BIG DIVE

DATA SCIENCE & ANALYTICS

Modelli Statistici

DAVIDE PASSARETTI

Organized by top ix torino piemonte internet exchange

Designed for INTESA SANPAOLO

In collaboration with aizoon technology consulting, ISI Foundation, TODO

La statistica: cos'è e perché è utile

- Se abbiamo i **dati** e la **necessità di interpretare** la *complessità* del mondo, dobbiamo ricorrere alla **statistica**.
- Esempi di questioni nell'ambito socio–economico:
  - Gli studenti universitari provenienti da diverse parti del mondo percepiscono l'etica del business in modo diverso?
  - Qual è l'effetto della pubblicità sulle vendite?
  - I fondi comuni aggressivi hanno davvero rendimenti più alti rispetto a quelli più conservativi?
  - Vi è una stagionalità nei profitti della tua azienda?
  - Qual è la relazione tra la vendita dei cereali e il loro posizionamento sullo scaffale del supermercato?
  - Quanto sono affidabili le previsioni trimestrali per la tua azienda?
  - Ci sono caratteristiche comuni che descrivono i vostri clienti ed il motivo per cui scelgono i vostri prodotti? – e, soprattutto, queste caratteristiche sono proprie anche di coloro che non sono vostri clienti?
- Rispondere a queste domande con un approccio *data–driven* vuol dire interpretare la **variabilità** caratterizzante ciascun fenomeno.

Macroaree nella statistica

- La più marcata suddivisione in statistica è quella tra statistica **descrittiva** e statistica **inferenziale**. Un ramo della inferenziale è quello **predittivo** (obiettivo differente).
- Nella statistica *descrittiva*, l'obiettivo è *descrivere* tramite misure di sintesi un insieme di unità statistiche di cui abbiamo **conoscenza totale** (le unità che osserviamo sono il nostro oggetto di studio).
- La statistica *inferenziale* introduce una differenza tra *popolazione* e *campione*. Il *campione osservato* è il sottoinsieme di unità statistiche che abbiamo a disposizione per *fare inferenza* sulla popolazione di riferimento (campione rappresentativo). Abbiamo dunque una **conoscenza parziale** del fenomeno, indipendentemente dall'ampiezza campionaria. L'inferenza avviene tramite *stima* o *test di ipotesi*.
- La statistica *predittiva* è ancora inferenziale perché parte da un campione, tuttavia ha lo scopo di *prevedere* un outcome, più che spiegare/inferire l'esistenza di eventuali relazioni fra variabili.

Statistica descrittiva

Tipi di variabile

- Una variabile statistica (o carattere) è un attributo descrivente un'unità statistica oggetto di osservazione.
- I modi in cui una variabile può presentarsi sono chiamati modalità (o anche *valori* nel caso di variabili numeriche).
- Le modalità identificano il tipo di variabile in base a *discriminazione*, *ordine* e *distanza*:
  - le variabili categoriche nominali **discriminano** in base a caratteristiche qualitative (es: blu, verde, castano per il *colore degli occhi*),
  - le variabili categoriche ordinali sono ancora qualitative, ma aggiungono un **ordine** alla discriminazione (es: diploma, laurea, dottorato per il *titolo di studio*),
  - le variabili quantitative sono numeriche, pertanto, oltre ad ordinare, **quantificano una distanza** tra un valore e un altro (un individuo alto 1.60 m è più alto di 10 cm di uno alto 1.50 m). Sono divise in **discrete** (es: *numero di clienti*) e **continue** (es: *peso*, *altezza*, ecc. . .).

Indici di posizione – Tendenza centrale

Sia X una variabile quantitativa avente $n = 5$ valori: $X = \{3, 2, 5, 2, 8\}$.

Media aritmetica

La media aritmetica è uguale a $\mu_X = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{3+2+5+2+8}{5} = 4$.

Mediana

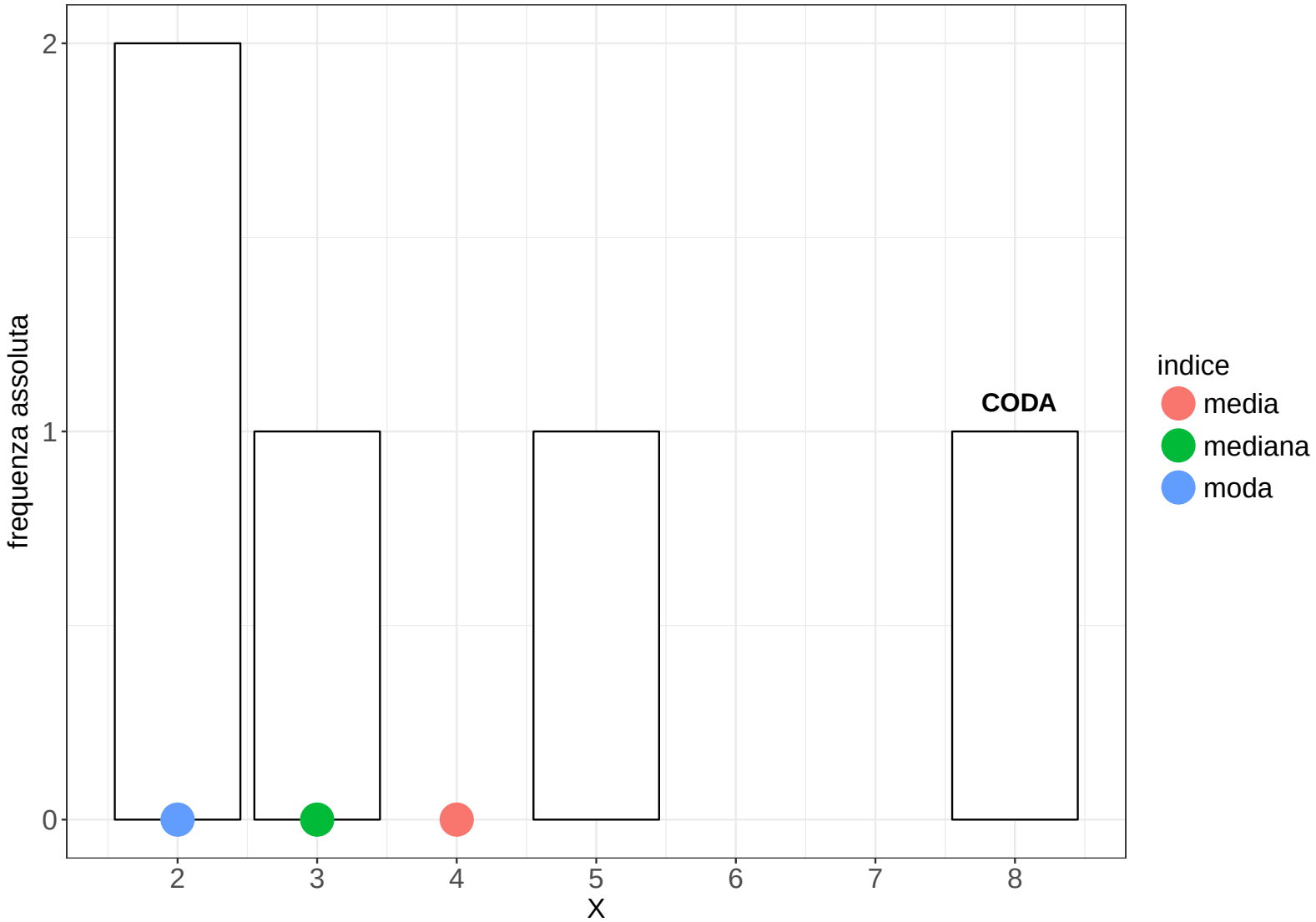
La mediana è il valore che lascia a sinistra (ma anche a destra) il 50% della distribuzione. Ordinando X, risulta che il valore centrale è $Me(X) = 3$.
È un indice **più robusto della media**.

Moda

La moda è il valore più frequente della distribuzione: $Mo(X) = 2$.
Se X è una variabile continua, ha più senso suddividerla in classi e trovarne la **classe modale**, ovvero la parte della distribuzione con più occorrenze.
È un indice **più robusto sia della media che della mediana**.

Indici di posizione – Idea di robustezza

La media insegue le code:



Indici di posizione – Quantili

Il quantile (o percentile) $q_X(\tau)$ è il valore che lascia a sinistra una frazione τ della distribuzione di X. Esempi più noti sono i quartili.

- Primo Quartile: $Q_1(X) \equiv q_X(0.25)$ lascia a sinistra $\frac{1}{4}$ della distribuzione di X. È il valore in posizione $\frac{n+1}{4}$ dei dati ordinati. Nel nostro esempietto numerico è il valore in posizione $\frac{5+1}{4} = 1.5$, cioè la media aritmetica tra 2 (in posizione 1) e 2 (in posizione 2) che è 2.
- Mediana (Secondo Quartile): $Me(X) \equiv Q_2(X) \equiv q_X(0.50)$ lascia a sinistra metà della distribuzione di X.
- Terzo Quartile: $Q_3(X) \equiv q_X(0.75)$ lascia a sinistra $\frac{3}{4}$ della distribuzione di X. È il valore in posizione $\frac{3(5+1)}{4}$ dei dati ordinati. Nel nostro esempietto numerico è il valore in posizione $\frac{18}{4} = 4.5$, cioè la media aritmetica tra 5 (in posizione 4) e 8 (in posizione 5) che è 6.5.

In statistica inferenziale si ragiona più su quantili estremi (0.9, 0.95, 0.975, 0.99, 0.995, ecc., e loro complementi a 1).

Indici di variabilità – In teoria

I seguenti indici di variabilità per variabili quantitative esprimono l'allontanamento della distribuzione da un dato indice di tendenza centrale.

Deviazioni dalla media aritmetica

- La varianza è la media degli scarti quadratici dei valori di X da μ_X: $\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2$.
- La deviazione standard σ_X è la radice quadrata della varianza ed ha il vantaggio di essere espressa su scala originaria.
- Il coefficiente di variazione è dato da $CV(X) = \frac{\sigma_X}{\mu_X}$ ed è un **indice relativo** (cioè senza unità di misura).

Esempio di deviazione dalla mediana

Il MAD (*Median Absolute Deviation*) è la mediana degli scarti in valore assoluto dalla mediana: $MAD(X) = Me\left(\left|x_i - Me(X)\right|\right), i = 1, \ldots, n$.
È un indice **più robusto**, ma meno utilizzato di σ_X.

Indici di variabilità – Piccolo esercizio

Abbiamo ancora la nostra variabile $X = \{3, 2, 5, 2, 8\}$.

$$\sigma_X^2 = \frac{(3-4)^2 + (2-4)^2 + (5-4)^2 + (2-4)^2 + (8-4)^2}{5} = \frac{26}{5} = 5.2$$

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{5.2} = 2.28$$

$$CV(X) = \frac{\sigma_X}{\mu_X} = \frac{2.28}{4} = 0.57$$

Il MAD si calcola come segue:

- gli scarti in valore assoluto dalla mediana sono:
  $|0|, |-1|, |2|, |-1|, |5| \longrightarrow 0, 1, 2, 1, 5.$
- la mediana di tali scarti è 1.

