Un corso di Probabilità e Statistica

Maurizio Pratelli

Anno Accademico 2021-22

Indice

Capitolo 1

Statistica descrittiva.

1.1 Introduzione

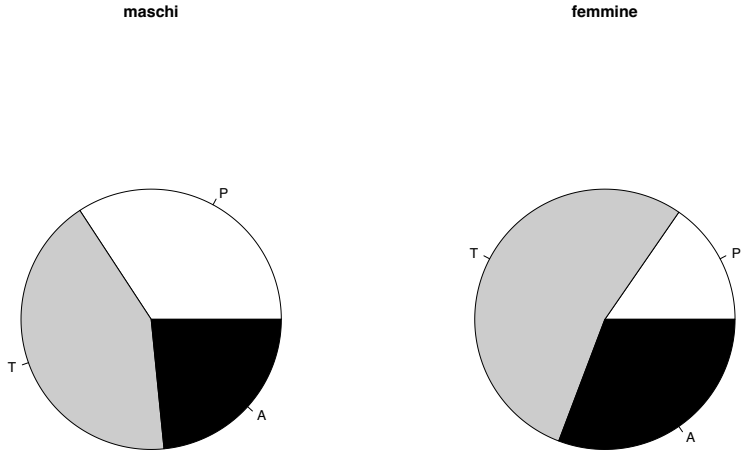
Si parla di *Statistica descrittiva* quando vengono analizzati dei dati, talvolta anche con dei metodi abbastanza sofisticati, ma senza l'interpretazione di un modello probabilistico; quando l'analisi dei dati verrà fatta utilizzando un modello probabilistico si parlerà invece di *Inferenza Statistica*.

Diamo per scontati i concetti di **popolazione** e di **campione**: spesso in statistica si parte da un campione per dedurne informazioni sull'intera popolazione.

Con il nome di *indagine statistica* si intende quindi una successione finita di dati, questi possono essere *qualitativi* o *quantitativi*. Sono qualitativi quando i dati dell'indagine sono un numero finito di *caratteri*, tali possono essere ad esempio gli esiti di n telefonate con lo scopo di effettuare una stima dell'orientamento di voto (in questo caso chiaramente si usa un *campione* per stimare l'orientamento di una *popolazione*).

La *frequenza* (assoluta) di un carattere è il numero di volte che questo carattere compare, mentre la *frequenza relativa* è la percentuale di volte nella quale questo carattere compare.

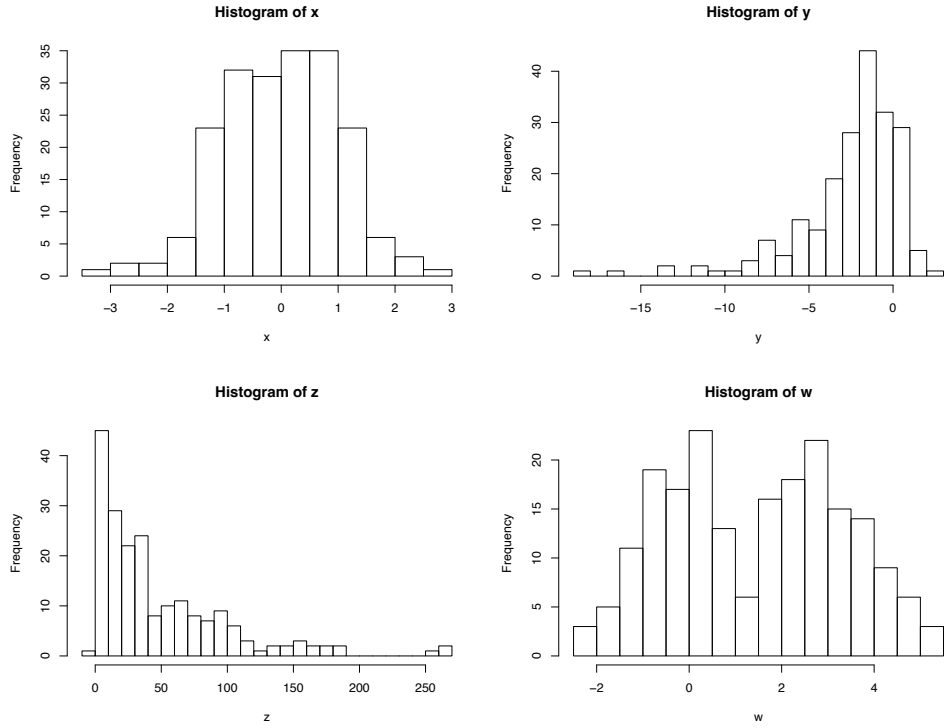
I dati di una indagine statistica qualitativa vengono spesso rappresentati con delle rappresentazioni grafiche intuitive (ad esempio i *diagrammi a torta*). Ne viene riportato sotto un esempio: è preso da un questionario proposto alle matricole di Ingegneria Gestionale (anno 2019) nel quale si chiedeva tra l'altro di indicare la provenienza, cioè dalla provincia di Pisa, dalla Toscana esterna a Pisa e dall'esterno della Toscana.



Quando invece i risultati sono n numeri, possibilmente tutti diversi, si parla di caratteri quantitativi: tali possono essere ad esempio i risultati di un esperimento fisico ripetuto più volte. Poiché questi numeri possono essere tutto diversi, non ha senso parlare di frequenza, invece la rappresentazione grafica più usata per rappresentare un campione numerico è l'*istogramma*, del quale il più comune è l'istogramma a bastoni (o a canne): l'altezza di una canna è proporzionale alla frequenza dei dati contenuti nella base della canna stessa.

L'istogramma è detto *normale* sei i dati sono a campana, abbastanza simmetrici rispetto al vertice; *"left skewed"* o *"right skewed"* se sono spostati rispettivamente a sinistra o a destra, *"bimodale"* se ci sono due (o eventualmente più) vertici.

Sono riportati qua sotto 4 istogrammi, rispettivamente *"normale", "left-skewed", "right-skewed" e "bimodale"*.



1.2 Dati numerici e grandezze che li sintetizzano

Supponiamo ora di avere un vettore $x = (x_1, \ldots, x_n)$ di dati numerici.

Definizione 1.2.1. Si chiama **media** (empirica) la media aritmetica dei numeri, cioè $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$.

Definizione 1.2.2. Si chiama **varianza campionaria** il numero $var(x) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}$ mentre si chiama **varianza empirica** la stessa somma divisa per n cioè $var_e(x) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}$.

Partiamo dall'eguaglianza $\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\,\overline{x}^2$ (questa eguaglianza di verifica immediata verrà utilizzata molte volte): dividendo per n si ottiene la formula $\boxed{var_e(x) = \sum_{i=1}^{n} \frac{x_i^2}{n} - \overline{x}^2}$.

La differenza tra varianza campionaria e varianza empirica sarà chiara più avanti (quando avremo introdotto le variabili aleatorie), per ora diciamo che la varianza empirica è più indicata quando si ha una *popolazione* mentre la varianza campionaria quando si ha un *campione*: poiché avremo quasi sempre a che fare con campioni ci concentriamo sulla varianza campionaria.

Definizione 1.2.3. La radice della varianza è chiamata **scarto quadratico medio** o anche **deviazione standard** (esiste dunque sia campionaria che empirica): la indichiamo $\sigma(x)$ (o $\sigma_e(x)$ quando è empirica) e si ha dunque $\sigma(x) = \sqrt{var(x)}$.

La varianza (che viene anche indicata $\sigma^2(x)$) è una misura della *dispersione dei dati* rispetto alla media: intanto è evidente che è eguale a 0 se e solo se i dati sono tutti eguali. Più in generale vale la seguente diseguaglianza: preso un campione di dati x ed un numero positivo d si ha

$$\boxed{\#\big\{x_i \,:\, |x_i - \overline{x}| > d\big\} \leq \frac{\sum_i (x_i - \overline{x})^2}{d^2}}$$

dove con il simbolo $\#A$ si indica la *cardinalità* dell'insieme A, cioè il numero dei suoi elementi. Vediamo la dimostrazione di questa diseguaglianza:

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 \geq \sum_{i\,:\,|x_i - \overline{x}| > d} (x_i - \overline{x})^2$$

$$\geq \sum_{i\,:\,|x_i - \overline{x}| > d} d^2 = d^2 \,.\, \#\big\{x_i \,:\, |x_i - \overline{x}| > d\big\}$$

Osserviamo poi che dividendo per n si ha la formula

$$\boxed{\frac{\#\big\{x_i \,:\, |x_i - \overline{x}| > d\big\}}{n} \leq \frac{var_e(x)}{d^2}}$$

Questa formula è un caso particolare della **diseguaglianza di Chebyshev** che vedremo più avanti, notiamo che il termine di sinistra è la *percentuale di dati* che differiscono da $\overline{x}$ più di d, ed è quindi evidente l'idea di varianza come misura della dispersione dei dati.

Quando l'istogramma dei dati è (ragionevolmente) normale, vale la seguente **regola empirica** sulla concentrazione dei dati:

- circa il 68% dei dati appartiene all'intervallo $\big[\overline{x} - \sigma(x)\,,\, \overline{x} + \sigma(x)\big]$
- circa il 95% dei dati appartiene all'intervallo $\big[\overline{x} - 2\,\sigma(x)\,,\, \overline{x} + 2\,\sigma(x)\big]$

- circa il 99.7% dei dati appartiene all'intervallo $\left[\overline{x} - 3\,\sigma(x)\,,\, \overline{x} + 3\,\sigma(x)\right]$

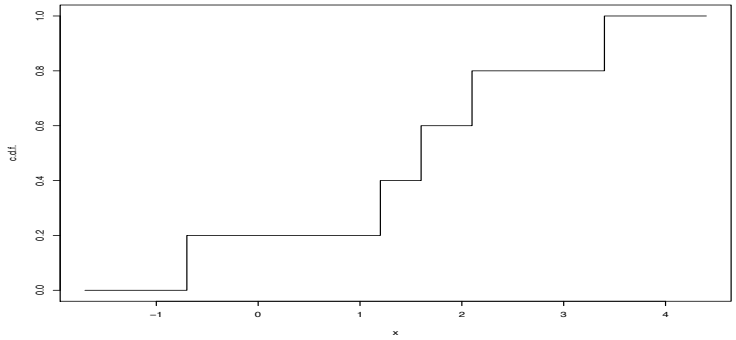
Si tratta appunto di una regola empirica che come tale non può essere dimostrata, ma quando studieremo più avanti le variabili Gaussiane (chiamate anche *normali*) forniremo una dimostrazione precisa di queste proprietà.

Ho utilizzato il punto (anziché la virgola) per indicare i decimali: il software **R** ammette solo il punto, mentre riserva la virgola per la separazione di dati diversi, per questo ci adeguiamo e indichiamo sistematicamente i decimali con il punto.

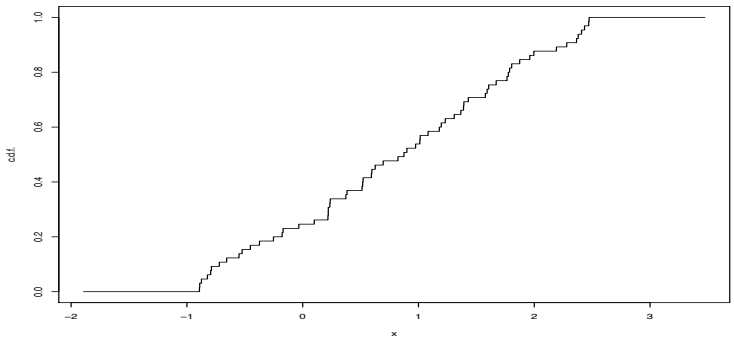
Torniamo a un campione di dati numerici x e definiamo la sua *Funzione di Ripartizione empirica* indicata anche *c.d.f.* (da Cumulative Distribution Function).

Prima di introdurre la definizione formale, vediamo come si ottiene in pratica la c.d.f : *ordiniamo* i numeri $x_1, \ldots, x_n$ in ordine crescente chiamandoli $x_{(1)}, \ldots, x_{(n)}$ (naturalmente è possibile che due o più siano eguali). La funzione $F_e(t)$ è eguale a 0 per $t < x_{(1)}$, poi fa un salto di ampiezza $1/n$ in corrispondenza di ognuno dei valori $x_{(i)}$ (se due dati sono eguali il salto è di ampiezza $2/n$) e poi è definitivamente eguale a 1 per $t \geq x_{(n)}$.

Nell'esempio qua sotto, considerando i dati (1.2, -0.7, 3.4, 1.6, 2.1), ne viene tracciata la c.d.f.



Quando i dati sono più numerosi, i "gradini" sono molto meno accentuati; nell'esempio successivo i dati sono 65.



Vediamo ora la definizione precisa:

Definizione 1.2.4. Si chiama **Funzione di ripartizione empirica** la funzione $F_e(.)$ definita su $\mathbb{R}$ dalla formula

$$F_e(t) = \frac{\#\{x_i \,|\, x_i \leq t\}}{n}$$

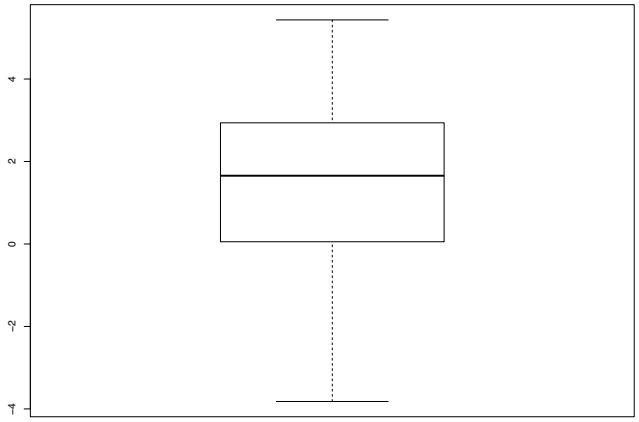
Sia ora k un numero con $0 < k < 100$: intuitivamente il *k-mo percentile* è il più piccolo numero che supera il $k\,\%$ dei dati, ma la definizione è più elaborata:

Definizione 1.2.5. Si chiama **k-mo percentile** un numero t tale che almeno $\frac{k}{100}$ dei dati sono inferiori o eguali a t e almeno $1 - \frac{k}{100}$ dei dati sono superiori o eguali a t.

Non sempre questo percentile è univocamente determinato, a volte sono due punti e in tal caso per convenzione si usa prendere il punto di mezzo. Talvolta si preferisce considerare il numero $\beta = \frac{k}{100}$ (ovviamente $0 < \beta < 1$) e si parla di β**-quantile**; quando introdurremo le variabili aleatorie si parlerà sempre di quantili.

In particolare lo 0.25-quantile è chiamato *primo quartile*, lo 0.50-quantile è chiamato *secondo quartile o mediana* e lo 0.75-quantile è detto *terzo quartile*.

Il **box-plot** dei dati è una rappresentazione grafica che evidenza i quartili: si ottiene sovrapponendo a una linea che va dal minimo al massimo dei dati un rettangolo che va dal primo al terzo quartile, con una linea che lo divide al livello della mediana. Qui sotto è riportato un esempio di box-plot.



1.3 Dati multipli

Ci limitiamo al caso di *coppie di dati* (o dati bivariati), ma quello che diremo si adatta facilmente a più dati.

I dati dell'indagine statistica sono un insieme di *coppie* di numeri $(x, y) = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ (ad esempio potrebbero essere le misurazioni di temperatura e pressione arteriosa presi dalle cartelle cliniche di un certo numero di pazienti).

Definizione 1.3.1. Si chiama **covarianza campionaria** il numero $cov(x) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})\,(y_i - \overline{y})}{n-1}$, mentre si chiama **covarianza empirica** la stessa somma divisa per n cioè $cov_e(x) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})\,(y_i - \overline{y})}{n}$.

Partendo dall'eguaglianza $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}$ e dividendo per n, si ottiene la formula $\boxed{cov_e(x, y) = \sum_{i=1}^{n} \frac{x_i\, y_i}{n} - \overline{x}\,\overline{y}}$.

Definizione 1.3.2. Supponiamo $\sigma(x) \neq 0$ e $\sigma(y) \neq 0$: si chiama **coefficiente di correlazione** tra x e y il numero

$$r(x, y) = \frac{cov(x, y)}{\sigma(x)\,\sigma(y)}$$

Si può osservare che la definizione non cambia se si sostituiscono covarianza e deviazione standard con covarianza empirica e deviazione standard empirica, ed in realtà vale la formula

$$\boxed{r(x,y) = \frac{\sum_{i=1}^n (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\,\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}}$$

Dalla *diseguaglianza di Schwartz* in $\mathbb{R}^n$ si deduce

$$\sum_{i=1}^n \left|(x_i - \overline{x}).(y_i - \overline{y})\right| \leq \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\,\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}$$

e di conseguenza $|r(x,y)| \leq 1$.

Intuitivamente il coefficiente di correlazione misura il *legame di natura lineare* tra i dati x e y ma per comprendere bene questa affermazione occorre introdurre la **retta di regressione**. Vogliamo calcolare questo termine

$$\inf_{a,b\,\in\,\mathbb{R}^2} \sum_{i=1}^n \left(y_i - a - b\,x_i\right)^2$$

vogliamo cioè approssimare nel modo migliore i dati y_i con una combinazione lineare affine $a + b\,x_i$: quell'estremo inferiore in realtà è un **minimo** e vale il seguente risultato.

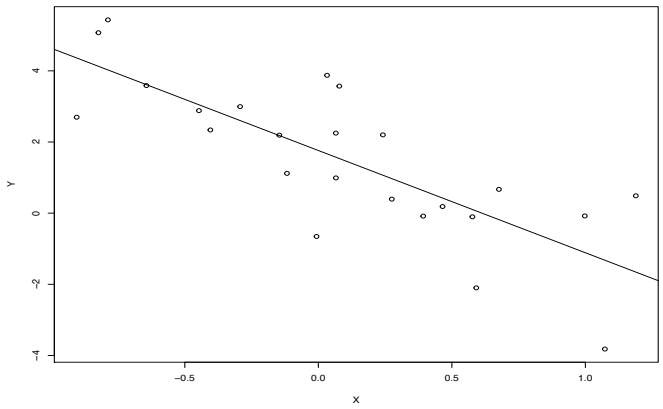
Teorema 1.3.3. *Il minimo al variare di* $(a,b) \in \mathbb{R}^2$ *della quantità* $\sum_{i=1}^n \left(y_i - a - b\,x_i\right)^2$ *si ottiene con* $b^* = \frac{cov(x,y)}{var(x)}$ *ed* $a^* = -b^*\,\overline{x} + \overline{y}$ *e vale l'eguaglianza*

$$\min_{a,b\,\in\,\mathbb{R}^2} \sum_{i=1}^n \left(y_i - a - b\,x_i\right)^2 = \sum_{i=1}^n \left(y_i - \overline{y}\right)^2 \left(1 - r(x,y)^2\right)$$

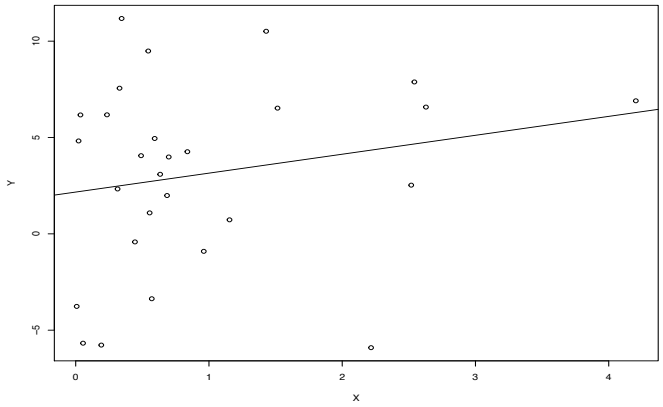
La retta $y = a^* + b^*\,x$ *è chiamata* **retta di regressione**.

