le due cose coincidono se ho campionato tutta la popolazione). Se per esempio analizzo un campione molto grande di donne e trovo che il 41.3 % di loro ha avuto un solo figlio, posso dire che se chiedo ad una donna scelta a caso quanti figli ha, la probabilità di avere come risposta “1” è pari a 0.413

- Stessa cosa vale per le distribuzioni di frequenza e le distribuzioni di probabilità a cui tendono quando il campione diventa così grande da essere la popolazione

- Esempi eventi compatibili

  - Lancio due dadi: gli eventi "1 nel primo lancio” e "1 nel secondo lancio" non sono incompatibili

  - Estraggo una carta da un mazzo: l’evento “è rossa” e “è diversa da un asso” non sono incompatibili

  - Osservo un animale: l'evento "vola" e "non ha le penne" non sono incompatibili

- Come le frequenze relative, le probabilità non possono mai essere inferiori a 0 o superiore a 1, e la somma delle probabilità associate a tutti i possibili risultati (eventi) diversi incompatibili (ovvero che non si possono verificare insieme) è per forza di cose pari a 1

- Esempi eventi incompatibili

  - Lancio un dado: l'evento "1" e l'evento "4" sono incompatibili, cioè non si possono verificare simultaneamente (ci sono eventi compatibili nel lancio di un dado?)

  - Campiono un individuo e considero come "evento" la sua altezza (nello spazio di tutti gli eventi possibili che sono tutte le altezze)

DISTRIBUZIONI DI FREQUENZA E DISTRIBUZIONI DI PROBABILITA’ (già visto in parte)

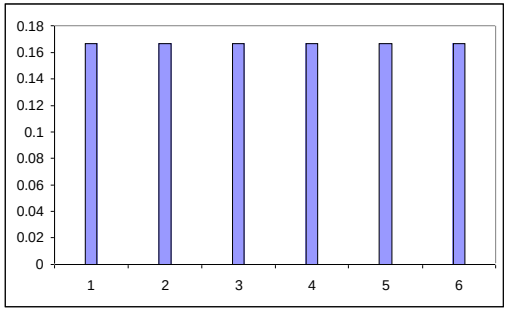
- **Distribuzione di frequenza**: ricostruita a partire dai dati campionati

- **Distribuzione di probabilità**: ricostruita a partire dai dati di tutta la popolazione

- **Distribuzione *teorica* di probabilità**: è definita da una funzione matematica di cui conosco le caratteristiche e che mi permette di calcolare una probabilità associata a ciascun valore o intervallo di valori

DISTRIBUZIONI DI PROBABILITA' DISCRETE

- Per variabili di tipo discreto
- La distribuzione fornisce la probabilità che il valore assuma uno specifico valore (ogni possibile valore è uno dei risultati incompatibili della prova casuale)



Esempio: distribuzione uniforme discreta

Conosco in questo caso la funzione matematica:

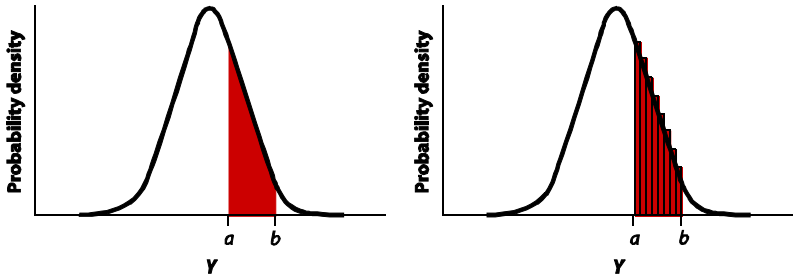
$$f(x)=\frac{1}{k}$$

- Lancio di una moneta equilibrata (k=2)
- Lancio di un dado equilibrato (k=6)
- Frequenza attesa di cattura in *4* tipi trappole ugualmente efficienti (k=4)
- La distribuzione di probabilità discreta è comunque (con o senza una funzione matematica) una lista di probabilità di ciascun evento: la somma di questa lista di probabilità deve fare 1

DISTRIBUZIONI DI PROBABILITA' CONTINUE

- Per variabili di tipo continuo, che possono quindi assumere infiniti possibili valori
- L'altezza (asse delle y) di queste distribuzioni non fornisce la probabilità di osservare un valore (che è, per definizione, pari a 0)
- L'altezza della curva è invece una densità di probabilità (una probabilità divisa per un intervallo), e infatti si dovrebbero chiamare più precisamente *distribuzioni di densità*

- Ciò che conta è che possiamo usarle per ottenere qualcosa di più interessante, ossia la probabilità di avere un'osservazione entro un certo intervallo. Di queste curve ci interessa quindi l'integrale (...pensate all'area delle barre negli istogrammi)

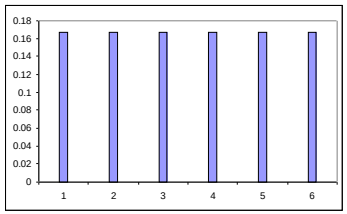


- In altre parole, quando ci serviremo di distribuzioni teoriche di probabilità per variabili continue, sarà l'area sottesa dalla curva, e non il valore di Y, a corrispondere alla probabilità.

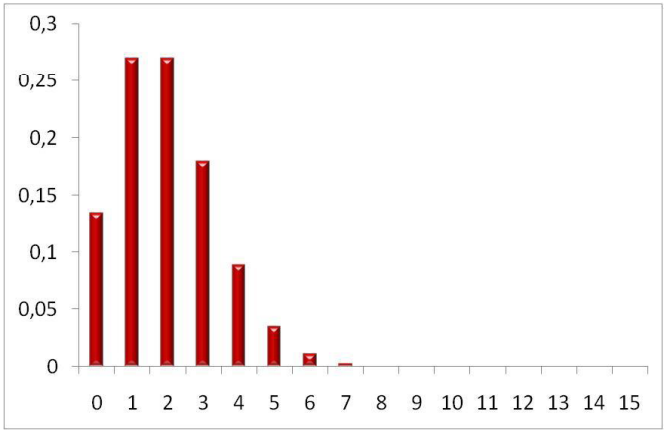
- La normale, o gaussiana, è una delle distribuzioni teoriche continue più usate in biostatistica

- **La regola della somma**: se due eventi sono incompatibili, la probabilità che se ne verifichi uno <u>oppure</u> l'altro è dato dalla somma delle loro rispettive probabilità

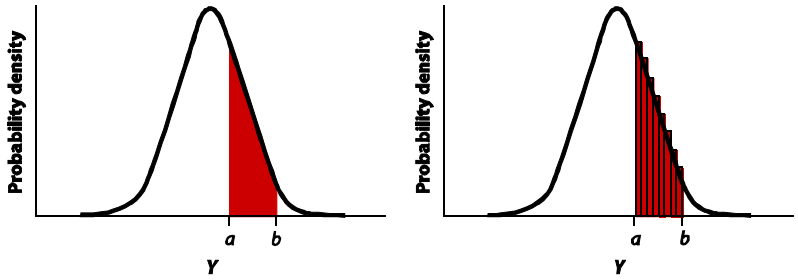
[se non sono incompatibili, la regola non vale]



Probabilità di ottenere 1 o 4 lanciando un dado?



Probabilità di ottenere avere meno di 2 figli (0 o 1) se questa è la distribuzione di probabilità del numero di figli nella popolazione?



- Probabilità di avere un peso tra ***a*** e ***b***?
- Probabilità di avere un peso inferiore ad ***a*** o superiore a ***b***?

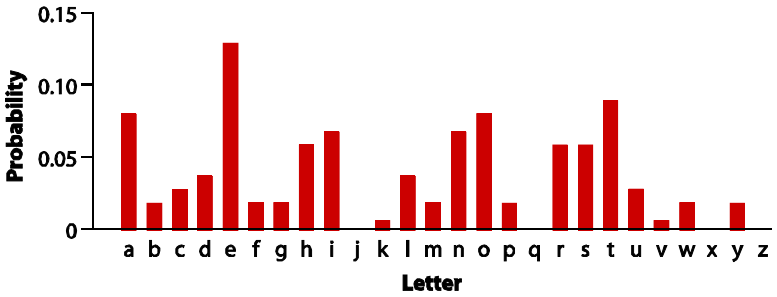
- **La regola del prodotto:** se due eventi sono indipendenti, la probabilità che si verifichino entrambi (uno e l'altro) è data dal prodotto delle due probabilità

[due eventi sono indipendenti se il verificarsi di uno non influenza il verificarsi dell'altro]
[se non sono indipendenti, la regola non vale]

- Esempio: Probabilità di ottenere 1 e 1 nel lancio di due dadi?
- Esempio: 5.6A (pg 66)

Probabilità di numeri doppi nel lancio di due dadi?

Problema 4, pg. 73



LA VERIFICA (TEST) DELLE IPOTESI

- Nella stima ci chiediamo quale sia un intervallo di valori verosimili per il parametro o per un certo effetto

- Nella verifica delle ipotesi ci chiediamo se il parametro, o l'effetto, differiscano *significativamente* da un valore previsto dall'ipotesi nulla

- Come già detto, nella statistica inferenziale classica vengono sempre confrontate due ipotesi: l'ipotesi nulla e l'ipotesi alternativa

- Come già detto, in realtà questo confronto non è diretto: quello che si confronta realmente sono i dati con l'ipotesi nulla

- Se i dati sono "troppo insoliti" per quanto ci aspetteremmo se fosse vera l'ipotesi nulla, scartiamo questa ipotesi a favore dell'ipotesi alternativa

- Fondamentale in questa operazione è il concetto di probabilità

- Esempio con studio sull'efficacia del vaccino antipolio

In altre parole:

1. Si cerca di prevedere come potrebbero essere i dati se fosse vera l'ipotesi nulla

2. Se i dati osservati sono molto diversi da quelli si potrebbero ottenere se fosse vera l'ipotesi nulla, allora l'ipotesi nulla VIENE RIFIUTATA (e di conseguenza, si accetta l'ipotesi alternativa)

3. Se invece i dati osservati non sono troppo diversi da quelli che si potrebbero ottenere se fosse vera l'ipotesi nulla, allora l'ipotesi nulla NON VIENE RIFIUTATA (ovvero, si dice che i dati osservati sono compatibili con l'ipotesi nulla)

- La probabilità ci servirà per quantificare il concetto di "molto diversi"
- L'ipotesi nulla non viene mai accettata!

- **Ipotesi nulla, o H_0**
  - E' un enunciato specifico che riguarda un parametro nella popolazione (o nelle popolazioni)
  - E' l'ipotesi che, tutto sommato, sarebbe interessante rifiutare
  - E' generalmente il punto di vista scettico

- **Ipotesi alternativa, o H_a o H_1**
  - Rappresenta tutte le altre ipotesi riguardo al parametro non specificate dall'ipotesi nulla; non è quindi specifica
  - E' l'ipotesi che generalmente viene formulata prima di fare un test, l'idea cioè che ha avuto il ricercatore e che lo ha indotto a fare un esperimento o a raccogliere dei dati sul campo (e che quindi sarebbe interessante, soprattutto per il ricercatore che l'ha avuta, poter dimostrare alla fine dello studio)

Esempi di H_0: qual è la corrispondente H_1 ?

- La densità di delfini nelle zone in cui la pesca viene effettuata con le reti a deriva è uguale alla densità di delfini nelle aree in cui la pesca viene effettuata senza queste reti

- Gli effetti antidepressivi della sertralina non differiscono da quelli dell'amitriptilina

- Genitori con occhi marroni, ciascuno dei quali ha avuto un genitore con occhi azzurri, hanno figli con occhi marroni e figli con occhi azzurri in un rapporto 3:1 (rapporto predetto da una teoria)

- La crescita media tra il terzo e quarto mese di un bambino allattato con latte artificiale è pari a 2.2 cm

Esempio di verifica delle ipotesi: i rospi destrimani e mancini e l'analisi di una proporzione

Di cosa abbiamo bisogno?

1. Formulare correttamente le ipotesi H_0 e H_A
   [ipotesi alternative bilaterali e unilaterali]

2. Una statistica test
   [numero osservato di rospi destrimani nell'esempio]

3. La distribuzione nulla
   [di solito la parte difficile...]

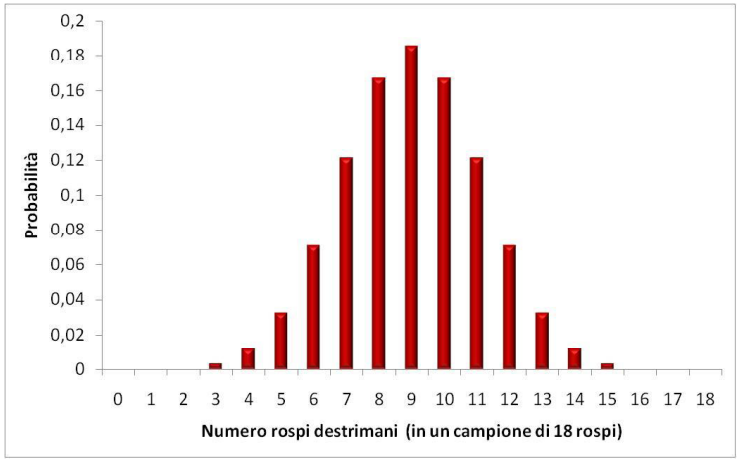
Esercizio 3 (pg 87)

La distribuzione nulla è la distribuzione campionaria della statistica test sotto H_0 (ovvero, assumendo che sia vera l'ipotesi nulla)

[cos'è una distribuzione campionaria?]

Nell'esempio dei rospi dobbiamo quindi trovare la distribuzione campionaria del numero di rospi destrimani (la statistica test) in campioni di 18 rospi, assumendo vera l'ipotesi che nella popolazione ci siano metà rospi mancini e metà rospi destrimani

0	0,000004
1	0,000069
2	0,000584
3	0,003113
4	0,011673
5	0,032684
6	0,070816
7	0,121399
8	0,166924
9	0,185471
10	0,166924
11	0,121399
12	0,070816
13	0,032684
14	0,011673
15	0,003113
16	0,000584
17	0,000069
18	0,000004

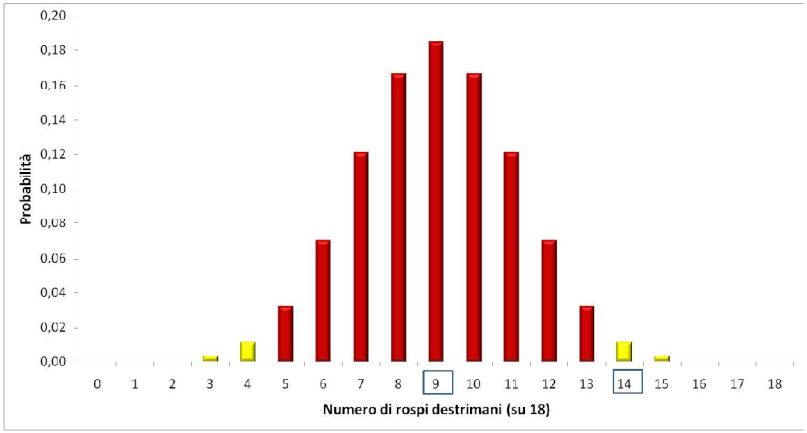


Cosa ci dice questa distribuzione nulla?

Quantificare l'incertezza nel test delle ipotesi: il P-value

- Conoscendo la distribuzione nulla, possiamo ora quantificare probabilisticamente se i dati osservati sono "troppo insoliti" oppure no

- Ricorriamo al P-value: è la probabilità di ottenere la statistica test calcolata o un valore più estremo se fosse vera l'ipotesi nulla

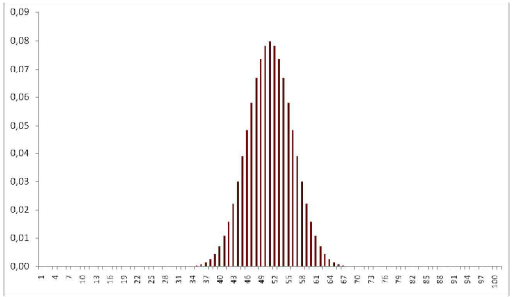


In giallo, le probabilità che contribuiscono al P-value

Nell'esempio, P-value = 2x(Pr[14]+ Pr[15]+ Pr[16]+ Pr[17]+ Pr[18]) = 0.031

Attenzione al significato del P-value

- **Il P-value non è la probabilità di osservare la statistica test assumendo vera l'ipotesi nulla**
  - la probabilità di un singolo valore della statistica è la probabilità di osservare uno specifico data set con quella statistica test, ma non la misura di compatibilità (o incompatibilità) dei dati con l'ipotesi nulla usata nella statistica inferenziale classica



- **Il P-value non è la probabilità dell'ipotesi nulla di essere vera (e 1 - P-value non è la probabilità dell'ipotesi alternativa di essere vera)**

  - Supponiamo di osservare 12 rospi destrimani su 18 (o 12 volte prevediamo correttamente il risultato del lancio di una moneta su 18 lanci)

  - Il P-value è in questo caso pari a 0.24. Potremmo concludere che l'ipotesi alternativa è tre volte più probabile della nulla? NO! I dati sono infatti ancora ampiamente compatibili con l'ipotesi nulla (0.24 è una probabilità piuttosto alta)

  - Le ipotesi nulla e alternativa non si confrontano direttamente e non sono eventi casuali (approccio frequentista). Sono i dati che vengono confrontati con quanto previsto dall'ipotesi nulla.

La significatività statistica

- Ora si tratta di decidere quando il P-value è piccolo abbastanza per rifiutare l'ipotesi nulla

- Questa soglia è il livello di significatività α, la possiamo scegliere noi, ma in genere viene fissata a 0.05 o 0.01

- E' soggettiva e in alcuni casi non deve essere considerata come un limite definitivo

L'approccio delle regioni di accettazione e rifiuto di H_0
Esempio 6.4, pg.83

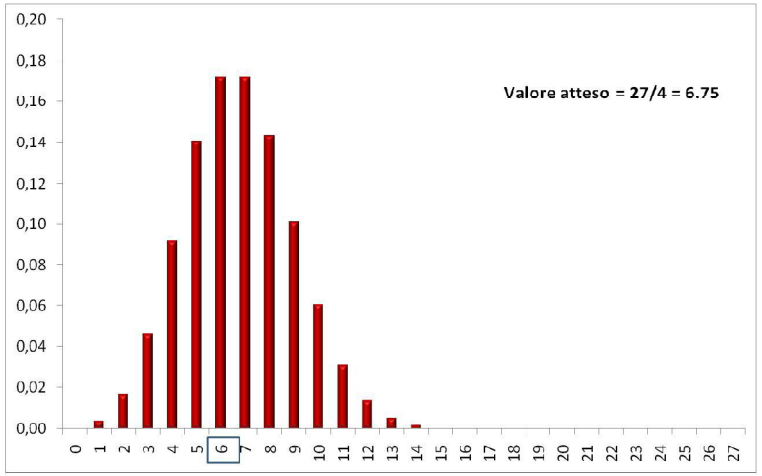
<u>Vediamo prima l'approccio del P-value appena visto</u>

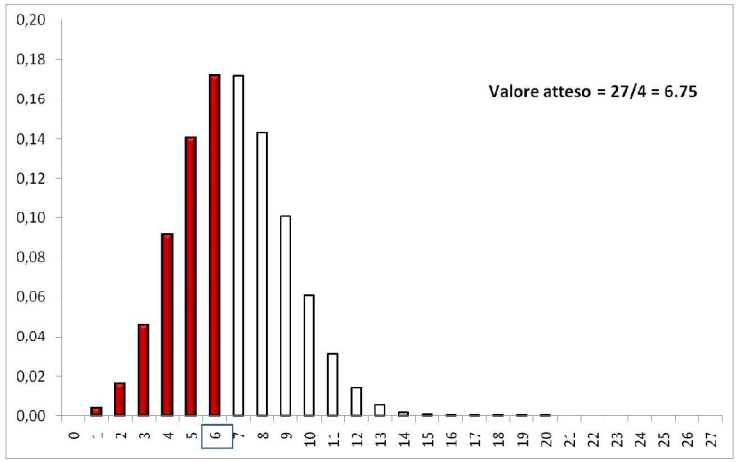
- I dati: dall'incrocio di F1 tra loro si ottengono 6 individui con fiori sinistrorsi e 21 con fiori destrorsi

- H_0: p=1/4
  H_A: p≠1/4

- Statistica test: numero discendenti sinistrorsi

- Distribuzione nulla

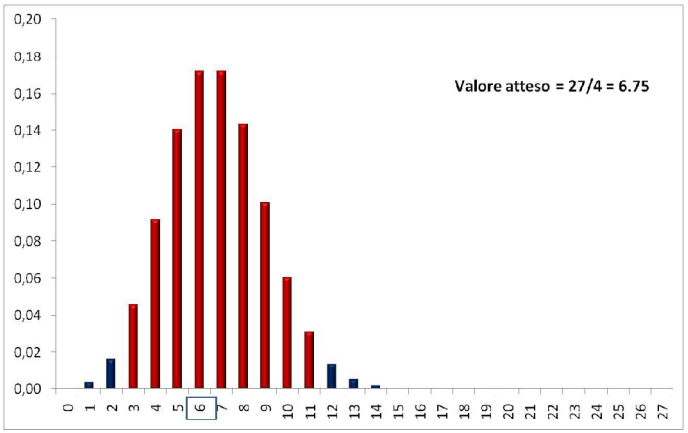




In rosso, le probabilità che contribuiscono al P-value per valori uguali all'osservato o più distanti da H_0 in una direzione
(deve poi essere moltiplicato per 2)

L'approccio delle regioni di accettazione e rifiuto di H_0

0	0,0004
1	0,0038
2	0,0165
3	0,0459
4	0,0917
5	0,1406
6	0,1719
7	0,1719
8	0,1432
9	0,1008
10	0,0605
11	0,0312
12	0,0138
13	0,0053
14	0,0018
15	0,0005
16	0,0001
17	0,0000
18	0,0000
19	0,0000
20	0,0000
21	0,0000
22	0,0000
23	0,0000
24	0,0000
25	0,0000
26	0,0000
27	0,0000



In blu, le probabilità che contribuiscono alla regione di rifiuto, il rosso quelle che definiscono la regione di accettazione

Gli errori nella verifica delle ipotesi

- Nella statistica inferenziale si cerca di dire qualcosa di valido in generale, per la popolazione o le popolazioni, attraverso l'analisi di uno o più campioni
- E' chiaro però che esiste comunque la possibilità di giungere a conclusioni errate, appunto perché i dati raccolti rappresentano solo una parte dell'evento che sto analizzando
- Nella verifica delle ipotesi si possono commettere due tipi di errori, quelli di primo tipo (tipo I) e quelli di secondo tipo (tipo II)

	H_0 è vera	**H_1 è vera**
Non escludo H_0	Decisione corretta (P = 1-α = livello di protezione)	Decisione incorretta, compio un errore di II tipo (P = β)
Rifiuto H_0	Decisione incorretta, compio un errore di I tipo (P = α = livello di significatività)	Decisione corretta (P = 1- β = potenza)

Errore di primo tipo

- La probabilità di compiere un errore di primo tipo è data dal livello di significatività α prescelto
- E' la frazione di volte che viene rifiutata un ipotesi nulla vera se ripetessi tante volte il test su campioni diversi (presi dalla stessa, o dalle stesse, popolazione/i)
- Scegliendo in anticipo α, definiamo il rischio che siamo disposti ad accettare di compiere un errore di primo tipo
- Alla fine del test, se le evidenze saranno a favore dell'ipotesi alternativa, non sapremo ovviamente se avremo commesso un errore di primo tipo oppure no. Potremo solo dire che la probabilità di averlo commesso, se fosse vera l'ipotesi nulla, è molto bassa (e pari ad α)

- La probabilità complementare (1- α) viene chiamata livello di protezione di un test, ed è appunto la probabilità di non rifiutare l'ipotesi nulla quando l'ipotesi nulla è vera. Un test con un altro livello di protezione è detto conservativo
  - Un test molto conservativo può essere visto come un test che vuole rischiare molto poco di fare un errore di primo tipo, che sappiamo essere un errore molto grave perché rifiutare l'ipotesi nulla è una decisione forte (come condannare un imputato) mentre non rifiutarla non significa in realtà accettarla (ma solo dire che i dati sono compatibili con essa)
- Da notare che nel calcolo degli intervalli di confidenza (utilizzati nella stima di un parametro, **non** nella verifica di ipotesi), il termine 1- α prende il nome di grado di confidenza

Errore di secondo tipo

- La probabilità di commettere un errore di secondo tipo viene generalmente indicato con il simbolo β
- La probabilità complementare, (1- β), ossia la probabilità di rifiutare correttamente un ipotesi nulla falsa, si chiama potenza del test
  - Maggiore è la potenza di un test, maggiore sarà la possibilità del test di identificare come corretta l'ipotesi alternativa quando questa è effettivamente vera

- La probabilità di fare un errore di secondo tipo, ovvero il rischio di non rifiutare un'ipotesi nulla falsa, e di conseguenza la potenza di un test, non si può stabilire a priori
  - Dipende infatti dalla vera distanza tra ipotesi nulla e alternativa, distanza che è ignota
  - Dipende dalla varianza delle variabili in gioco, che non può essere modificata
- La probabilità di fare un errore di secondo tipo, però, dipende anche dal numero di osservazioni e dal livello di significatività α prescelto. Quindi:
  - è possibile ridurre l'errore di II tipo (e quindi aumentare la potenza) aumentando la dimensione campionaria
  - è possibile ridurre l'errore di II tipo (e quindi aumentare la potenza) aumentando il livello di significatività α (ma questa scelta ci espone a maggiori rischi di errore di tipo I)

Test a due code e a una coda

- L'ipotesi alternativa è specificata diversamente

- I test a una coda sono da utilizzare in casi molto particolari e con prudenza (e la scelta se fare un test a una o due code deve sempre precedere l'analisi dei dati): aumentano la potenza di un test ma se usati quando non dovrebbero aumentano l'errore di tipo I

- Esempio pg. 84
  - Test sulla somiglianza padri-figlie. A 18 individui vengono presentati 18 set diversi di tre fotografie. Ogni set è costituito dalla foto di una ragazza, di suo padre, e di un altro uomo. Ad ogni individuo viene richiesto di identificare il padre. 13 individuano correttamente il padre, 5 indicano l'altro uomo. L'ipotesi nulla è $p = 0.5$, ovvero non esiste somiglianza e l'indicazione di un uomo rispetto ad un altro è casuale. L'ipotesi alternativa è che ci sia somiglianza, e quindi è che $p > 0.5$ (ovvero che l'identificazione sia corretta in più del 50% dei casi). L'ipotesi alternativa $p < 0.5$ non ha senso.