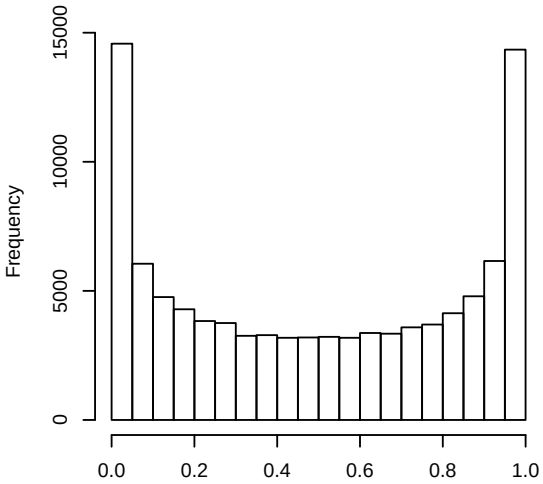
Indici di asimmetria

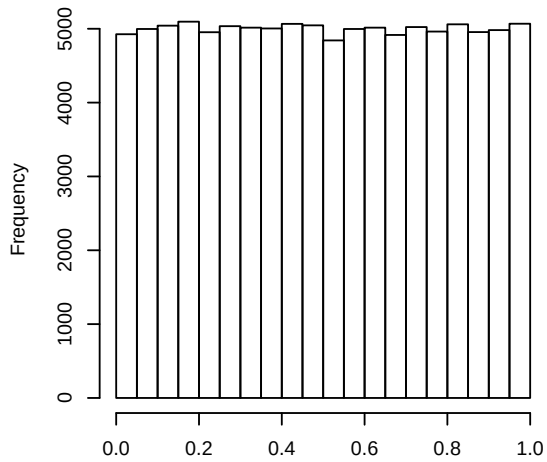
- La asimmetria di una distribuzione è solitamente misurata ragionando sul maggior peso che può avere una delle due code rispetto all'altra.
- L'indice è positivo se la coda destra è più "pesante", negativo se la coda sinistra è più pesante, uguale a zero se le due code si equivalgono in peso o la distribuzione è simmetrica.
- Se la distribuzione è simmetrica, $\mu_X = Me(X)$ e tutti gli indici di asimmetria saranno uguali a zero.
- Due esempi di indici sono presentati di seguito:
  - l'indice di Fisher $\mathcal{A}_F = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu_X}{\sigma} \right)^3$ determina quali tra gli scarti positivi e gli scarti negativi dalla media abbiano maggior peso;
  - l'indice di Hotelling-Solomon: $\mathcal{A}_{HS} = \frac{\mu - Me}{\sigma}$ riflette sul fatto che la media insegua più la coda di quanto faccia la mediana.
    È un **indice normalizzato** tra -1 e 1.

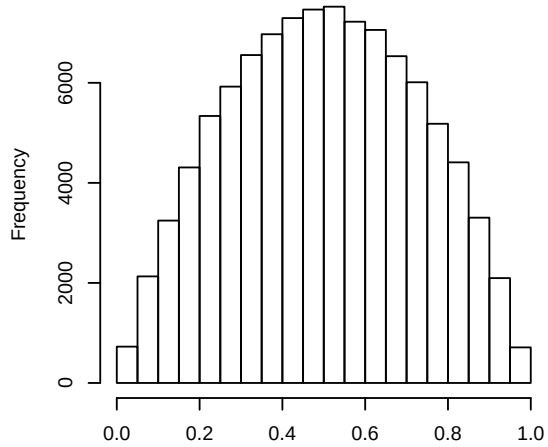
Sintesi a 5 e boxplot

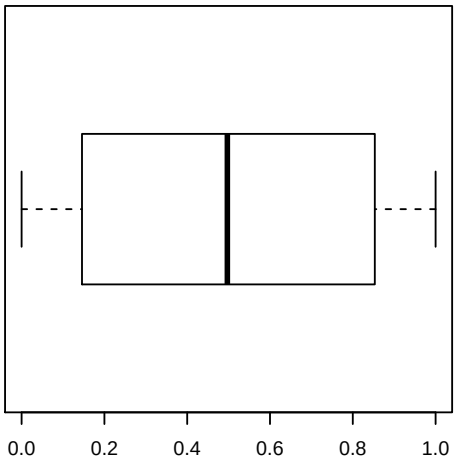
- Il campo di variazione interquartile è una misura di variabilità per una variabile quantitativa X: $\Delta(X) = Q_3(X) - Q_1(X)$.
- Se un valore della distribuzione è minore di $Q_1(X) - \frac{3}{2}\Delta(X)$ o maggiore di $Q_3(X) + \frac{3}{2}\Delta(X)$ è considerato anomalo (*outlier*).
- Il minimo relativo $Min_R(X)$ è il valore non anomalo più piccolo della distribuzione.
- Il massimo relativo $Max_R(X)$ è il valore non anomalo più grande della distribuzione.
- La "sintesi a 5" è formata da:
  $Min_R(X)$, $Q_1(X)$, $Me(X)$, $Q_3(X)$, $Max_R(X)$.
- Il boxplot rappresenta Q_1 e Q_3 come estremi di una scatola ed il minimo e massimo relativi come baffi. All'interno della scatola vi è $Me(X)$. Gli outliers sono rappresentati con dei puntini al di là dei baffi.
- Il boxplot ci dice molto sulla variabilità e la forma di una distribuzione.

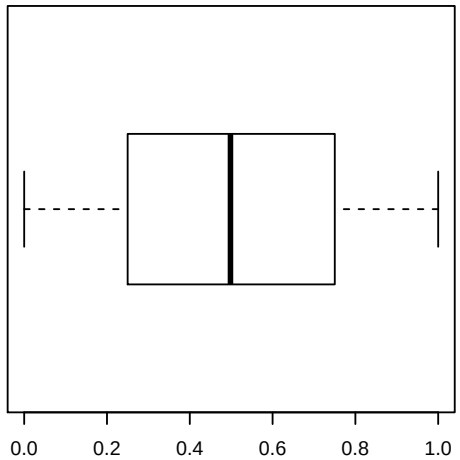
Boxplot e distribuzioni simmetriche

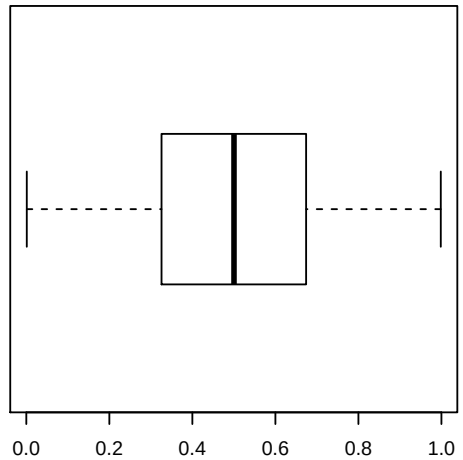




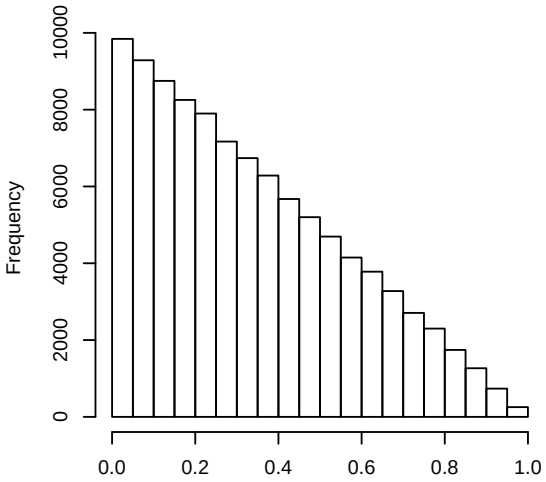


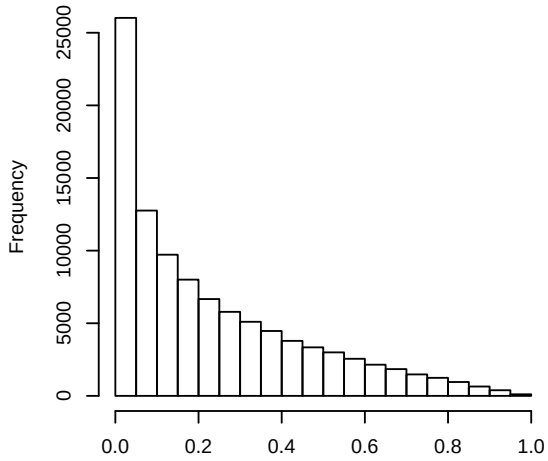


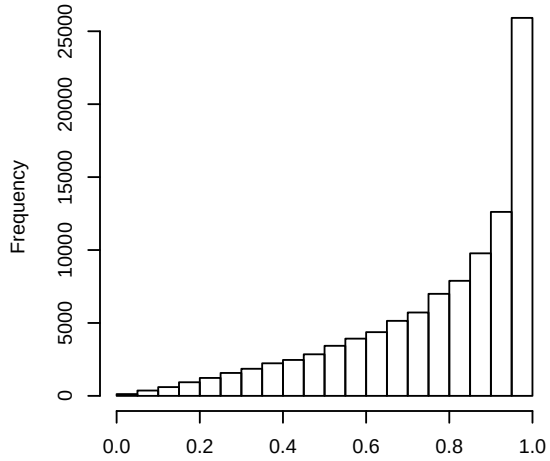


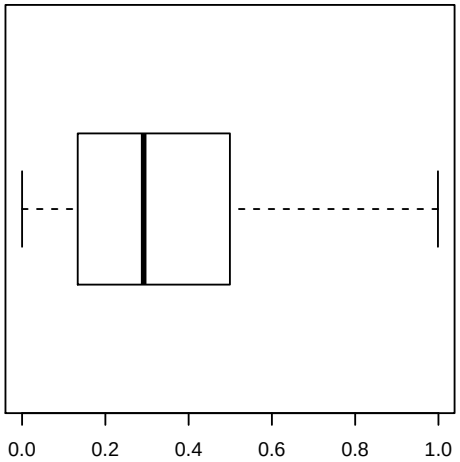


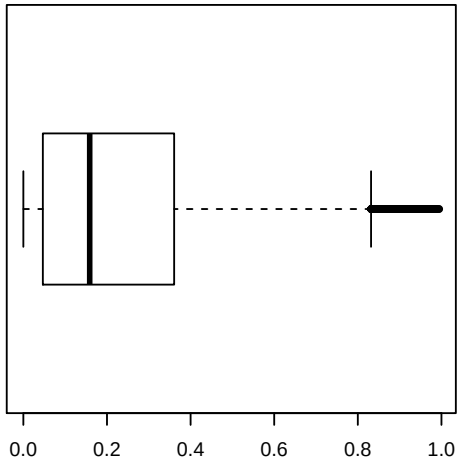
Boxplot, distribuzioni asimmetriche, outliers