Di conseguenza, quanto più $r(x,y)$ è vicino a 1 in valore assoluto, tanto più i punti tendono ad essere allineati; e in particolare $|r(x,y)| = 1$ se e solo se i punti *si trovano tutti su una stessa retta*. Inoltre $r(x,y)$ è positivo o negativo se il coefficiente angolare della retta è rispettivamente positivo o negativo; notiamo infine che nella formula di b^* è indifferente considerare varianza e covarianza empiriche o campionarie.

Nella figura che segue sono rappresentate delle coppie di dati ed è tracciata la retta di regressione; in questo particolare esempio si ha $r(x,y) \approx -0.78$, $a^* \approx 1.76$ e $b^* \approx -2.87$.



Nell'esempio successivo il coefficiente di correlazione è ≈ 0.2422 ed i dati sono molto più sparpagliati.



Vediamo ora la dimostrazione del teorema 1.3.3.

Dimostrazione. Cominciamo a vedere il principio della dimostrazione: consideriamo la funzione $Q(a,b) = \sum_{i=1}^{n} \left(y_i - a - b\, x_i\right)^2$ ed è immediato constatare che è una funzione continua (addirittura C^∞) che tende a $+\infty$ quando $|a|, |b| \to \infty$ e pertanto ha un minimo. Nel punto di minimo si devono annullare le derivate parziali: scrivendo le equazioni $\frac{\partial Q}{\partial a} = 0$ e $\frac{\partial Q}{\partial b} = 0$ si trova che queste sono risolte solo dai punti a^* e b^* scritti sopra che sono pertanto l'unico punto di minimo. Il valore minimo è pertanto a eguale a $Q(a^*, b^*)$.

Si tratta ora di tradurre questo principio in calcoli, un poco noiosi ma del tutto elementari. Le equazioni $\frac{\partial Q}{\partial a} = 0$ e $\frac{\partial Q}{\partial b} = 0$ diventano

$$\begin{cases} \sum_i y_i - n\, a - b \sum_i x_i = 0 \\ \sum_i x_i y_i - a \sum_i x_i - b \sum_i x_i^2 = 0 \end{cases}$$

e dividendo per n si arriva alle equazioni

$$\begin{cases} a + b\,\overline{x} = \overline{y} \\ a\,\overline{x} + b \sum_i \frac{x_i^2}{n} = \sum_i \frac{x_i\,y_i}{n} \end{cases}$$

Si risolve questo come un normale sistema di due equazioni in due incognite, trovando la soluzione

$$\begin{cases} a^* = -b^*\,\overline{x} + \overline{y} \\ b^* = \dfrac{\sum_i \frac{x_i\,y_i}{n} - \overline{x}\,\overline{y}}{\sum_i \frac{x_i^2}{n} - \overline{x}^2} = \dfrac{cov_e(x,y)}{var_e(x)} \end{cases}$$

Inserendo questi valori, con dei conti più lunghi ma sempre elementari, si trova appunto $Q(a^*, b^*) = \sum_{i=1}^n \left(y_i - \overline{y}\right)^2 \left(1 - r(x,y)^2\right)$

□

Capitolo 2

Breve introduzione al Calcolo delle Probabilità.

2.1 Prime definizioni

Si parla di probabilità di un evento quando ci si trova di fronte a un evento il cui esito non è sicuro, ma del quale vogliamo in qualche modo misurare la fiducia che questo possa realizzarsi.

Si pongono allora due problemi:

- come si possono rappresentare gli eventi?
- quali proprietà deve soddisfare la probabilità associata a questi eventi?

Per quanto riguarda la prima domanda, si parte usualmente da alcune *affermazioni* legate tra loro dai connettivi logici "**o**", "**e**", "**non**": è facile convincersi che la situazione si può rappresentare con una famiglia di sottinsiemi (chiamati *eventi*) di un opportuno insieme Ω. L'insieme Ω, che rappresenta tutti i possibili esiti, è chiamato usualmente *spazio fondamentale* o *universo*, o anche (soprattutto in Statistica) *spazio dei campioni*.

Se vogliamo rappresentare il risultato di un giro alla *roulette*, lo spazio naturale è l'insieme $\Omega = \{0, 1, \ldots, 36\}$ (ricordiamo che la ruota della roulette contiene 37 caselle, numerate da 0 a 36, e che lo 0 non è considerato né pari né dispari).

Le affermazioni *"è uscito un numero pari"* oppure *"è uscito un numero maggiore di 10"* si traducono con i sottinsiemi $A = \{2, 4, \ldots, 36\}$ e $B = \{11, 12, \ldots, 36\}$; l'affermazione *"***non** *maggiore di 10"* diventa $B^c = \{0, 1, \ldots, 9\}$, l'affermazione *"pari* **e** *maggiore di 10"* diventa $A \cap B = \{12, 14, \ldots, 36\}$.

Allo stesso modo *"pari* **oppure** *maggiore di 10"* diventa $A \cup B$; l'intero spazio Ω diventa l'evento che sicuramente si è realizzato mentre l'insieme vuoto $\emptyset$ rappresenta un evento impossibile.

Siamo stati dunque condotti a considerare un insieme Ω ed una famiglia di sottinsiemi (chiamati *eventi*): tale famiglia contiene l'insieme vuoto e l'intero spazio, ed è stabile per le operazioni di *unione* (finita), *intersezione* e *complementazione*. Una tale famiglia di insiemi si chiama un'**algebra di parti**.

Il *grado di fiducia* che un evento si realizzi (chiamato *probabilità*), è rappresentato da un numero compreso tra 0 e 1; inoltre è intuitivo supporre che se due eventi sono incompatibili (cioè hanno intersezione vuota) la probabilità che si realizzi uno qualsiasi dei due debba essere la somma delle probabilità dei singoli eventi. Questo equivale a dire che la probabilità è una funzione d'insieme *(finitamente) additiva*.

Cominciamo a dare la prima definizione (provvisoria):

Definizione 2.1.1 (**Probabilità finitamente additiva).** Si chiama probabilità (finitamente additiva) una funzione $\mathbf{P}$ definita su un'algebra di parti di un insieme Ω e a valori in [0,1] che soddisfa le proprietà seguenti:
a) se $A, B \in \mathcal{F}$ e $A \cap B = \emptyset$, allora $\mathbf{P}\big(A \cup B\big) = \mathbf{P}\big(A\big) + \mathbf{P}\big(B\big)$;
b) $\mathbf{P}(\Omega) = 1$.

Si chiama **trascurabile** un evento A tale che $\mathbf{P}(A) = 0$ e si chiama **quasi certo** un evento A tale che $\mathbf{P}(A) = 1$.

Vediamo alcune conseguenze immediate che si possono provare facilmente per esercizio:

1. $\mathbf{P}(\emptyset) = 0$;
2. $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$;
3. se $B \subset A$, $\mathbf{P}\big(A \setminus B\big) = \mathbf{P}(A) - \mathbf{P}(B)$, dove si è posto $A \setminus B = A \cap B^c$;
4. $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$;
5. $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(A \cap B) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) + \mathbf{P}(A \cap B \cap C)$, e così via ...

Le definizioni sopra riportate, oltre ad essere molto intuitive, sono supportate da valide argomentazioni logiche, tuttavia dal punto di vista matematico presentano una difficoltà: la **additività semplice** non consente di *andare al limite*, e di conseguenza di *calcolare degli integrali*. La buona proprietà per

poter effettuare queste operazioni è la **additività numerabile**, detta anche **σ-additività**.

Inoltre la famiglia di parti sulla quale possa essere definita una funzione σ-additiva è opportuno che sia stabile per *unione numerabile* e *intersezione numerabile*: un'algebra di parti stabile anche per unione e intersezione numerabile è chiamata **σ-algebra**.

Per questo motivo, seguendo quella che è chiamata la *definizione assiomatica di Probabilità secondo Kolmogorov*, sostituiamo alla precedente questa definizione:

Definizione 2.1.2 (Probabilità). Assegnato un insieme Ω ed una σ-algebra $\mathcal{F}$ di parti di Ω, si chiama probabilità una funzione $\mathbf{P} : \mathcal{F} \longrightarrow [0,1]$ tale che

a) se $(A_n)_{n=1,2,\dots}$ è una successione di elementi di $\mathcal{F}$ a due a due disgiunti, si ha $\mathbf{P}\big(\bigcup_{n=1}^{+\infty} A_n\big) = \sum_{n=1}^{+\infty} \mathbf{P}(A_n)$;

b) $\mathbf{P}(\Omega) = 1$.

Una terna $\big(\Omega, \mathcal{F}, \mathbf{P}\big)$ formata da un insieme Ω, una σ-algebra $\mathcal{F}$ di parti di Ω ed una probabilità $\mathbf{P}$ definita su $\mathcal{F}$ viene chiamata *spazio di Probabilità*.

Rimane da spiegare perché la σ-additività permette di *"andare al limite"*. Consideriamo una successione di eventi *crescente* nel senso che $A_n \subseteq A_{n+1}$ e sia $A = \bigcup_{n=1}^{+\infty} A_n = \lim_{n\to\infty} A_n$: vale la formula

$$\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n)$$

Vediamo come si dimostra questa formula: poniamo $B_1 = A_1$, $B_n = A_n \backslash A_{n-1}$ per $n > 1$. Gli insiemi $(B_n)_{n\geq 1}$ sono a due a due disgiunti e per l'additività finita si ha $\mathbf{P}(B_n) = \mathbf{P}(A_n) - \mathbf{P}(A_{n-1})$.

Poiché $\bigcup_{n\geq 1} A_n = \bigcup_{n\geq 1} B_n$, si ha $\mathbf{P}(A) = \sum_{n=1}^{+\infty} \mathbf{P}(B_n) =$
$= \lim_{n\to\infty} \sum_{h=1}^{n} \mathbf{P}(B_h) = \lim_{n\to\infty} \mathbf{P}(A_n)$.

Una proprietà perfettamente analoga vale per le successioni *decrescenti* di insiemi; cioè se $B_n \supseteq B_{n+1}$, posto $B = \bigcap_{n=1}^{+\infty} B_n$, si ha $\mathbf{P}(B) = \lim_{n\to\infty} \mathbf{P}(B_n)$. La dimostrazione segue dalla precedente passando al complementare.

Si può anche dimostrare questo fatto: *se la funzione* $\mathbf{P}$ *è finitamente additiva e* **passa al limite**, *allora è σ-additiva.*

Nell'enunciato precedente non importa se $\mathbf{P}$ passa al limite sulle successioni crescenti o su quelle decrescenti, perché queste due proprietà sono equivalenti.

È naturale a questo punto chiedersi perché in alcune situazioni la probabilità è assegnata solo su *alcuni* e non *tutti* i sottinsiemi dello spazio Ω : il motivo di questo è una difficoltà esclusivamente di ordine matematico, che verrà ripresa più avanti.

2.2 Il caso di uno spazio finito: elementi di Calcolo Combinatorio

La difficoltà enunciata alla fine del paragrafo precedente (cioè l'impossibilità, in alcuni casi, di estendere la probabilità a *tutti* i sottinsiemi dell'insieme Ω) non si pone se Ω è un insieme finito (cioè $\Omega = \{\omega_1, \dots, \omega_n\}$).

In tal caso la probabilità è univocamente determinata dai numeri $p_i = \mathbf{P}(\{\omega_i\})$, $(p_i \geq 0\,,\, p_1 + \cdots + p_n = 1)$; per ogni evento $A \subset \Omega$ si ha $\mathbf{P}(A) = \sum_{\omega_i \in A} p_i$.

La stessa cosa vale se l'insieme Ω è numerabile ($\Omega = \{\omega_1, \omega_2, \dots\}$) : vale la formula appena scritta, dove la somma finita diventa la somma di una serie se l'evento A è un insieme di cardinalità infinita.

Nel caso in cui Ω sia un insieme finito e gli *eventi elementari* ω_i siano equiprobabili, si parla di *distribuzione uniforme di probabilità su* Ω; naturalmente non esiste una distribuzione uniforme di probabilità su un insieme Ω numerabile ma infinito.

Tornando al caso di Ω finito e distribuzione uniforme di probabilità, si ottiene la formula

$$\mathbf{P}(A) = \frac{\#A}{\#\Omega} = \frac{\text{“casi favorevoli”}}{\text{“casi possibili”}}$$

dove con $\#A$ si indica la *cardinalità* (o numero degli elementi) dell'insieme A. La formula sopra scritta è anche chiamata *rapporto tra casi favorevoli e casi possibili* e talvolta ad essa ci si riferisce indicandola come la *definizione classica di Probabilità*.

In questo ambito, i problemi diventano molto spesso problemi di *calcolo combinatorio*: delle varie formule riportate dai libri (talvolta con nomi diversi da un libro all'altro) bisogna, a mio avviso, conoscerne soltanto tre. Tutte le altre si possono dedurre da queste come esercizio.

Prima di riportare queste formule premettiamo una comoda notazione: dato un intero n, anziché dire *“un insieme di n elementi”*, scriveremo più brevemente $\{1, \dots, n\}$.

Proposizione 2.2.1. *Siano k ed n due interi: il numero di applicazioni da $\{1, \dots, k\}$ a $\{1, \dots, n\}$ è n^k*

Proposizione 2.2.2 (Permutazioni). *Il numero di modi in cui si possono ordinare gli elementi di $\{1, \dots, n\}$ è $n!$*

Questa formula, così come la precedente, si dimostra per induzione.

Proposizione 2.2.3 (**Coefficiente binomiale**). *Siano* $0 \leq k \leq n$ *: il numero di sottinsiemi di* $\{1, \ldots, n\}$ *formati da* k *elementi è*

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n.(n-1)\ldots(n-k+1)}{k!}$$

Anche questa formula si dimostra per induzione, a scelta su k o su n.

Ricordiamo che questo coefficiente è chiamato coefficiente binomiale perché interviene nella *Formula del binomio di Newton*:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}$$

Vediamo ora, a titolo d'esempio e lasciando la dimostrazione per esercizio, due formule che si possono dedurre dalle precedenti.

Esempio 2.2.4. Siano $0 \leq k \leq n$: il numero di sottinsiemi *ordinati* di $\{1, \ldots, n\}$ formati da k elementi è $\frac{n!}{(n-k)!}$

Notiamo che questo numero coincide anche con il numero delle *applicazioni iniettive* da $\{1, \ldots, k\}$ in $\{1, \ldots, n\}$.

Esempio 2.2.5. Siano $k_1, \ldots, k_h$ interi con $k_1 + \cdots + k_h = n$: il numero di modi in cui si possono scegliere h sottinsiemi di $\{1, \ldots, n\}$ formati rispettivamente da $k_1, \ldots, k_h$ elementi è

$$\frac{n!}{k_1! \ldots k_h!}$$

2.3 Probabilità condizionata e indipendenza

Quando si è a conoscenza della realizzazione di un evento, cambia la valutazione di probabilità di ogni altro evento: ad esempio se si sa che il numero uscito su un giro della roulette è un numero *pari*, la probabilità che sia uscito il numero 16 non è più $\frac{1}{37}$ ma $\frac{1}{18}$.

Se si è realizzato l'evento $B = \{2, 4, \ldots, 36\}$ (cioè è uscito un numero pari) sono rimasti 18 *casi possibili* dei quali uno è *favorevole*: se indichiamo con $A = \{16\}$, notiamo che la nuova probabilità che è stata attribuita ad A verifica la formula $\dfrac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$.

Si possono fornire diversi esempi simili che sempre verificano la formula sopra riportata: queste considerazioni sono all'origine della definizione che segue.

Definizione 2.3.1. Assegnato uno spazio di probabilità $\big(\Omega, \mathcal{F}, \mathbf{P}\big)$ ed un evento B *non trascurabile*, si chiama *probabilità condizionata* di A rispetto a B il numero

$$\mathbf{P}\big(A\big|B\big) \;=\; \frac{\mathbf{P}\big(A \cap B\big)}{\mathbf{P}\big(B\big)}$$

Essa indica la probabilità che viene associata all'evento A, coerentemente con la valutazione precedentemente assegnata, in seguito all'informazione che si è realizzato l'evento B.

È facile verificare che, fissato B non trascurabile, la funzione $A \longrightarrow \mathbf{P}(A|B)$ è effettivamente una probabilità.

Dati due eventi A e B non trascurabili, è immediato constatare che vale la formula $\mathbf{P}(A \cap B) = \mathbf{P}(A|B)\,.\,\mathbf{P}(B) = \mathbf{P}(B|A)\,.\,\mathbf{P}(A)$.

Proposizione 2.3.2. *Siano $A_1, \ldots, A_n$ eventi, e supponiamo che $A_1 \cap \ldots \cap A_{n-1}$ sia non trascurabile: vale la formula*

$$\mathbf{P}\big(A_1 \cap \ldots \cap A_n\big) \;=\; \mathbf{P}\big(A_1\big).\mathbf{P}\big(A_2\big|A_1\big) \ldots \mathbf{P}\big(A_n\big|A_1 \cap \ldots \cap A_{n-1}\big)$$

La dimostrazione si ottiene immediatamente scrivendo i vari termini; si noti che, se $1 \leq k < n-1$, anche $A_1 \cap \ldots \cap A_k$ è non trascurabile. Questa formula è chiamata talvolta *formula del condizionamento ripetuto.*

Definizione 2.3.3 (**Sistema di alternative).** Si chiama *sistema di alternative* una partizione di Ω in n eventi non trascurabili $B_1, \ldots, B_n$.

Ricordiamo che *partizione* significa che gli insiemi B_i sono a due a due disgiunti e che la loro unione è l'intero insieme Ω.

Proposizione 2.3.4 (**Formula di Bayes).** *Sia $B_1, \ldots, B_n$ un sistema di alternative: assegnato una qualunque evento A non trascurabile, valgono le formule*

$$\mathbf{P}(A) \;=\; \sum_{i=1}^{n} \mathbf{P}\big(A\big|B_i\big)\,\mathbf{P}\big(B_i\big)$$

$$\mathbf{P}\big(B_i\big|A\big) \;=\; \frac{\mathbf{P}\big(A\big|B_i\big)\,\mathbf{P}\big(B_i\big)}{\mathbf{P}\big(A\big)} = \frac{\mathbf{P}\big(A\big|B_i\big)\,\mathbf{P}\big(B_i\big)}{\sum_{j=1}^{n} \mathbf{P}\big(A\big|B_j\big)\,\mathbf{P}\big(B_j\big)}$$

Dimostrazione. Per quanto riguarda la prima formula, si noti che
$A = (A \cap B_1) \cup \ldots \cup (A \cap B_n)$ e questi eventi sono a due a due disgiunti: si ha pertanto

$$\mathbf{P}(A) = \sum_{i=1}^{n} \mathbf{P}\big(A \cap B_i\big) \;=\; \sum_{i=1}^{n} \mathbf{P}\big(A\big|B_i\big)\,\mathbf{P}\big(B_i\big)$$

La seconda formula ne è una conseguenza immediata. □

In realtà si da il nome di *formula di Bayes* (chiamata talvolta anche *formula delle probabilità delle cause*) alla seconda delle formule della Proposizione 2.3.4 mentre la prima è chiamata *formula di fattorizzazione*.