Confronto tra verifica delle ipotesi e IC

- In alcuni casi, il calcolo di IC permette anche un test di ipotesi

- In generale, è più diretta per giungere a conclusioni riguardo le ipotesi scientifiche (quale favorire?)

- La stima dei parametri e il calcolo dell'IC è comunque fondamentale per capire la "dimensione dell'effetto"

Alcuni punti molto importanti nella verifica delle ipotesi

1. Inferenza statistica e cautela verso le "novità"

- La verifica di ipotesi è forse lo strumento statistico più importante per il processo conoscitivo scientifico

- Considerando che H_0 tendenzialmente definisce la situazione sperimentale "conservatrice" e H_1 quella che porta ad una scoperta nella ricerca, si capisce come la logica dell'inferenza statistica abbia un carattere di cautela verso l'innovazione: consente di rifiutare l'ipotesi nulla solo se i dati sono veramente incompatibili con essa (α è in genere fissato al 5%)

- Possiamo pensare alla verifica di ipotesi come ad un processo
  - L'imputato è il parametro sotto test, H_0 corrisponde all'innocenza
  - L'assoluzione corrisponde a non rifiutare H_0
  - La sentenza di colpevolezza corrisponde a rifiutare H_0 e favorire H_A

- Il sistema legislativo consente di condannare solo nel caso di forti evidenze di colpevolezza, nel caso cioè in cui la probabilità che l'imputato (il parametro) sia innocente (assumo H_0), sia molto bassa (minore di α). In questo caso ci garantiamo di non condannare quasi mai un innocente (*errore di primo tipo*), errore ben più grave di assolvere un colpevole (*errore di secondo tipo*).

2. L’ipotesi nulla non viene mai accettata

- Un risultato non significativo indica solo che non si è in grado di rifiutare l’ipotesi nulla

- Potrei per esempio avere una proporzione di rospi destrimani nella popolazione leggermente diversa da 0.5 (per esempio pari a 0.53) ma i dati se il campione è piccolo e il test poco potente risultano ancora compatibili con l’ipotesi nulla

- L’evidenza in favore dell’ipotesi alternativa non è sufficientemente forte per escludere l’ipotesi nulla di partenza (che rappresenta in genere un modello o un idea più semplice)

- Niente esclude che in un successivo esperimento questa differenza diventi evidente

- Volendo continuare con l'analogia del processo, questo corrisponde al fatto che l'imputato non viene mai assolto in modo definitivo, ma all'eventuale presenza di nuove prove di colpevolezza, il processo verrebbe riaperto (si eseguirebbe di nuovo il test con i nuovi dati raccolti)

- E' importante calcolare gli IC per capire la precisione della stima e se l'ipotesi nulla non viene rifiutata perché l'informazione raccolta è forse insufficiente

3. Il livello di significatività non corrisponde alla dimensione dell’effetto (Scheda 3)

- Lo stesso effetto diventa più o meno significativo semplicemente in funzione del numero di dati disponibili: avere più dati, significa avere maggiori informazioni, per cui anche l'effetto più piccolo diventa significativo con un adeguato numero di osservazioni.

- Un risultato significativo non significa un risultato importante ci indica solo quanto poco probabile è che un certo effetto sia dovuto al caso

- Interpretare la "dimensione", e quindi l'importanza del risultato, è compito dello studioso.

- Per esempio, potrebbe risultare, sulla base di un campione di 10000 persone che fanno jogging regolarmente, che il loro rischio di infarto è statisticamente maggiore rispetto a chi non lo pratica (favorendo cioè l’ipotesi alternativa). Se però questo rischio aumenta, pur se in maniera statisticamente significativa, solo dello 0.01% , questo risultato potrebbe non avere una grande importanza sociale o comunque biologica.

PROBLEMI: 11, 12, 13, 14, 15, 16, 17, 18

L'analisi di proporzioni e la distribuzione binomiale

- Abbiamo già visto un esempio di verifica dell'ipotesi su una proporzione, ma la distribuzione nulla era ottenuta in un modo piuttosto complicato
- Vediamo la distribuzione teorica discreta di probabilità che ci serve: la distribuzione binomiale
- Vedremo anche che la distribuzione binomiale è anche la distribuzione campionaria di una proporzione

La distribuzione binomiale

- Supponiamo di compiere un esperimento con due soli risultati possibili
  - Lancio una moneta: ottengo testa o croce?
  - Faccio un figlio: sarà maschio o femmina?
  - Provo un esame: viene superato oppure no?
  - Misuro la temperatura: e' < 36.5 oppure ≥36.5 ?
  - Estraggo a caso un individuo dalla popolazione: fuma oppure no?
  - Estraggo a caso un individuo dalla popolazione: ha una mutazione A→C in posizione 56 nel gene per l'emoglobina?

- Un esperimento di questo tipo è detto **esperimento bernoulliano**

- Chiamiamo uno dei due eventi *successo* (**S**) e l'altro (l'evento complementare) *insuccesso* (**I**)
  - Non importa quale dei due viene chiamato successo e quale insucceso, è una scelta arbitraria; per esempio
    - testa = successo; croce = insuccesso
    - fumatore = successo; non fumatore = insuccesso
    - la mutazione A→C in posizione 56 nel gene per l'emoglobina è presente = successo; la mutazione A→C in posizione 56 nel gene per l'emoglobina è assente = insuccesso

- Chiamiamo ora
  - p = probabilità dell'evento **S** (successo) in una singola prova
  - $(1-p)$ = probabilità dell'evento **I** (insuccesso) in una singola prova

- Supponiamo ora invece di ripetere l'esperimento bernoulliano **2 volte** e di essere interessati al numero di successi
  - Il numero di ripetizioni, o prove, detto anche *numero di prove*, si indica con n
  - In questo caso $n = 2$

- Esempi
  - Lancio due monete (o due volte la stessa moneta) e registro il numero di teste
  - Estraggo due individui a caso da una popolazione, chiedo se fumano, e registro il numero di fumatori

$$\mathbf{P(SS) = p^2}$$
$$\mathbf{P(SI) = p(1-p)}$$
$$\mathbf{P(IS) = (1-p)p}$$
$$\mathbf{P(II) = (1-p)^2}$$

e anche

$$\mathbf{P(X=0) = (1-p)^2}$$
$$\mathbf{P(X=1) = 2p(1-p)}$$
$$\mathbf{P(X=2) = p^2}$$

dove X è la variabile " numero di successi in un campione casuale di n prove", o, brevemente, "numero di successi in n prove"

Questi sono già i termini della distribuzione binomiale per n=2. Se chiamo q=1-p, si vede chiaramente che i termini sono dati dall'espansione del binomio $(p+q)^2 = p^2 + 2pq + q^2$

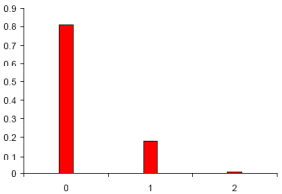
- I risultati possibili con n=2
  - **SS** (prima prova = successo; seconda prova = successo)
  - **SI** (prima prova = successo; seconda prova = insuccesso)
  - **IS** (prima prova = insuccesso; seconda prova = successo)
  - **II** (prima prova = insuccesso; seconda prova = insuccesso)

- Se
  - il risultato della prima prova non influenza il risultato della seconda prova, e
  - le probabilità di successo p è la stessa in ogni prova

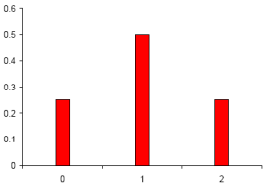
ALLORA

Due esempi con n=2 e due diversi valori di *p* (indicato con π)

x = numero di successi	P($X = x$) = P(x) = Probabilità di ottenere x successi se π = 0.1
0	$(1-\pi)^2 = 0.81$
1	$2(1-\pi)\pi = 0.18$
2	$\pi^2 = 0.01$



x = numero di successi	P($X = x$) = P(x) = Probabilità di ottenere x successi se π = 0.5
0	$(1-\pi)^2 = 0.25$
1	$2(1-\pi)\pi = 0.5$
2	$\pi^2 = 0.25$



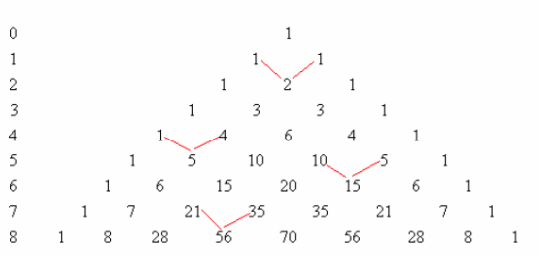
- Aumentando il numero di prove, e ragionando quindi per dimensioni campionarie maggiori, i calcoli non si complicano molto

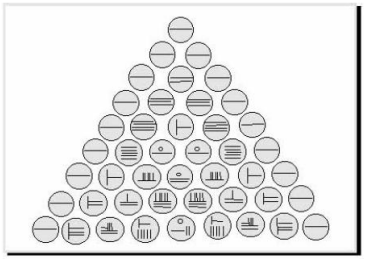
- Vediamo per n = 3

Risultato	x = numero di successi	P(*X* = *x*) = P(*x*) = Probabilità di ottenere *x* successi
III	0	$(1-\pi)^3$
IIS oppure ISI oppure SII	1	$3(1-\pi)^2\pi$
ISS oppure SIS oppure SSI	2	$3(1-\pi)\pi^2$
SSS	3	π^3

- Ovvero, l'espansione del binomio $(p+q)^3$

- Per n maggiori, si può ricorrere al triangolo di Tartaglia, o meglio, a triangolo di Chu Shin-Chieh, per trovare i coefficienti dei diversi termini, ma per fortuna c'e' anche la funzione matematica della distribuzione binomiale





Chu Shin-Chieh (1303)

- Ma per fortuna esiste l'equazione matematica della distribuzione binomiale

$$P(X\ successi\ in\ n\ prove) = \binom{n}{X} p^X (1-p)^{n-X}$$

dove

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

[questo termine si chiama *coefficiente binomiale*]

- 0! = 1
- n! = n x [(n-1)]! = n x (n-1)x [(n-2)]! = ecc.
  - Es: 5! = 5 x 4 x 3 x 2 x 1 = 120
- $\binom{n}{0} = \binom{n}{n} = 1$

Perché è giusto aspettarsi che per X = 0 o X = *n* il coefficiente binomiale sia pari a 1 e per X = 1 o X = *n*-1 il coefficiente binomiale sia pari a *n* ?

ESEMPIO: Qual è la probabilità di osservare solo alberi sani in un campione casuale di 10 alberi presi da una popolazione in cui la proporzione (= probabilità) di alberi con una certa malattia è pari a 0.1?

- Successo: albero sano
- Insuccesso: albero malato
- P(Successo nella singola prova) = *p* = 0.9

$$P(10\ successi\ in\ 10\ prove) = \binom{10}{10} 0.9^{10}(0.1)^{10-10}$$

$$P(10\ successi\ in\ 10\ prove) = 1 \times 0.9^{10} \times 1 = 0.35$$

In questo caso mi bastava in realtà la regola del prodotto…

Ma se cercavo invece la probabilità di osservare 4 alberi sani e 6 malati in un campione casuale di 10 alberi presi da una popolazione in cui la proporzione (= probabilità) di alberi con una certa malattia è pari a 0.1?

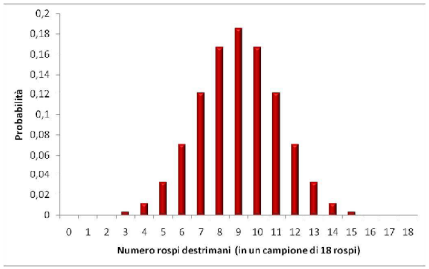
$$P(4\ successi\ in\ 10\ prove) = \binom{10}{4} 0.9^{4}(0.1)^{6}$$

$$P(4\ successi\ in\ 10\ prove) = \frac{10!}{4!\,6!} 0.9^{4}(0.1)^{6}$$

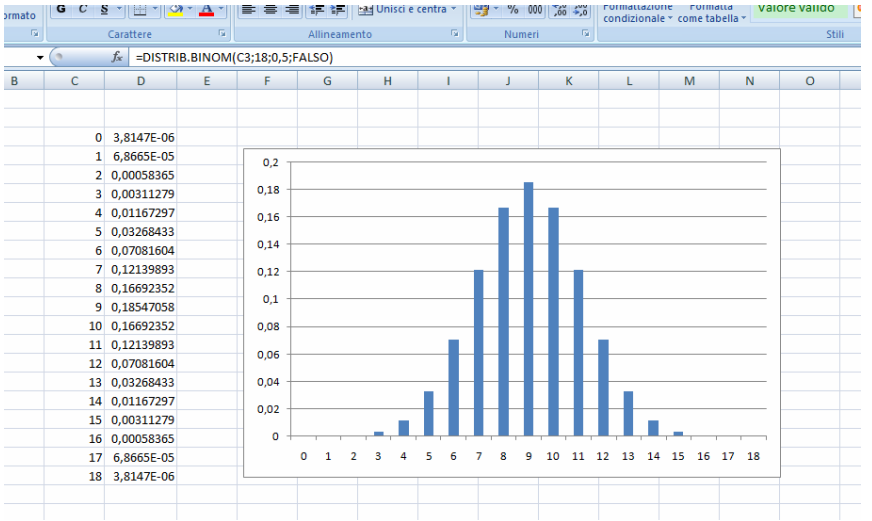
$$P(4\ successi\ in\ 10\ prove) = \frac{10!}{4!\,6!} 0.9^{4}(0.1)^{6} = 1.4 \cdot 10^{-4}$$

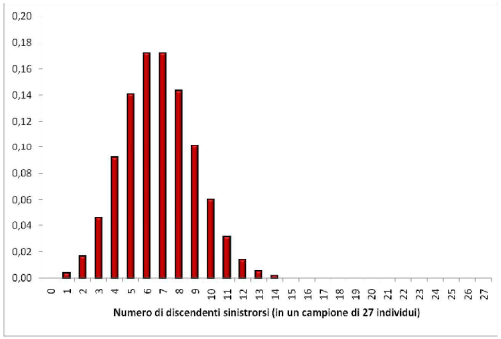
[provate con Excel, funzione Distrib.Binom()]

➢ Ovviamente ora siamo in grado, utilizzando più volte la funzione della distribuzione binomiale, di costruire la distribuzione nulla per il test sui rospi, sui fiori asimmetrici, ecc.



$$P(X\ rospi\ destrimani\ in\ 18\ prove) = \binom{18}{X} 0.5^{X}(0.5)^{18-X}$$





$$P(X\ discendenti\ sinistrorsi\ in\ 27\ prove) = \binom{27}{X} 0.25^X (0.75)^{27-X}$$

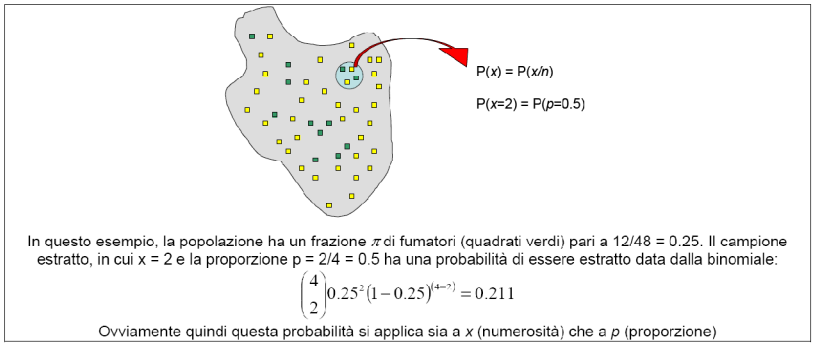
- Quindi, siamo ora in grado di svolgere un TEST ESATTO BINOMIALE su una proporzione
- Infatti, possiamo costruire la distribuzione nulla del numero di successi in n prove per qualsiasi valore di *p* (il parametro previsto dall'ipotesi nulla) e di *n* (la dimensione campionaria)
- Vediamo ora qualche proprietà della statistica che possiamo usare per stimare *p*
- Questa statistica è semplicemente la proporzione nel campione

- In ogni data set con X individui in una certa categoria (successi) su un totale di n osservazioni (prove), la miglior stima della proporzione nella popolazione (che è poi la probabilità di avere un successo nella singola prova) è dato da

$$\hat{p} = \frac{X}{n}$$

- Come è fatta la distribuzione campionaria di $\hat{p}$?

La distribuzione campionaria di una proporzione è binomiale perché lo è la distribuzione campionaria del numero di successi in *n* prove



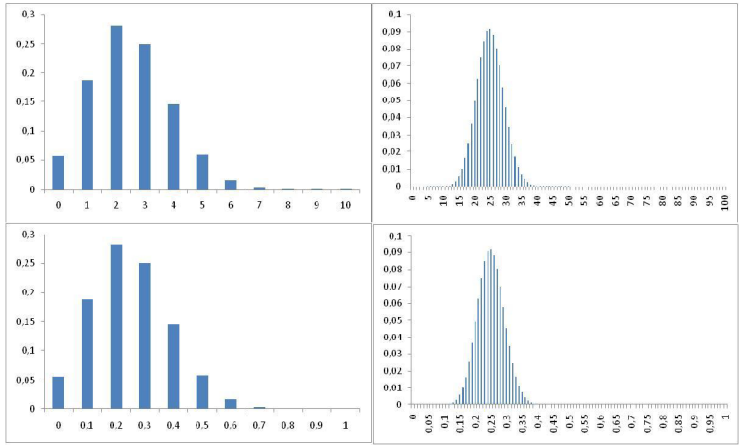
In questo esempio, la popolazione ha un frazione π di fumatori (quadrati verdi) pari a 12/48 = 0.25. Il campione estratto, in cui x = 2 e la proporzione p = 2/4 = 0.5 ha una probabilità di essere estratto data dalla binomiale:

$$\binom{4}{2} 0.25^2 (1-0.25)^{(4-2)} = 0.211$$

Ovviamente quindi questa probabilità si applica sia a *x* (numerosità) che a *p* (proporzione)

Distribuzione di X (sopra) e $\hat{p}$ (sotto) per p=0.25 e n=10

Distribuzione di X (sopra) e $\hat{p}$ (sotto) per p=0.25 e n=100



Due cose importanti sulla <u>distribuzione campionaria di $\hat{p}$</u>

1. La media è centrata su p (in altre parole, $\hat{p}$ è una stima non distorta)

2. La deviazione standard (= errore standard della stima) è pari a

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

Una stima di questa grandezza, che misura la precisione di una proporzione stimata, è data da

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

- Ora possiamo calcolare l'IC di una proporzione
- Metodo di Wald (molto diffuso ma impreciso soprattutto quando n è piccolo e p è vicino a 0 o a 1)

$$IC_{95\%}[proporzione] = \hat{p} \pm 1.96 \cdot ES_{\hat{p}}$$

- Metodo di Agresti e Coull (da preferire, ma anche questo è impreciso per n piccoli)

$$IC_{95\%}[proporzione] = p' \pm 1.96 \cdot \sqrt{\frac{p'(1-p')}{n+4}}$$

Dove p' = (X+2)/(*n*+4)

Esempio 7.3

Dati: 30 figli di radiologi su 87 sono maschi
Calcolare ES e IC della proporzione e commentare

$\hat{p} = \frac{X}{n} = \frac{30}{87} = 0.345$

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

$IC_{95\%}[proporzione] = p' \pm 1.96 \cdot \sqrt{\frac{p'(1-p')}{n+4}}$ [Metodo di Agresti e Coull]

p' = (X+2)/(n+4) = 32/91=0.352

$IC_{95\%}[proporzione] = 0.352 \pm 1.96 \cdot \sqrt{\frac{0.352(1-0.352)}{87+4}}$ $IC_{95\%}$ = 0.254<p<0.450

Esercizio: i geni per la spermatogenesi si trovano soprattutto sul cromosoma X?

Sesso e cromosoma X

Uno studio condotto su 25 geni coinvolti nella spermatogenesi (il processo di maturazione degli spermatozoi a partire dagli elementi germinali precursori) ha identificato la loro posizione nel genoma del topo (*Mus musculus*). L'obiettivo della ricerca era di verificare una previsione della teoria evoluzionistica secondo la quale tali geni dovrebbero essere sovrarappresentati (ossia presenti con una frequenza sproporzionatamente più elevata) sul cromosoma X.[3] È risultato che 10 dei 25 geni per la spermatogenesi (il 40%) risiedono sul cromosoma X (Wang et al., 2001) (*vedi* Figura 7.2-1). Se i geni per la spermatogenesi fossero distribuiti casualmente in tutto il genoma, allora ci attenderemmo che soltanto il 6,1% di essi risieda sul cromosoma X, dato che il cromosoma X contiene il 6,1% del totale dei geni. Questi risultati avvalorano l'ipotesi che i geni per la spermatogenesi si trovano preferenzialmente sul cromosoma X? ■

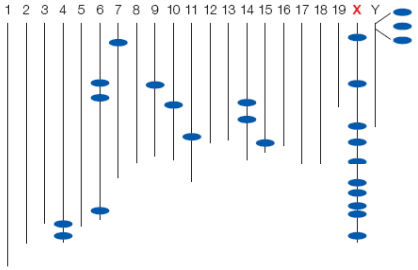
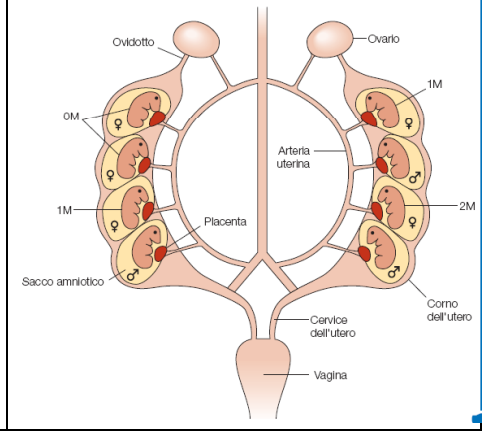


Figura 7.2-1
Rappresentazione schematica del genoma dei topi. Ogni linea verticale rappresenta un cromosoma, e la sua lunghezza è una misura dell'estensione relativa. Ogni punto su una linea indica un singolo gene coinvolto nella spermatogenesi. Si noti l'abbondanza di questi geni sul cromosoma X.

Esercizio: la scelta dei maschi nel topo dipende dalla posizione fetale delle femmine?

I topi hanno figliate costituite da parecchi piccoli. Quando questi sono ancora nell'utero materno, vengono disposti in fila, in modo che ogni feto che non è alle estremità della fila si trovi tra due fratelli. È stato dimostrato che i feti di sesso femminile localizzati tra due feti di sesso maschile (2M) sono soggetti a livelli di testosterone più elevati di quelli a cui sono soggetti i feti di sesso femminile che non sono adiacenti a feti di sesso maschile (0M), perché l'ormone prodotto dai maschi si diffonde attraverso le membrane fetali e attraverso il liquido amniotico, raggiungendo le femmine adiacenti. È noto che i livelli fetali di testosterone più elevati esercitano parecchi effetti sulle femmine in una fase più avanzata della vita; questi effetti comprendono un aumento dei livelli di aggressività, un diverso tasso di crescita e persino la determinazione dell'occhio che si apre per primo. Uno studio si era proposto di misurare se i livelli fetali di testosterone influenzassero il grado di attrattività delle femmine di topo sui maschi nell'età adulta (vom Saal & Bronson, 1980). 24 maschi di topo sono stati messi in condizione di scegliere tra una femmina 0M e una femmina 2M. (Maschi e femmine erano stati scelti a caso.) Ogni maschio è stato collocato su una piattaforma, da cui poteva saltare nella gabbia della femmina preferita. Dei 24 maschi, 19 hanno scelto la femmina 0M.

a. Questo risultato indica che la scelta dei maschi è influenzata dal posizionamento fetale delle femmine?



Ulteriori problemi: 14, 16, 18, 22 (21)

Scheda 4. Correlazione, relazione di causa ed effetto, variabili di confondimento, studi osservazionali e studi sperimentali

L' adattamento di una distribuzione teorica prevista da un modello ad una distribuzione di frequenza osservata: il test del χ^2

Il test del χ^2

- E' un test di goodness-of-fit
- Possiamo usarlo anche in semplici studi già visti al posto del test binomiale esatto, ma solo solo se certe assunzioni sono soddisfatte
- Può essere usato in molte altre situazioni in cui il numero di categorie è >2. Esempio con il modello proporzionale

Non si nasce nel week-end?