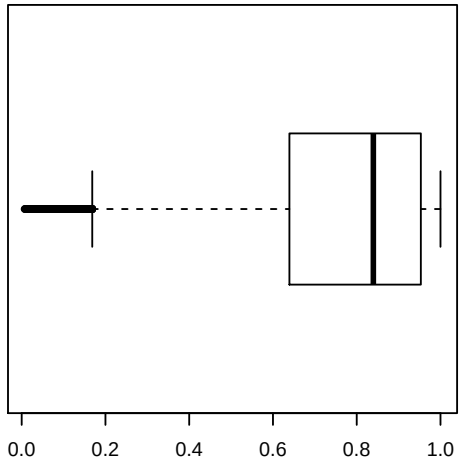












Covarianza e correlazione lineari

- La covarianza misura *simmetricamente* come due variabili numeriche X e Y **varino insieme linearmente**. Essa è la media del prodotto degli scarti di X da μ_X e di Y da μ_Y:

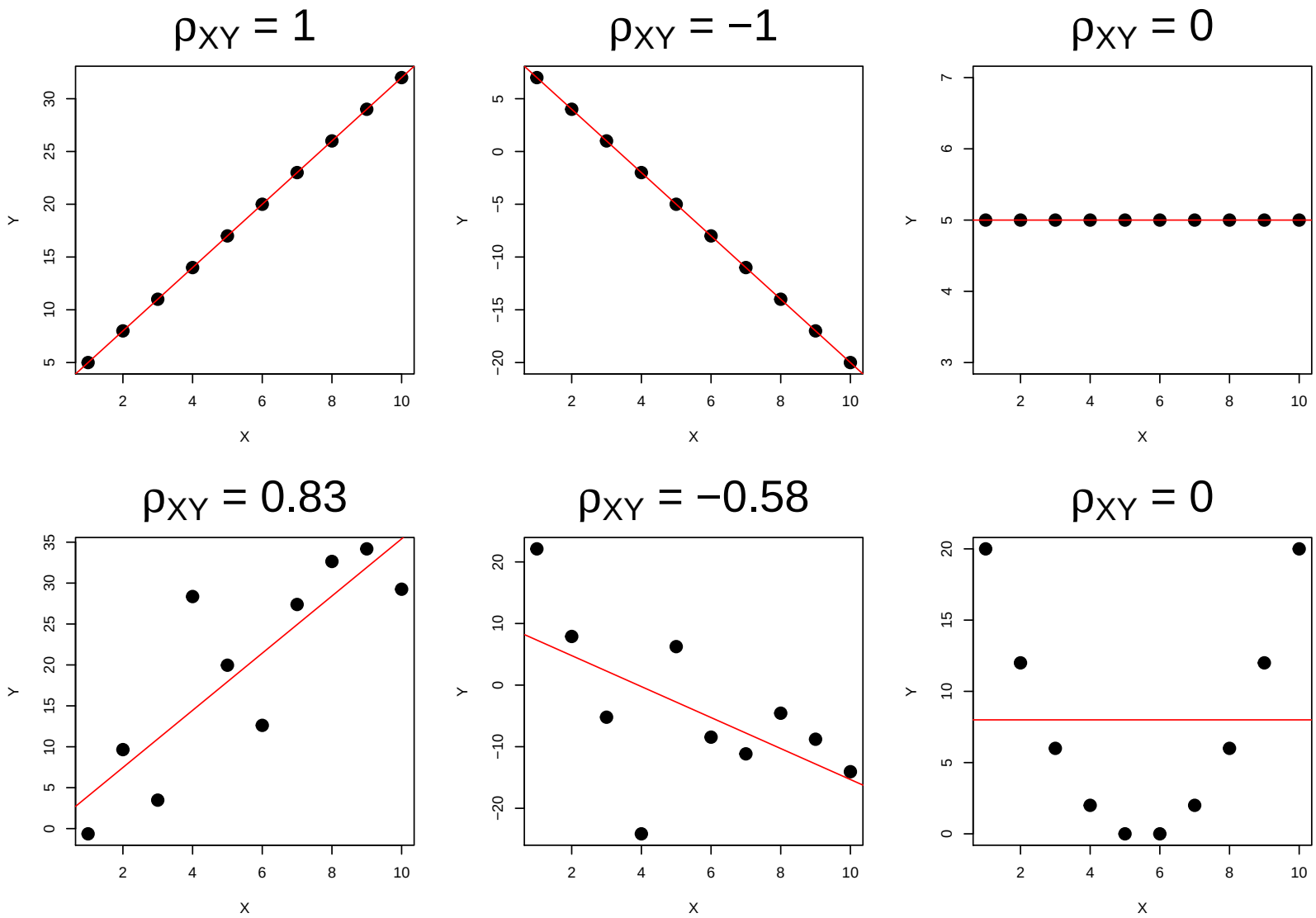
$$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)$$

- La correlazione lineare ρ_{XY} è una normalizzazione tra -1 e 1 della covarianza e pertanto ne conserva il segno. Essa è ottenuta dividendo σ_{XY} per il prodotto delle deviazioni standard di X e Y:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y}$$

- Un ρ_{XY} uguale a 1 [-1] indica perfetta correlazione lineare positiva [negativa]. Se $\rho_{XY} = 0$ vi è assenza di correlazione **lineare**.
  Non è comunque escludibile un'associazione non lineare (non monotòna) tra X e Y come nel caso di un legame parabolico.

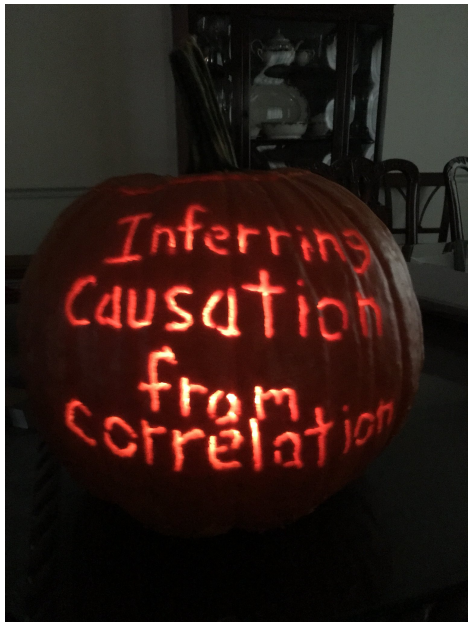
Esempi di correlazione lineare

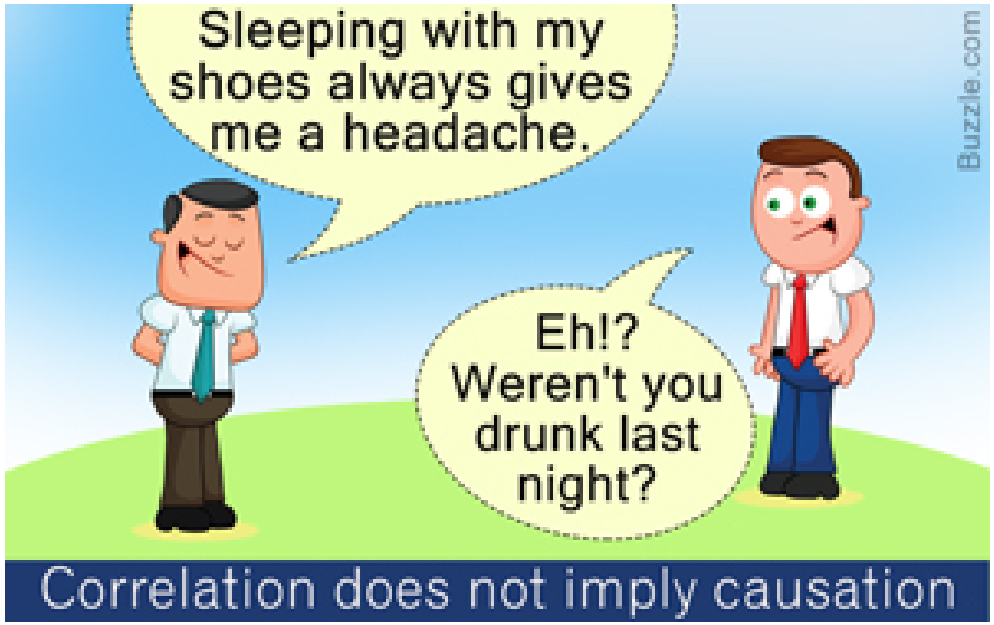


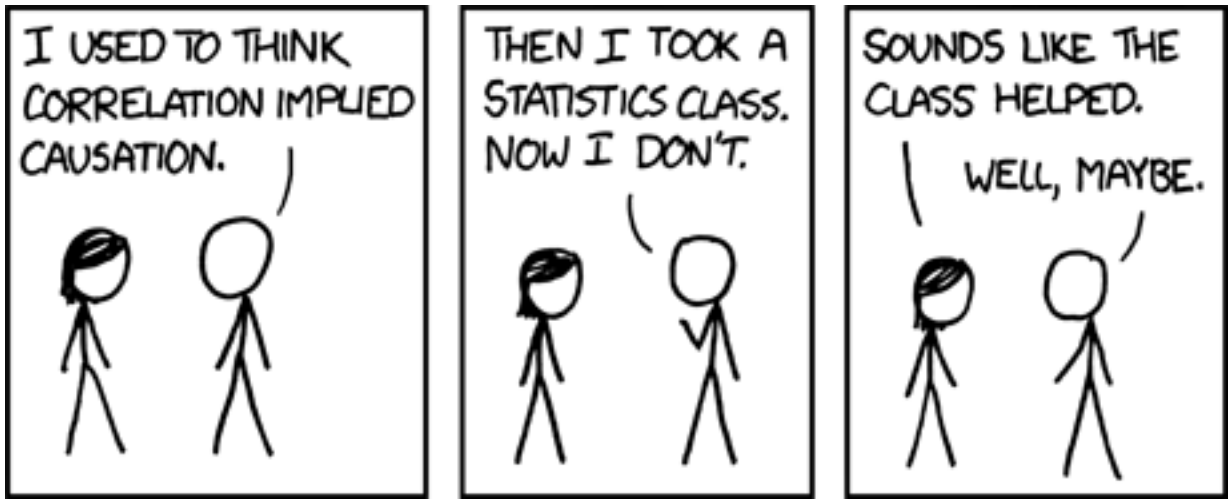
Correlazione e causalità

- La correlazione è una condizione di interdipendenza tra X e Y, senz'altro meno forte di un nesso di causalità.
- Affinchè si possa dire che X abbia un *effetto causale* su Y bisogna che:
  - le due variabili siano **correlate**,
  - vi sia un esperimento controllato in cui X **preceda** Y,
  - nell'esperimento, bisogna controllare/**escludere** possibili effetti su Y derivanti da **fattori confondenti**.
- A volte una variabile Z può causare entrambe X e Y. Esempio: correlazione positiva tra numero di pompieri (X) e numero di vittime (Y) in un incendio. Ma ciò è dovuto alle dimensioni dell'incendio (Z), non certo ad un nesso causa–effetto tra X e Y.
- Soprattutto nelle serie storiche, si osservano spesso dei pattern molto simili di variabili che non hanno nulla a che fare l'una con l'altra. Esempi su questo link: http://www.tylervigen.com/spurious-correlations.