Le formule della Proposizione 2.3.4 sono valide anche se il sistema di alternative anziché essere finito è numerabile, naturalmente sostituendo alle somme finite le somme di una serie.

È istruttivo risolvere il facile esercizio che segue in due modi, utilizzando prima il calcolo combinatorio e in seguito la formula 2.3.2.

Esercizio 2.3.5. Qual è la probabilità che, in una estrazione del lotto, tutti e 5 i numeri estratti siano inferiori o eguali a 20?

Introduciamo ora il concetto di *indipendenza* (stocastica): vogliamo tradurre con una formula matematica l'idea che la conoscenza che si è realizzato l'evento A non modifica la valutazione di probabilità di B e viceversa.

A tale scopo consideriamo due eventi A e B (non trascurabili) e proviamo a scrivere le eguaglianze $\mathbf{P}(A) = \mathbf{P}(A|B)$ e $\mathbf{P}(B) = \mathbf{P}(B|A)$: un esame immediato mostra che queste sono equivalenti tra loro ed equivalenti all'eguaglianza $\mathbf{P}(A \cap B) = \mathbf{P}(A).\mathbf{P}(B)$.

A differenza delle due precedenti, quest'ultima è simmetrica rispetto ai due eventi ed ha senso anche se uno dei due (o anche tutti e due) sono trascurabili: ne segue che questa è la *buona* definizione di indipendenza.

Definizione 2.3.6 (Indipendenza stocastica). Due eventi A e B sono detti indipendenti se vale l'eguaglianza

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\,.\,\mathbf{P}(B)$$

È un facile esercizio provare le seguenti affermazioni:

- se A e B sono indipendenti, sono indipendenti anche A^c e B, A e B^c, A^c e B^c;
- se $\mathbf{P}(A) = 0$ oppure $\mathbf{P}(A) = 1$, A è indipendente da qualsiasi altro evento;
- due eventi *incompatibili* (cioè che hanno intersezione vuota) non possono essere indipendenti, a meno che uno dei due sia trascurabile.

Vediamo ora come si estende questa definizione al caso di più eventi, cominciando con tre eventi. Affinché A, B e C siano indipendenti, occorre innanzi tutto che siano soddisfatte le tre eguaglianze $\mathbf{P}(A\cap B) = \mathbf{P}(A).\mathbf{P}(B)$, $\mathbf{P}(A\cap C) = \mathbf{P}(A).\mathbf{P}(C)$ e $\mathbf{P}(B\cap C) = \mathbf{P}(B).\mathbf{P}(C)$ (cioè che siano *a due a due*

indipendenti), ma deve anche esse soddisfatta l'eguaglianza $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A).\mathbf{P}(B).\mathbf{P}(C)$.

Non basta che siano a due a due indipendenti, come è mostrato nell'esempio seguente: sull'insieme $\Omega = \{1, 2, 3, 4\}$ munito della distribuzione uniforme di probabilità, gli eventi $A = \{1, 2\}$, $B = \{1, 3\}$ e $C = \{2, 3\}$ sono *a due a due* indipendenti, ma non sono *globalmente* indipendenti.

È facile trovare un controesempio simile per mostrare che non basta neppure che valga la sola eguaglianza $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A).\mathbf{P}(B).\mathbf{P}(C)$.

A questo punto si indovina facilmente quale sarà la definizione di indipendenza per un numero maggiore di eventi, e la definizione formale è la seguente:

Definizione 2.3.7 (**Indipendenza di più eventi).** Assegnati n eventi $A_1, \ldots, A_n$, questi si dicono *indipendenti* se per ogni intero k con $2 \leq k \leq n$ e per ogni scelta di interi $1 \leq i_1 < i_2 < \ldots < i_k \leq n$, vale l'eguaglianza

$$\mathbf{P}\big(A_{i_1} \cap \cdots \cap A_{i_k}\big) \;=\; \mathbf{P}\big(A_{i_1}\big) \,.\, \mathbf{P}\big(A_{i_k}\big)$$

Quando il numero di eventi cresce, il numero di eguaglianze da verificare sale notevolmente: per chi si vuole cimentare nel calcolo, per n eventi le eguaglianze da verificare sono $2^n - n - 1$ (per $n = 8$ sono 247, per $n = 10$ più di 1000). Ma questa in pratica non è una preoccupazione seria, perché c'è un caso tipico di indipendenza e quasi sempre si farà riferimento a quello.

Osservazione 2.3.8. Sono indipendenti gli esiti di *prove ripetute nelle medesime condizioni*: ad esempio sono indipendenti i risultati di successivi lanci di monete o successivi giri della ruota della roulette, ma *non sono indipendenti* ad esempio i risultati delle 5 estrazioni nel lotto.

2.4 Due esempi di probabilità sulla retta reale

Precisiamo subito che quelli che introduciamo in questo paragrafo non sono gli unici due casi di probabilità sulla retta reale $\mathbb{R}$, sono solo due esempi anche se nelle applicazion quasi sempre ci si riduce ad uno di questi.

Il primo esempio è quello di una **probabilità discreta** (talvola chiamata anche *atomica*) e sostanzialmente lo abbiamo già incontrato: la probabilità è *concentrata* su una successione (finita o numerabile) di punti $x_1, x_2, \ldots ...$ e (posto $p(x_i) = \mathbf{P}(\{x_i\})$) per ogni sottinsieme $A \subseteq \mathbb{R}$ vale la formula

$$\mathbf{P}(A) = \sum_{x_i \in A} \mathbf{P}(x_i) = \sum_{x_i \in A} p(x_i)$$

Abbiamo scritto $\mathbf{P}(x_i)$ invece del più formale $\mathbf{P}(\{x_i\})$ (d'ora innanzi faremo sempre così).

Definizione 2.4.1. Si chiama **funzione di massa** o anche **densità discreta** della probabilità discreta $\mathbf{P}$ la funzione $p(x_i) = \mathbf{P}(x_i)$

Come si è visto, per conoscere una tale probabilità basta conoscere i valori sui quali è concentrata (cioè $x_1, x_2, \dots$) ed i numeri $p_i = p(x_i)$: naturalmente $p_i \geq 0$ e $\sum_{i=1,2,\dots} p_i = 1$. Si può anche pensare di definire la *funzione di massa* $p(x) = \mathbf{P}(x)$ su tutta la retta (naturalmente sarà eguale a 0 fuori dei punti $x_1, x_2, \dots$) e allora la sola funzione di massa identifica la probabilità.

Un punto importante è il seguente: *una probabilità discreta può sempre essere definita su tutti i sottinsiemi di* $\mathbb{R}$.

Prima di affrontare il secondo caso, partiamo con un esempio: vogliamo *scegliere a caso un numero compreso tra* 0 *e* 1.

Lo spazio più naturale è $\Omega = [0, 1]$ e ad un intervallo $[a, b]$ (non importa se questo intervallo è aperto o chiuso) appare naturale attribuire come probabilità la sua lunghezza $(b - a)$; inoltre è ovvio supporre che la probabilità sia *invariante per traslazioni (modulo 1)*, cioè $\mathbf{P}(A) = \mathbf{P}(A + c)$, dove con $A + c$ si intende il traslato di A (modulo 1).

Innanzi tutto questo esempio mostra in modo naturale la differenza tra evento trascurabile ed evento impossibile: consideriamo un numero $x \in [0, 1]$, poiché per ogni n si ha $\{x\} \subseteq [x - \frac{1}{n}\, ,\, x + \frac{1}{n}]$, si ha $\mathbf{P}(x) \leq \frac{2}{n}$ e necessariamente $\mathbf{P}(x) = 0$. Quindi *ogni insieme* $\{x\}$ *è trascurabile, ma non è impossibile* perché evidentemente un numero viene scelto.

Ma soprattutto vale il seguente risultato: **non è possibile estendere la probabilità sopra definita a tutti i sottinsiemi dell'intervallo** $[0, 1]$.

La prova di questo fatto è un risultato avanzato di matematica, il famoso *controesempio di Vitali di un insieme non misurabile* (questo risultato non può essere esposto in un corso introduttivo come questo).

Sono detti *misurabili* gli insiemi sui quali può essere costruita la probabilità sopra esposta, ma questo non è un problema: per esibire un insieme non misurabile bisogna ricorrere all'assioma della scelta e non capiterà mai negli esempi e nelle applicazioni di scontrarsi con un insieme non misurabile.

Vediamo adesso un secondo esempio di probabilità sulla retta, le probabilità definite da una densità.

Definizione 2.4.2. Si chiama **densità di probabilità** sulla retta reale un funzione a valori positivi $f : \mathbb{R} \to [0, +\infty[$, integrabile e tale che

$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$: ad essa è associata una probabilità mediante la formula

$$\mathbf{P}(A) = \int_A f(x)\,\mathrm{d}x$$

Vediamo perché è una probabilità: innanzi tutto $\mathbf{P}(\mathbb{R}) = \int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$, poi se $A \cap B = \emptyset$ si ha

$$\mathbf{P}(A \cup B) = \int_{A\cup B} f(x)\,\mathrm{d}x = \int_A f(x)\,\mathrm{d}x + \int_B f(x)\,\mathrm{d}x = \mathbf{P}(A) + \mathbf{P}(B)$$

Per controllare che è *numerabilmente additiva* è più agevole verificare che *passa al limite* e questa è una conseguenza di un risultato di integrazione sul quale sorvoliamo (teorema di Beppo Levi).

Notiamo che l'esempio precedente (un numero scelto a caso tra 0 e 1) è un caso particolare di probabilità associata a una densità, nel quale la densità è definita da

$$f(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{altrove} \end{cases}$$

La probabilità definita da una densità *non può essere costruita su tutti i sottinsiemi di* $\mathbb{R}$ *ma solo sugli insiemi misurabili* ma questo come abbiamo visto non è un problema, perché *non può capitare nelle applicazioni e negli esempi di incontrare un insieme non misurabile*; quindi d'ora innanzi scriveremo A, $B, \ldots$ senza mai specificare che devono essere misurabili. Possiamo pensare a livello intuitivo che gli insiemi misurabili siano gli insiemi che hanno una struttura sufficientemente semplice perché si possa calcolare la loro probabilità.

Come ultima osservazione, notiamo che se la probabilità $\mathbf{P}$ è definita dalla densità f, ogni punto x è trascurabile: si ha infatti $\mathbf{P}(x) = \int_{\{x\}} f(t)\,\mathrm{d}t = 0$.

2.5 Introduzione alle variabili aleatorie

Supponiamo di aver puntato alla roulette 1 € sul numero 28 ed 1 € sul pari: possiamo domandarci qual è la probabilità di vincere più di 10 €, oppure la probabilità di perdere. Lo spazio naturale per descrivere l'esito di un giro della roulette è l'insieme $\Omega = \{0, 1, \ldots, 36\}$ munito della distribuzione uniforme di probabilità, ma le domande scritte sopra non corrispondono direttamente a sottinsiemi di Ω.

Siamo naturalmente portati a introdurre una funzione $X : \Omega \to \mathbb{R}$ (la funzione *vittoria netta*) che in questo esempio risulta essere così definita:

$$X(\omega) \;=\; \begin{cases} 36 & \omega = 28 \\ 0 & \omega \text{ pari, } \omega \neq 28 \\ -1 & \omega = 0 \\ -2 & \omega \text{ dispari} \end{cases}$$

La risposta alla prima domanda diventa
$\mathbf{P}\big\{\omega_i \big|\, X(\omega_i) \geq 10\big\} = \mathbf{P}\big(X^{-1}\big([10, +\infty[\big)\big) = \frac{1}{37}$ e la risposta alla seconda è $\mathbf{P}\big\{\omega_i \big|\, X(\omega_i) < 0\big\} = \mathbf{P}\big(X^{-1}\big(]-\infty, 0[\big)\big) = \frac{19}{37}$.

In definitiva, abbiamo naturalmente introdotto una funzione $X : \Omega \to \mathbb{R}$ ed abbiamo *trasportato* la probabilità dai sottinsiemi di Ω ai sottinsiemi di $\mathbb{R}$ mediante la formula

$$\mathbf{P}_X\big(A\big) \;=\; \mathbf{P}\big(X^{-1}(A)\big)$$

È bene familiarizzarsi subito con la notazione $X^{-1}(A) = \big\{X \in A\big\}$.

Definizione 2.5.1 (**Variabile aleatoria).** Assegnato uno spazio Ω sul quale è definita una probabilità $\mathbf{P}$, si chiama *variabile aleatoria* reale una funzione $X : \Omega \to \mathbb{R}$; si chiama **legge di probabilità** (o anche *distribuzione di probabilità*) della v.a. X la probabilità definita sui sottinsiemi di $\mathbb{R}$ dalla formula

$$\mathbf{P}_X\big(A\big) \;=\; \mathbf{P}\big(X^{-1}(A)\big)$$

È immediato constatare che la funzione sopra definita è effettivamente una probabilità: se $(A_n)_{n\geq 1}$ è una successione di sottinsiemi di $\mathbb{R}$ a due a due disgiunti, anche le immagini inverse sono disgiunte e si ha

$$\mathbf{P}_X\big(\bigcup_{n=1}^{+\infty} A_n\big) = \mathbf{P}\big(X^{-1}\big(\bigcup_{n=1}^{+\infty} A_n\big)\big) = \sum_{n=1}^{+\infty} \mathbf{P}\big(X^{-1}\big(A_n\big)\big) = \sum_{n=1}^{+\infty} \mathbf{P}_X\big(A_n\big)$$

Quando due variabili aleatorie hanno la stessa legge di probabilità sono dette **equidistribuite** (o anche *isonome*).

Nella maggior parte delle applicazioni in realtà non si incontra lo spazio Ω e la variabile $X : \Omega \to \mathbb{R}$, ma solo la legge di probabilità $\mathbf{P}_X$, ma questo non è un problema: *infatti si può mostrare che, assegnata una probabilità* $\mathbf{Q}$ *sui sottinsiemi di* $\mathbb{R}$*, è possibile costruire un insieme* Ω*, una probabilità* $\mathbf{P}$ *sui sottinsiemi di* Ω *ed una variabile* $X : \Omega \to \mathbb{R}$ *tale che* $\mathbf{P}_X = \mathbf{Q}$.

Notiamo ancora che non a caso indichiamo (e indicheremo sempre) le variabili aleatorie con lettere maiuscole X,Y,T ... e le variabili reali con lettere minuscole: sembra una pignoleria ma in realtà è importante per evitare confusioni.

Ai due esempi di probabilità su $\mathbb{R}$ introdotti nel paragrafo precedente, corrispondono due tipi di variabili aleatorie: anche qui è importante chiarire subito che non tutte le v.a. appartengono a uno di questi due tipi, ma sono i casi di gran lunga più usati nelle applicazioni.

Definizione 2.5.2 (**Variabile aleatoria discreta).** Una variabile aleatoria è detta *discreta* se la sua immagine è un sottinsieme finito o numerabile di $\mathbb{R}$, o (ciò che è lo stesso) se la sua legge di probabilità è discreta.

Dunque conoscere la legge di probabilità di una v.a. discreta equivale a conoscere quali sono i *valori* $\{x_1, x_2, \dots\}$ che prende la variabile, ed i numeri $p(x_i) = \mathbf{P}\{X = x_i\}$ (cioè la sua funzione di massa o densità discreta). Preso $A \subseteq \mathbb{R}$ vale la formula

$$\mathbf{P}_X(A) = \mathbf{P}\big\{X \in A\big\} = \sum_{x_i \in A} p_X(x_i)$$

Definizione 2.5.3 (**Variabile aleatoria con densità).** Una variabile aleatoria è detta *con densità* se la sua legge di probabilità è definita da una densità f , cioè se esiste una densità di probabilità f tale che valga la formula

$$\mathbf{P}_X(A) = \mathbf{P}\big\{X \in A\big\} = \int_A f(x)\,\mathrm{d}x$$

Le variabili con densità sono anche chiamate *variabili continue*, ma non è una buona dizione: l'origine di questo nome sarà chiarito nel capitolo successivo.

2.6 Appendice

2.6.1 Brevi richiami sulle serie numeriche

Data una successione di numeri reali $a_1, a_2, \dots$, posto $s_n = a_1 + \cdots + a_n$, si chiama *somma della serie* il limite (se esiste) della successione $(s_n)_{n\geq 1}$, e si dice che *la serie converge* se questo limite esiste. Più precisamente, per definizione

$$\sum_{n=1}^{+\infty} a_n \;=\; \lim_{n\to\infty} \sum_{k=1}^{n} a_k \;=\; \lim_{n\to\infty} s_n$$

Se la serie converge, la successione $(a_n)_{n\geq 1}$ è infinitesima (infatti si ha $a_n = s_n - s_{n-1}$) , ma non è vero il viceversa.

Vediamo ora alcune proprietà importanti delle serie a termini positivi (cioè $a_n \geq 0$, qualunque sia n): in tal caso la successione delle somme parziali $(s_n)_{n\geq 1}$ è *monotona crescente* e pertanto esiste comunque (finito o infinito) il limite. Ha sempre senso quindi scrivere $\sum_{n=1}^{+\infty} a_n \in [0, +\infty]$.

Vediamo subito un esempio: preso un numero positivo a si ha

$$\sum_{n=1}^{+\infty} \frac{1}{n^a} = \begin{cases} +\infty & \text{se} \quad a \leq 1 \\ < +\infty & \text{se} \quad a > 1 \end{cases}$$

Questo risultato è perfettamente analogo al risultato seguente per gli integrali, che sarà visto nell'appendice 3.8.1

$$\int_1^{+\infty} \frac{1}{x^a}\,\mathrm{d}x = \begin{cases} +\infty & \text{se} \quad a \leq 1 \\ \frac{1}{a-1} & \text{se} \quad a > 1 \end{cases}$$

Le serie a termini di segno positivo hanno interessanti proprietà, in particolare si può *cambiare l'ordine della somma* e *vale la proprietà associativa*: di seguito vediamo gli enunciati precisi nelle due seguenti proposizioni, nelle quali si suppone che la successione $(a_n)_{n\geq 1}$ sia formata da termini positivi.