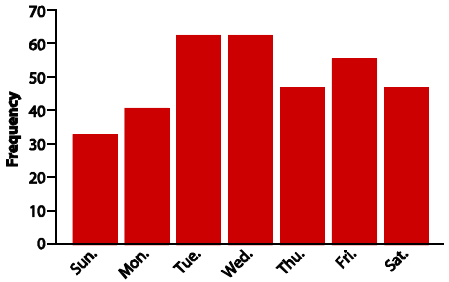


Tabella 8.1-1

Giorno della settimana per 350 nascite negli Stati Uniti nel 1999.

Giorno	Numero di nati
Domenica	33
Lunedi	41
Martedi	63
Mercoledi	63
Giovedi	47
Sabato	56
Domenica	47
Totale	350

Tabella 8.2-1

Frequenza attesa di nati in ogni giorno della settimana nel 1999 assumendo il modello proporzionale.

Giorno	Numero di giorni nel 1999	Proporzione di giorni nel 1999	Frequenza attesa di nati
Domenica	52	52/365 = 0,142	49,863
Lunedi	52	0,142	49,863
Martedi	52	0,142	49,863
Mercoledi	52	0,142	49,863
Giovedi	52	0,142	49,863
Sabato	53	53/365 =0.145	50,822
Domenica	52	0,142	49,863
Somma	365	1	350

H_0 : La probabilità di nascita è la stessa in tutti i giorni della settimana
H_A: La probabilità di nascita non è la stessa in tutti i giorni della settimana

[Gli attesi dipendono anche da quanti L,Ma, Me,...ci sono in un anno: è il modello proporzionale]

- A questo punto mi serve una statistica test

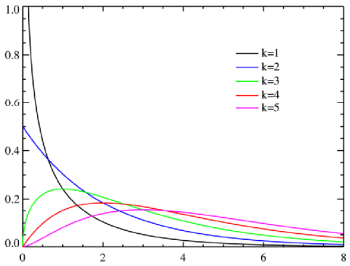
$$\chi^2 = \sum_i \frac{(Osservati_i - Attesi_i)^2}{Attesi_i} = \sum \frac{(O-A)^2}{A}$$

- Intuitivamente, questa statistica test misura la deviazione tra numerosità osservate e quelle previste dall'ipote nulla (il modello proporzionale in questo caso)
- Come sempre, ci serve la distribuzione nulla della statistica test, cioè la distribuzione campionaria di χ^2 se fosse vera l'ipotesi nulla

- Per simulazione è possibile, ma lungo e non pratico

- Esiste però la distribuzione teorica di χ^2
  - Continua, sempre positiva, dipende dai *gradi di libertà*

- I gradi di libertà devono essere determinati per il data set che si sta analizzando



gdl = df = gradi di libertà (*degrees of freedom*) = numero di categorie - 1 - (numero di parametri stimato in base ai dati)

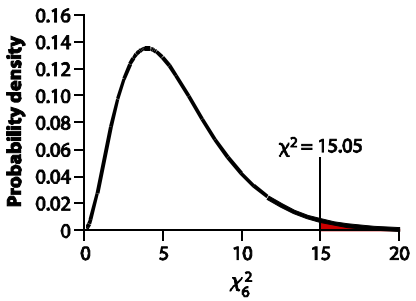
[numero di categorie indipendenti]

- Come posso calcolare il P-value o determinare le regioni di accetazione/rifiuto? Integrale molto complesso, quindi uso un PC oppure le tavole statistiche del χ^2

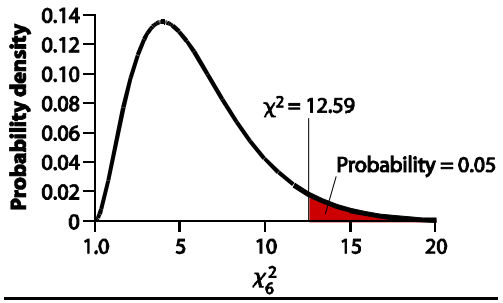
	α									
df	0,999	0,995	0,99	0,975	0,95	0,05	0,025	0,01	0,005	0,001
1	0,0000016	0,0000393	0,00016	0,00098	0,00393	3,84	5,02	6,63	7,88	10,83
2	0,002	0,01	0,02	0,05	0,10	5,99	7,38	9,21	10,60	13,82
3	0,02	0,07	0,11	0,22	0,35	7,81	9,35	11,34	12,84	16,27
4	0,09	0,21	0,30	0,48	0,71	9,49	11,14	13,28	14,86	18,47
5	0,21	0,41	0,55	0,83	1,15	11,07	12,83	15,09	16,75	20,52
6	0,38	0,68	0,87	1,24	1,64	12,59	14,45	16,81	18,55	22,46
7	0,60	0,99	1,24	1,69	2,17	14,07	16,01	18,48	20,28	24,32
8	0,86	1,34	1,65	2,18	2,73	15,51	17,53	20,09	21,95	26,12
9	1,15	1,73	2,09	2,70	3,33	16,92	19,02	21,67	23,59	27,88
10	1,48	2,16	2,56	3,25	3,94	18,31	20,48	23,21	25,19	29,59
11	1,83	2,60	3,05	3,82	4,57	19,68	21,92	24,72	26,76	31,26
12	2,21	3,07	3,57	4,40	5,23	21,03	23,34	26,22	28,30	32,91
13	2,62	3,57	4,11	5,01	5,89	22,36	24,74	27,69	29,82	34,53
14	3,04	4,07	4,66	5,63	6,57	23,68	26,12	29,14	31,32	36,12
15	3,48	4,60	5,23	6,26	7,26	25,00	27,49	30,58	32,80	37,70

- La tabella è a una coda, ma il test è a due code. Perchè?

La distribuzione teorica del χ^2 con df = 6



In rosso il P-value per l'esempio 8.1; il valore di χ^2 indicato è quello calcolato



In rosso l'area di rifiuto per $\alpha = 0.05$; il valore di χ^2 calcolato è il valore critico per $\alpha = 0.05$

Assunzioni del test χ^2:

- Campione casuale
- Nessuna categoria deve avere una frequenza attesa minore di 1
- Non più del 20% delle categorie deve avere frequenze attese minori di 5

Test esatto binomiale

- Successo: nascita nel w/e
- H_0 : p= 2/7
  H_A: p≠ 2/7
- Statistica test: numero di nascite nel w/e
- Distribuzione nulla: binomiale
- P-value: 2 x Probabilità (calcolata con la binomiale per n=716 e p = 0.286) di avere 216 o meno successi [*perché meno?*]

Esercizio 15, pag. 122

- Con due categorie conviene applicare il test esatto binomiale (è esatto!), ma non sempre è possibile se *n* è molto grande
- Dati esempio: 216 nascite nei week-end, 716 nei giorni feriali
- Verificare l'ipotesi nulla che il tasso di natalità sia lo stesso nei giorni feriali e nei giorni del w/e
- Vediamo i due approcci possibili, visto che ci sono solo due categorie

- Ottenuto con Excel: P-value = 0.000117
- Per α=0.05, ma anche per α=0.01, P-value<α, quindi

  Rifiuto H_0

Test del Chi-quadrato (χ^2)

- H_0 : p= 2/7
  H_A: p≠ 2/7
- Statistica test: χ^2
- Distribuzione nulla: distribuzione teorica del χ^2 con df =1
- Calcolo i valori attesi nelle due categorie
  - Nascite attese nel w/e: (2/7)x932 = 266.29
  - Nascite attese nei giorni fer.: (5/7)x932 = 665.71
- Calcolo la statistica test

$$\chi^2 = \frac{(216-266.29)^2}{266.29} + \frac{(716-665.71)^2}{665.71}$$

$$\chi^2 = 9.50 + 3.80 = 13.30$$

- Consulto la tabella della distribuzione nulla
  - Il valore critico con α=0.05 è pari a 3.84
  - Il valore critico con α= 0.01 è pari a 6.63
- Il valore calcolato (13.30) è superiore a quello critico, anche scegliendo un α=0.01

Rifiuto H_0

[P-value <0.001 (tabella); P-value = 0.000265 (Excel), impreciso perché Chi-quadro non è un test esatto]

Calcolo ES e IC della proporzione di nascite nel w/e

$\hat{p} = \frac{X}{n} = \frac{216}{932} = 0.232$

$ES_{\hat{p}} = \sqrt{\frac{0.232(1-0.232)}{931}} = 0.014$

$IC_{95\%}[proporzione] = p' \pm 1.96 \cdot \sqrt{\frac{p'(1-p')}{n+4}}$ [Metodo di Agresti e Coull]

p' = (X+2)/(n+4) = 218/936=0.233

$IC_{95\%}[proporzione] = 0.233 \pm 1.96 \cdot \sqrt{\frac{0.233(1-0.233)}{932+4}} = 0.233 \pm 1.96 \cdot 0.014$

$IC_{95\%}$ = 0.206<p<0.260 **[non include p=2/7=0.286 !]**

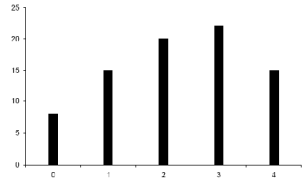
L' adattamento di una distribuzione binomiale ai dati osservati [uso ancora il test del χ^2 per verificare le ipotesi]

- Esistono molti altri modelli, oltre a quello proporzionale
- Tra questi, siamo spesso interessati ad un modello che prevede che i dati si distribuiscano in categorie secondo una distribuzione binomiale: è quello che mi aspetto quando ogni categoria corrisponde a numeri di successi in n prove, e non interviene nulla d'altro che il caso nel determinare in natura in quale categoria viene inserita un'osservazione
- **Attenzione**: usiamo la binomiale ma non come distribuzione nulla e per trovare il P-value, ma come modello per trovare i valori attesi
- Il test statistico per verificare le ipotesi sarà un test del χ^2

Esempio: stiamo studiando la sopravvivenza dei piccoli di una certa specie di uccelli, e ci concentriamo solo su nidi con 4 uova

I nidi studiati sono 80. Per ciascuno di essi, determino quanti sono i piccoli che sono sopravvissuti e hanno quindi lasciato il nido (la variabile X). La tabella di frequenza è:

x	n_i	n_i x
0	8	0
1	15	15
2	20	40
3	22	66
4	15	60



Distribuzione di frequenza della variabile *numero di piccoli sopravvissuti per nido*

- Devo calcolare i valori attesi, ossia quelli previsti dalla binomiale

- Il parametro *n* della binomiale è noto (4, attenzione che con *n* adesso indico due cose diverse), ma non conosco *p*, ossia la probabilità di successo (sopravvivenza) nella singola prova (uovo) in un 4 prove (un nido)

Attenzione (ancora): ogni singolo nido potrebbe corrispondere al campione di rospi nel test esatto binomiale. In ogni nido ci sono un numero di prove fisso (n=4), con una certa probabilità di successo e insuccesso in ogni prova. Qui però non siamo interessati ad un singolo nido e alla proporzione di sopravvissuti in quel nido, ma abbiamo molti nidi e siamo interessati alla distribuzione del numero di sopravvissuti

- Se in ogni nido la probabilità di sopravvivere fosse uguale e indipendente per ogni piccolo, quale distribuzione mi aspetterei del numero di sopravvissuti?

La distribuzione binomiale, è questo quindi il nostro modello nullo di partenza, che prevede l'azione casuale della mortalità sui piccoli; non avrebbe senso un modello proporzionale

- Stimo p (che userò poi direttamente nella binomiale) dai dati

$$\hat{p} = \frac{numero\ totale\ di\ piccoli\ sopravvissuti}{numero\ totale\ di\ uova} = \frac{181}{320} = 0.566$$

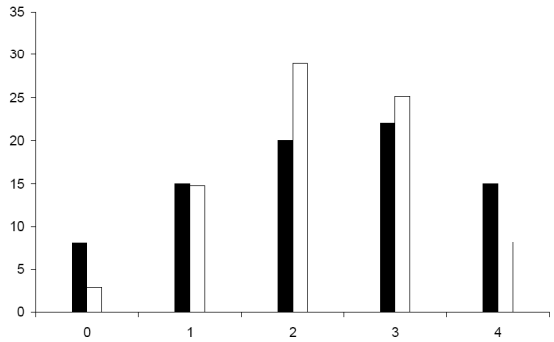
- A questo punto calcolo le frequenze attese in ogni classe

$$P(x=0) = \binom{4}{0} 0.5656^0 (1-0.5656)^4 = (1-0.5656)^4 = 0.03561$$

$$P(x=1) = \binom{4}{1} 0.5656^1 (1-0.5656)^3 = 4\left[0.5656^1 (1-0.5656)^3\right] = 0.18545$$

x	Numerosità osservata	P(*x*) con π = 0.5656	Numerosità attesa = 80(P(x))	Valore di Chi-quadrato $(O-A)^2/A$
0	8	0.03561	2.85	9.315
1	15	0.18545	14.84	0.002
2	20	0.36220	28.98	2.781
3	22	0.31440	25.15	0.395
4	15	0.10234	8.187	5.669

- Il calcolo della statistica test χ^2 (somma ultima colonna) porta al valore di 18.16
- I df sono 5-1-1=3
- Cosa concludo?



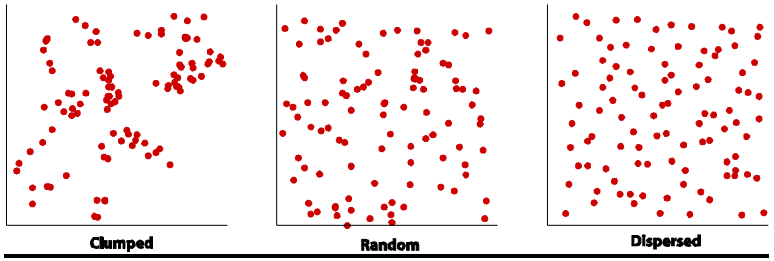
Distribuzione osservata (in nero) e attesa (in bianco). Sono significativamente diverse, con α = 0.01. La distribuzione osservata sembra più "contagiosa" (numerosità elevate agli estremi) di quella attesa

- Da notare. Il parametro della distribuzione binomiale di confronto potrebbe anche essere dato (cioè non da stimare dai dati). In questo caso il numero di gradi di libertà sarebbe pari al numero di classi meno 1.

L' adattamento di una distribuzione <u>di Poisson</u> ai dati osservati [uso ancora il test del χ^2 per verificare le ipotesi!]

- **Cos'è la distribuzione di Poisson?**

<u>Casualità nello spazio o nel tempo: la distribuzione teorica di Poisson</u>



- Cosa sono i punti?

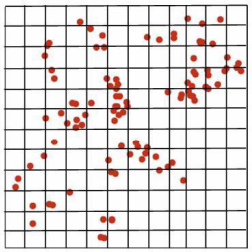
o Organismi e loro posizione. Domanda (per esempio):
  - Come vengono colonizzate nuove aree disponibili? Ovvero, gli organismi sono distribuiti casualmente nell'area analizzata?

o Osservazioni lungo una linea, in un volume, nel tempo. Domanda (per esempio):
  - Gli uccelli si distribuiscono a caso lungo un cavo?
  - Ci sono disomogeneità non casuali in diverse unità di volume?
  - La distribuzione degli eventi nel tempo è casuale?

- La distribuzione di Poisson descrive il numero di successi in intervalli (o blocchi) spaziali, volumetrici, o temporali quando
  - i successi avvengono indipendentemente l'uno dall'altro
  - i successi hanno la stessa probabilità di verificarsi in ogni punto dello spazio, di volume, o di tempo

- Possiamo quindi usare questa distribuzione teorica di probabilità come modello per predire se le osservazioni che abbiamo fatto (nel tempo, nello spazio) sono compatibili con il semplice effetto del caso

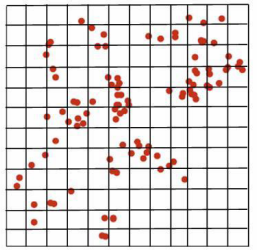
  [a differenza della binomiale, il numero di prove non è noto!]



Numero di piante	*Frequenza (numero di quadrati)*
0	69
1	44
2	21
3	5
4	3
5	1
6	1
	144

- Un esempio con blocchi nello spazio: osservazioni relative a presenze di piante in una superficie, che noi suddividiamo il quadrati di uguale dimensione

- I dati sono 144 valori; ogni valore è il numero di piante in ciascun quadrato; posso costruire la tabella di frequenza (numero di quadrati con 0 piante, numero di quadrati con 1 pianta, numero di quadrati con 2 piante, ecc.)



La distribuzione di Poisson

$$P(X;\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

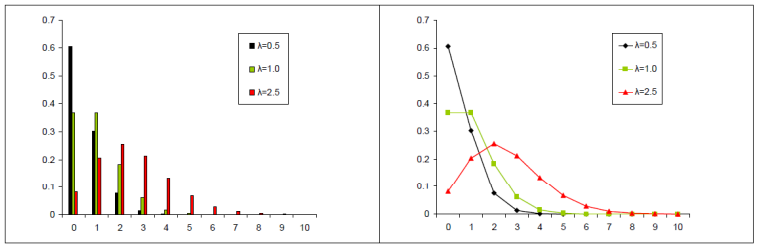
X = numero di eventi in ogni blocco (intervallo) di tempo o spazio

λ = media di eventi per blocco, ossia numero totale di eventi/numero totale blocchi

Nell'esempio delle piante, c'erano in tutto 124 piante in 144 blocchi, ovvero potevo stimare χ con la media $\bar{x}$ = 124/144 = 0.86

- La distribuzione di Poisson è discreta e con un solo parametro, la media λ. La varianza è uguale alla media

- E' il limite della binomiale quando il numero di prove tende a infinito e quando la probabilità di successo nella singola prova tende a 0 (ogni blocco può essere pensato come se fosse suddiviso in infiniti sotto-blocchi)



Distribuzioni di Poisson con diversi valori di λ
(le tre distribuzioni rappresentate sono identiche nelle due diverse rappresentazioni grafiche riportate a destra e a sinistra)

- Utile ricordare che

$$P(X;\lambda)=\frac{\lambda^x}{x!}e^{-\lambda}$$

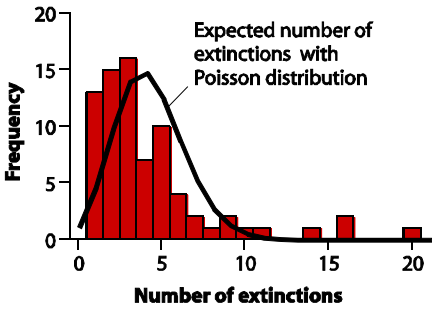
$$P(x=0)=e^{-\lambda}$$
$$P(x=1)=\lambda e^{-\lambda}=P(0)\lambda$$
$$P(x=2)=\frac{\lambda^2}{2!}e^{-\lambda}=P(1)\frac{\lambda}{2}$$
$$P(x=3)=\frac{\lambda^3}{3!}e^{-\lambda}=P(2)\frac{\lambda}{3}$$
$$P(x=k)=\frac{\lambda^k}{k!}e^{-\lambda}=P(k-1)\frac{\lambda}{k}$$

3. Estinzioni di massa: un test sul modello di distribuzione poissoniano (pg 116)

Tabella 8.6-3

La frequenza osservata e la frequenza attesa di intervalli di tempo con un dato numero di estinzioni di famiglie di invertebrati marini.

Numero di 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
Totale	76	76



Ulteriori esempi sulla bontà di adattamento di una distribuzione osservata alla distribuzione teorica binomiale

Esempio 1

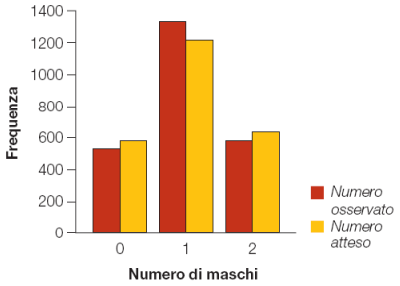
La mortalità in pesci in acquario dipende soprattutto dal caso (la scelta casuale di quale pesce finisce in quale acquario, e altri eventi che agiscono con uguale probabilità su ciascun pesce) o forse dalla diffusione di malattie contagiose?

- In 120 acquari vengono inseriti 6 pesci di una certa specie, scelti a caso da una vasca grande. Da quel momento in poi, non si interviene più sugli acquari e dopo un mese si contano i pesci sopravvissuti per ogni vasca. I risultati, come numero di vasche con 0,1,2,3,4,5,6 pesci sopravvissuti, è il seguente: 12,26,32,29,10,7,4.

Esempio 2

Il rapporto tra i due sessi nelle famiglie è casuale?

Nel Capitolo 5 abbiamo affermato che nella specie umana il sesso dei figli di una stessa coppia è una variabile indipendente. Per esempio, il fatto di avere già generato un maschio non fa variare la probabilità che anche il figlio successivo sia un maschio. Quindi, in assenza di complicazioni, ci attendiamo che il numero di maschi e il numero di femmine nati in famiglie con 2 figli siano conformi a una distribuzione binomiale, con $n = 2$ e p uguale alla probabilità di avere un maschio in ogni singola prova. È quanto si osserva? Rodgers e Doughty (2001) hanno verificato questa ipotesi usando i dati provenienti dal National Longitudinal Survey of Youth (NLSY), dove vengono registrati i dati sul sesso dei figli in campioni casuali di famiglie che variano nel numero dei componenti. ■



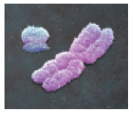
La distribuzione di frequenza del numero di maschi in famiglie con 2 figli.

Numero di maschi	Numero osservato di famiglie
0	530
1	1332
2	582
Totale	2444

Ulteriore esempio di test del chi-quadrato con due categorie (Esempio 8.4, pg 112)

Il contenuto genico del cromosoma X umano

I cromosomi sessuali vengono ereditati con un meccanismo diverso rispetto a quello degli altri cromosomi, e ciò influenza in molti modi la loro evoluzione. Questi cromosomi sono inoltre anche sotto altri aspetti? Per esempio, sul cromosoma X umano esistono tanti geni quanti ci si aspetterebbe sulla base delle sue dimensioni? Il Progetto Genoma Umano ha identificato nel cromosoma X dell'uomo 751 geni dei 20 290 geni trovati finora nell'intero genoma.[5] Sappiamo anche che il cromosoma X contiene circa il 5,2% del DNA totale. Assumendo vero il modello proporzionale, quindi, ci aspettiamo che il 5,2% dei geni sia localizzato sul cromosoma X. È ciò che osserviamo effettivamente? ■



Numero di geni sul cromosoma X umano e sul resto del genoma.

Cromosoma	Osservato	Atteso
X	781	1055
Non X	19 509	19 509
Totale	20 290	20 290

Attenzione, questa NON è una tabella di contingenza!

Risultato: χ^2 calcolato = 75.1. Questo valore è nettamente superiore al valore critico con 1 gdl. L'ipotesi nulla è rifiutata. Nel genoma umano. il numero di geni sul cromosoma X è significativamente minore di quello che ci aspetteremmo sulla base delle sue dimensioni.

Cosa fare quando le assunzioni richieste dal test del chi quadrato non vengono soddisfatte?

- Alcune soluzioni
  - Ricorrere ad un altro test che non necessiti della distribuzione teorica nulla del χ^2
    - χ^2 come "goodness-of –fit test" ma con la distribuzione nulla ricostruita per simulazione
  - Raggruppare alcune categorie
    - E necessario che le nuove categorie abbiano una logica e un significato
  - Eliminare alcune categorie
    - Attenzione: il data set si riduce. Attenzione anche al fatto che l'interpretazione finale non si applicherà ai dati originali ma a quelli ottenuti dopo l'eliminazione

Scheda 5: pianificare l'esperimento

- Meglio pianificare piuttosto che ritrovarsi con dati inutili o inutilizzabili!
- Enunciare bene la domanda scientifica e chiedersi se è interessante (importante) trovare una risposta e se è già stata affrontata da altri
- Pensate nei dettagli all'esperimento o alla raccolta di dati, senza aggiungere inutili complicazioni
- Ragionate su possibili problemi (pseudoreplicazione, campioni non casuali, variabili confondenti)
- La dimensione campionaria è sufficiente?
- Pensare ai risultati possibili: sarebbero sufficienti per trarre una conclusione?
- Discutete con altri colleghi

Lo studio dell'associazione tra variabili categoriche

- Esempio: la variabile esplicativa *colore* determina una diversa probabilità di essere predati (misurata dalla variabile risposta *predato/non predato*)?

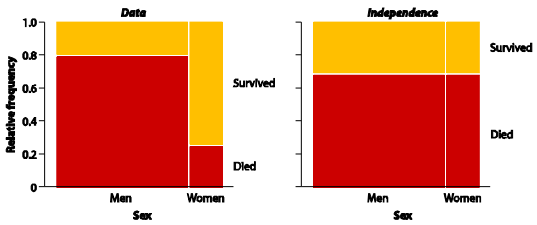
- Esempio: la variabile esplicativa *fumatore/non fumatore* determina una diversa probabilità di sviluppare un tumore (misurata dalla variabile risposta *tumore/non tumore*)?