Correlation does not imply causation







Statistica inferenziale

Concetti base di probabilità

- Probabilità dell'evento E compresa tra 0 e 1:
  - $\mathbf{P}(E) = 0 \Longrightarrow E$ è (quasi) impossibile
  - $\mathbf{P}(E) = 1 \Longrightarrow E$ è (quasi) certo
- Lo spazio campionario Ω è l'insieme di tutti gli eventi elementari
- Interpretazione *classica* della probabilità di E: $\mathbf{P}(E) = \frac{\text{n. casi favorevoli}}{\text{n. casi possibili}}$
- Dati due eventi E ed F:
  - $\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F) - \mathbf{P}(E \cap F)$
  - E ed F incompatibili $\Longleftrightarrow \mathbf{P}(E \cap F) = 0$
  - $\mathbf{P}(E \big| F) = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(F)}$, dunque $\mathbf{P}(E \cap F) = \mathbf{P}(E \big| F) \times \mathbf{P}(F)$
  - E ed F indipendenti $\Longleftrightarrow \mathbf{P}(E \big| F) = \mathbf{P}(E)$ e $\mathbf{P}(F \big| E) = \mathbf{P}(F)$
- Data una partizione di Ω formata dagli n eventi $E_1, \ldots, E_n$:
  - $\mathbf{P}(F) = \sum_{j=1}^{n} \mathbf{P}(F \cap E_j) = \sum_{j=1}^{n} \left[\mathbf{P}(F \big| E_j) \times \mathbf{P}(E_j) \right]$
  - $\mathbf{P}(E_j \big| F) = \frac{\mathbf{P}(F \cap E_j)}{\mathbf{P}(F)} = \frac{\mathbf{P}(F \Big| E_j) \times \mathbf{P}(E_j)}{\sum_{h=1}^{n} \left[\mathbf{P}(F \Big| E_h) \times \mathbf{P}(E_h) \right]}$

Variabili casuali – pmf, cdf

Una variabile casuale è una variabile i cui possibili valori sono gli outcome numerici di un fenomeno aleatorio, ciascuno avente probabilità nota.

Variabili casuali discrete

Le v.c. discrete assumono un insieme finito di valori (es: 0 e 1 rispettivamente per insuccesso e successo) o un'infinità numerabile di valori (es: numero di clienti in un bar).

- La `pmf` (funzione di probabilità) di una v.c. discreta X in un punto x indica la probabilità che X assuma valore x (es: per il lancio di un dado, $\mathbf{P}(X = 3)$ indica la probabilità che esca la faccia 3).
- La `cdf` (funzione di ripartizione) di una v.c. discreta X in un punto x è la somma dei valori della `pmf` fino a x (es: per il lancio di un dado, $\texttt{cdf}(3)$ indica la probabilità che esca la faccia 1 o 2 o 3).

Variabili casuali – pdf, cdf

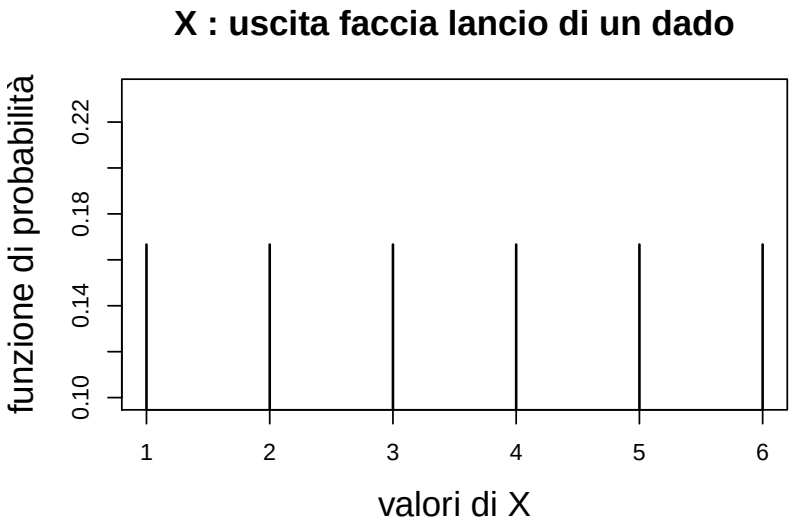
Una variabile casuale è una variabile i cui possibili valori sono gli outcome numerici di un fenomeno aleatorio, ciascuno avente probabilità nota.

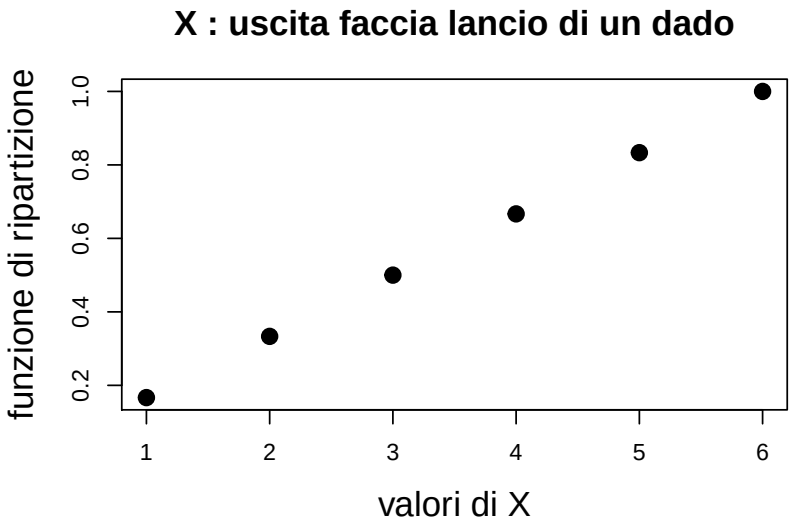
Variabili casuali continue

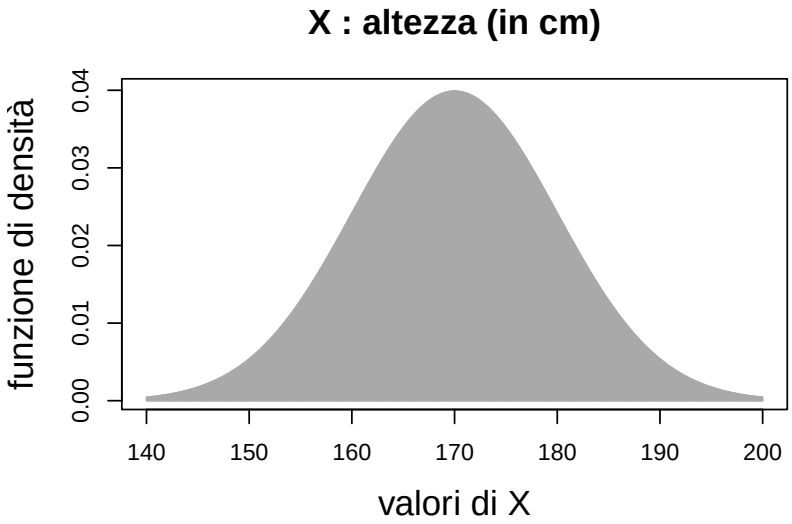
Le v.c. continue assumono un insieme infinito di valori in quanto misurate su scala continua (es: peso, altezza, distanza, ecc. . .).

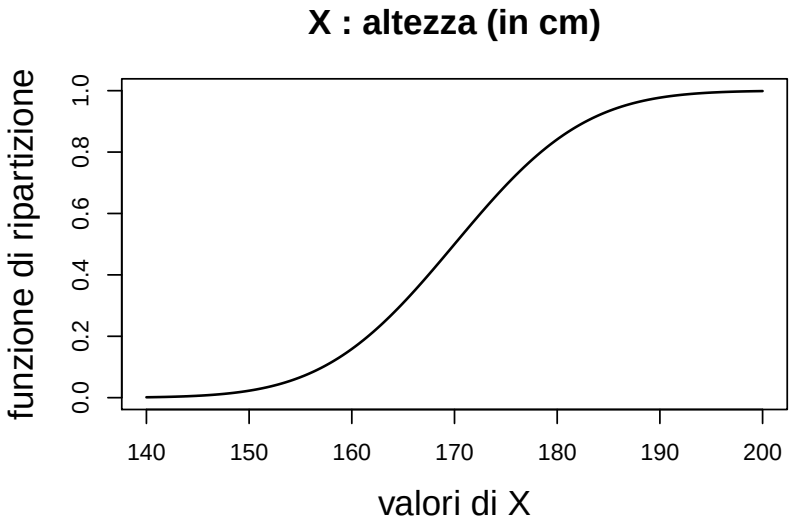
- La `pdf` (funzione di densità di probabilità) di una v.c. continua X in un punto x è l'equivalente della `pmf` nel caso continuo.
  Essa esprime solo un densità e non una "reale" probabilità, in quanto nel continuo ogni punto ha probabilità zero.
- La `cdf` (funzione di ripartizione) di una v.c. continua X in un punto x è l'integrale della `pdf` da $-\infty$ fino a x
  (es: per l'altezza, $\texttt{cdf}(170 \text{ cm}) \equiv \mathbf{P}(X \leq 170 \text{ cm})$).

Variabili casuali – `pmf`, `pdf`, `cdf` (esempi grafici)









Variabili casuali discrete – v.c. binomiale

Variabile casuale di Bernoulli

La v.c. di Bernoulli X descrive una prova con outcome dicotomico (1 = successo, 0 = insuccesso) ed ha come unico parametro $\pi = \mathbf{P}(X = 1)$. Formalmente: $X \sim Ber(\pi)$

Variabile casuale binomiale

La v.c. binomiale X è la somma di n v.c. bernoulliane con stesso parametro π. Essa conta quindi il numero di successi x (prove aventi outcome $= 1$) in un totale di n prove indipendenti. Formalmente: $X \sim Bin(n, \pi)$

La `pmf` è uguale a:

$$\mathbf{P}(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

dove $\binom{n}{x}$, detto coefficiente binomiale, fornisce il numero delle combinazioni semplici di n elementi di classe x.

Variabili casuali discrete – v.c. multinomiale

La v.c. multinomiale è una estensione della v.c. binomiale a $G \geq 2$ gruppi (per $G = 2$ riabbiamo la v.c. binomiale). In n prove indipendenti, essa conta dunque il numero di successi x_g in ciascun gruppo $g = 1, \ldots, G$, con g avente probabilità di successo π_g. Ovviamente, $\sum_{g=1}^{G} x_g = n$ e $\sum_{g=1}^{G} \pi_g = 1$.