Proposizione 2.6.1. *Sia $v : \mathbb{N} \to \mathbb{N}$ una applicazione biunivoca: allora*

$$\sum_{n=1}^{+\infty} a_n = \sum_{n=1}^{+\infty} a_{v(n)}$$

Proposizione 2.6.2. *Sia $A_1, A_2, \ldots$ una partizione di $\mathbb{N}$ (non importa se formata di insiemi finiti o infiniti): vale la formula*

$$\sum_{n=1}^{+\infty} a_n = \sum_{n=1}^{+\infty} \sum_{k\in A_n} a_k$$

Le dimostrazioni degli enunciati precedenti sono lasciate ad un corso di analisi. È importante sottolineare il fatto che queste due proprietà si estendono alle serie *assolutamente convergenti*: ricordiamo che una serie numerica è detta **assolutamente convergente** se

$$\sum_{n=1}^{+\infty} \left|a_n\right| < +\infty$$

Una serie assolutamente convergente è convergente (questo è un esercizio immediato) ma non è vero il viceversa, e non è difficile trovare degli esempi di serie convergenti ma non assolutamente convergenti che non verificano 2.6.1 e 2.6.2. Vale anzi il seguente risultato

Proposizione 2.6.3. *Supponiamo che la successione* $(a_n)_{n\geq 1}$ *sia tale che la seria ad essa associata* converga ma non converga assolutamente*: assegnato un qualsiasi* $l \in [-\infty, +\infty]$ *, è possibile determinare una funzione biunivoca* $v : \mathbb{N} \to \mathbb{N}$ *tale che si abbia*

$$\lim_{n\to\infty} \sum_{k=1}^{n} a_{v(k)} \;=\; l$$

Sorvoliamo sulla prova di questo fatto, ma da esso appare evidente che *le buone proprietà di sommabilità di una serie si hanno quando la serie è assolutamente convergente.*

Terminiamo con due esempi, che incontreremo più avanti. Il primo riguarda la **serie geometrica**: se $|a| < 1$, si ha $\sum_{h=0}^{+\infty} a^h = \frac{1}{1-a}$. La dimostrazione è abbastanza semplice: la formula $\displaystyle\sum_{h=0}^{n} a^h \;=\; \frac{a^{n+1}-1}{a-1}$ è un tipico esercizio sul principio di induzione, e poi si va al limite.

Il secondo invece, la cui prova è lasciata a un corso di analisi, è la **serie esponenziale**: più precisamente per ogni numero reale x vale lo sviluppo in serie

$$\mathrm{e}^x = \sum_{n=0}^{+\infty} \frac{x^n}{n!}$$

Capitolo 3

Variabili aleatorie, valori attesi e momenti

3.1 Variabili aleatorie: densità, funzione di ripartizione, quantili.

Nel capitolo precedente sono state definite le variabili aleatorie (discrete e con densità) e le loro leggi di probabilità: insistiamo sul fatto che è sostanzialmente equivalente avere una *variabile aleatoria* (cioè una funzione $X : \Omega \to \mathbb{R}$) oppure una *distribuzione di probabilità* sulla retta reale $\mathbb{R}$.

Introduciamo ora la *funzione di ripartizione* indicata anche brevemente **c.d.f.** (da *"Cumulative Distribution Function"*).

Definizione 3.1.1 (Funzione di ripartizione). Si chiama *Funzione di ripartizione* della v.a. X la funzione $F_X : \mathbb{R} \to [0, 1]$ definita da

$$F_X(x) = \mathbf{P}\{X \leq x\}$$

È evidente dalla definizione che in realtà $F_X(.)$ dipende solo dalla *legge di probabilità* della v.a. X; quando non c'è pericolo di confusione scriviamo più semplicemente F.

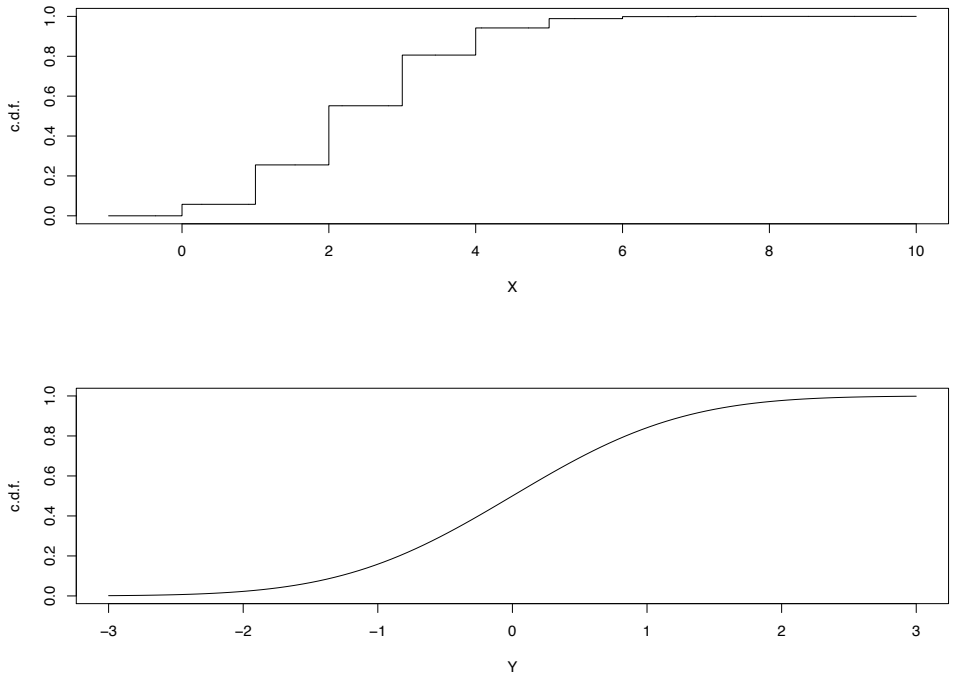
Quando la v.a. X è discreta, la sua funzione di ripartizione prende la forma $F(x) = \sum_{x_i \leq x} p(x_i)$, ed ha un tipico andamento a gradini.

Nel primo capitolo è stata definita la *funzione di ripartizione empirica* di un insieme di dati $x = (x_1, \ldots, x_n)$: ora possiamo notare che questa può essere interpretata come la funzione di ripartizione di una v.a. X che prende i valori $x_1, \ldots, x_n$ tutti con la stessa probabilità (cioè $1/n$).

Quando la variabile ha densità, la sua funzione di ripartizione diventa $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ ed è ovviamente *continua*.

Per questo motivo le variabili con densità sono anche chiamate *variabili continue*, ma non è una buona dizione: è possibile infatti costruire un esempio di funzione di ripartizione continua la cui v.a. associata non ha densità (l'esempio è piuttosto sofisticato e riguarda le cosiddette *misure cantoriane* quindi è fuori della portata di questo corso). Tuttavia la dizione *variabili continue* è largamente entrata nell'uso ... e quindi ci rassegnamo ad accettarla.

Sono riportati qua sotto due tipici grafici delle c.d.f. di una v.a. X discreta e di una Y con densità.



Vediamo quali sono le proprietà di una *Funzione di ripartizione*:

- è (debolmente) crescente, cioè se $x < y$ allora $F(x) \leq F(y)$;
- $\lim_{x\to-\infty} F(x) = 0$ e $\lim_{x\to+\infty} F(x) = 1$;
- F è continua a destra cioè $F(x) = \lim_{y\to x, y>x} F(y)$.

A parte la prima proprietà che è sostanzialmente evidente, tutte le altre sono conseguenza della *continuità della probabilità*: vediamo ad esempio $\lim_{x\to+\infty} F(x) = 1$.

Prendiamo una successione $x_1 < x_2 < x_3 \dots$ con $\lim_{n\to\infty} x_n = +\infty$ e sia $A_n = \{X \le x_n\}$: evidentemente si ha $A_n \subseteq A_{n+1}$, inoltre $\bigcup_n A_n = \bigcup_n \{X \le x_n\} = \Omega$.

Si ha allora $\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(\Omega) = 1$. Le altre proprietà si dimostrano allo stesso modo.

Ma quello che è più importante è che quanto detto sopra può essere in un certo senso invertito, cioè *assegnata una funzione* $F : \mathbb{R} \to [0,1]$ *con le proprietà sopra elencate, esiste una ed una sola probabilità* $\mathbf{P}$ *sui sottinsiemi di* $\mathbb{R}$ *tale che* F *sia la c.d.f. di una variabile la cui legge di probabilità sia* $\mathbf{P}$.

Come conseguenza *due v.a. che hanno la stessa funzione di ripartizione sono equidistribuite* e dalla funzione di ripartizione si può ricostruire la legge di probabilità: la dimostrazione di queste affermazioni richiede nozioni di teoria della misura, tuttavia nei due casi particolari delle variabili *discrete* e *con densità* si ricostruisce facilmente la distribuzione di probabilità della variabile.

Nel caso delle variabili **discrete**, abbiamo visto che la c.d.f. è *a gradini*, vi è un salto in corrispondenza di ogni punto x tale che $\mathbf{P}\{X = x\} > 0$ e la probabilità è proprio l'ampiezza del salto: si ha pertanto

$$\mathbf{P}\{X = x\} = F(x) - F_-(x)$$

dove con $F_-(x)$ si intende il limite sinistro di F nel punto x.

Viceversa per le variabili **con densità**, partendo dalla formula $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$, si ottiene (in tutti i punti in cui f è continua) $f(x) = \dfrac{\mathrm{d}F(x)}{\mathrm{d}x}$: la densità è quasi sempre *continua a tratti* (se non proprio continua) e quindi può essere ottenuta derivando la c.d.f.

Vediamo ora la definizione di *quantile*: intuitivamente, preso $0<\beta<1$, il β-quantile di una v.a. X è il numero r_β tale che $\mathbf{P}\{X \le r_\beta\} = F(r_\beta) = \beta$. Come si vede il quantile dipende solo dalla c.d.f. della variabile aleatoria o, ciò che è lo stesso, dalla sua legge di probabilità. Questa definizione però non va bene per due motivi: non sempre esiste un numero r tale che $F(r) = \beta$ e non è detto che tale numero sia univocamente determinato. La definizione allora deve essere modificata:

Definizione 3.1.2 (Quantile). Assegnata una v.a. X ed un numero β con $0<\beta<1$, si chiama β-*quantile* un numero r tale che si abbia $\mathbf{P}\{X \le r\} \ge \beta$ e $\mathbf{P}\{X \ge r\} \ge 1-\beta$.

Vediamo come si calcola in pratica nel caso delle v.a. discrete o con densità.

Prendiamo una v.a. X discreta, siano $x_1, x_2, \ldots$ i valori che assume e supponiamo di poterli ordinare $x_1 < x_2 < \ldots$ (non sempre è possibile ma nella maggior parte degli esempi è così): se non esiste alcun x_j tale che $F(x_j) = \beta$ allora r_β è il più piccolo x_j tale che $F(x_j \geq \beta)$, ma se esiste un x_j tale che $F(x_j) = \beta$, ogni numero r con $x_j \leq r \leq x_{j+1}$ soddisfa la definizione precedente. Di solito per convenzione si prende $r_\beta = \dfrac{x_j + x_{j+1}}{2}$.

Molto più facile è il caso delle v.a. *con densità*, soprattutto se F non ha un intervallo nel quale è costante: in tal caso semplicemente $r_\beta = F^{-1}(\beta)$.

3.2 Primi esempi di variabili aleatorie

3.2.1 Esempi di variabili discrete

Esempio 3.2.1 (**Variabile Binomiale).**

Consideriamo n prove ripetute di un esperimento che ha solo due esiti, chiamiamo successo uno di questi e sia p con $0 < p < 1$ la probabilità del successo: la variabile X *conta il numero dei successi.*

È evidente che i valori possibili sono $0, 1, \ldots, n$ e, per $0 \leq h \leq n$ si ha

$$p(h) = \mathbf{P}\{X = h\} = \binom{n}{h} p^h (1-p)^{n-h}$$

Infatti, considerando una sequenza di risultati con h successi ed $(n-h)$ insuccessi, la sua probabilità è $p^h.(1-p)^{n-h}$ e gli h successi si possono disporre in $\binom{n}{h}$ modi.

La variabile Binomiale è indicata $B(n,p)$, quando n è eguale ad 1, è chiamata di **Bernoulli** di parametro p.

Esempio 3.2.2 (**Variabile Geometrica).**

Questa volta l'esperimento viene ripetuto *fino a quando avviene il successo* (ad esempio si lancia un dado fino a quando esce il numero 6) e la v.a. X conta il numero di tentativi che è stato necessario effettuare: se p è la probabilità del successo, i valori possibili di X sono 1, 2, 3, ... e si ha $p(h) = \mathbf{P}\{X = h\} = (1-p)^{h-1}p$.

La variabile geometrica ha una interessante proprietà che è chiamata *assenza di memoria*: presi h, n interi positivi, si ha

$$\mathbf{P}\big\{X = n + h \,\big|\, X > n\big\} = \mathbf{P}\{X = h\}$$

la prova di questa proprietà è un facile esercizio; è meno facile provare che se X è a valori interi positivi ed è *senza memoria*, necessariamente è geometrica (con un opportuno parametro p).

Esempio 3.2.3 (**Variabile di Poisson).**

Preso un parametro $\lambda > 0$, si dice che X è una variabile di Poisson di parametro λ se i suoi valori sono tutti i naturali 0, 1, 2, ... e si ha $p(h) = \mathbf{P}\{X = h\} = \mathrm{e}^{-\lambda}\,\dfrac{\lambda^h}{h!}$.

Questa volta occorre provare che è una *buona definizione*, cioè che vale l'eguaglianza $\sum_{h=0}^{+\infty} p(h) = 1$, e questa è una conseguenza dello sviluppo esponenziale $\mathrm{e}^{\lambda} = \displaystyle\sum_{h=0}^{+\infty} \frac{\lambda^h}{h!}$.

3.2.2 Esempi di variabili con densità

Esempio 3.2.4 (**Densità uniforme).**

È la densità più semplice: presi due numeri $a < b$ la densità *uniforme sull'intervallo* [a,b] è costante su $[a, b]$ e nulla fuori dell'intervallo. Allora necessariamente

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{altrove} \end{cases}$$

L'esempio precedentemente visto del numero preso a caso tra 0 e 1 corrisponde alla densità uniforme su $[0, 1]$.

Esempio 3.2.5 (**Densità esponenziale).**

La densità esponenziale di parametro λ ($\lambda > 0$) è così definita:

$$f(x) = \begin{cases} \lambda\,\mathrm{e}^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

È facile controllare che è effettivamente una densità, infatti la *primitiva* di $\lambda\,\mathrm{e}^{-\lambda x}$ è proprio $-\mathrm{e}^{-\lambda x}$ e dunque $\int_0^{+\infty} \lambda\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x = -\mathrm{e}^{-\lambda x}\big|_0^{+\infty} = 1$.

Poiché la densità è diversa da 0 solo per x positivo, la variabile X prende solo valori positivi (nel senso che $\mathbf{P}\{X \le 0\} = 0$); inoltre anche la variabile esponenziale è *senza memoria* e questo significa che è soddisfatta l'eguagliana (per s, t positivi)

$$\mathbf{P}\big\{X \le s + t \,\big|\, X > s\big\} = \mathbf{P}\{X \le t\}$$

La verifica di questa eguaglianza è un facile esercizio, mentre è più impegnativo provare che se la v.a. X ha densità, prende solo valori positivi ed è senza memoria, necessariamente è esponenziale.

Nel paragrafo successivo verrà presentata la densità gaussiana, di importanza fondamentale in inferenza statistica.

Ma prima facciamo una osservazione: supponiamo che X abbia densità f e sia $Y = h \circ X$ e ci domandiamo se Y ha densità e come possiamo calcolarla. Innanzi tutto non è detto che Y abbia densità (dipende dalla funzione h) e poi non esistono regole precise: la strada da percorrere è calcolare per prima cosa la c.d.f. di Y cioè

$$F_Y(y) = \mathbf{P}\{Y \leq y\} = \mathbf{P}\{h(X) \leq y\}$$

ed esaminarla. Se è derivabile (o almeno derivabile a tratti), derivando si ottiene la densità di Y.

Ma c'è una situazione nella quale si può dare una formula precisa: supponiamo che h sia biunivoca, derivabile con inversa derivabile, e sia $Y = h(X)$. Cominciamo a supporre che h sia *crescente*, allora

$$F_Y(y) = \mathbf{P}\{Y \leq y\} = \mathbf{P}\{h(X) \leq y\} = \mathbf{P}\{X \leq h^{-1}(y)\} = F_X(h^{-1}(y))$$

e derivando si ottiene la formula $f_Y(y) = f_X(h^{-1}(y)).\dfrac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y}$.

Se h è *decrescente*, si ha $\mathbf{P}\{h(X) \leq y\} = \mathbf{P}\{X \geq h^{-1}(y)\} = 1 - F_X(h^{-1}(y))$ ma poi la derivata di h^{-1} è negativa; inoltre non è detto che la densità sia diversa da 0 su tutta la retta.

Possiamo sintetizzare tutto in un risultato generale esposto sotto, precisando una cosa: col termine generico di *intervallo aperto* si può intendere sia un intervallo limitato del tipo $]a, b[$ che una semiretta $]-\infty, b[$ oppure $]a, +\infty[$, che tutta la retta $]-\infty, +\infty[$.

Proposizione 3.2.6. *Supponiamo che X abbia densità f_X diversa da 0 su un intervallo aperto A, sia $h : A \to B$ (dove B è un altro intervallo aperto) biunivoca, derivabile con inversa derivabile e sia $Y = h \circ X$: allora Y ha densità f_Y data dalla formula*

$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)).\left|\frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}\,y}\right| & y \in B \\ 0 & y \notin B \end{cases}$$

La formula proposta sopra è molto comoda, ma bisogna essere sicuri che possa essere applicata: è istruttivo al riguardo svolgere l'esempio che segue.

Esempio 3.2.7. Siano rispettivamente X con densità uniforme su $[-1, 2]$ e Y con densità esponenziale di parametro 2, e consideriamo per entrambe il quadrato, cioè $Z = X^2$ e $W = Y^2$: calcolare le densità di Z e di W esaminando in particolare se può essere utilizzata la formula 3.2.6

3.3 La densità gaussiana o normale

Tra tutte le densità, la densità **Gaussiana** (chiamata anche *Normale*) occupa un posto di assoluta rilevanza: in realtà è stata introdotta da De Moivre e Laplace, ma poi estensivamente usata da Gauss soprattutto nella teoria degli errori e da questo deriva il suo nome.

Consideriamo la funzione $f(x) = \mathrm{e}^{-\frac{x^2}{2}}$: è una funzione regolarissima, va a 0 molto in fretta per $|x| \to \infty$ e quindi è integrabile, ma *non è possibile scrivere la sua primitiva in termini di funzioni elementari.* Quindi ad esempio non è possibile calcolare $\int_1^3 \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x$ se non ricorrendo ad approssimazioni numeriche.

Tuttavia Laplace è riuscito a fare questo calcolo:

$$\int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = \sqrt{2\,\pi}$$

La dimostrazione di questo calcolo richiede integrali doppi e passaggio alle coordinate polari per gli integrali e pertanto è rinviata all'appendice; per chi ha dimestichezza con gli strumenti appena elencati in realtà non è affatto difficile ... naturalmente ci voleva il colpo d'ingegno di Laplace per arrivarci. La conseguenza immediata però è che dividendo la funzione sopra scritta per $\sqrt{2\,\pi}$ si ottiene una *densità di probabilità.*

Definizione 3.3.1 (Densità Gaussiana standard). Si chiama densità Gaussiana (o Normale) standard, indicata $N(0,1)$, la funzione

$$\varphi(x) = \frac{1}{\sqrt{2\,\pi}}\,\mathrm{e}^{-\frac{x^2}{2}}$$

La sua *funzione di ripartizione* è

$$\Phi(x) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$$

È talmente centrale nell'inferenza statistica che riserviamo le lettere φ, Φ e q_α per indicare rispettivamente la densità, la c.d.f. e lo α-quantile della variabile $N(0,1)$. Notiamo che, come conseguenza del fatto che la densità ϕ è una funzione *pari* (cioè $\varphi(x) = \varphi(-x)$), valgono le seguenti relazioni per la c.d.f. e per i quantili: presi $x \in \mathbb{R}$ e $0 < \alpha < 1$, si ha

$$\Phi(-x) = 1 - \Phi(x) \qquad\qquad q_{1-\alpha} = -q_\alpha$$

È più facile convincersi delle eguaglianze sopra scritte con una rappresentazione grafica piuttosto che darne una dimostrazione formale (peraltro non

difficile); di nuovo insistiamo sul fatto che queste eguaglianze non sono specifiche della variabile gaussiana standard ma conseguenza del fatto che la densità è una funzione pari.