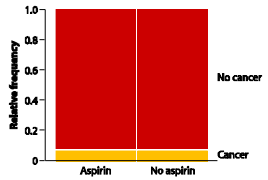
- Un generale, la proporzione di animali predati è significativamente diversa in gruppi di animali di diversi colori?

- In generale, le due variabili sono associate? Ovvero, la probabilità di sviluppare un tumore dipende dal fatto che una persona fumi oppure no?

- La descrizione dei dati in questi studi l'abbiamo già vista (tabelle di contingenza, diagrammi e grafici, pg 21-22)

[attenzione: ogni variabile può avere più in 2 categorie]





- Ora vedremo l'analisi inferenziale, rispondendo alla domanda: "*l'associazione osservata è significativamente superiore a quanto potrebbe generarsi solo per caso in seguito all'errore dovuto campionamento?"*

- L'ipotesi nulla sarà quindi *assenza di associazione tra le due variabili categoriche*, quella alternativa sarà *presenza di associazione tra le due variabili categoriche*

- Prima di fare il test statistico (che sarà basato sula distribuzione del chi quadrato) vediamo l'**odds** (pronostico, quota) e l'**odds ratio** (rapporto di odds, rischio relativo), e l'IC dell' odds reatio

$$O = \frac{p}{1-p}$$

- E' un rapporto tra la probabilità di un evento e la probabilità che l'evento non si verifichi

- Se O = 1 (scritto anche 1:1), probabilità di successo e insuccesso sono uguali (ogni successo si verifica un insuccesso)

- Se O = 10 (scritto anche 10:1) la probabilità di successo è 10 volte più grande di quella dell'insuccesso

- Se misura l'odds in due gruppi, e ne faccio il rapporto, ottengo l'odds ratio: è un rischio che avvenga un evento in un gruppo relativamente al rischio nell'altro gruppo

	Sani	Malati (Leucemia)	Totale
Molta attività sociale	5343	1020	6363
Poca attività sociale	895	252	1147
Totale	6238	1272	7510

[i gruppi sono definiti dalla condizione di salute]

a) Studio sperimentale o osservazionale?

b) Proporzione di bambini con attività sociale: 5343/6238 = 85.7% nei sani; 1020/1272 = 80.2% nei malati di leucemia

c) Odds di attività sociale nei malati: 0.802/0.198 = 4.0; nei sani tale odds è pari a 0.857/0.143 = 6.0

d) Odds ratio di malati vs. sani: $\widehat{OR}$ = 4/6 = 0.67 (sarebbe 0.68 considerando più decimali)

- Se l'odds ratio è vicino a uno, la probabilità che si verifichi l'evento è molto simile nei due gruppi: variabile risposta (evento studiato) e quella esplicativa (quella che distingue i gruppi) non sono associate

- Al contrario (OR diverso da 1), nei due gruppi esistono rischi diversi

- Esempio 23 (pg 139) e calcolo IC

e) Calcolo IC per sapere se 1 (no associazione) è incluso

1) Calcolo l'ES del logaritmo naturale dell'odds ratio

$$ES\left[\ln\left(\widehat{OR}\right)\right] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$ES\left[\ln\left(\widehat{OR}\right)\right] = \sqrt{\frac{1}{5343}+\frac{1}{895}+\frac{1}{1020}+\frac{1}{252}} = 0.08$$

2) Se l'intervallo di confidenza è al 95%, posso usare la regola già vista: sottraggo e aggiungo alla stima 1.96 volta l'errore standard

$$IC_{95\%}\left[\ln\left(\widehat{OR}\right)\right] = \ln\left(\widehat{OR}\right) \pm 1.96 \cdot ES_{\ln\left(\widehat{OR}\right)}$$

$$IC_{95\%}\left[\ln\left(\widehat{OR}\right)\right] = \ln(0.67) \pm 1.96 \cdot 0.08$$

$$IC_{95\%}\left[\ln\left(\widehat{OR}\right)\right] = -0.4 \pm 0.16$$

$$IC_{95\%}\left[\ln\left(\widehat{OR}\right)\right] => tra - 0.56\ e - 0.24$$

3) Calcolo i limiti (con e^x) di OR e ottengo

$$IC_{95\%}\left[\left(\widehat{OR}\right)\right] => tra\ 0.57\ e\ 0.79$$

f) Cosa concludiamo sui dati ottenuti? C'e' un'associazione tra socialità e sviluppo della leucemia?

g) Quali potrebbero essere altre spiegazioni? (salute pre-socializzazione? Altre variabili confondenti associate alla scelta asilo/no asilo che potrebbero influire sulla salute? Ricchezza famiglie?)

Il test del χ^2 per l'analisi delle tabelle di contingenza

- L'associazione tra variabili categoriche può essere testata ancora utilizzando la statistica test χ^2 e la distribuzione teorica del χ^2
  - E' appunto il test del χ^2 per l'analisi di una tabella di contingenza
  - E' anche un test per confrontare due distribuzioni di frequenza osservate
  - E' anche un test per confrontare una distribuzione di frequenza in due dimensioni con la distribuzione attesa in caso di indipendenza tra le due dimensioni/categorie (modello teorico)

- Ipotesi nulla: le due variabili categoriche sono indipendenti
- Ipotesi alternativa: le due variabili categoriche sono associate
- Le due varaibili categoriche possono avere più di due categorie
- Esempio con tabella di contingenza 2x2 (Es. 7, pg 136)

7. In uno studio di Doll et al. (1994) è stata esaminata la relazione tra assunzione moderata di alcolici e rischio di cardiopatie. Sono stati monitorati in tutto 420 uomini, 209 «astemi» e 201 «bevitori moderati», per un periodo di 10 anni, ed è il stato confrontato il numero di individui che hanno subito arresto cardiaco nei due gruppi. Tutti gli uomini avevano 40 anni di età all'inizio dell'esperimento. Alla fine del periodo considerato, 12 astemi e 9 bevitori moderati avevano avuto un arresto cardiaco.

Osservato			
	Arresto cardiaco	Nessun arresto cardiaco	Totale
Astemi	12	197	209
Bevitori	9	192	201
Totale	21	389	410

[i due gruppi seguiti sono gli astemi e i bevitori]

- Domanda: la frequenza di arresti cardiaci è diversa nei due gruppi
- Lo studio può essere visto come un confronto tra due proporzioni, tra due distribuzioni di frequenza osservate in due gruppi, o come un test sull'associazione tra bere alcool e rischio di arresto cardiaco
- Dobbiamo cercare le frequenze attese se fosse vera l'ipotesi nulla di indipendenza. Cosa abbiamo imparato sulla probabilità che si verifichino eventi indipendenti?

La regola del prodotto: se due eventi sono indipendenti, la probabilità che si verifichino entrambi (uno e l'altro) è data dal prodotto delle due probabilità

$$\Pr[A \text{ e } B] = \Pr[A] \times \Pr[B]$$

- Qual è nel nostro campione una stima della probabilità di essere astemi ? 209/410 = 0.5097
- Qual è nel nostro campione una stima della probabilità di aver subito un arresto cardiaco ? 21/410 = 0.0512
- Se le due variabili categoriche fossero indipendenti, allora una stima della probabilità di essere astemi e aver subito un arresto cardiaco è data da (regola del prodotto) 0.5097*0.0512=0.026
- Frequenza attesa di essere astemi e aver subito un arresto cardiaco se è vera l'ipotesi nulla di indipendeza = 0.026*410 = 10.7
- Ripeto il ragionamento per ognuna delle 4 celle

Atteso			
	Arresto cardiaco	Nessun arresto cardiaco	Totale
Astemi	10,7	198,3	209
Bevitori	10,3	190,7	201
Totale	21	389	410

- Regoletta più semplice:

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

- Attenzione; i totali di riga e di colonna nelle tabelle degli attesi e degli osservati devono essere gli stessi!

- A questo punto, calcoliamo 4 elementi di χ^2, (4 categorie, combinando le due variabili categoriche), li sommiamo per ottenere la statistica test, da confrontare con il valore critico

$$\frac{(Osservato - atteso)^2}{atteso}$$

	Arresto cardiaco	Nessun arresto cardiaco	Totale
Astemi	0,16	0,01	
Bevitori	0,16	0,01	
Totale			0,34

- Quanti df?

$$df = (r-1)(c-1)$$

- Perche?

- Cosa concludiamo riguardo a alcool e arresto cardiaco?

- Attenzione che è uno studio osservazionale: gli individui non sono stati assegnati a caso ai trattamenti bere/non bere; che è astemio ora potrebbe essere stato un gran bevitore

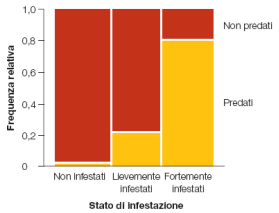
- Ricordarsi le assunzioni del test del χ^2 e cosa fare se tali assunzioni non sono soddisfatte

- Vediamo un altro esempio con una tabella 2x3

L'incredibile verme che convince un pesce a farsi mangiare

Molti parassiti hanno più di una specie ospite e il singolo parassita deve trasferirsi da un ospite a un altro per portare a termine il suo ciclo biologico. I trematodi della specie *Euhaplorchis californiensis* utilizzano tre ospiti durante il loro ciclo biologico. Raggiungono la maturità negli uccelli, dove depongono le uova che raggiungono l'ambiente esterno attraverso le feci dell'ospite. Il gasteropode *C. californica* ingerisce queste uova, che si schiudono e crescono fino a raggiungere un altro stadio del ciclo biologico all'interno del mollusco. Durante questa fase, *Cerithidea californica* perde la capacità di riprodursi a causa della presenza del parassita. Successivamente, quando un gasteropode infestato viene ingerito dal pesce *Fundulus parviponnis*, il parassita si sviluppa fino a raggiungere il successivo stadio del suo ciclo biologico e si incista nella scatola cranica del pesce. Infine, quando il pesce viene ingerito da un uccello, il parassita diventa un adulto maturo e ricomincia il ciclo.

È stato osservato che il pesce infestato trascorre un tempo eccessivamente lungo in prossimità della superficie dell'acqua, dove è più vulnerabile dalla predazione da parte degli uccelli. Questo comportamento è certamente vantaggioso per il parassita, perché aumenta la possibilità di essere ingerito da un uccello, l'ospite successivo. Lafferty e Morris (1996) hanno verificato l'ipotesi che l'infestazione influenzi il rischio di predazione da parte degli uccelli. In una grande vasca all'aperto sono stati posti tre diversi gruppi di pesci della specie F. parvipennis: non infestati, lievemente infestati e fortemente infestati. La vasca è stata lasciata accessibile a varie specie di uccelli, specialmente aironi e garzette. La Tabella 9.3-1 riporta il numero di pesci ingeriti a seconda del loro livello di infestazione. ■



La tabella di contingenza

Frequenze osservate di pesci predati e non predati da parte di uccelli secondo il livello di infestazione da parte di trematodi.

	Non infestati	Lievemente infestati	Fortemente infestati	Totali delle righe
Mangiati dagli uccelli	1	10	37	48
Non mangiati dagli uccelli	49	35	9	93
Totali delle colonne	50	45	46	141

Verificare se livello di infestazione e rischio di predazione sono associati oppure no

Il test esatto di Fisher per tabelle di contingenza 2x2

I dati (Problema 16, Pg 138)

	Emisf. sx. inattivato	Emisf. dx. inattivato	Totale
Scelta propria immag.	5	1	6
Scelta altra immag.	0	4	4
Totale	5	5	10

La tabella ossservata e tutte le tabelle possibili con le loro probabilità

5	1
0	4

P = 0.02381

[

Distribuzione ipergeometrica

$$P\left(\begin{matrix} a & b \\ c & d \end{matrix}\right) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,b!\,d!\,n!}$$

$$P\left(\begin{matrix} 5 & 1 \\ 0 & 4 \end{matrix}\right) = \frac{(6)!\,(4)!\,(5)!\,(5)!}{5!\,4!\,10!} = 120/5040$$

]

I valori attesi (e sotto ancora i dati)

	Emisf. sx. inattivato	Emisf. dx. inattivato	Totale
Scelta propria immag.	3.0	3.0	6
Scelta altra immag.	2.0	2.0	4
Totale	5	5	10

	Emisf. sx. inattivato	Emisf. dx. inattivato	Totale
Scelta propria immag.	5	1	6
Scelta altra immag.	0	4	4
Totale	5	5	10

5	1
0	4

P = 0.02381

4	2
1	3

P = 0.2381

3	3
2	2

P = 0.4762

2	4
3	1

P = 0.2381

1	5
4	0

P = 0.02381

Pvalue = 0.04762

Esempio 9.4

I dati

	Femmine in estro	Femmine non in estro	Totale
Morse da vampiri	15	6	21
Non morse da vampiri	7	322	329
Totale	22	328	350

I dati e i valori attesi

	Femmine in estro	Femmine non in estro	Totale
Morse da vampiri	15	6	21
Non morse da vampiri	7	322	329
Totale	22	328	350

	Femmine in estro	Femmine non in estro	Totale
Morse da vampiri	1.3	19.7	21
Non morse da vampiri	20.7	308.3	329
Totale	22	328	350

Le tabelle più estreme

16	5
6	323

17	4
5	324

18	5
4	325

19	2
3	326

20	1
2	327

21	0
1	328

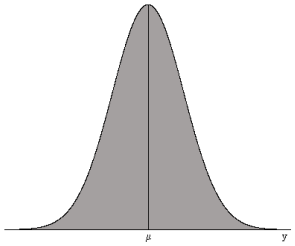
Sito per svolgere il test esatto di Fisher per tabelle 2x2

//research.microsoft.com/en-us/um/redmond/projects/mscompbio/FisherExactTest/

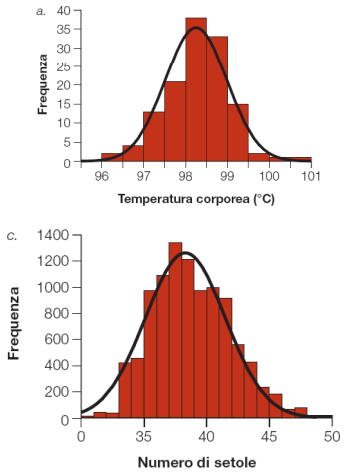
La distribuzione normale (o gaussiana)

- Importante: tendono a seguire questa distribuzione teorica continua gli errori di misura, molte variabili biologiche, molte stime e molte statistiche test. E molti test statistici assumono che la variabile analizzata abbia una distribuzione normale

$$f(y;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(y-\mu)^2}{2\sigma^2}} \qquad -\infty < y < +\infty$$

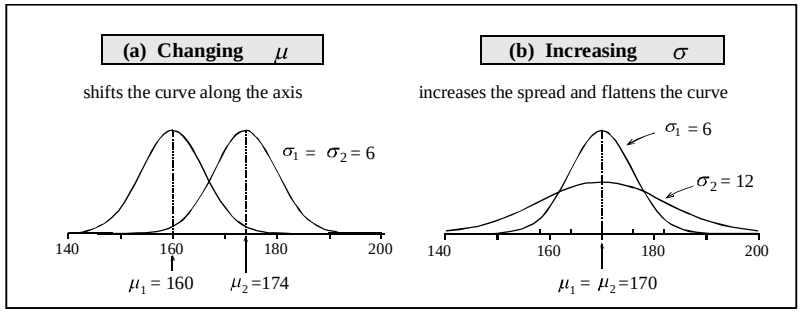


VARIABILI BIOLOGICHE E DISTRIBUZIONE NORMALE

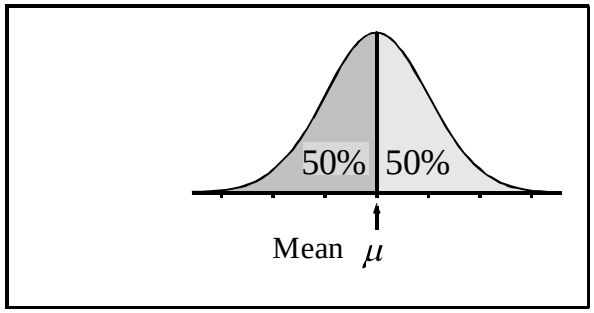


Alcune proprietà della distribuzione normale

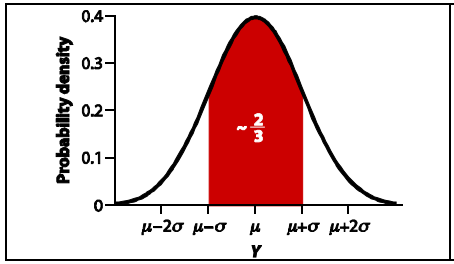
- E' definita da due parametri: la media μ (posizione) e la varianza σ^2 (dispersione e quindi forma)

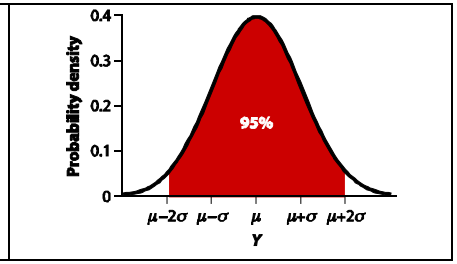


- Unimodale simmetrica, centrata sulla media (media, moda e mediana coincidono)



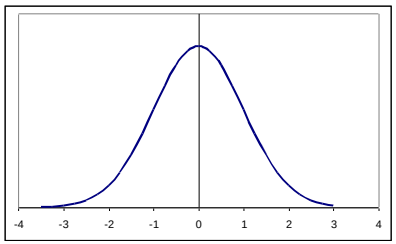
- L'area sottostante somma a 1 (come tutte le distribuzioni di probabilità)





- Area compresa tra la media e ± 1 deviazione standard = 68.3 %
- Area compresa tra la media e ± 2 deviazione standard = 95.4 %
- Area compresa tra la media e ± 3 deviazione standard = 99.7 %
- Area compresa tra la media e ± 1.96 deviazione standard = 95 %

La distribuzione normale standardizzata: è una distribuzione normale con media = 0 e varianza = 1



- Area compresa tra 0 e ± 1 = 68.3 %
- Area compresa tra 0 e ± 2 = 95.4 %
- Area compresa tra 0 e ± 3 = 99.7 %
- Area compresa tra la 0 e ± 1.96 = 95 %

La tavola statistica per la variabile Z (con Z indichiamo una variabile con dist. norm. st.)

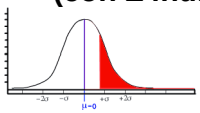


Tabella 10.4-1

Probabilità di osservare valori di Z > a,bc nella curva normale standardizzata. La cifra immediatamente prima della virgola decimale e quella immediatamente dopo (cioè, a,b) sono date nella prima colonna, e la seconda cifra dopo la virgola decimale (cioè, c) è data nella prima riga. Tabella estratta dalla Tavola Statistica B.

Prime due cifre di *a,bc*	Seconda cifra dopo la virgola decimale (*c*)									
	0	1	2	3	4	5	6	7	8	9
...										
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,0110

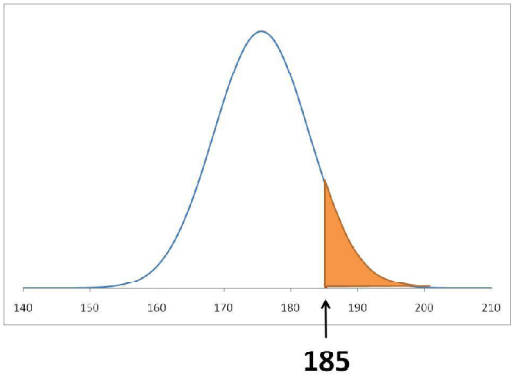
- Possiamo trovare questi numeri in Excel?
  - Funzione distrib.norm.st (attenzione: è cumulativa!)

- Ogni variabile Y con distribuzione normale diventa una variabile con distribuzione normale standardizzata Z se calcoliamo lo scarto standardizzato per ogni valore

$$Z = \frac{Y - \mu}{\sigma}$$

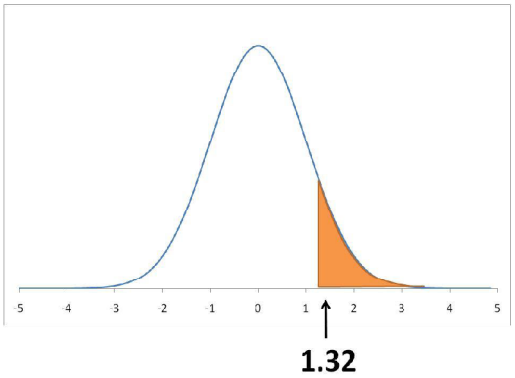
(Z è la distanza tra un valore e la media in unità di deviazioni standard)

- Quindi, la tabella di Z può essere usata per ogni variabile con distribuzione normale, dopo aver standardizzato il valore di interesse



Domanda: quale frazione di osservazioni saranno maggiori di 185 se la media è pari a 175.6 e la deviazione standard è pari a 7.1 (nella popolazione)?

$$Z = \frac{185 - 175.6}{7.1} = 1.32$$



ESEMPIO 10.4

La NASA esclude dai corsi per diventare astronauti chiunque sia più alto di 193.0 cm o più basso di 148.6 cm.
Negli uomini (popolazione USA), l'altezza media è 175.6 cm, con s = 7.1 cm.
Nelle donne (popolazione USA), l'altezza media è 162.6 cm, con s = 6.4.

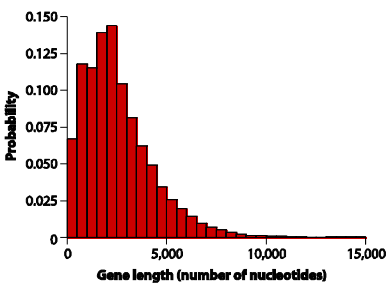
Calcolare le frazioni di popolazione, separatamente per maschi e femmine, esclusi dai programmi NASA. Discutere i risultati.

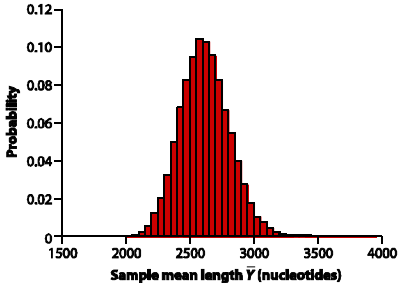
La distribuzione della media campionaria

[è una stima: ci interessano sempre
le distribuzioni campionarie delle stime!]

- Torniamo brevemente al capitolo 4: teoria del campionamento; distribuzione campionaria della media, ovvero della variabile $\overline{y}$ [nell'esempio, la "media campionaria della lunghezza genica in campioni con *n* = 100"]

- Usiamo un computer (avendo a disposizione la popolazione) avevamo capito diverse cose

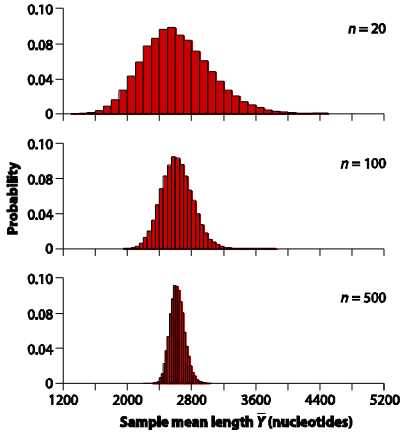




- Ha una forma a campana (diversa dalla distribuzione di y)

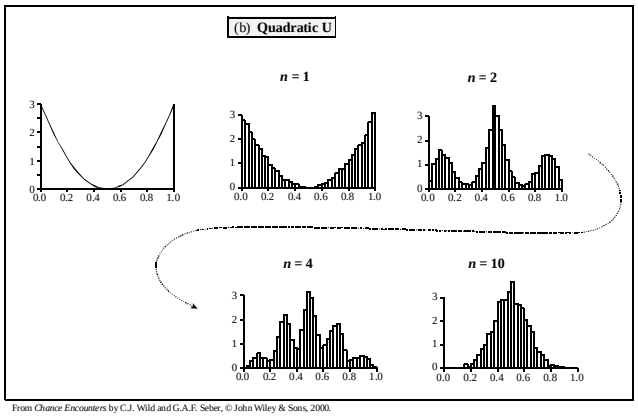
- E centrata sulla media della popolazione μ (quindi $\overline{y}$ è una stima corretta, ossia non distorta)

- Ha un'ampiezza inferiore all'ampiezza della distribuzione di y



Aumentando la dimensione del campione (*n*) si riduce la dispersione della distribuzione campionaria: l'errore di campionamento si riduce e aumenta la precisione della stima

E ancora una caratteristica della distribuzione di $\overline{y}$



Pensiamo anche per esempio alla media dei punteggi lanciando 2, oppure 5, oppure 10 dadi

[vedi anche esempio 10.6]

Figura 10.6-1
La distribuzione dei tempi di risposta [in millisecondi (ms)] per premere un pulsante dopo l'accensione di una luce.