Formalmente: $X \sim Multinom(n, \pi_1, \ldots, \pi_g, \ldots, \pi_{G-1})$

La `pmf` è uguale a:

$$\mathbf{P}(X = \{x_1, \ldots, x_g, \ldots, x_G\}) = \binom{n}{x_1, \ldots, x_g, \ldots, x_G} \prod_{g=1}^{G} \pi_g^{x_g}$$

dove $x_G = n - \sum_{g=1}^{G-1} x_g$, $\pi_G = 1 - \sum_{g=1}^{G-1} \pi_g$ e $\binom{n}{x_1,\ldots,x_g,\ldots,x_G} = \frac{n!}{x_1! \cdots x_g! \cdots x_G!}$ è detto coefficiente multinomiale e conta il numero di possibili sequenze risultanti nel vettore di successi $\{x_1, \ldots, x_g, \ldots, x_G\}$.

Variabili casuali discrete – v.c. di Poisson

La v.c. di Poisson X conta il numero di eventi x in intervalli di tempo/spazio uguali e tra loro indipendenti (ossia: il verificarsi di un evento in un intervallo non incide sulla probabilità dell'evento in un altro intervallo). È dunque una variabile casuale per **dati di conteggio** (*count data*).
Essa è descritta da un unico parametro λ che coincide sia con il valore atteso $\mathrm{E}(X)$ (numero medio di eventi nell'intervallo) che con la varianza $\mathrm{Var}(X)$ (misura di allontanamento dal valore atteso). Formalmente: $X \sim Poi(\lambda)$.
La `pmf` è uguale a:

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

La v.c. di Poisson è anche una buona approssimazione della v.c. binomiale per un π molto piccolo e un numero di prove bernoulliane n molto grande. Pertanto è chiamata anche *legge degli eventi rari*.

Variabili casuali discrete – v.c. binomiale negativa

L'obiettivo *standard* della v.c. binomiale negativa è il conteggio del numero di fallimenti x necessari prima di osservare l'r–esimo successo in prove bernoulliane caratterizzate da una stessa probabilità di successo π. Formalmente: $X \sim NB(\pi, r)$. La `pmf` è uguale a:

$$\mathbf{P}(X = x) = \binom{r + x - 1}{x} \pi^r (1 - \pi)^x$$

Nel modellare i *count data*, la v.c. binomiale negativa è utile per gestire l'*overdispersion* (varianza dei conteggi osservati superiore al valore atteso). Infatti, se introduciamo aleatorietà nel λ di una v.c. di Poisson Y rendendo λ una v.c. gamma di parametri k e β, la `pmf` risultante (dimostrazione al paragrafo 4 di queste note) sarà quella della v.c. binomiale negativa $Y \sim NB\left(\pi = \frac{1}{\beta+1}, r = k\right)$:

$$\mathbf{P}(Y = y) = \binom{k + y - 1}{y} \left(\frac{1}{\beta + 1}\right)^k \left(1 - \frac{1}{\beta + 1}\right)^y$$

Variabili casuali continue – v.c. normale (gaussiana)

La v.c. normale ha una forma campanulare, in quanto i valori si addensano vicino alla media e decrescono simmetricamente verso le code. Essa è descritta dai parametri media e varianza: $X \sim N(\mu, \sigma^2)$.
La sua `pdf` in x è la seguente:

$$\texttt{pdf}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Il 68.3%, 95.4%, 99.7% dei valori sono concentrati rispettivamente in 1, 2, 3 deviazioni standard da μ.
- La **standardizzazione** di X: $\mathcal{Z} = \frac{X-\mu}{\sigma}$ fa sì che $\mu_{\mathcal{Z}} = 0$ e $\sigma^2_{\mathcal{Z}} = 1$. Ciò è conveniente soprattutto perché $\mathbf{P}(X \leq x) = \mathbf{P}(\mathcal{Z} \leq \frac{x-\mu}{\sigma})$.
- La normale ricopre grande importanza soprattutto per il **Teorema del Limite Centrale** (TLC): *una combinazione lineare di variabili casuali indipendenti ed identicamente distribuite converge ad una gaussiana*.

Nozione di test statistico e relativi errori

- Eseguire un test statistico vuol dire avvalersi di un campione per verificare un'ipotesi riguardante una caratteristica della popolazione.
- L'ipotesi da cui si parte è detta *ipotesi nulla* (H_0), cui viene contrapposta un'ipotesi alternativa (H_1) che prendiamo in considerazione solo se vi è forte evidenza empirica.
- In un test statistico si possono commettere 2 errori:
  - rifiutare H_0 quando H_0 è vera (**errore di primo tipo**)
  - non rifiutare H_0 quando H_1 è vera (**errore di secondo tipo**)
- La probabilità di commettere l'errore del I tipo è detta α ed è quella che di solito ci preme tenere ad un valore basso (max 10%).
- La probabilità di commettere l'errore di II tipo è detta β ed è quella che solitamente non controlliamo e che decresce al crescere dell'ampiezza campionaria.
- La *potenza del test* è la probabilità di rifiutare correttamente H_0 ed è uguale a $1 - \beta$.

Passi di un test statistico e *p*-value

Seguono i passi di un test statistico (es: verificare che un parametro θ sia uguale a θ_0):

1. Stabilire le ipotesi:
   - $H_0 : \theta = \theta_0$
   - $H_1 : \theta \neq \theta_0$ (test a due code) oppure $H_1 : \theta \gtrless \theta_0$ (test a una coda)
2. Fissare α.
3. Determinare la **statistica test**, ovvero una funzione del campione che, partendo da assunzioni iniziali, segue (o è ben approssimata da) una distribuzione di probabilità nota sotto H_0.
4. Calcolare il **valore osservato** della statistica test e determinare quanto sia probabile osservare un valore così estremo o più estremo sotto l'ipotesi nulla. Questa probabilità è detta ***p*-value**.
5. Se il *p*-value è **minore** dell'α fissato, allora c'è davvero **evidenza di rifiuto** di H_0.

Confronto tra gruppi in media

- Date due popolazioni M ed F (es: maschi e femmine), vogliamo confrontare i rispettivi redditi medi annuali μ_M e μ_F.
- M ed F hanno rispettive varianze σ^2_M e σ^2_F, a noi comunque ignote.
- Estraiamo un campione di n_M unità da M ed uno di n_F unità da F.
- $H_0 : (\mu_M - \mu_F) = 0$ $\qquad$ $H_1 : (\mu_M - \mu_F) \neq$ oppure $>$ oppure < 0
- Se le due ampiezze campionarie non sono esigue, possiamo sfruttare il TLC senza il bisogno di assumere gaussianità delle popolazioni.
- Calcolate le medie campionarie $\bar{x}_M$ e $\bar{x}_F$ e le varianze campionarie corrette s^2_M e s^2_F, ove $s^2_j = (n_j - 1)^{-1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$, $j = \{M, F\}$, determiniamo il valore osservato della seguente statistica test:

$$T = \frac{\bar{x}_M - \bar{x}_F}{\sqrt{\frac{s^2_M}{n_M} + \frac{s^2_F}{n_F}}}$$

- T si distribuisce come una *t* di Student con g.d.l. determinabili tramite *Welch–Satterthwaite equation* (dettagli qui).

Confronto tra gruppi in proporzione di successi

- Date due popolazioni bernoulliane M ed F (es: maschi e femmine), vogliamo confrontare le rispettive percentuali di adesione ad una campagna antidroga, π_M e π_F.
- Estraiamo un campione di n_M unità da M ed uno di n_F unità da F.
- $H_0 : (\pi_M - \pi_F) = 0$ $\qquad$ $H_1 : (\pi_M - \pi_F) \neq$ oppure $>$ oppure < 0
- Per $j = \{M, F\}$, calcoliamo il numero campionario di successi $S_j = \sum_{i=1}^{n_j} x_{ij}$, la proporzione campionaria di successi $\hat{\pi}_j = \frac{S_j}{n_j}$ e uno stimatore congiunto delle due proporzioni campionarie $\hat{\pi} = \frac{S_M + S_F}{n_M + n_F}$.
- La seguente statistica test $\mathcal{Z}$ è asintoticamente distribuita come una normale standardizzata (n_M ed n_F non piccoli):

$$\mathcal{Z} = \frac{\hat{\pi}_M - \hat{\pi}_F}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_M} + \frac{1}{n_F}\right)}} \to N(0,1)$$

Associazione tra variabili categoriche

- Agli ospiti di tre diversi hotel situati nelle isole tropicali è stato chiesto tramite sondaggio anonimo se ritorneranno o meno.

Ritorna/Hotel	Golden Palm	Palm Royale	Palm Princess	TOT riga
SI	128	199	186	513
NO	88	33	66	187
TOT col.	216	232	252	700

- In caso gli hotel non differiscano in termini di consensi, la percentuale di ospiti soddisfatti non dovrebbe variare da hotel a hotel.
- Se così fosse (**caso di indipendenza** tra le due variabli categoriche "Ritorna" e "Hotel"), già conoscendo il totale degli ospiti soddisfatti (513) ed il numero di ospiti in ciascun hotel ($\{216, 232, 252\}$), potremmo determinare i valori nelle celle della distribuzione congiunta delle due variabili (**conteggi attesi**).

Associazione tra variabili categoriche – Test χ^2

- In caso di indipendenza, la probabilità di soddisfazione sarebbe costante e uguale a $\pi_{\text{SI}} = \frac{n_{1\bullet}}{n_{\bullet\bullet}} = \frac{513}{700} = 0.733$.
- Per esempio, gli ospiti soddisfatti di Golden Palm dovrebbero essere $\hat{n}_{11} = n_{\bullet 1} \times \pi_{\text{SI}} = 216 \times 0.733 = 158$.
- L'obiettivo è capire quanto la tabella osservata si discosti da quella attesa sotto ipotesi di indipendenza. A tal fine, per ciascuna cella calcoliamo la distanza chi-quadrato degli n_{ij} osservati dagli $\hat{n}_{ij}$ attesi:

$$d_{ij} = \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

- L'ipotesi nulla del test è l'indipendenza, cioè che le d_{ij} siano 0.
- La statistica test è uguale alla somma delle distanze trovate e si distribuisce come una v.c. Chi quadro con $g = (r-1) \times (c-1)$ g.d.l., ove r è il numero di righe della tabella e c il numero di colonne:

$$\sum_{i=1}^{r} \sum_{j=1}^{c} d_{ij} \sim \chi^2(g)$$

Associazione tra variabili categoriche – Test χ^2 – esempio

- Determiniamo la tabella delle d_{ij} per il nostro esempio pratico:

Ritorna/Hotel	Golden Palm	Palm Royale	Palm Princess
SI	5.8	4.94	0.009
NO	15.9	13.55	0.026

- Il valore osservato della statistica test χ^2_{oss} è uguale a $\sum d_{ij}$, ossia 40.2.
- χ^2_{oss} va ricercato su una distribuzione χ^2 con $(2-1)(3-1) = 2$ g.d.l.
- Il test è ad una coda, essendo la v.c. χ^2 sempre positiva.
- Il *p*-value è $\mathbf{P}(\chi^2 \geq \chi^2_{oss})_{H_0} \approx 0$, dunque rifiutiamo H_0 per qualsiasi α: **vi è associazione** tra le due variabili categoriche "Ritorna" e "Hotel".