È uno strumento preziosissimo (e di facile uso) la **tavola della variabile N(0,1)** : questa fornisce il valore di $\Phi(x)$ per $0 < x < 4$: per $x \geq 4$ si ha $\Phi(x) \approx 1$ e per x negativo il valore si ottiene dall'eguaglianza $\Phi(-x) = 1 - \Phi(x)$.

Inoltre leggendo la tavola *all'incontrario* si ricavano gli α-quantili per $\frac{1}{2} < \alpha < 1$ e per $0 < \alpha < \frac{1}{2}$ si usa l'eguaglianza $q_{1-\alpha} = -q_\alpha$.

In particolare, utilizzando appunto la tavola N(0,1), si trova che, se X è gaussiana standard

- $\mathbf{P}\{-1 \leq X \leq 1\} \approx 0.68$
- $\mathbf{P}\{-2 \leq X \leq 2\} \approx 0.94$
- $\mathbf{P}\{-3 \leq X \leq 3\} \approx 0.997$

(osserviamo che $\mathbf{P}\{-1 \leq X \leq 1\} = \Phi(1) - \Phi(-1) = 2\,\Phi(1) - 1$)

Passiamo ora alle densità gaussiane più generali: partiamo da X gaussiana standard, siano $\sigma > 0$ e $m \in \mathbb{R}$ e consideriamo la v.a. $Y = \sigma\, X + m$. Cominciamo a calcolare la sua funzione di ripartizione

$$F_Y(y) = \mathbf{P}\big\{Y \leq y\big\} = \mathbf{P}\Big\{\sigma\, X + m \leq y\Big\} = \mathbf{P}\Big\{X \leq \frac{y-m}{\sigma}\Big\} = \Phi\big(\frac{y-m}{\sigma}\big)$$

Per ottenere la densità di Y , chiamata densità $N(m, \sigma^2)$, si può o derivare $F_Y(.)$ o applicare direttamente la formula 3.2.6.

Definizione 3.3.2 (**Densità Gaussiana generale).** Si chiama densità gaussiana $N(m, \sigma^2)$ la funzione

$$f(x) = \frac{1}{\sqrt{2\,\pi}\,\sigma}\, \mathrm{e}^{-\frac{(x-m)^2}{2\,\sigma^2}}$$

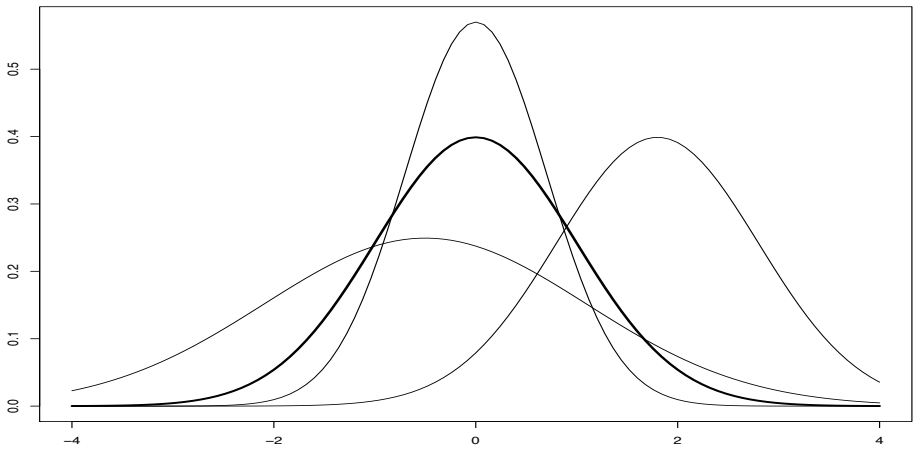
Non c'è bisogno di tavole per un variabile Gaussiana generica perché *ci si riporta sempre alla variabile Gaussiana standard*: una variabile con densità $N(m, \sigma^2)$ si può sempre pensare della forma $(\sigma X{+}m)$ con X gaussiana standard, o (ciò che è lo stesso) presa Y con densità $N(m, \sigma^2)$, la variabile $\dfrac{Y-m}{\sigma}$ è gaussiana $N(0, 1)$.

Le valutazioni scritte sopra per una v.a. X gaussiana standard, diventano se Y è gaussiana $N(m, \sigma^2)$

- $\mathbf{P}\{m - \sigma \leq Y \leq m + \sigma\} \approx 0.68$
- $\mathbf{P}\{m - 2\sigma \leq Y \leq m + 2\sigma\} \approx 0.94$
- $\mathbf{P}\{m - 3\sigma \leq Y \leq m + 3\sigma\} \approx 0.997$

e queste valutazioni forniscono una argomentazione precisa alla *regola empirica sulla dispersione dei dati* introdotta al primo capitolo quando l'istogramma dei dati è (abbastanza) normale.

Sono riportati qua sotto, sovrapposti, i grafici di 4 densità gaussiane con diversi m e σ^2 ; quella in grassetto è la densità Gaussiana standard.



3.4 Variabili aleatorie doppie

Ci limitiamo per semplicità al caso di due variabili, e sia $(X, Y) : \Omega \to \mathbb{R}^2$ una coppia di variabili aleatorie: la sua *legge di probabilità* è una probabilità sui sottinsiemi di $\mathbb{R}^2$.

La variabile doppia è detta *discreta* se la sua immagine è concentrata in un insieme finito o numerabile di punti (x_i, y_j) e la sua distribuzione di probabilità è identificata dalla *funzione di massa* $p(x_i, y_j) = \mathbf{P}\{X = x_i, Y = y_j\}$ e si ha, per $A \subseteq \mathbb{R}^2$

$$\mathbf{P}_{X,Y}(A) = \mathbf{P}\big\{(X, Y) \in A\big\} = \sum_{(x_i,y_j)\in A} p(x_i, y_j)$$

La variabile è invece *con densità* se esiste una funzione $f(x, y)$ definita su $\mathbb{R}^2$ a valori positivi, integrabile e con $\iint_{\mathbb{R}^2} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$ tale che valga,

per $A \subseteq \mathbb{R}^2$

$$\mathbf{P}_{X,Y}(A) = \mathbf{P}\big\{(X,Y) \in A\big\} = \iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

Perché non ci siano confusioni, precisiamo che la *virgola* sta per *intersezione*: ad esempio

$$\big\{X \in A\,, Y \in B\big\} = \big\{X \in A\big\} \cap \big\{Y \in B\big\} = X^{-1}(A) \cap Y^{-1}(B) = (X,Y)^{-1}\big(A \times B\big)$$

Presa (X,Y) discreta con funzione di massa $p(x_i, y_j)$, anche X e Y sono discrete con funzione di massa rispettive $p_1(.)$ e $p_2(.)$; allo stesso modo, se (X,Y) ha densità $f(x,y)$, anche X e Y hanno densità rispettivamente $f_1(.)$ e $f_2(.)$. È bene precisare che, mentre il fatto che X e Y siano discrete se lo è la coppia (X,Y) è un fatto evidente, il fatto che X e Y abbiano densità se ha densità la coppia (X,Y) non è proprio evidente ma ha bisogno di una dimostrazione.

Proposizione 3.4.1 (Formule per le densità marginali). *Se (X,Y) è discreta con funzione di massa $p(x_i, y_j)$, le funzioni di massa rispettivamente di X e di Y verificano le formule*

$$p_X(x_i) = \sum_{y_j} p(x_i, y_j) \qquad\qquad p_Y(y_j) = \sum_{x_i} p(x_i, y_j)$$

Se (X,Y) ha densità $f(x,y)$, anche X e Y hanno densità rispettivamente

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y \qquad\qquad f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x$$

Dimostrazione. Per quanto riguarda le variabili discrete, la prova è facile: notiamo che $\{X = x_i\} = \bigcup_j \{X = x_i, Y = y_j\}$ e questi ultimi insiemi sono a due a due disgiunti. Si ha pertanto

$$p_X(x_i) = \mathbf{P}\{X = x_i\} = \sum_{y_j} \mathbf{P}\{X = x_i, Y = y_j\} = \sum_{y_j} p(x_i, y_j)$$

e analoga è ovviamente l'altra eguaglianza. □

Passando alle variabili con densità, è intuitivo che la formula sia valida sostituendo alle somme gli integrali, ma la dimostrazione precisa richiede gli integrali doppi ed è pertanto rinviata all'appendice. Nel seguito di questo capitolo ci saranno diverse formule e dimostrazioni valide sia nel caso discreto

che in quello con densità: ci limitiamo a provarle nel caso discreto lasciando l'altro caso all'appendice.

Passiamo ora all'indipendenza di variabili aleatorie: cerchiamo una definizione che traduca l'idea che ogni informazione legata ad X è indipendente da ogni informazione legata ad Y.

Definizione 3.4.2 (**Indipendenza di variabili aleatorie).** Due variabili aleatorie X e Y sono dette indipendenti se, presi comunque A, B sottinsiemi di $\mathbb{R}$, gli eventi $X^{-1}(A)$ e $Y^{-1}(B)$ sono indipendenti, cioè se vale l'eguaglianza

$$\mathbf{P}\big\{X \in A\,,\, Y \in B\big\} = \mathbf{P}\big\{X \in A\big\}\,.\,\mathbf{P}\big\{Y \in B\big\}$$

È importante avere una caratterizzazione pratica dell'indipendenza, e questa è fornita dal seguente risultato

Proposizione 3.4.3 (**Caratterizzazione dell'indipendenza).** *Date due variabili discrete X e Y , queste sono indipendenti se e solo se vale l'eguaglianza tra le funzioni di massa*

$$p(x_i, y_j) = p_X(x_i)\,.\,p_Y(y_j)$$

Se (X, Y) ha densità, le variabili sono indipendenti se e solo se vale l'eguaglianza tra densità

$$f(x, y) \;=\; f_X(x)\,.\,f_Y(y)$$

Dimostrazione. Anche qui ci limitiamo al caso di variabili discrete. Da una parte, prendendo $A = \{x_i\}$ e $B = \{y_j\}$, si ha
$p(x_i, y_j) = \mathbf{P}\{X = x_i\,,\, Y = y_j\} = \mathbf{P}\{X = x_i\}\,\mathbf{P}\{Y = y_j\} = p_X(x_i)\,p_Y(y_j)\,.$
Viceversa, presi due sottinsiemi A e B di $\mathbb{R}$ si ha:

$$\mathbf{P}\big\{X \in A\,,\, Y \in B\big\} \;=\; \sum_{x_i \in A\,,\, y_j \in B} p(x_i, y_j) \;=\; \sum_{x_i \in A} \sum_{y_j \in B} p_X(x_i)\,p_Y(y_j) \;=$$

$$= \Big(\sum_{x_i \in A} p_X(x_i)\Big)\Big(\sum_{y_j \in B} p_Y(y_j)\Big) \;=\; \mathbf{P}\big\{X \in A\big\}\;\mathbf{P}\big\{Y \in B\big\}$$

□

Analoga sarà la definizione e caratterizzazione dell'indipendenza per più variabili, notiamo anche che *funzioni di variabili indipendenti sono indipendenti*, più precisamente prese X e Y indipendenti e due funzioni h e k, anche

la variabili $h(X)$ e $k(Y)$ sono indipendenti: questo è del tutto intuitivo ma anche facile da provare.

Più in generale, funzioni di più variabili indipendenti sono indipendenti *se la stessa variabile non compare in due funzioni diverse*: per fare un esempio, se X, Y e Z sono indipendenti, lo sono anche $\sqrt{X^2+Y^2}$ e Z^3 mentre *non sono indipendenti* $\sqrt{X^2+Y^2}$ e $\sqrt{X^2+Z^2}$.

Cominciamo ora con questo esempio:

Esempio 3.4.4. Se X e Y sono rispettivamente Binomiali $B(n,p)$ e $B(m,p)$ e sono indipendenti, $Z = X + Y$ è binomiale $B(n+m,p)$.

Per la dimostrazione è sufficiente supporre che Y sia di Bernoulli (e poi procedere per induzione): in tal caso

$$\mathbf{P}\{Z=h\} = \mathbf{P}\{X=h, Y=0\} + \mathbf{P}\{X=h-1, Y=1\} =$$

$$= \binom{n}{h} p^h (1-p)^{n-h}(1-p) + \binom{n}{h-1} p^{h-1}(1-p)^{n-h+1} p = \binom{n+1}{h} p^h (1-p)^{n+1-h}$$

Questo risultato suggerisce una formula generale:

Proposizione 3.4.5 (**Formula della convoluzione discreta).** *Supponiamo che X e Y siano a valori interi positivi e indipendenti, e sia $Z = X+Y$: dette p_X, p_Y e p_Z le funzioni di massa rispettivamente di X, Y e Z, si ha*

$$p_Z(n) = \sum_{h=0}^{n} p_X(h)\,.\,p_Y(n-h)$$

Questa formula è una conseguenza immediata dell'eguaglianza insiemistica

$$\{Z=n\} = \bigcup_{h=0}^{n} \big\{X=h\,,\,Y=n-h\big\}$$

e di conseguenza (tenendo conto dell'indipendenza) si ha

$$p_Z(n) = \mathbf{P}\{Z=n\} = \sum_{h=0}^{n} \mathbf{P}\{X=h\}\,\mathbf{P}\{Y=n-h\} = \sum_{h=0}^{n} p_X(h)\,p_Y(n-h)$$

Con questa formula si può provare ad esempio che se X è di Poisson di parametro λ, Y di Poisson di parametro μ e sono indipendenti, allora $X+Y$ è di Poisson di parametro $(\lambda+\mu)$, tuttavia evitiamo questo conto perché più avanti avremo uno strumento che ci permette di ottenere lo stesso risultato più facilmente.

La formula 3.4.5 suggerisce la formula seguente, di grande importanza.

Proposizione 3.4.6 (**Formula della convoluzione).** *Siano X e Y indipendenti, con densità rispettive $f_X(.)$ e $f_Y(.)$, e sia $Z = X + Y$: la variabile Z ha densità data da*

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x)\, f_Y(z-x)\, \mathrm{d}x = \int_{-\infty}^{+\infty} f_Y(y)\, f_X(z-y)\, \mathrm{d}y$$

La dimostrazione di questa formula richiede integrali doppi e derivazione sotto l'integrale.

Con questa formula si può provare ad esempio il seguente fondamentale risultato: se X e Y sono indipendenti e gaussiane rispettivamente $N(m_1, \sigma_1^2)$ e $N(m_2, \sigma_2^2)$, allora $Z = X+Y$ è gaussiana $N(m_1+m_2, \sigma_1^2+\sigma_2^2)$. Anche questo conto però non lo svolgiamo perché avremo uno strumento per ottenere il risultato con molta meno fatica.

3.5 Valori attesi e momenti

Se abbiamo un insieme di dati $x = (x_1, \ldots, x_n)$, abbiamo definito la media (empirica) $\overline{x} = \sum_{i=1}^n \frac{x_i}{n}$: cioè i numeri x_i sono *pesati* col numero $\frac{1}{n}$, sostanzialmente come se fossero tutti equiprobabili.

Ma se $x_1, \ldots, x_n$ sono i *valori* che assume una v.a. discreta X, come *media* di X sembra più naturale pesare questi valori con la funzione di massa, cioè $\sum_{i=1}^n x_i\, p(x_i)$. Se però i valori della variabile sono infiniti (come ad esempio per le variabili geometrica e di Poisson), occorre che la serie converga ed abbiamo visto che le buone proprietà di convergenza di una serie numerica si hanno quando la serie *converge assolutamente*.

Definizione 3.5.1 (**Valore atteso).** Sia X una variabile discreta con funzione di massa $p(.)$: si dice che X ha valore atteso se $\sum_i |x_i|\, p(x_i) < +\infty$ e in tal caso si chiama valore atteso il numero

$$\mathbf{E}[X] = \sum_i x_i\, p(x_i)$$

Sia X con densità $f(.)$: si dice che X ha valore atteso se $\int_{-\infty}^{+\infty} |x|\, f(x)\, \mathrm{d}x < +\infty$ e in tal caso si chiama valore atteso il numero

$$\mathbf{E}[X] = \int_{-\infty}^{+\infty} x\, f(x)\, \mathrm{d}x$$

Naturalmente esiste una definizione più generale valida per tutte le variabili aleatorie che contiene quelle sopra scritte come casi particolari, ma

necessita di elementi di teoria dell'integrazione generale. Il *valore atteso* è chiamato anche *speranza matematica* o *momento primo*; il termine inglese è *expectation*.

Notiamo che se X prende solo valori positivi, ha sempre senso scrivere $\mathbf{E}[X]$ che però potrebbe anche essere $+\infty$. Per le variabili discrete questo significa che i valori $x_1, x_2, \ldots$ sono tutti positivi e quindi ha comunque senso $\sum_{i=1}^{+\infty} x_i\, p(x_i)$; per le variabili con densità questo significa $f(x) = 0$ per $x < 0$ e quindi ha senso $\mathbf{E}[X] = \int_0^{+\infty} x\, f(x)\, \mathrm{d}x \in [0, +\infty]$.

Come conseguenza immediata, presa una generica v.a. X questa ha valore atteso se $\mathbf{E}[\,|X|\,] < +\infty$.

Vediamo che cosa succede quando abbiamo una *trasformazione* di una v.a., cioè una nuova v.a. Y della forma $g(X)$.

Proposizione 3.5.2. *Sia X discreta: la variabile $g(X)$ ammette valore atteso se $\sum_i |g(x_i)|\, p(x_i) < +\infty$, e in tal caso si ha*

$$\mathbf{E}[\,g(X)\,] = \sum_i g(x_i)\, p(x_i)$$

Sia X con densità $f(.)$: la variabile $g(X)$ ha valore atteso se $\int_{-\infty}^{+\infty} |g(x)|\, f(x)\, \mathrm{d}x < +\infty$, e in tal caso si ha

$$\mathbf{E}[\,g(X)\,] = \int_{-\infty}^{+\infty} g(x)\, f(x)\, \mathrm{d}x$$

Dimostrazione. Vediamo la dimostrazione nel caso discreto: cominciamo a supporre che i numeri $g(x_i)$ siamo tutti diversi, allora la la variabile $g(X)$ ha come valori $g(x_1)$, $g(x_2), \ldots$ con probabilità $p(x_1)$, $p(x_2), \ldots$ e la formula è dunque evidente. In generale però alcuni dei numeri $g(x_i)$ possono essere eguali e questo costringe a modificare la dimostrazione.

Cominciamo a supporre che la funzione g sia a valori positivi: la variabile $Y = g(X)$ prende un insieme (finito o) numerabile di valori $(y_1, y_2, \ldots)$. Consideriamo gli insiemi $A_i = \{j \,|\, g(x_j) = y_i\}$ e osserviamo che $p_Y(y_i) = \sum_{j \in A_i} \mathbf{P}\{X = x_j\} = \sum_{j \in A_i} p_X(x_j)$.