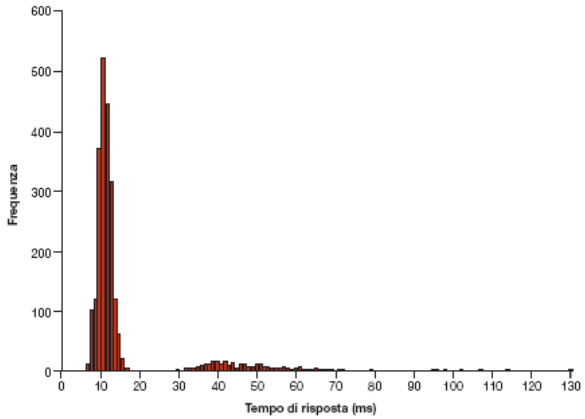
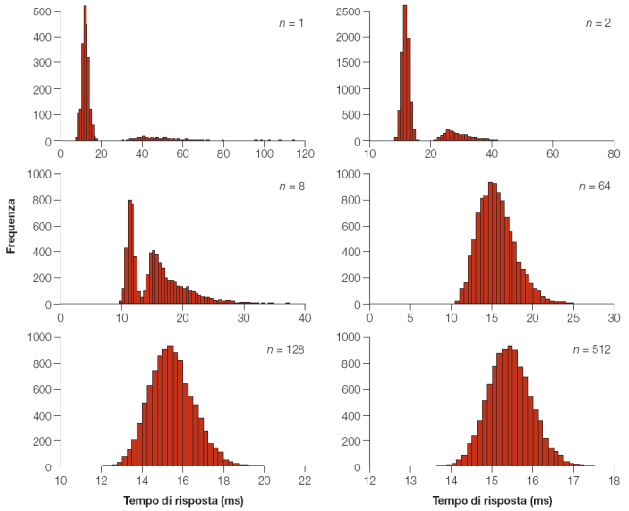


Figura 10.6-2
La distribuzione di frequenza delle medie campionarie in campioni di dimensione crescente. Ogni istogramma visualizza le medie di un gran numero di campioni ripetuti di dimensione *n* estratti dalla distribuzione dei tempi di risposta nella Figura 10.6-1. La scala e l'intervallo degli assi cambiano da diagramma a diagramma.

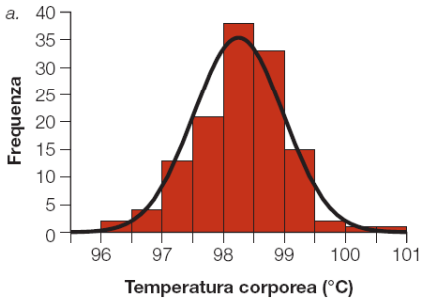


Possiamo ora essere più specifici riguardo la distribuzione della media campionaria

1. La distribuzione della media campionaria è normale se la variabile ha una distribuzione normale

2. Per il TLC (Teorema del Limite Centrale), la distribuzione della media campionaria (ma anche di una somma), è normale anche se la variabile non ha una distribuzione normale, basta che *n* sia sufficientemente grande

Queste distribuzioni hanno una motivazione!



Ogni variabile biologica dipende in realtà da molti fattori

Avevamo anche visto nel Capitolo 4…

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

[valida anche se Y non ha distribuzione normale]

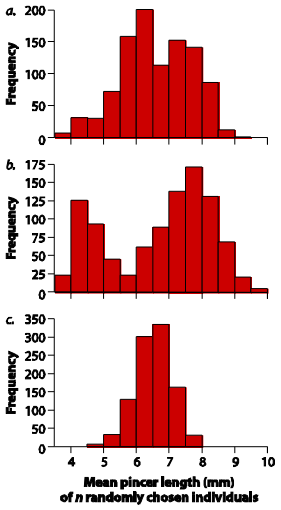
Quindi, posso applicare la normale standardizzata per capire alcune cose della media campionaria

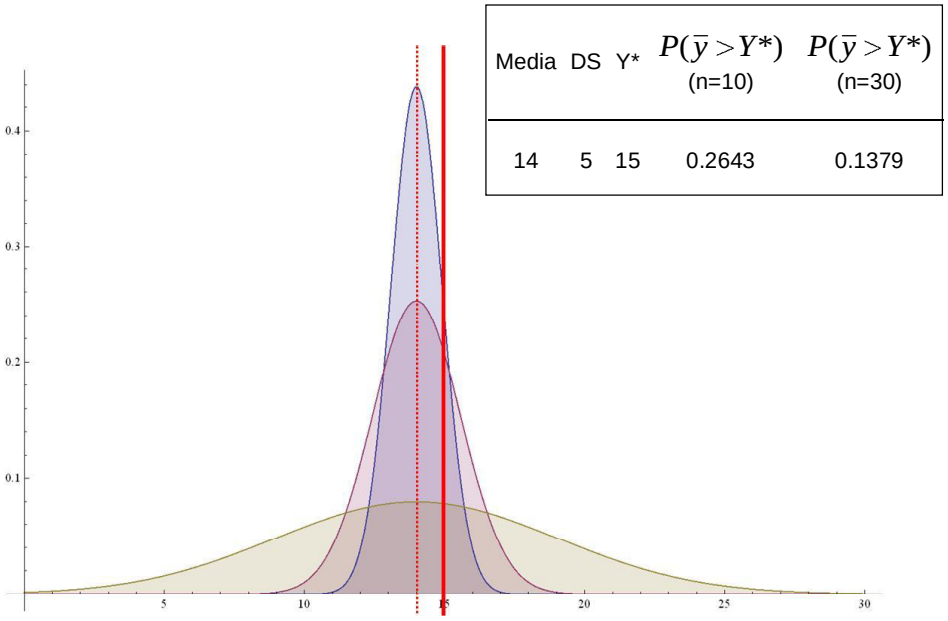
Esercizio 11, pg 155

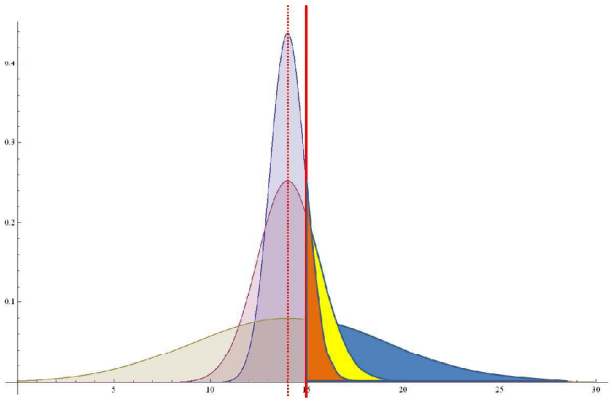
Media	DS	Y*	$P(\bar{y} > Y^*)$ (n=10)	$P(\bar{y} > Y^*)$ (n=30)
14	5	15	0.2643	0.1379
15	3	15.5		
-23	4	-22		
72	50	45		

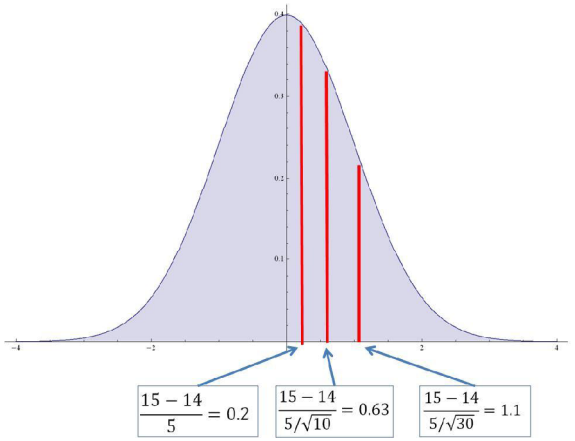
Attenzione: l'esercizio non implica il calcolo di frazioni di osservazioni, ma di frazioni di medie!

Esercizio 18, pg 155









Per tutto ciò che abbiamo detto, se $\bar{y}$ segue una distribuzione normale con media μ e deviazione standard $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, allora la variabile

$$Z = \frac{\bar{y} - \mu}{\sigma_{\bar{y}}}$$

segue una distribuzione normale standardizzata, e quindi, per esempio

$$P\left(-1.96 \le \frac{\bar{y} - \mu}{\sigma_{\bar{y}}} \le 1.96\right) = 0.95$$

$$P\left(-1.96 \le \frac{\bar{y} - \mu}{\sigma_{\bar{y}}} \le 1.96\right) = 0.95$$

Riarrangiando:

$$P\left(\left(\bar{y} - 1.96\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \left(\bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right)\right) = 0.95 = 95\%$$

- **Assumendo σ noto**, quindi, e dato un singolo valore di $\bar{y}$ in un campione, l'espressione qui sopra definisce un IC al 95% (e, sostituendo 1.96 con il valore appropriato in Tavola B, definisce un qualsiasi IC)

$$IC_{(1-\alpha)} => \bar{y} \pm z_{\alpha(2)} \cdot \sigma / \sqrt{n}$$

- In maniera simile, assumendo una certa H_0 che specifica un valore di $\mu = \mu_0$, la statistica test

$$Z = \frac{\bar{y} - \mu_0}{\sigma_{\bar{y}}}$$

permettere di testare H_0 (calcolo del P-value o approccio delle regioni di accettazione/rifiuto)

- **Ma σ non è quasi mai noto**, bisogna stimarlo

- Stimato σ con *s*, possiamo stimare l'errore standard con

$$ES_{\bar{y}} = s/\sqrt{n}$$

e a questo punto la variabile $\frac{\bar{y}-\mu}{ES_{\bar{y}}}$ non segue più una normale standardizzata (e l'intervallo di confidenza non è più calcolabile usando z)

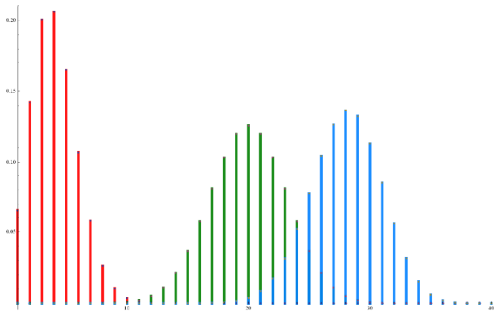
[Ecco perché la regoletta pratica, già vista nel Cap. 4

$$IC_{95\%} = \bar{y} \pm 2ES_{\bar{y}}$$

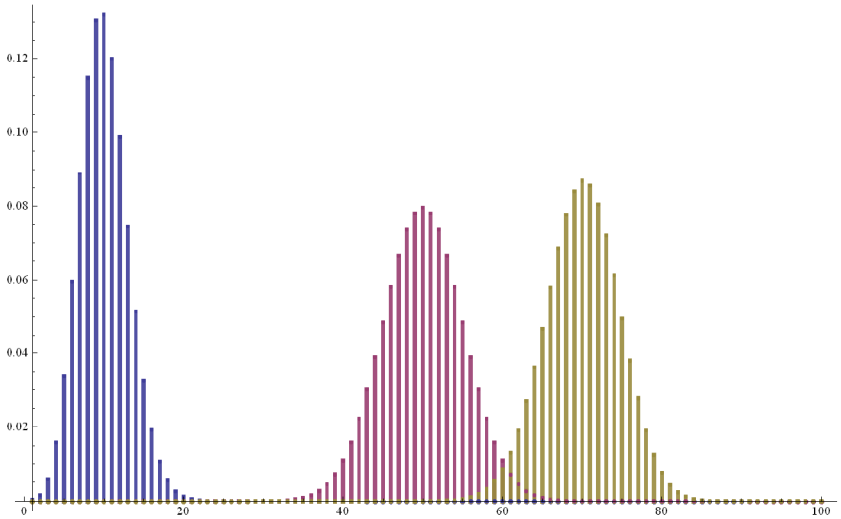
è solo una approssimazione]

Alcuni brevi note teoriche

1. La distribuzione normale approssima sufficientemente bene una distribuzione binomiale quando *np* e *n(1-p)* sono entrambi maggiori di 5



n = 40; p = 0.1; 0.5; 0.7



n = 100; p = 0.1; 0.5; 0.7

2. La distribuzione del Chi-quadrato ha simili assunzioni per le numerosità attese perché dipende dall'approssimazione normale della binomiale

Scheda 6: Controlli ed effetto placebo

- L'importanza del gruppo di controllo

- Spesso si migliora comunque, perché
  - da molte malattie si guarisce
  - i pazienti spesso vogliono "dare soddisfazione" al medico
  - i pazienti sono seguiti comunque,anche senza medicinali
  - effetto placebo ("piacerò")

- Il controllo deve necessariamente simulare in tutto e per tutto (tranne il farmaco o l'operazione) il trattamento: è un effetto reale in molti casi

- Il risultato eventuale deve essere "a parità di tutti gli altri fattori"

L'inferenza in una popolazione con distribuzione normale

$$Z = \frac{\bar{y} - \mu}{\sigma_{\bar{y}}}$$

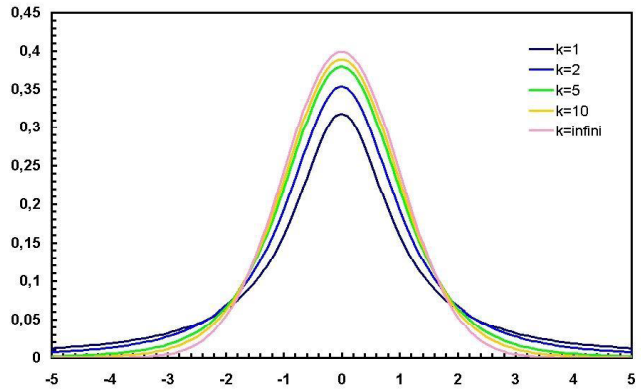
Questa variabile segue una distribuzione normale standardizzata

$$t = \frac{\bar{y} - \mu}{ES_{\bar{y}}}$$

Questa variabile (usata anche come statistica test) NON segue una distribuzione normale standardizzata,
ma, **se la variabile y è normale**, una distribuzione t con n-1 df

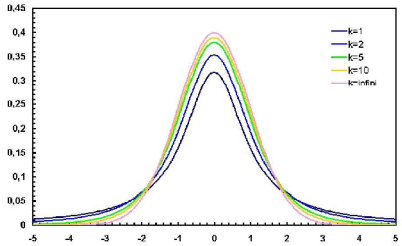
[importante ricordare bene i termini in queste espressioni]
[come cambia la forma della distribuzione,da Z a t? Perché?]

La distribuzione t



Caratteristiche principali

- Varia tra – infinito e + infinito
- Ha un parametro, i gradi di libertà (la normale standardizzata non ha parametri)
  - Per campioni di dimensioni diverse esistono quindi distribuzioni t diverse
- Media, moda, e mediana sono uguali
- Ha media pari a 0 e varianza maggiore di 1.
  - Se k è grande, la varianza tende a 1
- Rispetto alla normale standardizzata, ha code più *pesanti*
  - Maggiore concentrazioni di valori agli estremi, a causa della maggiore varianza rispetto alla normale standardizzata, dovuta all'errore nella stima di σ
- Diventa una distribuzione normale standardizzata quando i gradi di libertà (e quindi la numerosità del campione) tendono a infinito.



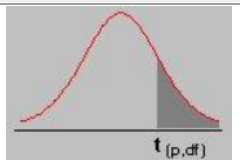
William Sealy Gosset (1876 - 1937) — Distribuzione t di Student

Ernst Karl Abbe (1840 - 1905) — Distribuzione del chi-quadrato

Carl Friedrich Gauss (1777 - 1855) — Distribuzione normale (Gaussiana)

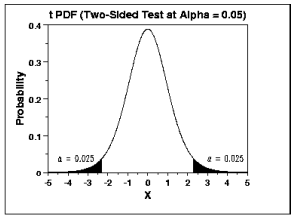
Jakob Bernoulli (1654 - 1705) — Distribuzione binomiale (esperimento bernoulliano)

Tavola della distribuzione t *ad una coda*



df	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

Tavola della distribuzione t *a due code*



Degrees of Freedom- 2 tails	0.01	0.02	0.05	0.10	0.20
1	63.66	31.82	12.71	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
10	3.169	2.764	2.228	1.812	1.372
15	2.947	2.602	2.132	1.753	1.341
25	2.787	2.485	2.060	1.708	1.316
∞	2.575	2.326	1.960	1.645	1.282

Tavola C, pg. 419-421
(unica tavola con valori critici a una e a due code)

	α(2) = 0.20	α(2) = 0.10	α(2) =0.05	α(2) =0.02	α(2) =0.01
df	α(1) = 0.10	α(1) = 0.05	α(1) = 0.025	α(1) = 0.01	α(1) = 0.005
1	3.077684	6.313752	12.70620	31.82052	63.65674
2	1.885618	2.919986	4.30265	6.96456	9.92484
3	1.637744	2.353363	3.18245	4.54070	5.84091
4	1.533206	2.131847	2.77645	3.74695	4.60409
5	1.475884	2.015048	2.57058	3.36493	4.03214
11	1.363430	1.795885	2.20099	2.71808	3.10581
12	1.356217	1.782288	2.17881	2.68100	3.05454
13	1.350171	1.770933	2.16037	2.65031	3.01228
inf	1.281552	1.644854	1.95996	2.32635	2.57583

Intervallo di confidenza della media

$$t = \frac{\bar{y} - \mu}{ES_{\bar{y}}}$$

$$P\left(-t_{\alpha(2),df} \le \frac{\bar{y} - \mu}{ES_{\bar{y}}} \le t_{\alpha(2),df}\right) = 1 - \alpha$$

$$P\left(\bar{y} - t_{\alpha(2),n-1} \cdot s/\sqrt{n} \le \mu \le \bar{y} + t_{\alpha(2),n-1} \cdot s/\sqrt{n}\right) = 1 - \alpha$$

$$IC_{(1-\alpha)} => \bar{y} \pm t_{\alpha(2),n-1} \cdot s/\sqrt{n}$$

Da occhio a occhio

Cyrtodiopsis dalmanni è un insetto dittero dall'aspetto bizzarro che vive nella giungla della Malesia. I suoi occhi si trovano alle estremità di lunghi peduncoli che sporgono dalla testa, e ci ricordano le creature della saga cinematografica *Guerre stellari*. Questi peduncoli oculari sono caratteristici di entrambi i sessi, ma sono particolarmente sviluppati nei maschi. L'ampiezza dei peduncoli oculari nei maschi aumenta la loro attrazione esercitata sulle femmine e il loro successo nei combattimenti contro altri maschi.

La distanza interoculare (tra occhio e occhio), espressa in millimetri, è stata misurata in un campione casuale di 9 maschi di *C. dalmanni*.[2] I dati sono i seguenti:

8,69 8,15 9,25 9,45 8,96 8,65 8,43 8,79 8,63

Possiamo usare questi valori per stimare la distanza interoculare media nella popolazione di *C. dalmanni* e quantificare l'incertezza della stima usando un intervallo di confidenza al 95%. Assumiamo che questa variabile abbia una distribuzione normale nella popolazione. ■

Il test *t* per un campione

La variabile $t = \frac{\bar{y}-\mu}{ES_{\bar{y}}}$ segue una distribuzione t con n-1 df

Quindi, assumendo una certa H_0 che specifica un valore di $\mu = \mu_0$, la statistica test

$$t = \frac{\bar{y} - \mu_0}{ES_{\bar{y}}}$$

ha una distribuzione nulla nota (appunto la distribuzione t con n-1 df)

Questo mi permette di fare un test su una media, usando l'approccio delle regioni di accettazione/rifiuto (basta la tavola C) oppure calcolando il P-value (ci vuole un computer).

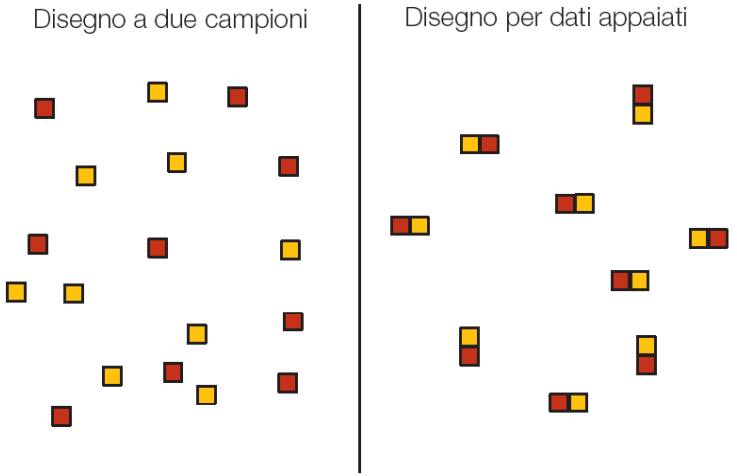
Esercizio 13, pag. 169

Assunzioni da verificare quando si utilizza la distribuzione t

- Campione casuale (come sempre)
- Variabile deve avere una distribuzione normale

➢ Il test è però abbastanza robusto alla non-normalità (per il TLC!)

Il confronto tra due medie



- **Disegno a due campioni indipendenti**
  - Tutte le unità campionarie sono indipendenti
  - Esempi:
    - Confronto l'altezza media in due popolazioni
    - Confronto la crescita media tra piante allevate in due condizioni diverse
    - Confronto la biodiversità in campioni provenienti da due habitat diversi

- **Disegno per dati appaiati**
  - Ad ogni unità campionaria si applicano entrambi i trattamenti
  - Esempi:
    - Confronto la biodiversità media trattando ogni plot in due modi diversi
    - Confronto l'ematocrito medio misurandolo in atleti prima e dopo una gara
    - Confronto il livello di ozono medio in due anni diversi misurandolo (nei due anni) nelle stesse località

Il confronto per dati appaiati può essere necessario, e spesso è conveniente

- Se i dati sono stati raccolti in modo appaiato, le osservazioni nelle coppie di valori non sono indipendenti e non posso quindi fare un confronto per due campioni indipendenti
- Se molti fattori sono responsabili dei valori che assume la variabile alla quale siamo interessati, e non solo il fattore che stiamo studiando, il test per dati appaiati permette di controllarli meglio e ridurre il rischio che mascherino l'effetto del fattore di interesse

Confronto per dati appaiati

- Calcolo inizialmente le differenze

Località	Concentrazione 1. Anno	Concentrazione 2. Anno	Differenza
Milano	400	345	55
Tokio	20	8	12
Berlino	24	29	-5
Roma	95	81	14
Parigi	228	204	24
Ferrara	116	140	-24
Bologna	65	36	29
Londra	112	75	37
Stoccolma	35	47	-12
Mosca	45	5	40
Palermo	81	65	16
New York	197	187	10
MEDIA			16.33
DEV ST			22.66

Da 2*n* osservazioni non indipendenti si è passati a *n* osservazioni indipendenti, ciascuna delle quali influenzata dal fattore di interesse e meno da altri fattori legati al fatto che le città sono diverse in aree diverse

- *d* è la nuova variabile "differenza tra coppie di osservazioni", $\bar{d}$ la media nel campione, e μ_d il corrispondente parametro

- Se d ha una distribuzione normale, posso calcolare il suo IC e svolgere un test t per un campione come già visto
  - Le ipotesi di partenza (sulle medie nei due gruppi)

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

diventano logicamente

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

- Calcolo IC

$$IC_{(1-\alpha)} = \bar{d} \pm t_{\alpha(2),df} ES_{\bar{d}}$$

$$IC_{(95\%)} = 16.33 \pm 2.20 \times \frac{22.66}{\sqrt{12}}$$

$1.94 \leq \mu_d \leq 30.79$ [e quindi?]

- Svolgo il test

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}} = \frac{\bar{d}}{s_d / \sqrt{n}} = 2.5$$

[e quindi?]

Esempio 12.2, pg 173 (dati appaiati dal ricercatore in fase sperimentale)

Esercizio 19, pg 190 (dati raccolti in natura)

la Repubblica.it
In diretta con l'Italia

Gli intention Poll Digis per ciascuna città Campione sono stati rilevati con 2.400 telefonate secondo il metodo di sistema di rilevazione Cati con margine di errore +/- 2% e livello di confidenza dei contatti del 95% (16.5.2011)

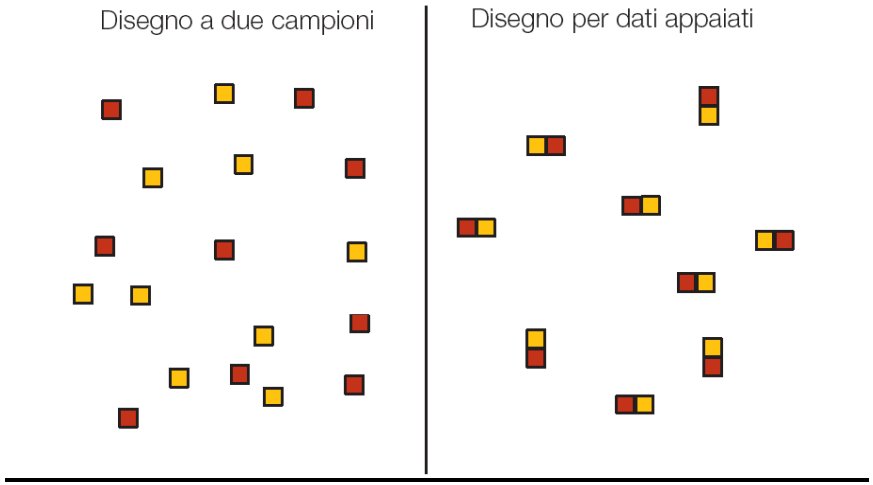
$$IC_{95\%}[proporzione] = \hat{p} \pm 1.96 \cdot ES_{\hat{p}}$$

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

$$ES_{\hat{p}} = \sqrt{\frac{0.25}{2399}} = 0.0102$$

$$IC_{95\%}[proporzione] = \hat{p} \pm 1.96 \cdot 0.0102 = \hat{p} \pm 0.0200$$

Il confronto tra le medie in DUE campioni <u>indipendenti</u>



Esempio 12.3 *Infilzare o essere infilzati*

Il frinosoma *Phrynosoma mcalli*, un rettile iguanide, ha molte caratteristiche insolite, a partire dalla capacità di spruzzare sangue dagli occhi. Come in tutti i frinosomi, la sua testa è circondata da una corona di aculei simili a corna (da cui il nome volgare inglese, *horned lizard*). Gli erpetologi hanno recentemente sottoposto a verifica statistica l'ipotesi che i lunghi aculei contribuiscano a proteggere questi animali dalla predazione. Per fare ciò, hanno sfruttato il comportamento brutale ma efficace di uno dei suoi principali predatori, l'averla stolida (*Lanius ludovicianus*), un piccolo uccello predatore che infilza le sue vittime sulle spine di piante o sulle punte del filo spinato e le divora in un secondo momento.