Alternative al test χ^2 per una tabella 2×2

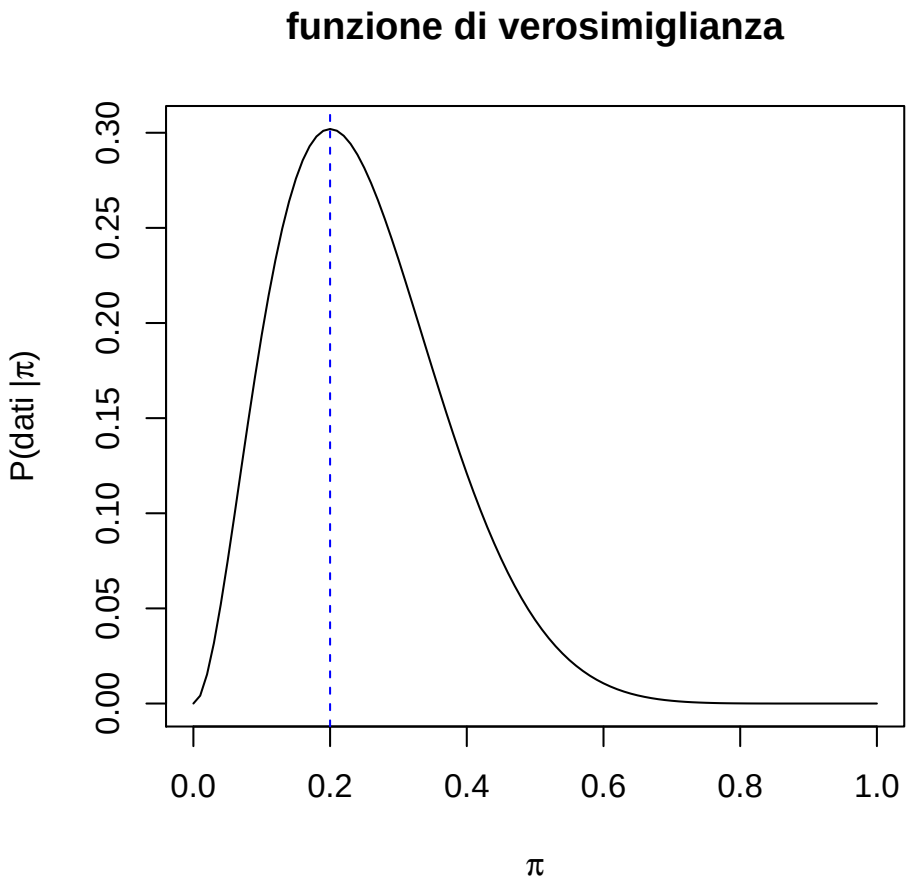
In una tabella 2×2 in cui non tutti gli n_{ij} siano superiori a 5, il test χ^2 è sconsigliato e viene generalmente sostuito dal **test esatto di Fisher**.
Ulteriori dettagli qui.

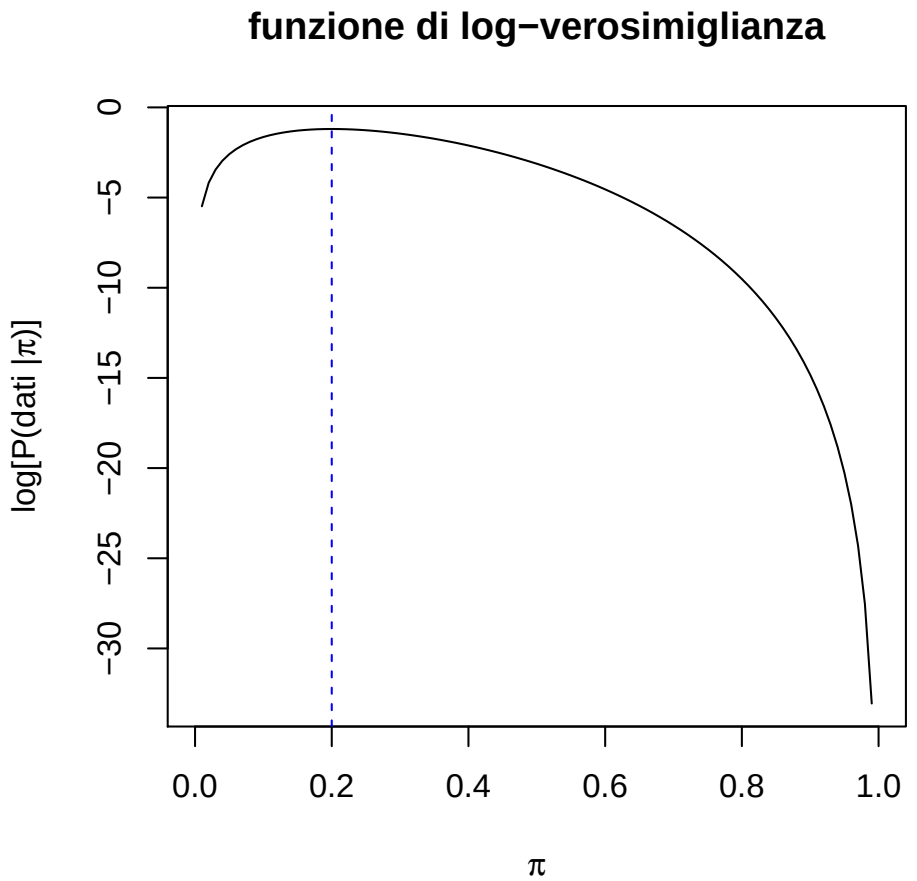
Nozione di verosimiglianza (*likelihood*)

- Finora conoscevamo la distribuzione di probabilità (es: v.c. binomiale) con i suoi parametri (es: n e π) e potevamo ottenere il valore della *funzione di probabilità* per $X = x$.
- Partiamo ora invece dai dati, ove osserviamo il valore di x (es: 2 successi) su n unità (es: $n = 10$). Conoscendo ancora la funzione di probabilità (es: v.c. binomiale), possiamo tracciare una funzione del parametro ignoto θ (nel nostro caso: $\theta = \pi \in [0, 1]$).
  Questa è la *likelihood function* di θ, in simboli $\mathcal{L}(\theta|\text{dati}) = \mathbf{P}(\text{dati}|\theta)$.
- La **stima di massima verosimiglianza** (*ML estimate*) è quel valore del parametro θ_{ML} (nel nostro caso: π_{ML}) che massimizza la probabilità di osservare i nostri dati.
- In genere, si lavora con la *log-likelihood function*, $l(.) = \log \mathcal{L}(.)$, poiché è massimizzata nello stesso punto di $\mathcal{L}(.)$ ed è più gestibile analiticamente (somme più agevoli di prodotti).

Verosimiglianza e log-verosimiglianza (esempio)

Abbiamo il nostro esempio di una popolazione bernoulliana da cui estraiamo $n = 10$ unità. Nel campione osserviamo $x = 2$ successi. La stima di massima verosimiglianza coincide con $\pi = \frac{2}{10}$, come visibile dai grafici:





Modelli lineari generalizzati (GLM)

Un modello lineare generalizzato mette in relazione una funzione del valore atteso della variabile dipendente Y con le variabili esplicative attraverso un'equazione lineare.
Esso è specificato da tre componenti:

- la componente aleatoria $(Y_1, \ldots, Y_n)$, costituita da un insieme di variabili casuali assunte reciprocamente indipendenti e con una distribuzione di probabilità appartenente alla **famiglia esponenziale**;
- la componente sistematica $(c + \sum_{j=1}^{p} \beta_j x_j)$, che specifica una combinazione lineare delle variabili esplicative nel modello;
- la funzione legame $(g(\mathrm{E}(Y_i)) = c + \sum_{j=1}^{p} \beta_j x_{ij})$, che mette in relazione la componente aleatoria e la componente sistematica del modello, specificando quale funzione g del valore atteso di Y_i dipende linearmente dalle variabili esplicative.

Inferenza sui GLM – Wald test asintotici

- I valori dei parametri di massima verosimiglianza di un GLM sono determinati tramite *iteratively reweighted least squares*.
- Questo processo di ottimizzazione restituisce anche l'*Hessiana* $(\mathbf{H})$, ossia la matrice delle derivate parziali seconde.
- Gli standard error delle stime sono ottenuti come radice quadrata degli elementi diagonali della matrice inversa dell'*Hessiana*:

$$\text{se}(\hat{\boldsymbol{\beta}}) = \sqrt{\text{diag}(\mathbf{H}^{-1})}$$

- Si può testare che β_j sia uguale ad un valore β_{j0} (es: $\beta_j = 0$ per i test di significatività) ricorrendo alla seguente statistica test asintoticamente distribuita come una normale standardizzata:

$$\frac{\hat{\beta}_j - \beta_{j0}}{\text{se}(\hat{\beta}_j)} \rightarrow \mathcal{N}(0, 1)$$

Inferenza sui GLM – *Likelihood ratio test*

- La devianza di un modello M (appartenente alla classe dei GLM) è il valore di massima log-verosimiglianza moltiplicato per -2.
- Un modello che stima tutti i parametri (modello *S*, cioè **saturated**) ha una devianza minore rispetto ad un suo *modello annidato* in cui alcuni parametri sono posti uguali a zero (modello *C*, cioè **constrained**).
- Tuttavia, il modello *C*, stimando meno parametri, è più stabile ed interpretabile. Esso ha infatti più gradi di libertà: g.d.l.$(M) = n - k$, ove n è l'ampiezza campionaria e k il numero di parametri stimati.
- Il test di rapporto tra verosimiglianze (*likelihood ratio test*) verifica che l'aumento in devianza nel passare da *S* a *C* sia giustificato dal minor numero di parametri stimati (ipotesi nulla). In caso lo sia, scegliamo *C*.
- La statistica test G si ottiene come differenza tra le devianze di *C* ed *S* e si distribuisce sotto H_0 come una v.c. Chi-quadro con g.d.l. uguali alla differenza in g.d.l. tra *C* ed *S*:

$$G = \left[-2\,\hat{l}(C)\right] - \left[-2\,\hat{l}(S)\right] \sim \chi^2\left(\text{g.d.l.}(C) - \text{g.d.l.}(S)\right)$$