Poiché quelle che seguono sono somme di serie a termini positivi, possiamo usare la proprietà associativa della somma: si ottiene pertanto

$$\mathbf{E}[\,Y\,] = \sum_i y_i\, p_Y(y_i) \;=\; \sum_i y_i \Big(\sum_{j \in A_i} p_X(x_j) \Big) =$$

$$= \sum_i \Big(\sum_{j \in A_i} g(x_j)\, p_X(x_j) \Big) \;=\; \sum_j g(x_j)\, p_X(x_j)$$

cioè l'eguaglianza desiderata.

Il caso generale si ottiene scrivendo la funzione g nella forma $g = g^+ - g^-$ e sommando i due integrali (ricordiamo che con $g^+(x) = \max\big(g(x), 0\big)$ e $g^-(x) = -\min\big(g(x), 0\big)$ intendiamo la *parte positiva* e *parte negativa* della funzione g). □

Quando si passa alle variabili con densità, una dimostrazione rigorosa è impegnativa ma può essere istruttivo vedere qualche facile esempio. Consideriamo ad esempio X con densità uniforme su $[0, 1]$ e $Y = X^2$: se si calcola la densità di Y si trova facilmente che si ottiene lo stesso risultato ponendo $\mathbf{E}[X^2] = \displaystyle\int_0^1 x^2 \, \mathrm{d}x = 1/3$, oppure $\mathbf{E}[Y] = \displaystyle\int_0^1 \frac{y}{2\sqrt{y}} \, \mathrm{d}y = 1/3$.

Proprietà analoghe si hanno per una v.a. Z della forma $Z = g(X, Y)$.

Vediamo le prime proprietà del valore atteso:

Proposizione 3.5.3. *Supponiamo che X ed Y abbiano valore atteso, allora $X+Y$ ha valore atteso e valgono le seguenti proprietà:*

- $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y] \qquad \mathbf{E}[aX + b] = a\,\mathbf{E}[X] + b$
- $\big|\,\mathbf{E}[X]\,\big| \le \mathbf{E}[\,|X|\,]$
- *se $X \ge 0$, allora* $\mathbf{E}[X] \ge 0 \qquad$ *quindi se $X \ge Y$, allora* $\mathbf{E}[X] \ge \mathbf{E}[Y]$.

Vediamo la dimostrazione della prima proprietà nel caso delle variabili discrete: intanto $X+Y$ ha valore atteso se $\sum_{x_i, y_j} |x_i + y_j| \, p(x_i, y_j) < +\infty$. Ma

$$\sum_{x_i, y_j} |x_i + y_j| \, p(x_i, y_j) \le \sum_{x_i, y_j} (|x_i| + |y_j|) \, p(x_i, y_j) =$$

$$= \sum_{x_i} |x_i| \sum_{y_j} p(x_i, y_j) + \sum_{y_j} |y_j| \sum_{x_i} p(x_i, y_j) =$$

$$= \sum_{x_i} |x_i| \, p_X(x_i) + \sum_{y_j} |y_j| \, p_Y(y_j) < +\infty$$

Poiché le serie convergono asolutamente, possiamo ripetere i passaggi togliendo i valori assoluti (e la diseguaglianza diventa eguaglianza) e si trova proprio $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.

In generale, se X e Y hanno valore atteso, non è detto che il prodotto $X.Y$ lo abbia, ma c'è un caso particolare nel quale siamo sicuri che esista $\mathbf{E}[XY]$ e si ha anche una *regola del prodotto.*

Proposizione 3.5.4. *Supponiamo che X e Y abbiano valore atteso a che siano* **indipendenti***: anche XY ha valore atteso e vale la formula*

$$\mathbf{E}[X\,Y] = \mathbf{E}[X]\,.\,\mathbf{E}[Y]$$

Dimostrazione. Di nuovo ci limitiamo al caso discreto.

$$\sum_{x_i,y_j} |x_i y_j|\, p(x_i,y_j) = \sum_{x_i,y_j} |x_i|\,|y_j|\, p_X(x_i)\, p_Y(y_j) =$$

$$= \Big(\sum_{x_i} |x_i|\, p_X(x_i)\Big).\Big(\sum_{y_j} |y_j|\, p_Y(y_j\Big) < +\infty$$

Abbiamo usato la *proprietà distributiva del prodotto ripestto alla somma*, che vale anche per una serie di termini positivi (ed anche se la serie è assolutamente convergente), ed è così provato che XY ha valore atteso. Ripetendo le stesse eguaglianze senza valore assoluto, si ottiene la formula $\mathbf{E}[X\,Y] = \mathbf{E}[X]\,.\,\mathbf{E}[Y]$. □

Parliamo ora dei momenti, considerando un intero positivo n:

Definizione 3.5.5 (Momenti). Si dice che la v.a. X ammette momento di ordine n se $\mathbf{E}\big[\,|X|^n\,\big] < +\infty$ e in tal caso si chiama *momento di ordine n* il numero $\mathbf{E}\big[\,X^n\,\big]$.

Naturalmente il momento primo è il valore atteso. Il risultato che segue mostra che se una v.a. possiede momento n-simo, possiede anche tutti i momenti di ordine inferiore.

Proposizione 3.5.6. *Siano $1 \le m < n$: se $\mathbf{E}\big[\,|X|^n\,\big] < +\infty$, anche $\mathbf{E}\big[\,|X|^m\,\big] < +\infty$.*

La dimostrazione più elegante segue da una diseguaglianza importante chiamata *diseguaglianza di Jensen* dalla quale si ricava un risultato più preciso, cioè la diseguaglianza $\mathbf{E}\big[\,|X|^m\,\big]^{1/m} \le \mathbf{E}\big[\,|X|^n\,\big]^{1/n}$.

Ma per i nostri scopi si può dimostrare 3.5.6 in modo molto più elementare, partendo dalla diseguaglianza valida per ogni numero t, $|t|^m \le |t|^n + 1$. Si ha pertanto

$$\sum_{x_i} |x_i|^m\, p(x_i) \le \sum_{x_i} (\,|x_i|^n + 1)\, p(x_i) = \sum_{x_i} |x_i|^n\, p(x_i) + \sum_{x_i} p(x_i) < +\infty$$

e quindi $\mathbf{E}[\,|X|^n] < +\infty \Rightarrow \mathbf{E}[\,|X|^m] < +\infty$.

3.6 Varianza, covarianza, correlazione

Cominciamo con due importanti diseguaglianze.

Proposizione 3.6.1 (**Disuguaglianza di Markov**). *Sia Y una variabile aleatoria a valori positivi ed $a > 0$: vale la disuguaglianza*

$$a\,\mathbf{P}\big\{Y \geq a\big\} \leq \mathbf{E}\,[Y]$$

Proposizione 3.6.2 (**Disuguaglianza di Schwartz**). *Siano X e Y due variabili aleatorie: vale la disuguaglianza*

$$\mathbf{E}\big[\,|XY|\,\big] \leq \sqrt{\mathbf{E}[X^2]}\,.\,\sqrt{\mathbf{E}[Y^2]}$$

Non si fa alcuna ipotesi sull'esistenza dei momenti, le diseguaglianze sopra scritte hanno senso e sono verificate anche se qualcuno dei valori attesi è infinito.

Cominciamo con 3.6.1 ed introduciamo una variabile ausiliaria Z così definita

$$Z(\omega) = \begin{cases} a & \text{se } Y(\omega) \geq a \\ 0 & \text{se } Y(\omega) < a \end{cases}$$

È evidente che vale la diseguaglianza $Z(\omega) \leq Y(\omega)$ e quindi $\mathbf{E}[Z] \leq \mathbf{E}[Y]$, e si ha $\mathbf{E}[Z] = a\,\mathbf{P}\{Y \geq a\}$.

Quanto alla diseguaglianza 3.6.2, il principio della dimostrazione è simile alla diseguaglianza classica di Schwartz su $\mathbb{R}^n$, sorvoliamo sui dettagli ma è importante sottolineare una conseguenza. Abbiamo detto che se X e Y hanno valore atteso non è detto che il prodotto XY abbia valore atteso (a parte il caso in cui sono indipendenti), tuttavia *se X e Y hanno momento secondo, il prodotto XY ha valore atteso.*

Per tutto questo paragrafo supponiamo che le variabili X, Y, ... considerate abbiano momento secondo.

Definizione 3.6.3 (**Varianza e scarto quadratico medio**). Si chiama *varianza* di una v.a. X il numero $Var(X) = \mathbf{E}\big[\,\big(X - \mathbf{E}[X]\big)^2\,\big]$. Si chiama *scarto quadratico medio* o anche *deviazione standard* la radice della varianza di X.

Sviluppando i calcoli si trova l'eguaglianza $Var(X) = \mathbf{E}[\,X^2\,] - \mathbf{E}[X]^2$ (di solito è più comodo calcolare la varianza in questo modo). Lo scarto quadratico medio è indicato $\sigma(X)$ e per questo la varianza è anche indicata $\sigma^2(X)$. Notiamo anche che si ha

$$Var(aX + b) = a^2\,.\,Var(X)$$

Se $x = (x_1, \ldots, x_n)$ è un insieme di dati, la *varianza empirica* di x definita nel primo capitolo coincide con la varianza di una v.a. discreta X che prende i valori $x_1, \ldots, x_n$ ciascuno con probabilità $1/n$.

Avevamo visto che la varianza empirica misura la dispersione dei dati intorno al loro valore medio, la stessa cosa succede con le variabili aleatorie grazie alla seguente disuguaglianza.

Proposizione 3.6.4 (**Disuguaglianza di Chebyshev**). *Preso $d > 0$ si ha*

$$\mathbf{P}\big\{\,\big|X - \mathbf{E}[X]\,\big| > d\big\} \leq \frac{Var(X)}{d^2}$$

Dimostrazione. Questa diseguaglianza deriva immediatamente dalla *disuguaglianza di Markov* 3.6.1 prendendo $Y = \big(X - \mathbf{E}[X]\,\big)^2$, $a = d^2$ e tenendo presente che

$$\mathbf{P}\big\{\,\big|X - \mathbf{E}[X]\,\big| > d\big\} = \mathbf{P}\big\{\,\big(X - \mathbf{E}[X]\,\big)^2 > d^2\big\}$$

□

Il caso limite si ha quando la varianza è 0: infatto *la varianza della variabile X è eguale a 0 se e solo se X è costante eccetto che su un insieme trascurabile, cioè* $\mathbf{P}\big\{\,\big|\,X - \mathbf{E}[X]\,\big| \neq 0\big\} = 0$.

Definizione 3.6.5 (**Covarianza**). Si chiama covarianza tra X e Y il numero

$$Cov(X,Y) = \mathbf{E}\big[\,\big(X - \mathbf{E}[X]\big)\,\big(Y - \mathbf{E}[Y]\big)\,\big] = \mathbf{E}\big[\,XY\,\big] - \mathbf{E}[X]\,\mathbf{E}[Y]$$

La covarianza soddisfa le proprietà sotto elencate, che è molto facile verificare direttamente:

- $Cov(aX + bY + c\,,\, Z) = a\,Cov(X,Z) + b\,Cov(Y,Z)$
- $Cov(X,Y) = Cov(Y,X) \qquad Var(X) = Cov(X,X)$
- $Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$

Il risultato che segue è sostanzialmente già dimostrato sopra.

Proposizione 3.6.6. *Prese due v.a. X e Y, sono equivalenti queste tre proprietà:*

- $\mathbf{E}[X\,Y] = \mathbf{E}[X]\,\mathbf{E}[Y]$
- $Cov(X,Y) = 0$

- $Var(X+Y) = Var(X) + Var(Y)$

Quando $Cov(X,Y) = 0$, le variabili sono dette **incorrelate** (o anche *scorrelate*); due variabili indipendenti sono incorrelate, ma non è vero il viceversa.

Definizione 3.6.7 (**Coefficiente di correlazione).** Se $Var(X) \neq 0$ e $Var(Y) \neq 0$, si chiama *coefficiente di correlazione* il numero

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\,\sigma(Y)}$$

Grazie alla diseguaglianza 3.6.2 si ha

$$\big|Cov(X,Y)\,\big| \leq \mathbf{E}\big[\,\big|X - \mathbf{E}[X]\big|\,\big|Y - \mathbf{E}[Y]\big|\,\big] \leq$$

$$\leq \sqrt{\mathbf{E}\big[\,(X - \mathbf{E}[X])^2\,\big]}\ \sqrt{\mathbf{E}\big[\,(Y - \mathbf{E}[Y])^2\,\big]} = \sigma(X)\,\sigma(Y)$$

e di conseguenza $|\rho(X,Y)| \leq 1$. Inoltre, con un procedimento sostanzialmente identico a quello usato nel primo capitolo per trovare la *retta di regressione*, si prova il seguente risultato

$$\min_{a,b\in\mathbb{R}^2} \mathbf{E}\big[\,\big(Y - a - bX\big)^2\,\big] = Var(Y)\,.\,\big(1 - \rho(X,Y)^2\big)$$

3.7 Alcuni esempi

Esempio 3.7.1 (**Variabile Binomiale).** Cominciamo con la variabile di Bernoulli: prende il valore 1 con probabilità p ed il valore 0 con probabilita $(1-p)$, ed è dunque immediato constatare che si ha $\mathbf{E}[X] = p$, $\mathbf{E}[X^2] = p$ e quindi $Var(X) = p - p^2 = p\,(1-p)$.

Poiché una variabile X *Binomiale* di parametri n e p può essere vista come somma di n variabili di Bernoulli di parametro p indipendenti, si ha $\mathbf{E}[X] = np$ e $Var(X) = np\,(1-p)$.

Esempio 3.7.2 (**Variabile di Poisson).** Presa X di Poisson di parametro λ, a priori non sappiamo se ha momento primo ma poiché prende solo valori positivi ha sicuramente senso scrivere

$$\mathbf{E}[X] = \sum_{h=0}^{+\infty} h\,\mathrm{e}^{-\lambda}\frac{\lambda^h}{h!} = \lambda\mathrm{e}^{-\lambda}\sum_{h=1}^{+\infty}\frac{\lambda^{h-1}}{(h-1)!} = \lambda\mathrm{e}^{-\lambda}\sum_{k=0}^{+\infty}\frac{\lambda^k}{k!} = \lambda$$

Con conti simili si prova $\mathbf{E}[X^2] = \lambda + \lambda^2$ e quindi $Var(X) = \lambda$.

Esempio 3.7.3 (**Densità uniforme).** Premettiamo una osservazione: se la densità di una v.a. X è diversa da 0 solo su un intervallo limitato, sicuramente la variabile *possiede tutti i momenti.*

Prendiamo ora X con densità uniforme su $[a,b]$:

$$\mathbf{E}[X] = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{x^2}{2\,(b-a)}\Big|_a^b = \frac{b^2-a^2}{2\,(b-a)} = \frac{a+b}{2}$$

$$\mathbf{E}[X^2] = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{x^3}{3\,(b-a)}\Big|_a^b = \frac{b^3-a^3}{3\,(b-a)} = \frac{a^2+ab+b^2}{3}$$

e quindi $Var(X) = \dfrac{(b-a)^2}{12}$.

Esempio 3.7.4 (**Densità esponenziale).** Anche di questa variabile a priori non sappiamo quali momenti possieda, tuttavia (essendo la densità diversa da 0 solo per x positivo) possiamo calcolare (con un cambio di variabili)
$\mathbf{E}[X] = \int_0^{+\infty} \lambda\, x\, \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}\int_0^{+\infty} t\,\mathrm{e}^{-t}\,\mathrm{d}t = \frac{1}{\lambda}$

Con conti simili si trova $\mathbf{E}[X^2] = \dfrac{2}{\lambda^2}$ e quindi $Var(X) = \dfrac{1}{\lambda^2}$.

Esempio 3.7.5 (**Variabili Gaussiane).** Cominciamo con X Gaussiana standard: intanto è facile convincersi del fatto che, qualunque sia n,

$$\frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{+\infty} |x|^n\,\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x < +\infty$$

e quindi la variabile *possiede tutti i momenti.* Inoltre se $n = 2h+1$ è dispari

$$\mathbf{E}[X^n] = \frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{+\infty} x^n\,\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = \lim_{M\to\infty}\frac{1}{\sqrt{2\,\pi}}\int_{-M}^{+M} x^n\,\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = 0$$

Quindi rimangono da calcolare gli n-momenti con n pari, cominciamo col momento secondo. Con una integrazione per parti (considerando che $-x\,\mathrm{e}^{-\frac{x^2}{2}}$ è la derivata di $\mathrm{e}^{-\frac{x^2}{2}}$) si ottiene

$$\mathbf{E}[X^2] = \frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{+\infty} x^2\,\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x =$$

$$= \frac{-x\,\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\,\pi}}\Big|_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = 0+1 = 1$$

e dunque anche $Var(X) = 1$.

Con conti analoghi si prova l'eguaglianza $\mathbf{E}\big[\,X^{2h+2}\,\big] = (2h+1)\,\mathbf{E}\big[\,X^{2h}\,\big]$ e quindi ad esempio $\mathbf{E}[X^4] = 3$, $\mathbf{E}[X^6] = 5\times 3 = 15$ e così via ...

Per quanto riguarda le variabili gaussiane più generali, una variabile Y con densità $N(m,\sigma^2)$ si scrive nella forma $(\sigma X + m)$ con X standard e quindi $\mathbf{E}[Y] = \mathbf{E}[\sigma X + m] = m$ e $Var(Y) = Var(\sigma X + m) = \sigma^2\, Var(X) = \sigma^2$.

In modo analogo si procede con i momenti di ordine superiore.

3.8 Appendice

3.8.1 Breve introduzione agli integrali impropri

Nel corso di Analisi del I anno è stato definito l'integrale di una funzione continua su un intervallo limitato: si parla di **integrale improprio** quando si integra su un *intervallo non limitato* (ad esempio una semiretta o tutta la retta) oppure quando la funzione *tende a infinito* ($\pm\infty$) in uno o più punti. Richiamo brevemente i punti essenziali della teoria senza dare dimostrazioni complete.

Cominciamo a considerare il caso nel quale **la funzione f da integrare è a valori positivi**: in questo caso se si integra su un intervallo infinito si considera il limite dell'integrale su intervalli finiti sempre più grandi o nel secondo caso il limite dell'integrale su intervalli che non contengono i punti sui quali la funzione tende a $+\infty$.

Ci spieghiamo meglio con due esempi, considerando $a > 0$:

$\int_1^{+\infty} \frac{1}{x^a}\,\mathrm{d}x = \lim_{M\to\infty} \int_1^M \frac{1}{x^a}\,\mathrm{d}x$

$\int_0^1 \frac{1}{x^a}\,\mathrm{d}x = \lim_{\varepsilon\to 0+} \int_\varepsilon^1 \frac{1}{x^a}\,\mathrm{d}x$

Notiamo che in entrambi i casi l'integrale è crescente (rispetto a $M\to\infty$ nel primo caso e rispetto a $\varepsilon\to 0+$ nel secondo caso) e quindi il limite esiste, però potrebbe anche essere $+\infty$. Esaminiamo in dettaglio i due casi:

$$\int_1^{+\infty} \frac{1}{x^a}\,\mathrm{d}x = \begin{cases} +\infty & \text{se} \quad a \le 1 \\ \frac{1}{a-1} & \text{se} \quad a > 1 \end{cases}$$

$$\int_0^1 \frac{1}{x^a}\,\mathrm{d}x = \begin{cases} +\infty & \text{se} \quad a \ge 1 \\ \frac{1}{1-a} & \text{se} \quad a < 1 \end{cases}$$

Vediamo il primo caso: la primitiva della funzione $\frac{1}{x^a} = x^{-a}$ è $\frac{x^{1-a}}{1-a}$ se $a \neq 1$, mentre è eguale a $\log x$ se $a = 1$.
Si ha pertanto nel primo caso (se $a \neq 1$):

$$\int_1^{+\infty} \frac{1}{x^a}\,\mathrm{d}x = \lim_{M\to\infty} \int_1^M \frac{1}{x^a}\,\mathrm{d}x = \lim_{M\to\infty} \frac{x^{1-a}}{1-a}\Big|_1^M = \begin{cases} +\infty & a < 1 \\ \frac{1}{a-1} & a > 1 \end{cases}$$

Quando $a = 1$ si ottiene $\int_1^{+\infty} \frac{1}{x}\,\mathrm{d}x = \lim_{M\to\infty} \log x\Big|_1^M = +\infty$.