Alcuni ricercatori hanno ritrovato i resti di 30 frinosomi che erano stati uccisi da averle e hanno misurato la lunghezza delle loro «corna» (Young et al., 2004). Come gruppo di confronto, i ricercatori hanno misurato lo stesso carattere in 154 frinosomi che godevano ancora di ottima salute. Poi hanno confrontato le lunghezze medie delle corna dei frinosomi morti con quelle degli animali vivi. Gli istogrammi delle misure sono rappresentati nella Figura 12.3-1. Le statistiche descrittive sono mostrate nella Tabella 12.3-1. ■

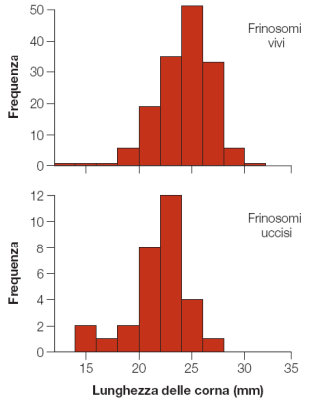


Figura 12.3-1
La distribuzione di frequenza della lunghezza delle corna di frinosomi vivi e di frinosomi uccisi. $n_1 = 154$ frinosomi vivi; $n_2 = 30$ frinosomi uccisi.

Tabella 12.3-1
Statistiche descrittive per le lunghezze delle «corna» nei frinosomi.

Gruppo di frinosomi	Media campionaria $\bar{Y}$ (mm)	Deviazione standard campionaria s (mm)	Dimensione campionaria *n*
frinosomi vivi	24,28	2,63	154
frinosomi uccisi	21,99	2,71	30

- Abbiamo quindi n_1 e n_2 osservazioni rilevate in campioni estratti da due popolazioni con medie μ_1 e μ_2 ignote

- La domanda alla quale siamo interessati è se le medie nelle due popolazioni sono diverse, e di quanto, e come sempre abbiamo solo dei campioni disponibili
  [non è diverso da chiedersi se esista o no, e quanto forte sia, l'associazione tra una variabile numerica e una variabile categorica]

- Per rispondere a questa domanda ci sono d'aiuto, come in altri casi già visti, l'IC e il test di ipotesi

- L'IC che ci interessa è l'intervallo di confidenza della differenza tra le due medie

- La differenza tra medie campionarie segue una distribuzione normale (come la media), e quindi la differenza standardizzata per l'ES stimato

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{ES_{\bar{Y}_1 - \bar{Y}_2}}$$

segue una distribuzione t con df = n_1+n_2-2 (la somma dei df nei due campioni), quando la variabile segue una distribuzione normale, con identica varianza, in entrambe le popolazioni

- Il test di ipotesi che ci interessa è

$H_0 : \mu_1 = \mu_2$ Le medie μ_1 e μ_2 sono uguali
$H_1 : \mu_1 \neq \mu_2$ Le medie μ_1 e μ_2 sono diverse
[se il test è bilaterale; altrimenti?]

- Seguendo gli stessi ragionamenti già visti per altri test, la statistica test t calcolata come

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2)}{ES_{\bar{Y}_1 - \bar{Y}_2}}$$

segue la distribuzione teorica t con df = n_1+n_2-2 se è vera
[$\mu_1 - \mu_2 = 0$ se è vera H_0]

Quindi:

$$-t_{\alpha(2),df} < \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{ES_{\bar{Y}_1 - \bar{Y}_2}} < t_{\alpha(2),df}$$

Riarrangiando

$$(\bar{Y}_1 - \bar{Y}_2) - t_{\alpha(2),df} ES_{\bar{Y}_1 - \bar{Y}_2} < (\mu_1 - \mu_2) < (\bar{Y}_1 - \bar{Y}_2) + t_{\alpha(2),df} ES_{\bar{Y}_1 - \bar{Y}_2}$$

Ovvero

$$IC_{1-\alpha}(\mu_1 - \mu_2) \rightarrow (\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha(2),df} ES_{\bar{Y}_1 - \bar{Y}_2}$$

Dove

$$ES_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}; \quad s^2{}_p = \frac{(n_1 - 1)s^2{}_1 + (n_2 - 1)s^2{}_2}{n_1 + n_2 - 2}$$

- Di conseguenza, posso verificare l'ipotesi nulla (con il *Pvalue* o l'approccio delle regioni di accettazione e rifiuto) calcolando proprio la statistica test

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2)}{ES_{\bar{Y}_1 - \bar{Y}_2}}$$

Torniamo ai frinosomi

Tabella 12.3-1

Statistiche descrittive per le lunghezze delle «corna» nei frinosomi.

Gruppo di frinosomi	Media campionaria $\bar{Y}$ (mm)	Deviazione standard campionaria s (mm)	Dimensione campionaria *n*
frinosomi vivi	24,28	2,63	154
frinosomi uccisi	21,99	2,71	30

Esempio

In due siti archeologici che si riferiscono a due diverse tribù di Indiani d'America, Apache e Cheyenne, vengono rinvenute delle punte di freccia, 8 nel primo sito e 7 nel secondo. Si vuole determinare se le due tribù utilizzassero frecce di dimensioni diverse. Assumiamo che le condizioni per poter applicare questo test (varianze uguali nelle due popolazioni, distribuzioni gaussiane della variabile nelle due popolazioni) siano soddisfatte

Guerrieri Apache

Indiano Cheyenne

DATI (lunghezze frecce in cm)
Apache (Tribù 1): 4.5; 5.2; 4.3; 4.7; 4.0; 3.9; 5.8; 2.8
Cheyenne (Tribù 2): 5.2; 5.7; 6.0; 6.7; 5.5; 5.4; 6.8

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

A partire dai dati calcolo:

$\overline{Y}_1 = 4.4$ $\overline{Y}_2 = 5.9$ $s^2_1 = 0.81$ $s^2_2 = 0.40$

La varianza comune è stimata con

$$s_p^2 = \frac{(8-1)\times 0.81-(7-1)\times 0.40}{8+7-2} = 0.62$$

Quindi

$$ES_{\overline{Y}_1-\overline{Y}_2} = \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$ES_{\overline{Y}_1-\overline{Y}_2} = \sqrt{0.62\left(\frac{1}{8}+\frac{1}{7}\right)} = 0.41$$

- Testo H_0

$$t = \frac{(\overline{Y}_1 - \overline{Y}_2)}{ES_{\overline{Y}_1-\overline{Y}_2}}$$

$$t = \frac{-1.5}{0.41} = -3.66$$

[e quindi?]

Pvalue=0.0029

- Per completezza, è interessante anche calcolare quanto più grandi erano le frecce dei Cheyenne (uso IC)

$$IC_{(1-\alpha)}(\mu_1 - \mu_2) \longrightarrow (-1.5) \pm t_{\alpha(2),df} \times 0.41$$

$$IC_{(1-0.01)}(\mu_1 - \mu_2) \longrightarrow (-1.5) \pm 3.01 \times 0.41$$

$$-2.73 < (\mu_1 - \mu_2) < -0.27$$

[e quindi?]

Alcuni punti importanti

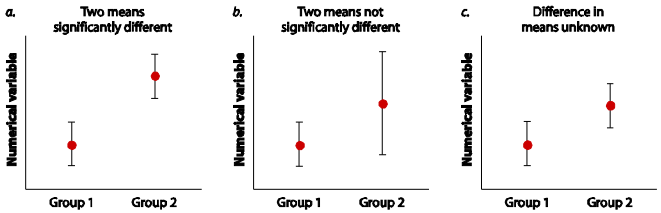
- Ricordarsi le assunzioni
  - Casualità, normalità, uguaglianza delle varianze
  - Robusto per deviazioni non eccessive dalla normalità
  - Robusto se varianze non troppo diverse (3x se disegno è bilanciato)

- Esistono test per confrontare le varianze: test F e test di Levene

(Howard Levene) 

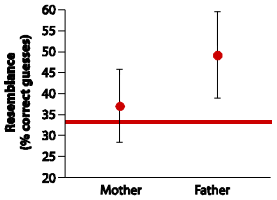
- Esistono alternative al test t per due campioni quando le varianze sono diverse. Uno di questi è il test di Welch

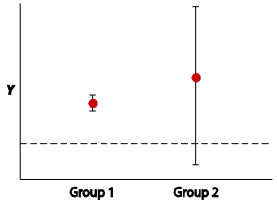
- Attenzione alla sovrapposizione tra IC di diverse medie



- Attenzione a definire le unità campionarie corrette (vedi esempio 12.4)

- Attenzione all'errore del confronto diretto





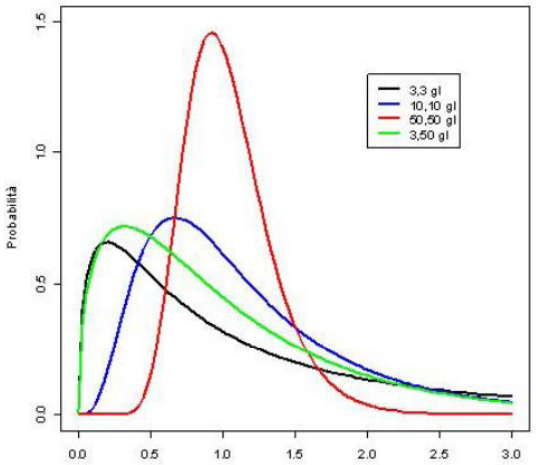
Confrontare due varianze

- Perché? Almeno tre motivi
  - Studiare se due popolazioni differiscono per i livelli di dispersione
  - Verificare se è possibile applicare il test t per due campioni
  - Fare un test sulle medie di più gruppi (ANOVA, da fare)

- Le ipotesi nulla e alternativa possono essere formalizzate come segue

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

- Come abbiamo sempre fatto in tutti i test statistici, dobbiamo trovare una statistica test la cui distribuzione teorica è nota quando è vera l'ipotesi nulla



- Nel caso di due varianze, la statistica test è data da F, il rapporto tra due varianze

$$F = \frac{s_1^2}{s_2^2}$$

- Se è vera l'ipotesi nulla che le due varianze nelle popolazioni sono uguali, e se la variabile segue una distribuzione normale in entrambe le popolazioni (**assunzione importante!**), il rapporto tra due varianze campionarie segue la distribuzione nulla di Fisher, detta anche distribuzione F (o F di Fisher)

Ronald Fisher (1890-1962)
Statistico, biologo evoluzionista, genetista
(Secondo alcuni, il più grande erede di Darwin)

- La distribuzione teorica F:
  - E' continua
  - Varia tra zero e infinito
  - Dipende dai gradi di libertà del numeratore ($gdl_1 = n_1-1$) e quelli del denominatore ($gdl_2 = n_2-1$)
  - E' circa centrata sul valore 1
  - Ci permette di definire le regioni di accettazione/rifiuto o il P-value per i test sulle varianze

Tabella della distribuzione F a una coda con α = 0.01

F - Distribution ($\alpha = 0.01$ in the Right Tail)

df_2 \ df_1	1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195
29	7.5977	5.4204	4.5378	4.0449	3.7254	3.4995	3.3303	3.1982	3.0920
30	7.5625	5.3903	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876
60	7.0771	4.9774	4.1259	3.6490	3.3389	3.1187	2.9530	2.8233	2.7185
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.7918	2.6629	2.5586
∞	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073

Denominator Degrees of Freedom

Numerator Degrees of Freedom

Attenzione! La struttura di questa tabella è diversa da tutte quelle viste finora
(ci sono due gradi di libertà da conoscere in ogni analisi,
e c'e' una tabella per ogni valore di P)

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_0 : \sigma_1^2 > \sigma_2^2$$

o, secondo alcuni testi,

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_0 : \sigma_1^2 > \sigma_2^2$$

$$F = \frac{s_1^2}{s_2^2} = \frac{3.6}{0.9} = 4.00$$

(se le femmine nello studio fossero state meno variabili, non avrei avuto bisogno del test)

F critico a una coda con α = 0.05 è pari a 3.18

[e quindi?]

Esercizio

Si misura la sopravvivenza (in giorni) in 10 maschi e 10 femmine di insetto stecco senza cibo e acqua.

	$\bar{Y}$	s^2
Femmine	8.5	3.6
Maschi	4.8	0.9

Considerando che l'ipotesi che le femmine siano meno variabili dei maschi possa essere esclusa a priori (e quindi una qualsiasi deviazione in tale direzione si dovrebbe considerare compatibile con l'ipotesi nulla), verificare l'ipotesi alternativa che le femmine siano più variabili dei maschi

Pvalue = P(F>4.00) = 0.0255

Se l'ipotesi alternativa fosse stata bilaterale, avrei avuto bisogno delle tavole per α = 0.025.

F critico a una coda con α = 0.025 (utili per testare un'ipotesi bilaterale con α=0.05) è pari a 4.03

[e quindi?]

Pvalue (due code) = 2x P(F>4.00) = 0.051

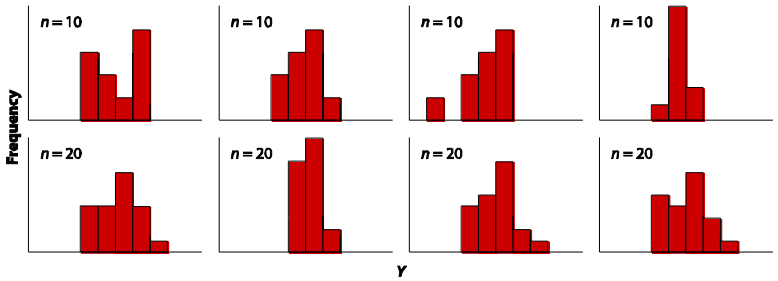
Scheda 7: Sintesi di molti test e guida pratica

Che tipo di dati ho di fronte?
Qual è il mio obiettivo?
➔ Quale test devo usare?

Tab1: singola variabile, singolo gruppi
Tab2: test di associazione
Tab3: confronto tra medie (una variabile, più gruppi)

Metodi grafici

Attenzione, se il campione è piccolo, non ci aspettiamo mai una distribuzione a campana anche se nella popolazione la distribuzione è normale



Cosa fare quando le assunzioni vengono violate? (normalità e uguaglianza di varianze)

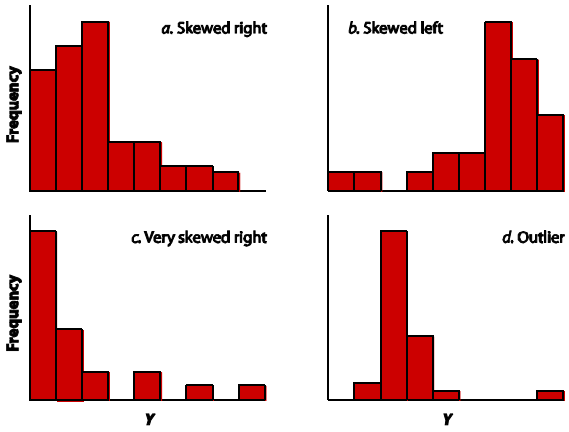
1. Ignorare le assunzioni (attenzione!)

2. Trasformare i dati

3. Utilizzare test non parametrici

Prima di tutto, come identificare deviazioni dalla normalità?

Se la numerosità inizia a diventare ragionevole, possiamo prevedere che distribuzioni di questo tipo non provengano da popolazioni normali



[il diagramma dei quantili normali può aiutare]

Test statistici per identificare la non-normalità

Chi Quadrato, Shapiro-Wilk

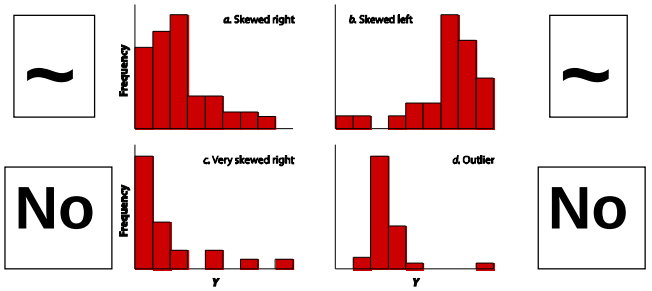
Ma:

- Se la dimensione del campione è piccola, la normalità non verrà quasi mai rifiutata
- Se la dimensione campionaria è grande, forse si rifiuterà la normalità anche quando questa deviazione non è cruciale (CLT)
- La semplice osservazione iniziale delle distribuzioni è quindi molto utile

Vediamo il test del Chi quadrato per identificare la non normalità

Ignorare le assunzioni?

- Per il TLC, nel caso si testino medie, deviazioni non eccessive dalla normalità (nella stessa direzione se si confrontano due gruppi) sono tollerabili



[c'e' sempre un po' di soggettività…]

Deviazioni dall'assunzione di uguaglianza di varianze

- Quando test t si può usare anche se se varianze non sono uguali?
- Quando si deve usare il test di Welch?
- Quando si devono usare test diversi (non parametrici)?

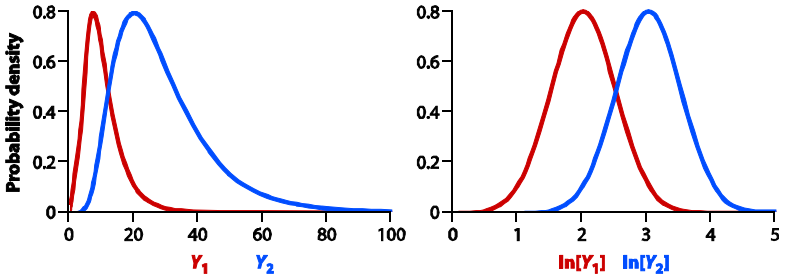
Trasformazione dei dati

- La trasformazione dei dati può essere applicata prima di svolgere un test statistico in modo da migliorare l'adattamento dei dati a una normale o da ridurre le differenze tra varianze
- Tutti i valori devono essere trasformati nello stesso modo
- Le trasformazioni più utilizzate sono quella logaritmica, arcoseno, e radice quadrata
- Il test statistico è semplicemente applicato e interpretato sui dati trasformati. Gli intervalli di confidenza devono essere retro-trasformati

- **Trasformazione logaritmica**: logaritmo naturale di ogni valore

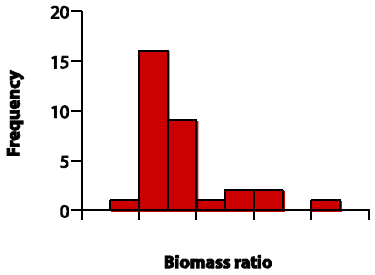
$$Y' = \ln[Y]$$

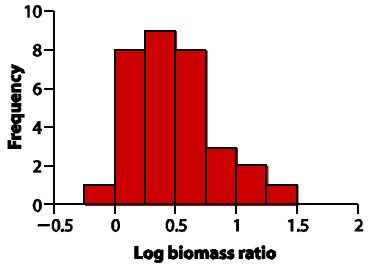
[se alcuni valori sono 0, $Y' = \ln[Y + 1]$]



- In genere utile quando
  - Le misure sono rapporti o prodotti
  - L'asimmetria dei dati originali è a destra
  - I dati coprono diversi ordini di grandezza
  - Nel confronto tra due gruppi, il gruppo con la maggiore media ha anche la maggiore deviazione standard

Esempio 13.1





[attenzione al calcolo degli IC!]

Trasformazione arcoseno: arcoseno della radice quadrata di ciascun valore

$$p' = \sin^{-1}\sqrt{p}$$

- Utile soprattutto per proporzioni

Trasformazione radice quadrata: radice di ciascun numero

$$Y' = \sqrt{Y + 1/2}$$

- Soprattutto per numerosità
- Simile alla trasformazione logaritmica

Altre trasformazioni

$Y' = \mathrm{Y}^2$ oppure $Y' = \mathrm{e}^{\mathrm{Y}}$ [se asimmetria è a sinistra]

$Y' = \frac{1}{Y}$ [se asimmetria è a destra]

Importante aspetto che riguarda l'"onestà statistica"

E' possibile trasformare i dati i molti modi, ma l'obiettivo non deve essere quello di raggiungere la significatività statistica!

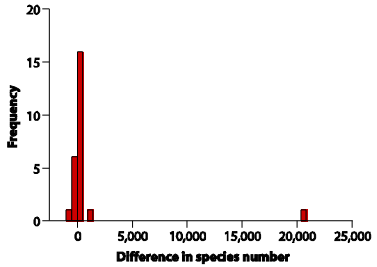
Alternative non-parametriche

- I test non parametrici (chiamati anche *distribution-free*) fanno meno assunzioni riguardo la distribuzione di probabilità della variabile (nella popolazione)
  - [Il test t è parametrico]
  - [Il test del chi-quadrato è non parametrico]
- Devono essere utilizzati quando la distribuzione della variabile studiata non è normale (e quando non si può ignorare questa violazione e le trasformazioni non funzionano)
  - Sono meno potenti dei test parametrici!
  - Quindi, se l'assunzione di normalità è verificata conviene usare i test parametrici
- I test non parametrici sono generalmente basati sui ranghi (ranks)
- I dati devo prima di tutto essere ordinati, e ad ogni osservazione viene assegnato un rango
- I test si basano quindi sui ranghi
- I ranghi fanno sì che tutte le distanze tra valori diventino uguali (a parte i casi di pareggio); la distribuzione dei ranghi è costante (non dipende dalla distribuzione dei dati); gli outliers non sono più outliers
- Attenzione ai pareggi (osservazioni con lo stesso valore)

Il test dei segno (alternative al test t per un campione)

- Il test sulla mediana
- Non è necessario identificare i ranghi, ma solo se ciascun valore è più grande o più piccolo della mediana ipotizzata da H_0
- Si svolge tecnicamente come un test binomiale (o chi quadro) sul numero di osservazioni maggiori della mediana specificata dall'ipotesi nulla (che se è vera l'ipotesi nulla devono essere il 50%)
- E' molto semplice e a volte l'unica possibilità
- Se n=5 (o minore), non si potrà mai rifiutare l'ipotesi nulla in un test dei segni a due code

[perché?]



La distribuzione delle differenze tra numero di specie in gruppi in cui le femmine si accoppiano più volte e il numero di specie in gruppi in cui le femmine si accoppiano una sola volta.
Normale?