Variabile casuale multinomiale – Verosimiglianza

Partiamo da una variabile $\mathbf{Y}$ con distribuzione multinomiale in cui l'unità i ha $y_{ig} = 1$ se appartiene a g e $y_{ig} = 0$ altrimenti. Poniamo $y_g = \sum y_{ig}$. Massimizzare la verosimiglianza di Y vuol dire trovare i valori delle probabilità di gruppo π_g, con $g = 1, \ldots, G$, che massimizzino la probabilità di osservare i nostri dati. Poiché il coefficiente multinomiale non contiene alcuna probabilità, il *core* della *likelihood* da massimizzare è:

$$\mathcal{L}(\pi_1, \ldots, \pi_g, \ldots, \pi_G) = \prod_{g=1}^{G} \pi_g^{y_g} = \prod_{g=1}^{G} \prod_{i=1}^{n} \pi_g^{y_{ig}} = \prod_{i=1}^{n} \mathcal{L}_i(\pi_1, \ldots, \pi_g, \ldots, \pi_G)$$

La corrispondente log-verosimiglianza da massimizzare è:

$$l(\pi_1, \ldots, \pi_g, \ldots, \pi_G) = \sum_{i=1}^{n} l_i(\pi_1, \ldots, \pi_g, \ldots, \pi_G) = \sum_{g=1}^{G} \sum_{i=1}^{n} y_{ig} \log(\pi_g)$$

Regressione logistica multinomiale – Verosimiglianza

In caso π_g sia legata a predittori $x_1, \ldots, x_p$ tramite una funzione legame appartenente alla famiglia esponenziale, avremmo un modello di regressione multinomiale. Nel caso più noto di regressione logistica multinomiale, la funzione legame è il **logit** (logaritmo degli *odds* di un gruppo g rispetto ad un altro di riferimento). L'inversa del logit è la funzione di ripartizione della distribuzione logistica ed è ciò che usiamo per la verosimiglianza:

$$\pi_g(x_1, \ldots, x_p) = \text{logit}^{-1}(c_g + \sum_{j=1}^{p} \beta_{gj} x_j) = \frac{\exp\left(c_g + \sum_{j=1}^{p} \beta_{gj} x_j\right)}{\sum_{h=1}^{G} \left[\exp\left(c_h + \sum_{j=1}^{p} \beta_{hj} x_j\right)\right]}$$

La log-verosimiglianza da massimizzare è quindi la seguente:

$$l(\substack{c_g, \beta_{g1}, \ldots, \beta_{gp} \\ \text{per i G gruppi}}) = \sum_{g=1}^{G} \sum_{i=1}^{n} y_{ig} \log(\pi_g(x_{i1}, \ldots, x_{ip}))$$

Regressione logistica multinomiale – Interpretazione

Nel modello logistico multinomiale i coefficienti di uno dei G gruppi (detto *baseline*) non vanno stimati. Essi sono posti uguali a zero. Nel caso la *baseline* sia il gruppo 1, il modello per il gruppo g è il seguente:

$$\text{logit}(\pi_g) = \log\left(\frac{\pi_g}{\pi_1}\right) = c_g + \sum_{j=1}^{p} \beta_{gj} x_j$$

- Dunque, l'interpretazione **in termine di logit** è la seguente:
  Per una variazione unitaria nel predittore x_j, il logaritmo degli *odds* del gruppo g rispetto al gruppo 1 varia di β_{gj}, tenendo costanti gli altri predittori.
- L'interpretazione **in termini di rapporto tra odds** è la seguente:
  Per una variazione unitaria nel predittore x_j, gli *odds* del gruppo g rispetto al gruppo 1 variano di un fattore pari a $\exp(\beta_{gj})$, cioè variano del $100 \times [\exp(\beta_{gj}) - 1]\%$, tenendo costanti gli altri predittori.

Variabile casuale di Poisson – Verosimiglianza

Partiamo da una v.c. $\mathbf{Y}$ con distribuzione di Poisson.

- Il parametro rispetto al quale massimizzare la funzione di verosimiglianza è λ, ossia il numero medio di eventi nell'intervallo.
- Essendo ogni y_i indipendente, avremo la seguente *likelihood*:

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda}\, \lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\, \lambda^{\sum_{i=1}^{n} y_i}}{y_1! \cdots y_n!}$$

  il cui denominatore non contiene il parametro e può essere trascurato.
- Ne consegue che il *core* della log-verosimiglianza è:

$$l(\lambda) = -n\lambda + \log(\lambda) \sum_{i=1}^{n} y_i$$

  il cui valore massimizzante è $\lambda_{ML} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

Regressione di Poisson – Verosimiglianza

- La regressione di Poisson è un GLM in cui il numero medio di eventi è spiegato da predittori tramite una funzione legame logaritmica:

$$\log(\lambda_i) = c + \sum_{j=1}^{p} \beta_j x_{ij}$$

- La log-verosimiglianza è la seguente:

$$l(c, \beta_1, \ldots, \beta_p) = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} [y_i \log(\lambda_i)]$$

che, sostituendo, risulta essere:

$$l(c, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left[-\exp\left(c + \sum_{j=1}^{p} \beta_j x_{ij} \right) + y_i \left(c + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right]$$

Regressione di Poisson - Interpretazione e probabilità predette

Interpretazione

L'interpretazione nel modello di Poisson deve considerare il fatto che la funzione legame è logaritmica (modello *log-lineare*):

- Per una variazione unitaria nel predittore x_j, il conteggio atteso varia di un fattore pari a $\exp(\beta_j)$, ossia del $100 \times [\exp(\beta_j) - 1]\%$, tenendo costanti gli altri predittori.

Probabilità predette

Possiamo calcolare la distribuzione dei conteggi attesi per determinati valori dei predittori $\{x_{1_0}, \ldots, x_{j_0}, \ldots, x_{p_0}\}$.

- Per un determinato conteggio m, la `pmf` condizionata è data da:

$$\mathbf{P}(m | x_{1_0}, \ldots, x_{j_0}, \ldots, x_{p_0}) = \frac{\mathrm{e}^{-\hat{\lambda}_0} \hat{\lambda}_0^m}{m!}$$

ove $\hat{\lambda}_0 = \hat{c} + \sum_{j=1}^{p} \hat{\beta}_{j_0} x_{ij_0}$.

Regressione di Poisson – *Overdispersion*

- Il modello di Poisson è descritto dal solo valor medio degli eventi nell'intervallo, condizionato ai valori dei predittori:
  $\lambda = g^{-1}\left(c + \sum_{j=1}^{p} \beta_j x_j\right)$, ove $g^{-1}(.) = \exp(.)$.
- In tale modello, **si assume** che λ rappresenti anche la varianza condizionata ai predittori.
- Nei casi reali, spesso osserviamo *overdispersion*, ossia una varianza dei conteggi osservati superiore alla media, la qual cosa incide soprattutto sulla affidabilità degli standard error (sottostimati).
- Si cerca spesso di ovviare all'*overdispersion* rimpiazzando la *likelihood* della Poisson con quella della v.c. binomiale negativa, in quanto sappiamo che quest'ultima può essere vista come una v.c. di Poisson con parametro λ non più fisso, ma distribuito come una v.c. gamma.
- L'aleatorietà di λ introduce **ulteriore variabilità** nel modello.

Regressione binomiale negativa

- Partiamo dall'introdurre un errore $\boldsymbol{\varepsilon}_i$ nel modello per la media:

$$\tilde{\lambda}_i = \exp\left(c + \sum_{j=1}^{p} \beta_j x_{ij} + \boldsymbol{\varepsilon}_i\right) = \exp\left(c + \sum_{j=1}^{p} \beta_j x_{ij}\right) \exp(\boldsymbol{\varepsilon}_i) = \lambda_i \underbrace{\exp(\boldsymbol{\varepsilon}_i)}_{\boldsymbol{\delta}_i}$$

- Si assume comunemente che $\boldsymbol{\delta}_i \sim Gamma(\boldsymbol{\alpha}^{-1}, \boldsymbol{\alpha}^{-1})$, con $\boldsymbol{\alpha} > 0$, ove $\boldsymbol{\alpha}$ è detto **parametro di dispersione**.
- La `pmf` condizionata è ora quella di una v.c. binomiale negativa:

$$\mathbf{P}(y_i | x_1, \ldots, x_p) = \frac{\Gamma(y_i + \boldsymbol{\alpha}_i^{-1})}{y_i! \, \Gamma(\boldsymbol{\alpha}_i^{-1})} \left(\frac{\boldsymbol{\alpha}_i^{-1}}{\boldsymbol{\alpha}_i^{-1} + \lambda_i}\right)^{\boldsymbol{\alpha}_i^{-1}} \left(\frac{\lambda_i}{\boldsymbol{\alpha}_i^{-1} + \lambda_i}\right)^{y_i}$$

- Inoltre risulta che:
  - $\mathrm{E}(y_i | x_1, \ldots, x_p) = \lambda_i$
  - $\mathrm{Var}(y_i | x_1, \ldots, x_p) = \lambda_i + \boldsymbol{\alpha}\, \lambda_i^2$
- L'addendo $\boldsymbol{\alpha}\, \lambda_i^2$ è l'ulteriore variabilità introdotta nel modello binomiale negativo rispetto al modello di Poisson.
- Possiamo verificare l'ipotesi nulla che $\alpha = 0$ ($\rightarrow$ ritorno al modello di Poisson) tramite *Likelihood ratio test* o Wald test.

Regressione *classica* (dati continui) – Stima

- Il modello di regressione lineare è un GLM con componente aleatoria gaussiana e con funzione legame $g(\mathrm{E}(Y)) = \mathrm{E}(Y)$ (funzione identità).
- È ampiamente dimostrato che i valori dei coefficienti di massima verosimiglianza di questo GLM sono gli stessi che **minimizzano** RSS, ossia la somma delle deviazioni (*residui* ε_i) al quadrato dei valori osservati di Y dalle le loro medie condizionate ai predittori:

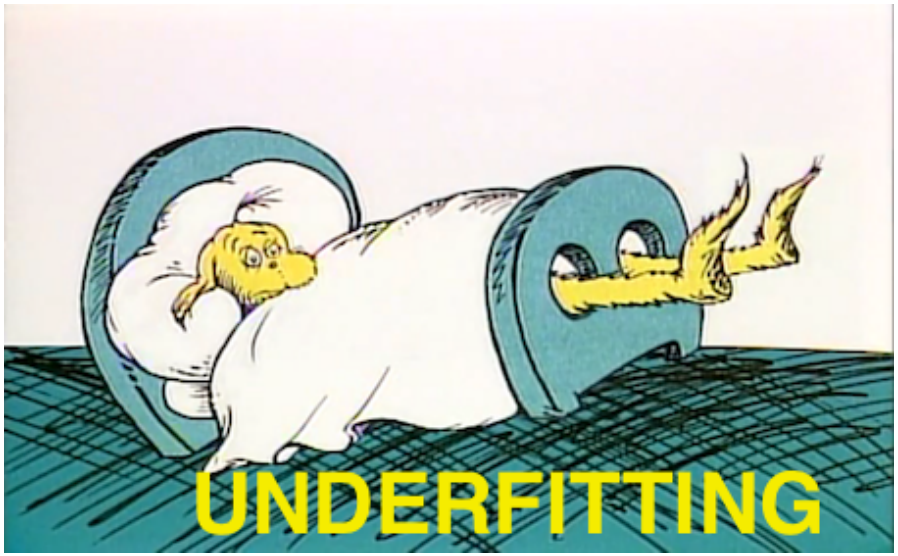
$$\mathrm{RSS} = \sum_{i=1}^{n} \left[\underbrace{Y_i - \mathrm{E}(Y_i)}_{\varepsilon_i} \right]^2 = \sum_{i=1}^{n} \left[Y_i - \left(c + \sum_{j=1}^{p} \beta_j \, x_{ij} \right) \right]^2$$