Nel secondo caso si ha (sempre se $a \neq 1$)

$$\int_0^1 \frac{1}{x^a}\,\mathrm{d}x = \lim_{\varepsilon\to 0+} \int_\varepsilon^1 \frac{1}{x^a}\,\mathrm{d}x = \lim_{\varepsilon\to 0+} \frac{x^{1-a}}{1-a}\Big|_\varepsilon^1 = \begin{cases} +\infty & a > 1 \\ \frac{1}{1-a} & a < 1 \end{cases}$$

A volte si scrive più brevemente ad esempio $\frac{x^{1-a}}{1-a}\big|_1^{+\infty}$ per indicare $\lim_{M\to\infty} \frac{x^{1-a}}{1-a}\big|_1^M$ (e in modo analogo $\frac{x^{1-a}}{1-a}\big|_0^1$).

Vediamo ora il caso nel quale **la funzione f da integrare prende anche valori negativi**: in questo caso non si può andare direttamente al limite come si è fatto in precedenza ma bisogna prima considerare l'integrale del valore assoluto $|f(x)|$ e controllare che questo integrale sia finito. Se $\int |f(x)|\,\mathrm{d}x < +\infty$ si può andare al limite (senza valore assoluto), altrimenti la funzione non è integrabile.

Questo è analogo a quanto si è visto per le serie numeriche, in effetti c'è similitudine tra somma (infinita) di una serie e integrale improprio.

Vediamo subito con un esempio perchè è fondamentale richiedere che l'integrale del valore assoluto sia finito: vogliamo esaminare $\int_{-\infty}^{+\infty} \frac{x}{1+x^2}\,\mathrm{d}x$.

Si ha $\int_{-\infty}^{+\infty} \frac{|x|}{1+x^2}\,\mathrm{d}x = 2\int_0^{+\infty} \frac{x}{1+x^2}\,\mathrm{d}x = \log(1+x^2)\big|_0^{+\infty} = +\infty$ e dunque la funzione non è integrabile.

Tuttavia, preso $M>0$, $\int_{-M}^{M} \frac{x}{1+x^2}\,\mathrm{d}x = 0$ e quindi siamo tentati di scrivere

$$\int_{-\infty}^{+\infty} \frac{x}{1+x^2}\,\mathrm{d}x = \lim_{M\to\infty} \int_{-M}^{M} \frac{x}{1+x^2}\,\mathrm{d}x = 0$$

Vediamo quali sarebbero i guai se fosse valida una definizione simile: per una generica funzione f integrabile, noi ci aspettiamo che si abbia $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{0} f(x)\,\mathrm{d}x + \int_0^{+\infty} f(x)\,\mathrm{d}x$.

Ma in questo caso $\int_{-\infty}^{0} \frac{x}{1+x^2}\,\mathrm{d}x = -\infty$ e $\int_0^{+\infty} \frac{x}{1+x^2}\,\mathrm{d}x = +\infty$ e prendendo la regola precedente avremmo $\int_{-\infty}^{+\infty} \frac{x}{1+x^2}\,\mathrm{d}x = -\infty+\infty$ che ovviamente non ha senso.

L'esempio precedente non è completamente a caso: si ha $\int_{-\infty}^{+\infty} \frac{1}{1+x^2}\,\mathrm{d}x = \arctan(x)\big|_{-\infty}^{+\infty} = \pi$ e dunque la funzione $f(x) = \frac{1}{\pi(1+x^2)}$ è una densità di probabilità (chiamata *densità di Cauchy*).

I calcoli precedenti mostrano che *una v.a. che ha densità di Cauchy non ha momento primo* (e a maggior ragione non ha momenti di ordine superiore).

La densità di Cauchy non ha un particolare interesse statistico, ci serve soprattutto come esempio di una variabile che non ha momenti. È anche facile provare che *la funzione generatrice dei momenti di una variabile con densità di Cauchy ha come dominio il solo punto 0.*

Vediamo un paio di maggiorazioni che ci saranno utili. La prima è questa: *preso $\lambda>0$, qualunque sia la potenza n si ha* $\int_0^{+\infty} x^n\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x < +\infty$.

Infatti, poiché $\lim_{x\to\infty} x^{n+2}\,\mathrm{e}^{-\lambda x} = 0$, esiste $C > 0$ tale che valga, per $x > C$

$$x^n\,\mathrm{e}^{-\lambda x} \le \frac{1}{x^2}$$

e di conseguenza

$$\int_0^{+\infty} x^n\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x \le \int_0^{C} x^n\,\mathrm{e}^{-\lambda x}\,\mathrm{d}x + \int_C^{+\infty} \frac{1}{x^2}\,\mathrm{d}x < +\infty$$

Come conseguenza *una variabile aleatoria con densità esponenziale di parametro λ possiede momenti di ogni ordine.*

Vediamo un'altra diseguaglianza: *qualunque sia la potenza n, si ha* $\int_{-\infty}^{+\infty} |x|^n\,\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = 2\int_0^{+\infty} x^n\,\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x < +\infty$.

La dimostrazione si fa in modo analogo a quanto appena visto, tenendo conto del fatto che la funzione $\mathrm{e}^{-\frac{x^2}{2}}$ tende a 0 (per $x \to \infty$) *molto più velocemente della funzione* $\mathrm{e}^{-\lambda x}$.

Come conseguenza *una variabile aleatoria con densità gaussiana N(0,1) possiede momenti di ogni ordine.*

3.8.2 Alcuni calcoli con gli integrali doppi

Dimostriamo l'uguaglianza (scoperta da Laplace) $\int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = \sqrt{2\pi}$.

$$\Big(\int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x\Big)^2 = \Big(\int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x\Big)\cdot\Big(\int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{y^2}{2}}\,\mathrm{d}y\Big) = \iint_{\mathbb{R}^2} \mathrm{e}^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y$$

e passando a coordinate polari ($x = r\cos(\theta)$, $y = r\sin(\theta)$) e usando la formula del cambio di variabili per integrali doppi si ottiene

$$\iint_{\mathbb{R}^2} \mathrm{e}^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y = \int_0^{2\pi}\mathrm{d}\theta\int_0^{+\infty} r\,\mathrm{e}^{-\frac{r^2}{2}}\,\mathrm{d}r = 2\pi$$

Diversi calcoli e formule in questo capitolo sono stati provati per variabili discrete ed enunciati per variabili con densità, sempre trasformando le somme in integrali: senza volerli ripetere tutti, vediamo ad esempio la *formula della convoluzione*. Siano dunque X e Y indipendenti, con densità rispettive f_X e f_Y, e sia $Z = X + Y$: partiamo dalla funzione di ripartizione

$$F_Z(z) = \mathbf{P}\big\{X + Y \le z\big\} = \iint_{\{(x,y)\,|\,x+y\le z\}} f_X(x)\,f_Y(y)\,\mathrm{d}x\,\mathrm{d}y =$$

$$= \int_{-\infty}^{+\infty} \mathrm{d}y \int_{-\infty}^{z-y} f_X(x)\, f_Y(z-x)\, \mathrm{d}x$$

e derivando dentro l'integrale rispetto a z si ottiene per la densità la formula

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y)\, f_Y(y)\, \mathrm{d}y$$

Capitolo 4

Complementi sulle variabili aleatorie e teoremi limite

4.1 La funzione generatrice dei momenti

La *Funzione generatrice dei momenti* è uno strumento non indispensabile, ma molto comodo per fare dei calcoli come vedremo poi negli esempi. Consideriamo una v.a. X e prendiamo e^{tX}: questa è una v.a. a valori positivi e quindi ha senso scrivere $\mathbf{E}\big[\,\mathrm{e}^{tX}\,\big]$ (numero positivo che può anche essere infinito).

Definizione 4.1.1 (**Funzione generatrice dei momenti).** Si chiama *Funzione generatrice dei momenti* della v.a. X la funzione

$$G_X(t) = \mathbf{E}\big[\,\mathrm{e}^{tX}\,\big] = \begin{cases} \sum_{x_i} \mathrm{e}^{tx_i}\, p(x_i) & X \text{ discreta} \\ \int_{\mathbb{R}} \mathrm{e}^{tx}\, f(x)\,\mathrm{d}x & X \text{ con densità} \end{cases}$$

Si chiama **dominio** di $G_X(.)$ l'insieme dei t tali che $G_X(t) < +\infty$.

Qualunque sia la variabile X si ha $G_X(0) = 1$; in alcuni casi il *dominio* di G_X è il solo punto 0, ad esempio con la *densità di Cauchy* definita da $f(x) = \dfrac{1}{\pi\,(1+x^2)}$. Quando il dominio non è il solo punto 0, è un *intervallo* in senso lato (può essere una semiretta oppure tutta la retta).

Il risultato fondamentale che riguarda la funzione generatrice è il seguente: *due variabili che hanno la stessa funzione generatrice dei momenti sono equidistribuite.* Bisogna però fare una precisazione: è possibile che il dominio di G_X e G_Y sia il solo punto 0 e che tuttavia X e Y non siano equidistribuite. Allora l'enunciato preciso è il seguente:

Teorema 4.1.2. *Supponiamo che i domini di $G_X(.)$ e di $G_Y(.)$ contengano un intervallo aperto $]a,b[$ e che si abbia $G_X(t) = G_Y(t)$ per ogni $t \in]\,a,b\,[$: allora X e Y sono* **equidistribuite**.

Omettiamo la dimostrazione, che è avanzata: limitatamente alle variabili con densità è una conseguenza della *trasformata di Laplace*. Vediamo viceversa alcune facili proprietà:

Proposizione 4.1.3. *Valgono le due seguenti regole di calcolo:*
a) $G_{aX+b}(t) = G_X(at)\,\mathrm{e}^{tb}$
b) se X e Y sono indipendenti, $G_{X+Y}(t) = G_X(t)\,G_Y(t)$

Dimostrazione. La proprietà a) è immediata: $G_{aX+b}(t) = \mathbf{E}\big[\,\mathrm{e}^{(aX+b)t}\,\big] = \mathbf{E}\big[\,\mathrm{e}^{(at)X}\,\mathrm{e}^{tb}\,\big] = \mathrm{e}^{tb}\,G_X(at)\,.$

Per quanto riguarda invece la b), si parte dall'osservazione che le variabili e^{tX} e e^{tY} sono indipendenti e si applica 3.5.4: si ha dunque

$$G_{X+Y}(t) = \mathbf{E}\big[\,\mathrm{e}^{t(X+Y)}\,\big] = \mathbf{E}\big[\mathrm{e}^{tX}\,\mathrm{e}^{tY}\big] = \mathbf{E}\big[\,\mathrm{e}^{tX}\,\big]\,\mathbf{E}\big[\,\mathrm{e}^{tY}\,\big] = G_X(t)\,G_Y(t)$$

□

Viene da domandarsi perché si chiama *Funzione generatrice dei momenti*: fermo restando che solitamente non conviene usare la funzione generatrice per calcolare i momenti (a parte casi particolari), il risultato che segue mostra che quando il dominio della funzione generatrice contiene un intorno di 0 è possibile utilizzarla per calcolare i momenti.

Teorema 4.1.4. *Supponiamo che $G_X(.)$ sia finita per t con $-\varepsilon < t < \varepsilon$: allora la variabile X possiede tutti i momenti e vale la formula*

$$\mathbf{E}\big[\,X^n\,\big] = \frac{\mathrm{d}^n G_X(t)}{\mathrm{d}t^n}\Big|_{t=0}$$

La dimostrazione rigorosa di questo teorema richiede proprietà di *derivazione sotto l'integrale* che non possediamo, accontentiamoci di una giustificazione intuitiva. Accettando che si possa *far entrare la derivata dentro il valore atteso*, si ottiene

$$\frac{\mathrm{d}\,G_X(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\big[\,\mathrm{e}^{tX}\big] = \mathbf{E}\big[\frac{\mathrm{d}\,e^{tX}}{\mathrm{d}t}\big] = \mathbf{E}\big[X\,\mathrm{e}^{tX}\big]$$

e, per $t = 0$, si ha proprio $G'_X(0) = \mathbf{E}[X]\,.$

Vediamo ora degli esempi, premettendo che tutti questi risultati possono essere ottenuti senza definire la funzione generatrice dei momenti ma facendo calcoli diretti o utilizzando la formula della convoluzione (discreta o per densità), tuttavia con questo procedimento tutto diventa più facile.

Esempio 4.1.5 (**Variabile geometrica).** Il valore atteso di una variabile geometrica di parametro p si può ottenere con calcolo diretto, ma bisogna *derivare per serie*; vediamo un altro modo (anche questo non rapidissimo) calcolando la funzione generatrice dei momenti.

$$G_X(t) = \sum_{k=1}^{+\infty} \mathrm{e}^{tk}(1-p)^{k-1}p = p\,\mathrm{e}^t \sum_{h=0}^{+\infty} \big(\mathrm{e}^t(1-p)\big)^h$$

Questa serie ha somma finita se $\big|\mathrm{e}^t(1-p)\big| < 1$, cioè $t < -\log(1-p)$, e in tal caso vale $\dfrac{p\,\mathrm{e}^t}{1-\mathrm{e}^t(1-p)}$; calcolando la derivata in 0 si ottiene $\mathbf{E}[X] = 1/p$.

Esempio 4.1.6 (**Variabile di Poisson).** Se X è di Poisson di parametro λ, si ha

$$G_X(t) = \sum_{k=0}^{+\infty} \mathrm{e}^{tk}\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\sum_{k=0}^{+\infty}\frac{(\mathrm{e}^t\,\lambda)^k}{k!} = \mathrm{e}^{\lambda(\mathrm{e}^t-1)}$$

Come conseguenza immediata, sfruttando 4.1.2 e 4.1.3, si trova che se X e Y sono indipendenti di Poisson rispettivamente di parametri λ e μ, allora $X+Y$ è di Poisson di parametro $\lambda+\mu$.

Esempio 4.1.7 (**Variabili Gaussiane).** Cominciamo a calcolare la funzione generatrice di una variabile gaussiana standard:

$$G_X(t) = \frac{1}{\sqrt{2\pi}}\int_{+\infty}^{+\infty} \mathrm{e}^{tx}\,\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = \frac{\mathrm{e}^{\frac{t^2}{2}}}{\sqrt{2\pi}}\int_{+\infty}^{+\infty}\mathrm{e}^{-\frac{(x-t)^2}{2}}\,\mathrm{d}x = \mathrm{e}^{\frac{t^2}{2}}$$

Di conseguenza, per una v.a. Y con densità $N(m,\sigma^2)$ (che possiamo pensare della forma $\sigma X + m$ con X standard), si ha $G_Y(t) = \mathrm{e}^{mt+\frac{\sigma^2t^2}{2}}$ e si trova il seguente risultato: se X e Y sono indipendenti e gaussiane rispettivamente $N(m_1,\sigma_1^2)$ e $N(m_2,\sigma_2^2)$, allora $Z = X+Y$ è gaussiana $N(m_1+m_2, \sigma_1^2+\sigma_2^2)$.

4.2 Due teoremi limite fondamentali

Cominciamo ad enunciare i due principali *teoremi limite*, e poi li discutiamo. Premettiamo una notazione: di fronte a una successione X_1, X_2, ... di variabili aleatorie *indipendenti* ed *equidistribuite*, scriveremo per brevità **i.i.d.** (da Independent Identically Distributed).

Teorema 4.2.1 (**Legge debole dei Grandi Numeri).** *Sia $X_1, X_2, \ldots$ una successione di variabili i.i.d. dotate di momento secondo, e sia $\mu = \mathbf{E}[X_i]$ il loro valore atteso: per ogni $\varepsilon > 0$ si ha*

$$\lim_{n\to\infty} \mathbf{P}\Big\{\Big|\frac{X_1+\cdots+X_n}{n} - \mu\Big| > \varepsilon\Big\} = 0$$

Teorema 4.2.2 (**Teorema Limite Centrale**). *Sia $X_1, X_2, \dots$ una successione di variabili i.i.d. con valore atteso $\mathbf{E}[X_i] = \mu$ e varianza $\sigma^2(X_i) = \sigma^2 > 0$: presi $-\infty \le a < b \le +\infty$ si ha*

$$\lim_{n\to\infty} \mathbf{P}\Big\{ a \le \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le b \Big\} = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-\frac{x^2}{2}}\, \mathrm{d}x = \Phi(b) - \Phi(a)$$

Cominciamo a guardare il primo risultato e a dare un senso alla convergenza, con questa definizione:

Definizione 4.2.3. Si dice che la successione di v.a. $X_1, X_2, \dots$ *converge in probabilità* alla v.a. X se, per ogni $\varepsilon > 0$ si ha

$$\lim_{n\to\infty} \mathbf{P}\big\{\, \big|X_n - X\big| > \varepsilon \big\} = 0$$

Il limite può anche essere una costante c; non ci dilunghiamo sulla convergenza in probabilità ma ci limitiamo a questo facile risultato che offre una *condizione sufficiente* di convergenza in probabilità.

Proposizione 4.2.4. *Supponiamo che si abbia*

$$\lim_{n\to\infty} \mathbf{E}\big[\, X_n \,\big] = c \qquad\qquad \lim_{n\to\infty} Var(X_n) = 0$$

allora la successione $(X_n)_{n\ge1}$ converge in probabilità alla costante c.

Dimostrazione. Si parte dalla disuguaglianza seguente (che si prova esattamente come la 3.6.4 facendola derivare da 3.6.1):

$$\mathbf{P}\big\{\, \big|X_n - c\big| > \varepsilon \big\} \le \frac{\mathbf{E}\big[\, (X_n - c)^2 \,\big]}{\varepsilon^2}$$

Ma $\mathbf{E}\big[\, (X_n - c)^2 \,\big] = \mathbf{E}\big[\, (X_n - \mathbf{E}[X_n] + \mathbf{E}[X_n] - c)^2 \,\big] =$
$\mathbf{E}[(X_n - \mathbf{E}[X_n])^2] + 2\,\mathbf{E}\big[X_n - \mathbf{E}[X_n]\,\big].\big(\mathbf{E}[X_n] - c\big) + \big(\mathbf{E}[X_n] - c\big)^2 =$
$= Var(X_n) + \big(\mathbf{E}[X_n] - c\big)^2$
Di conseguenza $\lim_{n\to\infty} \mathbf{E}\big[\, (X_n - c)^2 \,\big] = 0$ e questo conclude la dimostrazione. □

A questo punto è facile provare il teorema 4.2.1: detto infatti $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$ si ha

$$\mathbf{E}\big[\overline{X}_n\big] = \frac{\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]}{n} = \mu \qquad \text{e}$$

$$\lim_{n\to\infty} Var(\overline{X}_n) = \lim_{n\to\infty} \frac{Var(X_1) + \cdots + Var(X_n)}{n^2} = \lim_{n\to\infty} \frac{Var(X_1)}{n} = 0\,.$$

In linguaggio discorsivo si può dire che *la media empirica dei primi n termini converge alla media teorica.*

In realtà le ipotesi possono essere indebolite: è sufficiente che le variabili siano *incorrelate*, che abbiano tutte valore atteso μ e varianze *equilimitate* (cioè esiste C tale che $Var(X_i) \leq C$ per ogni i).