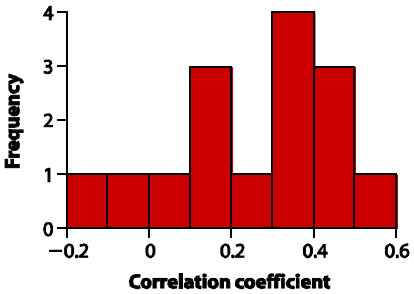
Esempio 13.4, pagina 203 (attenzione colonne in tabella)

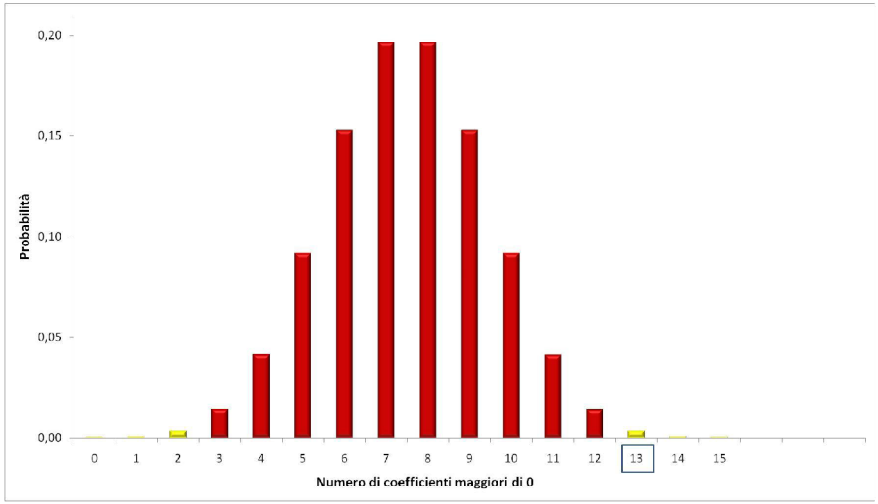
Table 1. Paired phylogenetic contrasts

Order	Polyandrous family	Polyandrous clade	No. of species	Monandrous family	Monandrous clade	No. of species
Coleoptera	Anobiidae	*Ernobius* spp.	53	Anobiidae	*Xestobium* spp.	10
	Dermestidae	*Dermestes* spp.	73	Dermestidae	*Trogoderma* spp.	120
	Elateridae	*Agriotes* spp.	228	Elateridae	*Selatosomus* spp.	74
Diptera	Muscidae	*Coenosia* spp.	353	Anthomyiidae	*Delia* spp.	289
	Cecidomyiidae	*Rhopalomyia* spp.	157	Cecidomyiidae	*Mayetiola* spp.	30
	Chironomidae	*Chironomus* spp.	>300	Chironomidae	*Pontomyia* spp.	4
	Chironomidae	*Stictochironomus* spp.	34	Chironomidae	*Clunio* spp.	18
	Drosophilidae	Total for family	3,400	Culicidae	Total for family	3,500
	Dryomyzidae	Total for family	20	Calliphoridae	Total for family	>1,000
	Tephritidae	*Anastrepha* spp.	196	Tephritidae	*Bactrocera* spp.	486
	Sciaridae	Total for family	1,750	Bibionidae	Total for family	660
	Scatophagidae	*Scatophaga* spp.	55	Muscidae	*Musca* spp.	63
Ephemeroptera	Siphlonuridae	*Siphlonurus* spp.	37	Caenidae	*Caenis* spp.	115
Homoptera	Psyllidae	*Cacopsylla* spp.	>100	Diaspididae	*Aonidiella* spp.	30
Lepidoptera	Noctuidae	Total for family	21,000	Psychidae	Total for family	600
	Tortricidae	*Choristoneura* spp.	37	Tortricidae	*Epiphyas* spp.	40
	Nymphalidae	*Eueides* spp. (aliphera clade)	7	Nymphalidae	*Eueides* spp. (vibilia clade)	5
	Nymphalidae	*Heliconius* spp. (silvaniform clade)	15	Nymphalidae	*Heliconius* spp. (sara/sapho clade)	7
	Nymphalidae	*Polygonia*/*Kaniska*/*Roddia* spp.	18	Nymphalidae	*Nymphalis* spp.	6
	Nymphalidae	*Acraea* spp.	240	Nymphalidae	*Cethosia* spp.	13
	Pieridae	*Dixeia* spp.	15	Pieridae	*Ascia* spp.	14
	Pieridae	*Colias*/*Zerene* spp.	77	Pieridae	*Phoebis* spp.	16
	Pieridae	*Euchloe* spp.	15	Pieridae	*Anthocaris* spp.	14
	Pieridae	*Eurema* spp.	85	Pieridae	*Gonepteryx* spp.	6
	Pieridae	*Dismorphia* spp.	86	Pieridae	*Leptidea* spp.	8

Problema 22, pagina 217

-0,16	-
-0,04	-
0,034	+
0,014	+
0,137	+
0,118	+
0,395	+
0,363	+
0,35	+
0,376	+
0,253	+
0,44	+
0,453	+
0,46	+
0,563	+





$$Pvalue = 2 \times [P(X = 13) + P(X = 14) + P(X = 15)]$$

Un test non parametrico per confrontare due gruppi: il test U di Mann-Whitney

- Verifica se la distribuzione di frequenza dei dati in due popolazioni è la stessa (H0)
- Se assumiamo in partenza che *la forma* (varianza e asimmetria) delle due distribuzioni sia la stessa, allora diventa un test sulla posizione

Come funzione il test U di Mann-Whitney:

1. Unire i dati dei due gruppi (mantenendo l'etichetta), ordinarli, e assegnare i ranghi

2. Calcolare la somma R_1 dei ranghi nel gruppo con n più piccolo

3. Calcolare U_1 come

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

4. Calcolare U_2 as

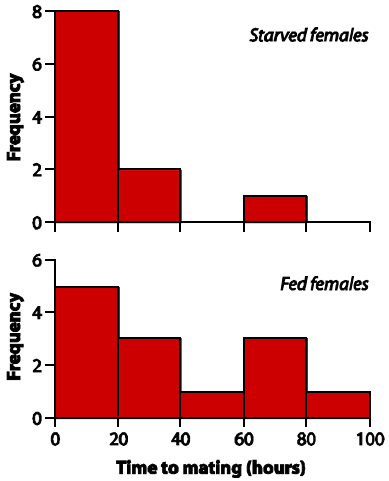
$$U_2 = n_1 n_2 - U_1$$

5. Usare come statistica test il più grande tra U_1 or U_2

6. Verificare se è possibile rifiutare H_0 in Tabella (o calcolo P-value con un programma al calcolatore)

[se i campioni hanno dimensioni grandi non in tabella, si può usare la trasformazione di U in una statistica test che segue la normale standardizzata]

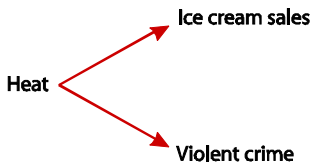
Esempio 13.5, pagina 205



Il disegno sperimentale

- Torniamo alla differenza tra studi osservazionali e studi sperimentali (Cap 1 e Scheda 5)
  - Solo negli studi sperimentali i trattamenti sono assegnati dallo sperimentatore
  - La conseguenza è che gli studi osservazionali possono identificare associazioni (per esempio tra trattamento e variabile risposta), ma gli studi sperimentali possono anche identificare relazioni di causa ed effetto

- La randomizzazione (assegnazione dei trattamenti a caso) minimizza l'effetto delle variabili di confondimento
  - Fumo e tumori: associazione o relazione causale?
  - Consumo di gelati e cimini violenti: associazione o relazione causale?
- La randomizzazione spezza la relazione che potrebbe esistere tra variabili di confondimento (per esempio, condizioni sociali e temperature nei due esempi) e trattamento



- Gli studi sperimentali sono quindi molto importanti e devono essere pianificati correttamente per eliminare (o almeno ridurre) la distorsione (della stima o del test) e ridurre l'errore di campionamento
- Alcune regole usate per pianificare correttamente un esperimento possono (e devono) essere utilizzate anche in uno studio osservazionale
  - Per esempio, la pianificazione della dimensione campionaria
- Attenzione sempre a non introdurre distorsioni di misura durante l'esperimento (freqeunti se le condizioni sono troppo "innaturali")

Le sperimentazioni cliniche: un buon esempio (quasi sempre)

- Sperimentazione clinica: studio sperimentale nell'uomo
- Definisce gli standard per la sperimentazione
- Esempio 14.2: studio sperimentale su un gel per ridurre la trasmissione dell'HIV

Come ridurre la distorsione

1. Analizzare sempre anche un gruppo di controllo
2. Randomizzazione
3. Cecità

Come ridurre l'errore di campionamento

1. Replicazione
2. Bilanciamento
3. Blocking

Come ridurre la distorsione
(dovuta a variabili di confondimento)

1. Il gruppo di controllo (anche Scheda 6)

Quale distorsione si avrebbe senza gruppo di controllo?

- I malati tendono spesso a migliorare comunque (tempo invece di trattamento)
- C'e' un impatto del trattamento, positive o negativo (somministrazione del trattamento invece del trattamento)
- Effetto placebo (risposta psicologica invece del trattamento)

Il gruppo di controllo appropriato deve

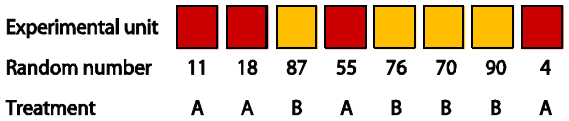
- Ricevere un placebo o il trattamento precedente
- Ricevere la stessa manipolazione se il trattamento richiede un intervento invasivo
- Negli esperimenti di campo, ricevere lo stesso distrurbo arrecato al gruppo trattato

2. Randomizzazione: I soggetti (o unità sperimentali) devono essere assegnati casualmente al gruppo di controllo o a gruppo trattato

- Le variabili di confondimento non sono eliminate, ma almeno il loro effetto è distribuito ugualmente nei due gruppi mediante la randomizzazione

  [esempio con pesci parassitati e non, predazione, e virus come var. di conf.]

- Semplice metodo di assegnazione randomizzata:



- Altri metodi di assegnazione non randomizzati (uomo):

  - Trattamenti A individui in clinica A e trattamenti B a individui in clinica B

    [problemi?]

  - Trattamenti assegnati alfabeticamente

    [problemi?]

3. Cecità: nessuno deve sapere chi riceve il trattamento e chi no

- Esperimento in singolo cieco: I soggetti (umani) non sanno cosa ricevono

- Esperimetno in doppio cieco: soggetti e ricercatori non sanno cosa ricevono

Questa accortezza è molto importante anche in esperimenti su soggetti non umani: è stato dimostrato gli effetti positivi di un trattamento sono sono molto più freqeunti in studi non ciechi

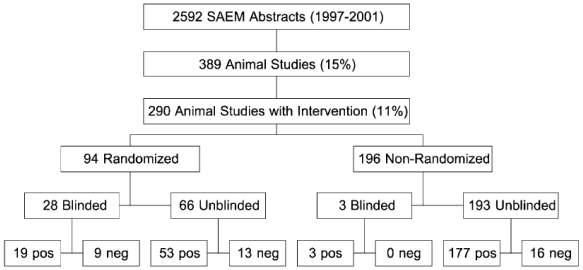


Figure 1. Classification of studies present at the Society for Academic Emergency Medicine (SAEM) annual meetings from 1997 to 2001.

Come ridurre l'influenza dell'errore di campionamento

- La variabile che misuriamo è influenzata da molti fattori, non solo quello al quale siamo interessati: esiste variabilità entro i gruppi che crea errore di campionamento (le stime non sono precise). Cosa possiamo fare?

- Fare in modo che tutte le variabili che possono influenzare la risposta all'esperimento abbiano valori simili in tutti i soggetti

[a volte impossibile, a volte svantaggioso]

1. Replicazione: ogni trattamento deve essere applicato a più soggetti (unità sperimentali) indipendenti

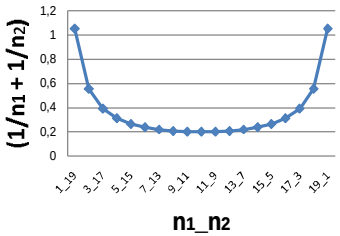
- Grandi campioni: piccolo errori di campionamento: più informazione: stime migliori e test più potenti

- Attenzione al problema della pseudo-replicazione (dati non indipendenti) (Scheda 2)

Bilanciamento: se i gruppi (per esempio, controllo e trattato) hanno la stessa dimensione campionaria, l'errore di campionamento si riduce

- Esempio: l'errore standard in un test t per due campioni dipende da $(1/n_1 + 1/n_2)$

- Il minimo di questa somma si ha, a partià di n totale, quando $n_1 = n_2$. Basta una media stimata male, per rendere la stima della differenza tra medie imprecisa

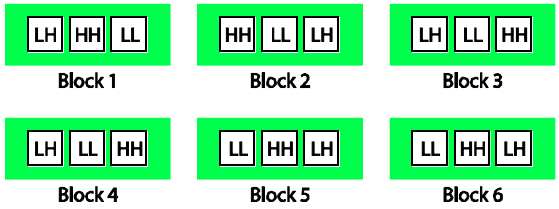


3. Blocking (raggruppamento in blocchi): le unità sperimentali che hanno caratteristiche simili sono raggruppate in blocchi, o strati; all'interno di ogni blocco, i trattamenti sono assegnati a caso alle diverse unità sperimentali

- E' una buona strategia per ridurre il rumore di fondo dovuto alla variabilità tra blocchi (esempio esperimento su HIV)

- Il disegno per dati appaiati è un esempio di blocking: ogni plot forestale (o ogni soggetto analizzato due volte) è un blocco, e la variazione tra blocchi non influisce sulla differenza tra trattamenti alla quale si è interessati

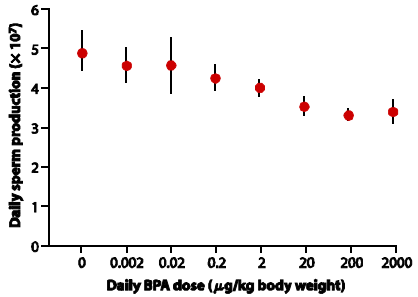
- Disegno a blocchi randomizzati (blocco randomizzato): trattamento applicato una volta in ogni blocco, ma più di due trattamenti



- Le differenze tra trattamenti sono misurate solo all'interno dei blocchi
  - Le differenze tra blocchi, non interessanti nell'esperimento, non "disturbano" l'analisi

- Il blocking è consigliato quando le differenze all'interno dei blocchi sono piccole (a parte ovviamente quelle eventualmente dovute al trattamento), ma sono grandi tra blocchi
  - Campi internamente omogenei
  - Acquari localizzati in un area omogenea della stanza
  - Pazienti di una stessa clinica
  - Esperimenti eseguiti lo stesso giorno

Trattamenti estremi: a volte un primo passo per cercare di evidenziare un effetto del trattamento (a dispetto di errori di campionamento) può essere quello di utilizzare trattamenti estremi, anche poco realistici(Esempio 14.4B)

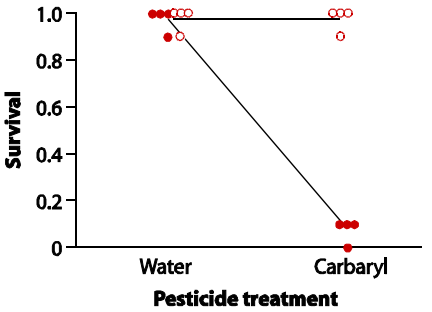


In questo caso, la tipica dose di esposizione nell'uomo era di 0.5-1.0 μg/kg; anche dosi superiori sono però analizzate nell'esperimento su ratti

Esperimenti multifattoriali

- Pianificazione di esperimenti multifattoriali
  - Rispondere a più domande con un solo esperimento
  - Studiare l'interazione
  - Definizione di esperimento fattoriale

Esempio 14.5



Alcune note sugli studi osservazionali

Si può ridurre distorsione e errore di campionamento con le stesse tecniche, tranne la randomizzazione

- Si può procedere anche ad **<u>appaiamento</u>** in studi osservazionali, cercando di ridurre l'effetto non solo dell'errore di campionamento, ma anche di variabili di confondimento **note** (o **probabili**)
  - Ogni individuo di un gruppo (per esempio, malati) viene appaiato con un individuo dell'altro gruppo (per esempio, sani) con caratteristiche molto simili per possibili variabili di confondimento

[oppure si usano gruppi che abbiano almeno distribuzioni simili nelle variabili di confondimento]

- Alternativamente, l'**<u>aggiustamento</u>** (per esempio con analisi della covarianza) cerca di considerare statisticamente (escludendolo) l'effetto delle variabili di confondimento

- Esempio: incidenza di una malattia in due gruppi (per esmepio, camminatori e non camminatori), ma nei due gruppi l'età dei soggetti non era la stessa. Posso statisticamente determinare la relazione tra età e incidenza della malattia, e fare in modo che i confronti siano "come se tutti avessero la stessa età"

Pianificare la dimensione dei campioni

- Pianificare la precisione
- Pianificare la potenza
- Pianificare per compensare la perdita di dati

<u>Pianificare la precisione</u>

- Gli IC diventano più piccolo quando *n* aumenta
- Ma quanto deve essere grande il campione (o i campioni) per ottenere una stima che, con una certa confidenza, non disti dal parametro più di un certo valore?
- Esempio con IC e incertezza per una media
- Esempio con IC e incertezza per una differenza tra medie

$IC_{(95\%)} => \bar{Y} \pm 2 \times ES = \bar{Y} \pm 2 \times \sigma / \sqrt{n}$

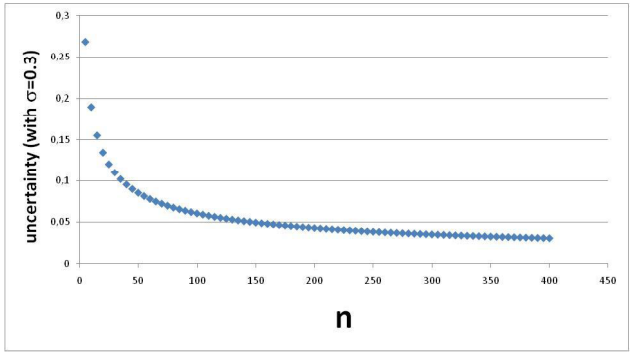
- Chiamiamo

Incertezza (massima con confidenza del 95%) = $2 \times \sigma / \sqrt{n}$

- Calcoliamo da qui il valore di *n* (minimo) che garantisce (al 95%) una stima di μ con una incertezza massima pari ad un certo valore prefissato

$$n = \left(\frac{2 \times \sigma}{\text{incertezza}} \right)^2 = 4 \left(\frac{\sigma}{\text{incertezza}} \right)^2$$

- Per esempio, se uno studio pilota ci ha suggerito che σ = 0.3 e vogliamo essere confidenti al 95% che l'incertezza nella stima della media non sia superiore a 0.05, la dimensione campionaria non dovrà essere inferiore a 144



- Molto importante: il decremento dell'incertezza è rapido per piccoli valori di *n* ma lento per grandi *n*

- Il vantaggio di grandi campioni si riduce progressivamente aumentando *n*

- Se invece

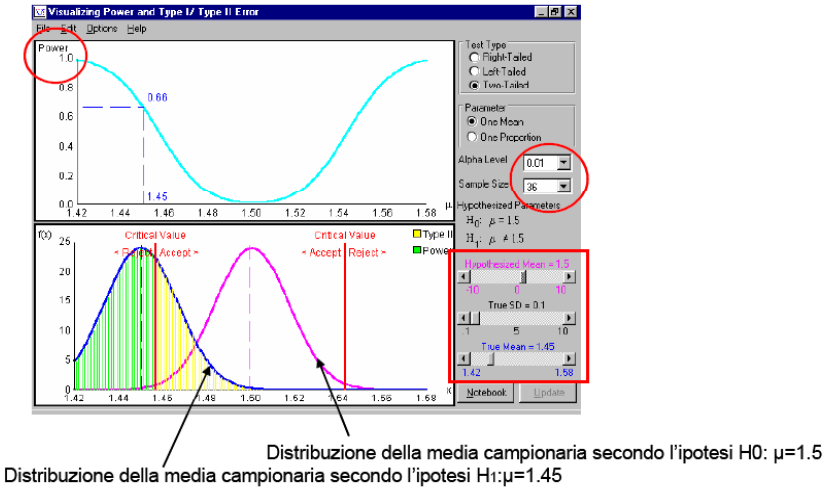
$IC_{(95\%)} => (\bar{Y}_1 - \bar{Y}_2) \pm 2 \times ES_{\bar{Y}_1 - \bar{Y}_2} = (\bar{Y}_1 - \bar{Y}_2) \pm 2 \times \sqrt{\sigma^2 \left(\frac{2}{n}\right)}$

Incertezza (massima con confidenza del 95%) = $2 \times \sqrt{\sigma^2 \left(\frac{2}{n}\right)}$

$$n = 8 \left(\frac{\sigma}{\text{incertezza}} \right)^2$$

- Per esempio, se uno studio pilota ci ha suggerito che σ = 0.3 e vogliamo essere confidenti al 95% che l'incertezza della stima della differenza tra due medie non sia superiore a 0.05, la dimensione campionaria in ognuno dei due gruppi non dovrà essere inferiore a 244

Pianificare la potenza



Distribuzione della media campionaria secondo l'ipotesi H1:μ=1.45

Distribuzione della media campionaria secondo l'ipotesi H0: μ=1.5

- Cosa emerge da questo grafico?

1. Maggiore è la distanza tra ipotesi nulla e ipotesi alternative, maggiore sarà il potere del test
   - Logico: è facile rifiutare l'ipotesi nulla se la realtà (ipotesi alternativa vera) è molto diversa dall'ipotesi nulla

2. Se la deviazione standard della variabile è bassa, sarà bassa la deviazione standard della stima (ES), e il potere aumenta
   - Logico: se tutti gli individui (in un campione, o all'interno di più campioni) sono simili, le stime dei parametri sono precise e anche piccole, ma reali, differenze sono identificate come significative

3. All'aumentare delle dimensioni campionarie, aumenta il potere statistico
   - Logico: se n è grande, ES è piccolo e, come sopra, le stime dei parametri sono precise e anche piccolo, ma reali differenze sono identificate come significative

- Conviene sempre fare un'analisi di potenza: con una certa dimensione campionaria, quale probabilità avrò di identificare un'ipotesi alternativa vera se la distanza tra ipotesi alternativa e nulla è almeno pari a un certo valore specificato D?

- Per esempio, sappiamo che la concentrazione media di una proteina in una pianta è 12 mg/l, con deviazione standard tra piante pari a 3

- Vogliamo fare un esperimento, trattando 5 piante, per vedere se la concentrazione media cambia, e ci interessa che una differenza pari ad almeno 2 mg/ml, se presente, venga identificata

- DOMANDA: quale sarà la probabilità che io riesca a identificare almeno questa differenza, se presente, con il mio campione di n = 5?

[risposta: intorno al 20%]

- Alternativamente, si può fissare un obiettivo di potenza (almeno 80%), e si calcolano le dimensioni campionarie minime che permettono di identificare una certa differenza specificata (D) con questa potenza

- Ogni test ha calcoli specifici. Per esempio, nel test t per due campioni, una formula approssimata per ottenere un potere dell'80% è data da

$$n \approx 16\left(\frac{\sigma}{D}\right)^2$$

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1: \text{almeno una media è diversa dalle altre}$$

- Equivalente a t test per due campioni se ci sono solo due gruppi

- ANOVA multifattoriale, ANOVA multivariata e ANCOVA: cosa sono?

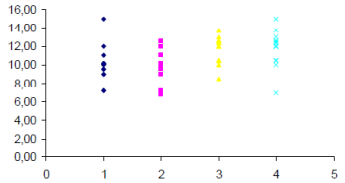
ANOVA: l'analisi della varianza per confrontare le medie di più di due gruppi

Premessa: L'ipotesi nulla è sulle medie, ma questa ipotesi è analizzata confrontando varianze

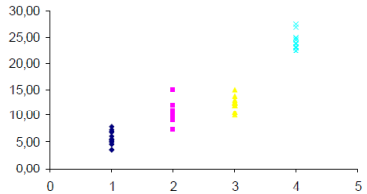
Un esempio sperimentale: 80 piante tutte con la stessa altezza iniziale sono distribuite a caso in 4 gruppi (20 piante per gruppo); le piante appartenenti a gruppi diversi sono irrigate con acqua a 4 diversi valori di ph; dopo un mese, si misura l'altezza di ogni pianta. **Domanda**: il ph influisce sulla crescita?

Un esempio osservazionale: 80 piante provenienti da 4 distante regioni geografiche (20 da ogni area) sono campionate e misurate per la loro concentrazione di cobalto; **Domanda**: la posizione geografica influisce sulla concentrazione di questo metallo pesante?

Esempio di un set di dati quando è probabilmente vera H_0



Esempio di un set di dati quando è probabilmente vera H_1



Attenzione! L'asse X non riporta necessariamente una variabile esplicativa numerica, ma semplicemente i gruppi (variabile esplicativa categorica)

- Prima di passare all'ANOVA, consideriamo il problema dei test multipli (data dredging)

[Scheda 8]

- Perchè non fare 6 test t se abbiamo 4 gruppi, o 10 test t se abbiamo 5 gruppi, o 45 test t se abbiamo 10 gruppi?

- Qual'è la probabilità di compiere almeno un errore di primo tipo (rifiuto di un'ipotesi nulla vera) facendo *c* test indipendenti?

$$P(almeno\ un\ errore\ di\ primo\ tipo\ in\ c\ test\ indip.) = 1 - (1-\alpha)^c$$

[se i test non sono indipendenti, questa P è un valore sovrastimato, ma prudente]

- Se c = 6 (e α = 0.05)

$$P(almeno\ un\ errore\ di\ primo\ tipo\ in\ c\ test\ indip.) = 0.26$$

- Se c = 10 (e α = 0.05)

$$P(almeno\ un\ errore\ di\ primo\ tipo\ in\ c\ test\ indip.) = 0.40$$

- Se c = 45 (e α = 0.05)

$$P(almeno\ un\ errore\ di\ primo\ tipo\ in\ c\ test\ indip.) = 0.90$$

- Approccio semplice ma molto conservative per i test multipli: la correzione di Bonferroni: adottare un livello di significatività $\alpha*$ nei singoli test *c* volte inferiore a quello α usuale

$$\alpha^* = \frac{\alpha}{c}$$

[usando $\alpha^* = \frac{\alpha}{c}$ in ogni test, $1 - \left(1 - \frac{\alpha}{c}\right)^c \approx \alpha$]

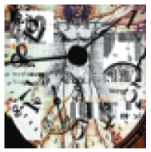
- Approcci più potenti: FDR (false discovery rate) e q-value (invece di P-value)

- In generale, meglio sviluppare analisi statistiche che non necessitino troppi test singoli: ANOVA, almeno nella sua fase iniziale di test di ipotesi nulla "generale", non richiede tanti test t ma un'unica verifica di ipotesi simultanea su tutte le medie

ANOVA: COME FUNZIONA? Partiamo da un esempio

Ginocchia che vedono la luce

L'attraversamento di vari fusi orari, come nel caso di un lungo viaggio aereo, può causare un disturbo noto come jet lag. Questo disturbo si supera quando l'alternarsi della luce e del buio percepito dagli occhi nel nuovo fuso orario sincronizza nuovamente e in maniera graduale l'orologio circadiano interno, che aveva subito un fenomeno di sfasamento. Campbell e Murphy (1998) presentarono uno studio che suggeriva che l'orologio circadiano umano potesse essere sincronizzato di nuovo anche con l'esposizione della superficie posteriore del *ginocchio* alla luce, un risultato che fu accolto con scetticismo da alcuni, ma che fu considerato un'importante scoperta da altri. In seguito, alcuni aspetti del disegno sperimentale di quello studio furono messi in discussione.[1] I dati in Tabella 15.1-1 provengono da un esperimento condotto da Wright e Czeisler (2002) che hanno riesaminato il fenomeno. Nel nuovo esperimento, il ritmo circadiano è stato misurato valutando il ciclo giornaliero di produzione di melatonina in 22 soggetti assegnati casualmente a uno di tre diversi trattamenti luminosi. I soggetti sono stati svegliati dal sonno e sottoposti a un singolo episodio di 3 ore di illuminazione intensa applicata soltanto agli occhi, soltanto alle ginocchia, e né agli occhi né alle ginocchia (il gruppo di controllo). Gli effetti del trattamento sul ritmo circadiano sono stati misurati due giorni dopo valutando l'entità dello sfasamento nel ciclo giornaliero di produzione di melatonina di ciascun soggetto. I risultati sono rappresentati graficamente in Figura 15.1-1. Una misura negativa indica un ritardo nella produzione di melatonina, che è l'effetto previsto del trattamento luminoso, mentre una misura positiva indica un anticipo. Il trattamento luminoso influenza lo sfasamento? ■



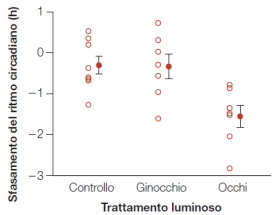
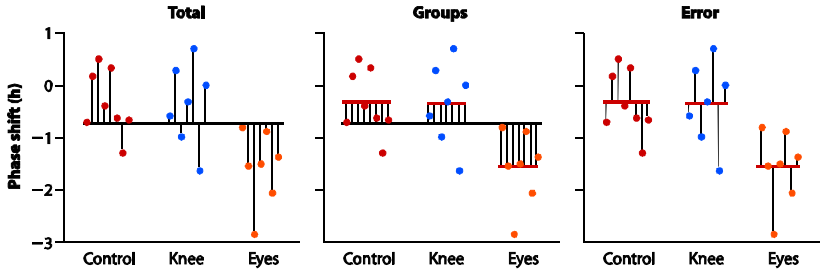


Figura 15.1-1
Diagramma a punti che visualizza lo sfasamento nel ritmo circadiano di produzione di melatonina in 22 soggetti a cui sono stati applicati trattamenti luminosi diversi (*circoletti vuoti*). I circoletti pieni e i segmenti verticali (barre di errore) rappresentano le medie dei gruppi ± l'errore standard.