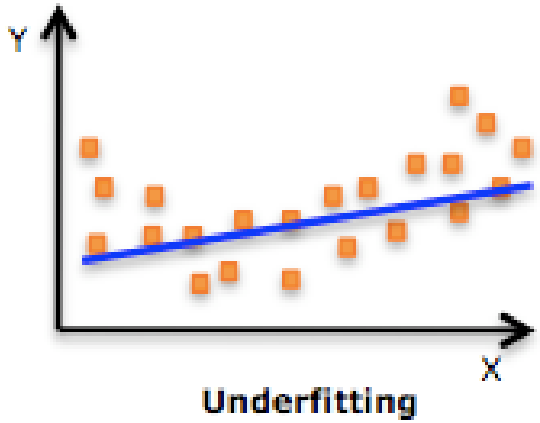
- Lo stimatore ordinario dei minimi quadrati (*OLS*) è lo stimatore lineare non distorto più efficiente (stimatore *BLUE*):
  - **non distorto**: il suo valore atteso coincide con il parametro della popolazione: $\mathrm{E}(\hat{\beta}_j) = \beta_j$;
  - **più efficiente**: dei lineari non distorti, è quello con varianza minore.

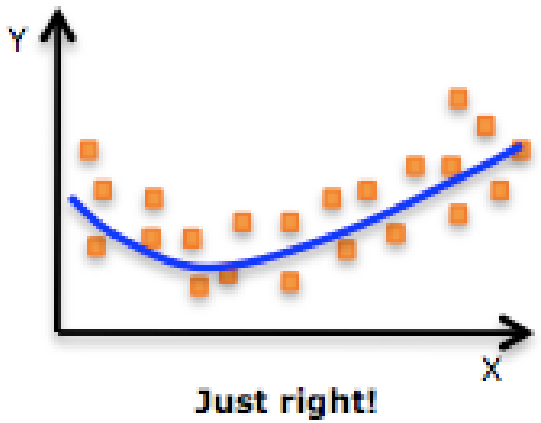
Regressione *classica* (dati continui) – Principali assunzioni

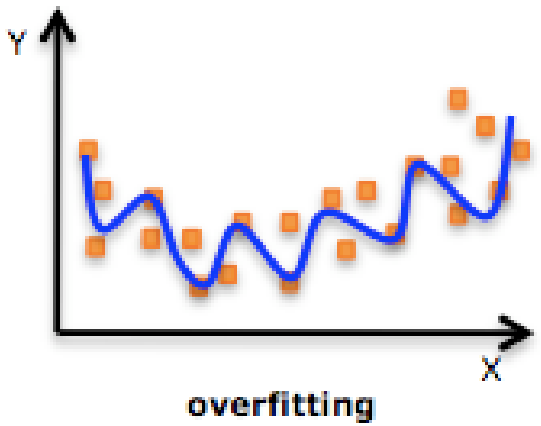
- Le assunzioni da cui partiamo per il modello di regressione $\mathrm{E}(Y_i|x_{i1}, \ldots, x_{ip})$ vengono verificate sui residui ε_i.
- Iniziamo con il dire che l'ipotesi di **normalità** dei residui è davvero necessaria ai fini dell'inferenza sui coefficienti solo nel caso l'ampiezza campionaria sia esigua, cioè quando il **TLC non è applicabile**.
- Un'assunzione fondamentale è quella della **linearità**, cioè che $\mathrm{E}(Y_i|x_{i1}, \ldots, x_{ip})$ sia una combinazione lineare dei predittori.
- Un'altra assunzione è quella dell'**omoschedasticità**, cioè che la varianza di Y sia costante ed indipendente dai predittori.
- L'assunzione di **non collinearità tra predittori** è essenziale ai fini della stabilità delle stime e dell'interpretazione degli effetti.
- Un'assunzione da considerare è che le Y_i siano v.c. indipendenti, il che implica che i residui siano tra loro **incorrelati**.

Underfitting ed overfitting









Variable selection – Overfitting e criteri di informazione

- La *variable selection* è utile quando si hanno molti predittori e bisogna predire nuove osservazioni scongiurando il rischio di **overfitting**.
- L'overfitting è la conseguenza di un modello troppo complesso (troppi parametri stimati) che ha un *fit* **troppo performante** sui dati su cui è stimato e **poco performante** su nuovi dati.
- La selezione dei predittori in un GLM può essere eseguita avvelendoci dei cosiddetti **criteri d'informazione** (*IC*, es: **AIC**, **BIC**, **HQC**).
- Un *IC* stima un compromesso tra la **bontà** (in termini di *fit*) e la **complessità** (in termini di numero di parametri k) del modello M. Infatti, la formula dell'*IC* di M è composta come segue:

$$IC(M) = \underbrace{-2\,\hat{l}(M)}_{\text{devianza di } M} + \underbrace{f(k)}_{\text{funzione crescente del n. di parametri}}$$

- La penalizzazione $f(k)$ è ciò che distingue i criteri di informazione.
- Scelto un *IC*, il modello con il **valore più basso è da preferire**.

Variable selection – *Best subset* e procedure *step-wise*

Best subset

- La procedura *best subset* calcola il valore dell'*IC* per ogni possibile sottoinsieme di predittori di ogni possibile cardinalità.
- Determina il migliore sottoinsieme, ma è sconveniente in termini di complessità computazionale (troppe combinazioni da provare).

Procedure *step-wise*

- La procedura **forward** parte dal *modello nullo* (contenente la sola costante) ed aggiunge, passo dopo passo, il predittore che dà il maggiore decremento dell'*IC*. Il processo termina quando nessun altro predittore può far ulteriormente decrescere l'*IC*.
- La procedura **backward** parte dal modello contenente tutti i predittori ed elimina, passo dopo passo, il predittore per cui vi è il maggiore decremento dell'*IC*. Il processo termina quando nessun'altra eliminazione può far ulteriormente decrescere l'*IC*.

Il trade-off tra *bias* e varianza

- Data una popolazione P, possiamo estrarre un campione di apprendimento $\mathcal{A}$ (osservazioni su cui stimare la funzione $\mathsf{E}(y) = f(x)$) ed un campione di *test* $\mathcal{T}$ (osservazioni *nuove* su cui testare $\hat{f}$).
- L'inaccuratezza di un modello è misurabile su una osservazione di *test* (x_0, y_0) tramite l'MSE (errore quadratico medio), ossia la differenza al quadrato tra l'osservazione reale y_0 e quella predetta $\hat{f}(x_0)$.
- Il valore atteso di $\mathsf{MSE}(x_0, y_0)$ è l'MSE medio nel punto (x_0, y_0) che risulterebbe dalle stime di f su $\mathsf{M} \to \infty$ campioni $\mathcal{A}_1, \ldots, \mathcal{A}_{\mathsf{M}}$ di P:

$$\mathsf{E}\underbrace{[y_0 - \hat{f}(x_0)]^2}_{\mathsf{MSE}(x_0, y_0)} = \underbrace{\mathsf{Var}\left(\hat{f}(x_0)\right)}_{\substack{\text{variabilità di } \hat{f} \\ \text{tra gli } \mathcal{A}_{\mathsf{M}} \text{ campioni}}} + \underbrace{[\mathsf{Bias}\left(\hat{f}(x_0)\right)]^2}_{\substack{\text{distorsione di } \hat{f} \\ \text{rispetto a } f}} + \mathsf{Var}(\underbrace{\varepsilon}_{\substack{\text{errore} \\ \text{irriducibile}}})$$

- Quando la varianza cresce, il *bias* decresce, e viceversa.
- L'obiettivo è minimizzare l'MSE atteso tramite la stima di un modello che dia il miglior compromesso (*trade-off*) tra *bias* e varianza di $\hat{f}$.

Regressione regolarizzata – Modelli di Ridge e Lasso

- La *variable selection* può evitare l'overfitting in un modello di regressione con molti predittori, poiché ne riduce la complessità e quindi la varianza. Ricordiamo che lo stimatore *OLS* non ha alcun *bias*.
- Un'altra possibilità è quella di introdurre una penalizzazione sui coefficienti del modello in modo da ridurre/regolarizzare l'effetto di ciascun predittore. Questa penalizzazione apporta *bias*.
  Si modifica quindi il criterio di minimizzazione degli RSS introducendo una componente penalizzante gestita da un parametro $\boldsymbol{\lambda} \in [0, +\infty)$:

$$\text{RSS}_{\boldsymbol{\lambda}} = \sum_{i=1}^{n} \left[Y_i - \left(c + \sum_{j=1}^{p} \beta_j \, x_{ij} \right) \right]^2 + \boldsymbol{\lambda} \, L(\beta_1, \ldots, \beta_p)$$

  ove $L(\beta_1, \ldots, \beta_p)$:
  - per il modello di Ridge è uguale a $\sum_{j=1}^{p} \beta_j^2$;
  - per il modello di Lasso è uguale a $\sum_{j=1}^{p} \left| \beta_j \right|$.

Regressione regolarizzata – Modelli di Ridge e Lasso

- Il parametro $\boldsymbol{\lambda}$ esprime l'ammontare di *bias* introdotto nello stimatore dei minimi quadrati al fine di far decrescere la varianza e di migliorare il *fit* del modello in termini di MSE.
- Il parametro $\boldsymbol{\lambda}$ è solitamente scelto per *cross-validation*, cioè suddividendo il campione di apprendimento in parti uguali ed usandone una per volta come campione di validazione (dettagli qui) per il calcolo dell'MSE. Il $\boldsymbol{\lambda}$ con il **minor valor medio dell'MSE è da preferire**.
- Chiaramente, se $\boldsymbol{\lambda} = 0$ riotteniamo le stime *OLS*.
- Per $\boldsymbol{\lambda} \longrightarrow \infty$, tutti i β_j vengono spinti verso lo 0.
- Il modello di Ridge non penalizza mai alcun coefficiente fino a zero, mentre il modello di Lasso riesce a farlo, funzionando dunque anche da **selettore di predittori**.
- Nessuno tra i due modelli di penalizzazione è superiore all'altro: la performance varia a seconda della *vera* funzione f della popolazione.

Grazie per l'attenzione e... ENJOY ΣTATS!!!

- **Per contattarmi:** passarettidav@gmail.com
- **Su di me:** Wordpress, Linkedin, Twitter, Stackoverflow