Tabella 15.1-1

Dati grezzi e statistiche descrittive dello sfasamento, in ore, per l'esperimento sul ritmo circadiano.

Trattamento	Dati (h)	$\bar{Y}$	s	n
Controllo	0,53, 0,36, 0,20, –0,37, –0,60, –0,64, –0,68, –1,27	–0,3088	0,6176	8
Ginocchia	0,73, 0,31, 0,03, –0,29, –0,56, –0,96, –1,61	–0,3357	0,7908	7
Occhi	–0,78, –0,86, –1,35, –1,48, –1,52, –2,04, –2,83	–1,5514	0,7063	7



- La variabilità totale viene suddivisa in due componenti: entro gruppi e tra gruppi: se è vera l'ipotesi nulla, queste due component non saranno mai tropo diverse

- La componente entro gruppi misura la distanza media delle osservazioni dalla media del rispettivo gruppo. E' detta anche media dei quadrati degli errori, varianza dell'errore, o varianza entro gruppi, o varianza residua. Non misura la deviazione da H_0

- La componente tra gruppi misura la distanza media tra le medie campionari e la media generale. E' detta anche media dei quadrati tra gruppi o varainza tra gruppi. Cresce al crescere della distanza tra H_1 e H_0

- Introduciamo abbreviazioni e vediamo come stimare queste due componenti della variabilità

- Media dei quadrati degli errori = *error mean square* (**MS_{errore}**) =

$$MS_{errore} = \frac{\sum s_i^2 (n_i - 1)}{N - k}$$

- Media dei quadrati tra gruppi = *group mean square* (**MS_{gruppi}**) =

$$MS_{gruppi} = \frac{\sum n_i (\bar{Y}_i - \bar{Y})^2}{k - 1}$$

Attenzione: $\bar{Y}$ è la media generale (non la media di medie); i numeratori sono la somma dei quadrati degli errori (o devianza entro gruppi o devianza dell'errore) e la somma dei quadrati tra gruppi (o devianza tra gruppi): SS_{errore} e SS_{gruppi}

- Queste due quantità misurano la stessa cosa se è vera l'ipotesi nulla: perchè?

- Sempre se è vera l'ipotesi nulla, la statistica test F calcolata come rapporto MS_{gruppi} / MS_{errore} si distribuisce secondo la distribuzione teorica F con (k-1) and (N-k) gradi di libertà: perché?

$$F = \frac{MS_{gruppi}}{MS_{errore}}$$

- La tabella dell'ANOVA

Fonte di variazione	df	Somma dei quadrati	Media dei quadrati	F	P-value
Gruppi	k-1	SS_{gruppi}	$MS_{gruppi} = SS_{gruppi}/(k-1)$	$F_{calc} = MS_{gruppi}/MS_{errore}$	$P(F>F_{calc})$
Errore	N-k	SS_{errore}	$MS_{errore} = SS_{errore}/(N-k)$		
Total	N-1	SS_{totale}			

- SS e df sono additivi

- Se è vera l'ipotesi alternativa, F assumerà valori significativamente **maggiori** di 1: perché?
  - Chiaramente, quanto grande deve essere la statistica F calcolata dai dati per rifiutare H_0 dipende dall'α prescelto e dai valori critici di F (o dal P-value calcolato).

- Le assunzioni dell'ANOVA sono quelle del t test per due campioni, ma estese ai k gruppi

- ANOVA è robusta a deviazioni di normalità se gli *n* non sono troppo piccoli, e deviazione da uguali varianze se gli n sono simili nei gruppi

Torniamo all'esempio

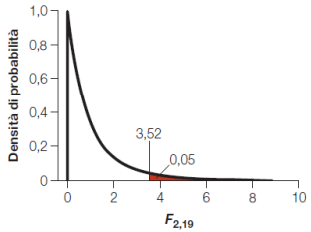


Figura 15.1-3
La distribuzione *F* con 2 e 19 gradi di libertà. Il valore di *F* varia da zero a più infinito. L'area sottesa dalla curva a destra del valore critico *F* = 3,52 (*ombreggiata*) è pari a 0,05. (Vedi Tavola Statistica D.)

Tabella 15.1-3
Tabella dell'ANOVA per i risultati dell'esperimento sul ritmo circadiano (Esempio 15.1).

Fonte di variazione	Somma dei quadrati	*df*	Media dei quadrati	Rapporto *F*	*P*
Gruppi (trattamento)	7,224	2	3,6122	7,29	0,004
Errore	9,415	19	0,4955		
Totale	16,639	21			

La variabilità spiegata

- Utilizzando le somme dei quadrati, e il fatto che godono della proprietà additiva, possiamo definire un utile indice, R^2

$$R^2 = \frac{SS_{gruppi}}{SS_{totale}}$$

- R^2 si può interpretare come la frazione di variabilità della variabile Y *spiegata* dalle differenze tra i gruppi

- R^2 ovviamente può assumere solo valori compresi tra 0 e 1

- Se R^2 è vicino a 0, solo una piccola frazione della variabilità nella variabile risposta (Y) è *spiegata* dai gruppi: la variabilità è soprattutto entro-gruppi, e il valore che assume Y in ogni osservazione dipende molto poco dal gruppo di appartenenza

- Se R^2 è vicino a 1, gran parte della variabilità nella variabile risposta (Y) dipende dal gruppo di appartenenza; la variabile esplicativa *spiega* perché i valori di Y sono diversi

- Nello studio sui ritmi circadiani, R^2 è pari a 0.43; il altre parole, il 43% della variabilità nello sfasamento del ritmo circadiano è spiegato dalle differenze tra trattamenti, il 57% non è spiegato dai trattamenti (e lo chiamiamo appunto "errore", ovvero variabilità all'interno dei gruppi)

Una alternativa non parametrica all'ANOVA unifattoriale è il test di Kruskal-Wallis;
è basato (come il test di Mann-Whitney) sui ranghi

Confronti pianificati e non pianificati

- Se ANOVA produce deviazioni significative da H_0, è interessante capire quali gruppi sono diversi da quali altri, e quanto sono grandi queste differenze

- E' importante distinguere tra confronti pianificati (detti a anche *a priori*) e confronti non pianificati (detti anche *a posteriori*)

Confronti pianificati (detti anche "a priori")

- Per buone ragioni (teoriche, studi precedenti) si è interessati solo ad un numero limitato dei confronti a coppie

- In questo caso, si può usare il test t per due campioni, ma con MS_{errore} al posto di s_p^2

- Stessa cosa nel calcolo dell'IC della differenza per il confronto pianificato

- Esempio per il confronto "*ginocchio*" - "*controllo*"

Confronti non pianificati (detti anche "a posteriori")

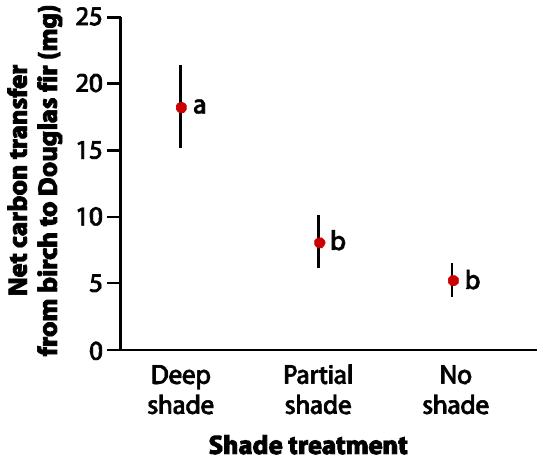
- Quando vengono analizzati tutti (o molti) confronti a coppie, il rischio di errori di primo tipo sale (data dredging), e quindi sono necessari specifici accorgimenti

- Tra i metodi specifici più usati c'e' il test di Tukey-Kramer

  - Tecnicamente, si devono svolgere tanti test t, ma utilizzando MS_{errore} al posto di s_p^2 e una distribuzione teorica diversa (non la distribuzione t ma la distribuzione q con parametri *k* e *N-k*)

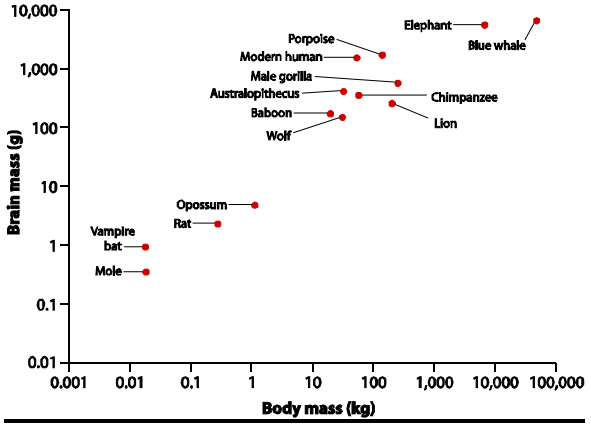
  - Questo garantisce che P(almeno un errore I tipo) ≤ α

  - Le medie sono ordinate, i confronti vengono testati a coppie, e la rappresentazione grafica è la seguente:



CORRELATION BETWEEN NUMERICAL VARIABLES

- The meaning of correlation



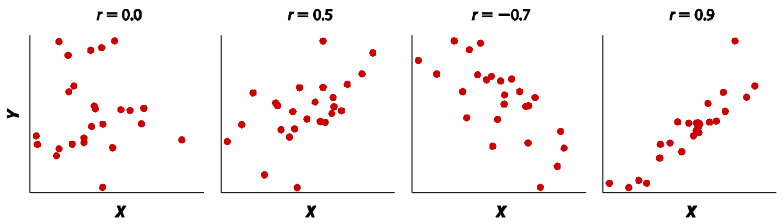
- As usual, we need to go from descriptive to inferential statistics (sample, and not populations, are available)

- The first step is computing a <u>linear</u> **correlation coefficient**, i.e. an index to measure the tendency of two numerical variable to linearly "co-vary"

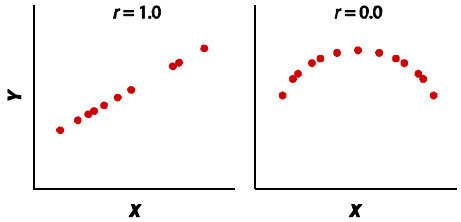
- The correlation coefficient in the population is indicated with ρ, the correlation coefficient computed form the sample data with r

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

- r varies between -1 and 1



- Pay always attention that this coefficient measures <u>linear</u> correlation!



- Now, as usual, we need a standard error, i.e. the standard deviation of the sampling distribution of r. An estimate of it is given by

$$SE_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

- SE can be used to test the null hypothesis that ρ =0

- [Computing the confidence interval requires an additional transformation, since the distribution of r is not normal (see page 436).

Testing the null hypothesis of zero correlation (see example 16.2, page 438)

- The data:

Inbreeding coefficient	Number of pups
0	6
0	6

0,13	7
0,13	5
0,13	4
0,19	8
0,19	7
0,19	4
0,25	6
0,24	3
0,24	3
0,24	3
0,24	3
0,24	2
0,24	2
0,27	3
0,3	5
0,3	3

0,3	2
0,3	1
0,36	3
0,4	3
0,37	2
0,22	4

- The scatter plot:

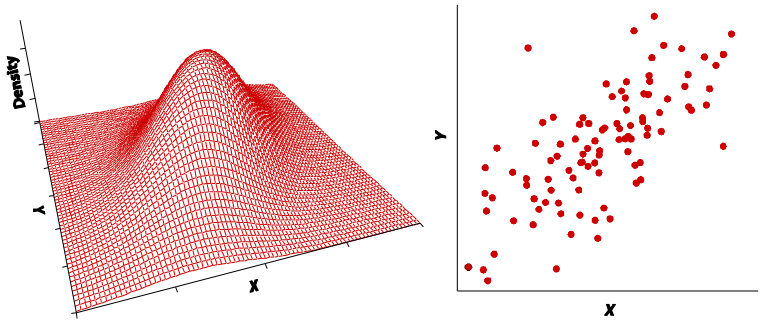
- r = -0.608 (verify it with Excel!)
- SE_r = 0.169 (verify it with Excel!)

$$t = \frac{r}{SE_r} = \frac{-0.608}{0.169} = -3.60$$

- Conclusion (based on the t distribution with df = n-2): the null hypothesis of zero correlation is rejected. Inbreeding is correlated with the number of surviving offspring in the analysed wlf population

Assumptions of the correlation analysis

- Random sampling
- Bivariate normal distribution
  - The relationship between X and Y is linear
  - The cloud of points in the scatterplot has a circular or elliptical shape
  - The frequency distributions of X and Y are normal

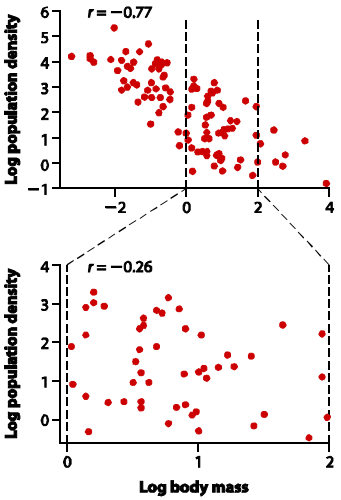


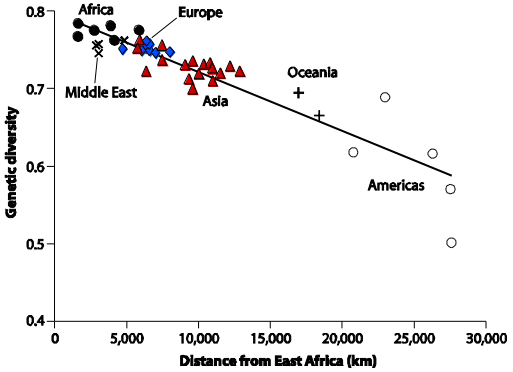
Data from three distributions that differ from bivariate normality

- Transforming the data (one or both variables) can produce new variables that match the assumptions

- The Spearman's rank correlation is the nonparametric equivalent to be used when transformations does not solve the problem of deviation from bivariate normality

A last warning: the correlation coefficient depends on the range





Linear regression

- Assuming that there is a straight line describing how the response variable is affected by the explanatory variable, linear regression is estimating the parameters of this line from a sample

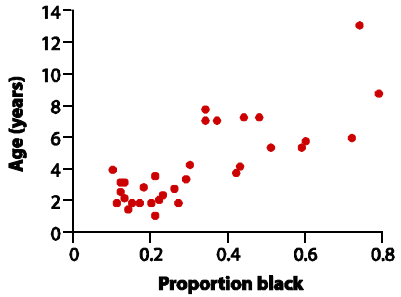
REGRESSION

- Regression is used to predict the value of one numerical variable, the response variable, from that of another, the explanatory variable, and to measure the rate of change of the response variable with the explanatory variable

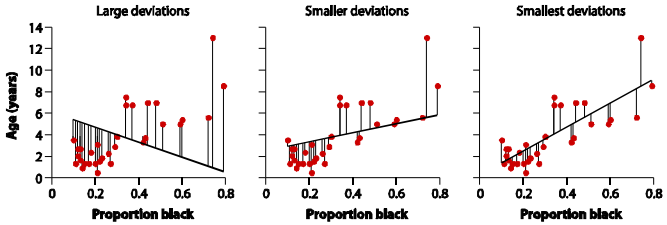
- Correlation measures the strength of the correlation (treating the two variables in the same way), regression measures how steeply the response variable changes with change in the explanatory variable

- As usual, the estimated line with its parameters is affected by sampling errors. Inferential statistics helps us to understand these errors computing confidence intervals and testing hypotheses

- The example of the lions: how can I predict the age of a male on the basis of the proportion of black on its nose?



- The method of least squares: the concept



- There is only one line (with a specific slope and intercept) which minimizes the squared deviation: this is the regression line

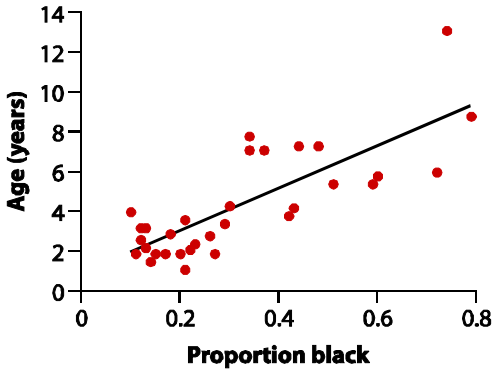
- Slope (b) and intercept (a) of the Y = a + bX regression line can be found as

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

- For the lions example, the regression line is Age = 0.88 +10.65(proportion of black)

- The slope (10.65) is the estimated change of age for a unit (1) of change in the proportion



- Two important points: 1. extrapolation based on the line may be dangerous; 2. The line is just a sample statistics: we need now inferences on the population, i.e. inferences on the parameters a (estimated with a) and b (estimated with b) [Warning! these Greek letters have been used before with very different meanings]

- If the assumptions of this analysis are met, the sampling distribution of *b* is normal with mean equal to β and **standard error** estimated by

$$SE_b = \sqrt{\frac{MS_{residual}}{\sum(X_i - \bar{X})^2}}$$

- The SE of b can be used to compute the confidence interval for the slope and testing a null hypothesis about a slope. Let´s see how.

- **Predicted values $\hat{Y}$**: using the regression line, we can compute, for any specific value of X, the mean value of Y predicted for all the individuals having that specific X value. This is called the "predicted value of Y"

- **Residuals**: the distance between observed and predicted values. For each *i* value, there is a *i* residual: $Y_i - \hat{Y}_i$. The spread of the scatter of points above and below the line is quantified by the residual mean square:

$$Residual\ mean\ square = MS_{residual} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$$

- **Confidence interval for the slope**

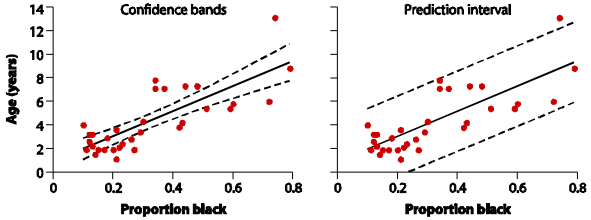
$$CI_{slope} = b \pm t_{\alpha(2),df} SE_b$$

where $t_{\alpha(2),df}$ is the two-tailed critical value of the t distribution with df = n-2

- **t test on the slope** (with H_0: $\beta = \beta_0$) (see Example 17.3)

$$t = \frac{b - \beta_0}{SE_b}$$

- **Confidence intervals for predictions:** the difference between confidence bands and prediction interval



ANOVA applied to the regression analysis

- The variation in the variable Y can be divided in two components: the residual component and the regression component

- The residual component is similar to the error component in the typical ANOVA analysis: is the variation of the observations around the value predicted y the regression

- The regression component is related to the distance between predicted values and the mean of Y

- If the null hypothesis that $\beta = 0$, the mean squares corresponding to these two components should be the same, and the ANOVA on the ratio can be used instead of the t-test (and computer programs usually produce an ANOVA table with the residual and the regression component, see page 479).

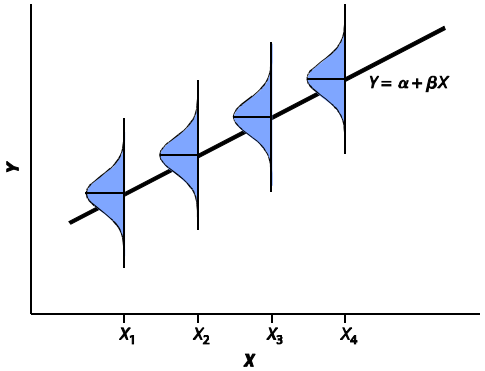
- The regression and total sum of squares ($SS_{regression}$ and SS_{total}) can be used to compute R^2, the fraction of variation in Y explained by the X (similarly to what was done in the typical ANOVA setting)

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

- If R^2 is close to 1, all data points will be very close to the line; if R^2 is close to 0, data points will be widely scattered above and below the regression line

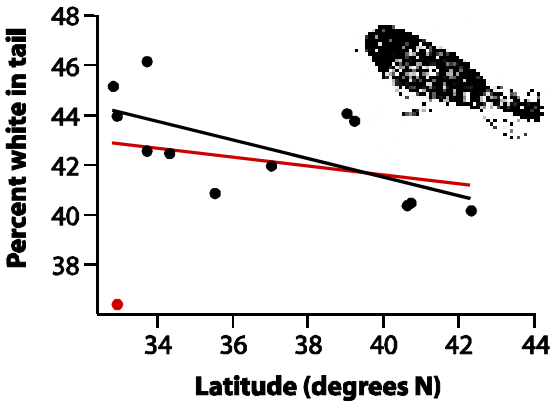
Assumptions of regression

- For each value of X, the mean of all possible Y values are along a straight line
- For each value of X, the distribution of Y is normal with the same variance
- For each value of X, the values of Y represent a random sample from the population
- No assumptions are made about X!



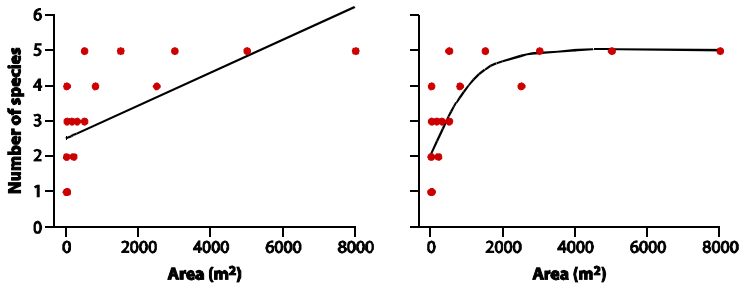
Outliers

- Outliers may have large impact on the regression line. Pay always attention to this point, comparing the regression line including and excluding them



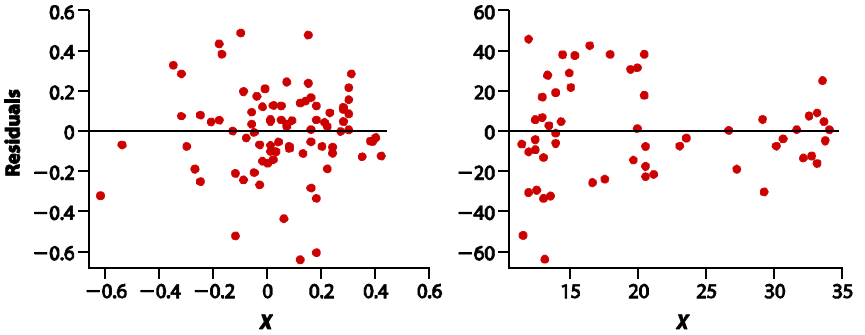
Detecting non-linearity

- Visual inspection of the data, possibly with a regression line on the scatter plot, can be very useful to detect deviations from linearity. Clearly, more formal methods are available to detect non-linearity and to identify non linear models that approximate better the trend



Detecting non-normality and unequal variances

- Very useful is the **residual plot.** It the assumptions of normality and equal variances are met, then the residual plot should have the following features: 1. Symmetric cloud above and below the line at 0; 2. Higher density close to the line at 0; 3. Almost no curvature; 4. Equal spread fro different X values



Non linear regression

- Many methods are available to analyse the data assuming a non-linear relationship. These methods implies different functions with different parameters to be estimated. They implies of course also a statistical inference on these parameters, and the comparison of hypotheses on the parameters and the comparison between models (i.e., different curves)

- It is important to remember , however, that overfitting has be avoided: it might be useless (in terms of predictions) and biologically meaningless