Il fatto che si chiami legge *debole* dei grandi numeri fa pensare che esistano delle versioni più generali: in effetti è così ma questo va al di là degli scopi di questo corso.

Veniamo ora al *teorema limite centrale*: il tipo di convergenza coinvolto in questo risultato si chiama *convergenza in distribuzione*, ma anche solo la definizione di questo tipo di convergenza è delicata. Questa definizione è rinviata all'appendice, insieme a qualche spunto sulla dimostrazione di 4.2.2.

Agli effetti pratici, possiamo dire che nelle ipotesi del teorema 4.2.2 la variabile

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}$$

quando n è grande, è *approssimativamente* Gaussiana standard.

Naturalmente *n grande* non è un termine preciso: deve essere almeno $n \geq 50$, ma per una buona approssimazione almeno $n \geq 80$.

Questo teorema è stato dimostrato da Laplace (con contributi precedenti di De Moivre) limitatatmente al caso in cui le X_i sono di *Bernoulli* di parametro p (e quindi $X_1 + \cdots + X_n$ risulta Binomiale di parametri n e p): in questo caso la dimostrazione segue dalla *Formula di Stirling*, non è difficile ma occorre fare dei conti piuttosto intricati. Ricordo che la formula di Stirling è una *valutazione approssimata* di $n!$, più precisamente si ha $n! \approx \sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n$.

La versione più generale del teorema che abbiamo enunciato è dovuta a Paul Lévy, e la sua dimostrazione richiede strumenti più avanzati, ma chiarisce bene il posto *assolutamente centrale* che le variabili Gaussiane ricoprono in statistica.

Uno degli usi più frequenti è l'approssimazione di variabili Binomiali con n grande: poiché possiamo pensare una variabile X Binomiale di parametri n e p come somma di n variabili di Bernoulli di parametro p , segue che la variabile $\dfrac{X - np}{\sqrt{n\,p(1-p)}}$ è approssimativamente $N(0,1)$. In questo caso poi ci sono valutazioni più precise della numerosità n : perché l'approssimazione sia buona, si deve avere $n\,p(1-p) \geq 15$, perché sia ottima si deve avere $n\,p(1-p) \geq 20$.

Esempio 4.2.5. Gli aerei della compagnia XYZ hanno 180 posti, ma la compagnia sa che in media quelli che si sono prenotati si presentano alla

partenza con probabilità 0.9 : per questo la compagnia mette in vendita 195 biglietti (*overbooking*). Ci domandiamo qual è la probabilità che qualche cliente rimanga a terra e di conseguenza la compagnia sia costretta a un risarcimento.

Il numero aleatorio X di clienti che si presentano è una variabile Binomiale di parametri 195 e 0.9, e bisogna calcolare $\mathbf{P}\{X \geq 181\}$: in base al teorema 4.2.2 la variabile

$$\frac{X - 195 \times 0.9}{\sqrt{195 \times 0.1 \times 0.9}} = \frac{X - 175.5}{4.18}$$

è approssimativamente gaussiana standard. Di conseguenza

$$\mathbf{P}\{X \geq 181\} = \mathbf{P}\Big\{\frac{X - 175.5}{4.18} \geq \frac{181 - 175.5}{4.18}\Big\} =$$

$$= \mathbf{P}\Big\{\frac{X - 175.5}{4.18} \geq 1.31\Big\} \approx 1 - \Phi(1.31) \approx 0.095$$

4.3 Altre densità di rilevante interesse in statistica

4.3.1 Densità Gamma

Cominciamo col definire la **funzione Gamma di Eulero**: è definita, per $r > 0$, da $\Gamma(r) = \int_0^{+\infty} x^{r-1}\,\mathrm{e}^{-x}\,\mathrm{d}x$. Questo integrale non si può calcolare direttamente (a parte quando r è intero) ma, se $r > 1$, vale la relazione $\Gamma(r) = (r-1)\,\Gamma(r-1)$: infatti integrando per parti si ottiene

$$\int_0^{+\infty} x^{r-1}\,\mathrm{e}^{-x}\,\mathrm{d}x = -x^{r-1}\,\mathrm{e}^{-x}\Big|_0^{+\infty} + (r-1)\int_0^{+\infty} x^{r-2}\,\mathrm{e}^{-x}\,\mathrm{d}x$$

Poiché $\Gamma(1) = 1$, ne segue che per n intero si ha $\Gamma(n) = (n-1)!$

Definizione 4.3.1 (**Densità Gamma).** Si chiama densità Gamma di parametri r e λ ($r > 0$, $\lambda > 0$), indicata $\Gamma(r, \lambda)$, la funzione così definita

$$f(x) = \begin{cases} \frac{1}{\Gamma(r)}\lambda^r\, x^{r-1}\,\mathrm{e}^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Verifichiamo che si tratta effettivamente di una densità: effettuando il cambio di variabili $\lambda x = t$ si ha infatti

$$\int_0^{+\infty} (\lambda x)^{r-1}\,\mathrm{e}^{-\lambda x}\,\lambda\,\mathrm{d}x = \int_0^{+\infty} t^{r-1}\mathrm{e}^{-t}\,\mathrm{d}t = \Gamma(r)$$

Si può osservare che la densità *esponenziale* di parametro λ è la densità $\Gamma(1,\lambda)$.

Una variabile Gamma (qualunque siano i parametri) *possiede tutti i momenti*: la formula che segue è uno strumento comodissimo per calcolare tutti i momenti di una variabile Gamma.

Proposizione 4.3.2. *Se X ha densità $\Gamma(r,\lambda)$, preso $\beta > 0$ si ha*

$$\mathbf{E}\big[\, X^\beta \,\big] = \frac{\Gamma(r+\beta)}{\Gamma(r)\,\lambda^\beta}$$

Dimostrazione. Cominciamo ad osservare che, poiché X prende solo valori positivi, ha senso scrivere

$$\mathbf{E}\big[\, X^\beta \,\big] = \frac{1}{\Gamma(r)} \int_0^{+\infty} x^\beta\, \lambda^r x^{r-1} \mathrm{e}^{-\lambda x}\, \mathrm{d}x =$$

$$= \frac{1}{\Gamma(r)\,\lambda^\beta} \int_0^{+\infty} \lambda^{r+\beta}\, x^{r+\beta-1}\, \mathrm{e}^{-\lambda x}\, \mathrm{d}x = \frac{\Gamma(r+\beta)}{\Gamma(r)\,\lambda^\beta}$$

□

Da questo calcolo segue che esistono tutti i momenti, inoltre il calcolo dei momenti diventa semplice infatti

$$\mathbf{E}[X] = \frac{\Gamma(r+1)}{\Gamma(r)\,\lambda} = \frac{r}{\lambda} \qquad \mathbf{E}[X^2] = \frac{\Gamma(r+2)}{\Gamma(r)\,\lambda} = \frac{(r+1)r}{\lambda^2}$$

di conseguenza $Var(X) = \frac{r}{\lambda^2}$ e così via ... notiamo che non era necessario calcolare $\Gamma(r)$ ma solo usare l'eguaglianza $\Gamma(r) = (r-1)\,\Gamma(r-1)$.

Vediamo ora il calcolo della *funzione generatrice dei momenti*

Proposizione 4.3.3. *Sia X con densità $\Gamma(r,\lambda)$: allora $G_X(t)$ è finito se $t < \lambda$ e in tal caso si ha*

$$G_X(t) = \big(\frac{\lambda}{\lambda - t}\big)^r$$

Dimostrazione. Infatti

$$\mathbf{E}\big[\mathrm{e}^{tX}\big] = \frac{1}{\Gamma(r)} \int_0^{+\infty} \mathrm{e}^{tx} \lambda^r\, x^{r-1} \mathrm{e}^{-\lambda x}\, \mathrm{d}x =$$

$$= \frac{\lambda^r}{\Gamma(r)\,(\lambda-t)^r} \int_0^{+\infty} (\lambda-t)^r\, x^{r-1} \mathrm{e}^{-(\lambda-t)x}\, \mathrm{d}x = \frac{\lambda^r}{(\lambda-t)^r}$$

□

Come conseguenza immediata, se X ha densità $\Gamma(r,\lambda)$, Y ha densità $\Gamma(s,\lambda)$ e sono indipendenti, allora $(X+Y)$ ha densità $\Gamma(r+s,\lambda)$ (si noti che il parametro λ deve essere lo stesso).

Una osservazione sui parametri: con la parametrizzazione che abbiamo scelto, il termine r è chiamato *"shape"* ed il termine λ è chiamato *"rate"* (questa è la parametrizzazione di default per il software **R**); alcuni autori però preferiscono mantenere la *"shape"* ma usare come secondo parametro la *"scale"* s dove $s = 1/\lambda$. Con questa parametrizzazione la densità diventa

$$f(x) = \begin{cases} \frac{1}{\Gamma(r)} \frac{1}{s} \left(\frac{x}{s}\right)^{r-1} \mathrm{e}^{-\frac{x}{s}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

4.3.2 Densità chi-quadro

Questa densità, come la successiva densità di Student, nasce da una precisa applicazione statistica che vedremo più avanti. Cominciamo con questo risultato:

Proposizione 4.3.4. *Siano $X_1, \ldots, X_n$ n variabili indipendenti gaussiane standard: la variabile $(X_1^2 + \cdots + X_n^2)$ ha densità $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$. A questa densità viene dato il nome di* **densità chi-quadro** *a n gradi di libertà, indicata $\chi^2(n)$.*

Dimostrazione. È sufficiente provare che, se X è gaussiana standard, X^2 è $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$, e questo si può fare in due modi: si può calcolare direttamente la densità di X^2 oppure provare che la *funzione generatrice dei momenti* di X^2 è eguale, per $t < 1/2$, a $\left(\frac{1/2}{1/2-t}\right)^{\frac{1}{2}} = \sqrt{\frac{1}{1-2t}}$. Infatti

$$G_{X^2}(t) = \mathbf{E}\left[\, \mathrm{e}^{tX^2} \,\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{tx^2}\, \mathrm{e}^{-\frac{x^2}{2}}\, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}(1-2t)}\, \mathrm{d}x$$

Ricordando che, preso $\sigma > 0$ si ha $\frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2\sigma^2}}\, \mathrm{d}x = 1$ e considerando che $\sigma = \frac{1}{\sqrt{1-2t}}$, si ottiene il risultato. □

La terminologia *gradi di libertà* è per il momento misteriosa, diventerà chiara quando ne vedremo le applicazioni statistiche. Lo α-quantile della variabile chi-quadro (che comparirà diverse volte più avanti) è indicato $\chi^2_{(\alpha,n)}$.

Poiché la variabile $\chi^2(n)$ è in realtà una $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, non c'è bisogno di nuovi conti per calcolare i momenti di questa variabile, così come è evidente che

somma di variabili chi-quadro indipendenti è ancora una variabile chi-quadro (per essere più precisi, se X ha densità $\chi^2(n)$, Y ha densità $\chi^2(m)$ e sono indipendenti, allora $(X+Y)$ ha densità $\chi^2(n+m)$).

Vediamo ora utili approssimazioni della variabile $\chi^2(n)$ quando n è grande (diciamo $n \geq 80$): partiamo dall'osservazione che, presa C_n con densità $\chi^2(n)$, si può pensare $C_n = X_1^2 + \cdots + X_n^2$ dove le X_i sono gaussiane standard indipendenti, inoltre $\mathbf{E}[X_i^2] = 1$ e $Var(X_i^2) = 2$. Valgono i seguenti risultati:

- per la *legge dei Grandi Numeri* 4.1.2 la successione $\frac{C_n}{n}$ converge a 1 in probabilità (per $n \to \infty$), e quindi $\dfrac{C_n}{n} \approx 1$
- per il *teorema Limite Centrale* 4.2.2 la successione $\dfrac{C_n - n}{\sqrt{2n}}$ converge in distribuzione alla variabile $N(0,1)$, e quindi $\dfrac{C_n - n}{\sqrt{2n}}$ è approssimativamente Gaussiana standard.

4.3.3 Densità di Student

Cominciamo subito con la definizione:

Definizione 4.3.5. Si chiama **densità di Student** a n gradi di libertà la densità della variabile

$$T_n = \frac{X}{\sqrt{\frac{C_n}{n}}} = \sqrt{n}\,\frac{X}{\sqrt{C_n}}$$

dove X è gaussiana $N(0,1)$, C_n ha densità $\chi^2(n)$ e sono indipendenti.

Questa definizione apparentemente misteriosa segue da una precisa applicazione statistica che vedremo, è possibile calcolare la densità di T_n con conti piuttosto lunghi ma in realtà non ce n'è bisogno: nelle applicazioni basterà conoscere la funzione di ripartizione ed i quantili di T_n, però qualcosa si può dire subito.

Proposizione 4.3.6. *La densità di T_n è una* funzione pari*; di conseguenza indicati con $F_n(x)$ e $\tau_{(\alpha,n)}$ rispettivamente la Funzione di ripartizione e lo α-quantile della variabile T_n, valgono le relazioni*

$$F_n(-x) = 1 - F_n(x) \qquad \tau_{(\alpha,n)} = -\tau_{(1-\alpha,n)}$$

Praticamente non c'è nulla da dimostrare: è infatti un facile esercizio provare che una v.a. X ha densità pari (cioè $f(x) = f(-x)$) se e solo se X e

$-X$ sono equidistribuite, e le variabili Gaussiane hanno densità pari quando $m = 0$ (dove m è la media). Poiché $-T_n = \sqrt{n}\,\frac{-X}{\sqrt{C_n}}$ e X è $N(0,1)$, il risultato è evidente. Quanto alle relazioni tra c.d.f. e quantili, quando le abbiamo dimostrate per le variabili gaussiane standard abbiamo precisato che erano conseguenza solo del fatto che la densità è una funzione pari.

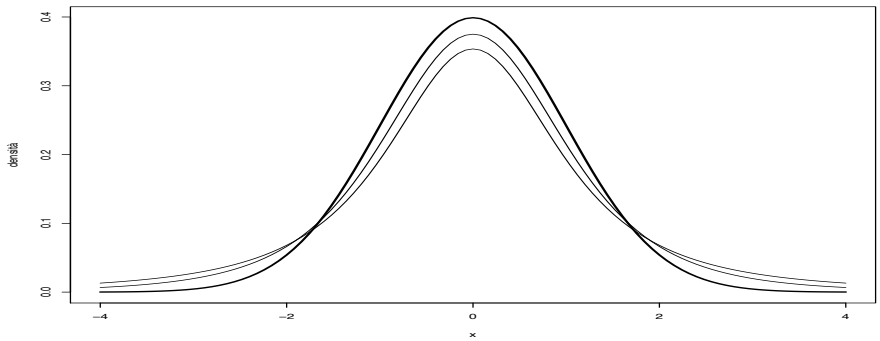
Abbiamo visto poco fa che, se n è grande e C_n è una variabile $\chi^2(n)$, $\frac{C_n}{n} \approx 1$; di conseguenza T_n è approssimativamente gaussiana standard. Volendo dare a questa affermazione pratica un enunciato preciso, vale il seguente risultato

Teorema 4.3.7. *Consideriamo, per ogni n, una variabile T_n di Student a n gradi di libertà: la successione* $(T_n)_{n\geq 1}$ converge in distribuzione *alla variabile* $N(0,1)$.

Non diamo di questo risultato una dimostrazione precisa; per quello che riguarda le applicazioni pratiche, quanto n è alto, nei conti la variabile di Student è sostituita dalla variabile gaussiana standard. Nelle applicazioni statistiche, della variabile di Student serve avere solo la funzione di ripartizione o il quantile, e questi usualmente vengono calcolati con un opportuno software; in passato esistevano le *tavole della variabile di Student* ma adesso sono superate.

L'espressione esplicita della densità di Student si può ottenere con dei conti piuttosto lunghi, giusto per curiosità è $f_n(x) = c_n\,(1+x^2)^{-(n+1)/2}$, dove la costante c_n ha un'espressione complicata che non riporto: da questa formula però si ottiene abbastanza facilmente che *la variabile T_n di Student a n gradi di libertà* ha momenti fino all'ordine $(n-1)$. Inoltre i momenti di ordine dispari, quando esistono, sono eguali a 0.

Rispetto alla densità N(0,1), la densità di Student va a 0 più lentamente per $|x| \to \infty$ (si parla di *"code pesanti"* o *"fat tails"*); sono tracciate sotto, sovrapposte, la densità $N(0,1)$ in grassetto e le densità di Student a 2 e a 4 gradi di libertà.



4.4 Appendice: convergenza in distribuzione e teorema limite centrale

La *definizione* di convergenza in distribuzione non è affatto semplice, si può esprimere facilmente se la variabile limite X ha funzione di ripartizione continua.

Definizione 4.4.1. Siano $(X_n)_{n\geq 1}$ un successione di v.a.r ed X una variabile aleatoria, siano rispettivamente F_n ed F le funzioni di ripartizione di X_n ed X e *supponiamo che* $F(.)$ *sia continua*: si dice che la successione $(X_n)_{n\geq 1}$ converge ad X in *distribuzione* se per ogni t si ha

$$\lim_{n\to\infty} F_n(t) = F(t)$$

Nel caso generale questa definizione deve essere modificata, ma a noi interessa in realtà il caso in cui la variabile limite è *Gaussiana* oppure *di Student* (quindi con funzione di ripartizione continua).

Tuttavia, tornando al Teorema 4.2.2 non è affatto facile provare che le c.d.f delle variabili approssimanti convergono puntualmente a $\Phi(.)$ e infatti si prendono altre strade: qui proviamo che, con le notazioni di 4.2.2, dette $Y_n = \dfrac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$ ed X una variabile gaussiana standard, si ha convergenza delle loro funzioni generatrici dei momenti, cioè per ogni t si ha

$$\lim_{n\to\infty} G_{Y_n}(t) = G_X(t) = \mathrm{e}^{\frac{t^2}{2}}$$

Se noi chiamiamo $Z_i = \dfrac{X_i - \mu}{\sigma}$, queste sono i.i.d. con media 0 e varianza 1, e quindi la loro funzione generatrice ammette lo sviluppo $G_{Z_i}(t) = 1 + \frac{t^2}{2} + o(t^2)$ e, poiché $Y_n = \dfrac{Z_1 + \cdots + Z_n}{\sqrt{n}}$ (e le funzioni generatrici delle variabili Z_i sono eguali) si ha

$$G_{Y_n}(t) = G_{X_1}\big(\frac{t}{\sqrt{n}}\big)\,..\,G_{X_n}\big(\frac{t}{\sqrt{n}}\big) = \Big(G_{X_1}\big(\frac{t}{\sqrt{n}}\big)\Big)^n =$$

$$= \Big(1 + \frac{t^2}{2\,n} + o(\frac{t^2}{n})\Big)^n$$

e a questo punto è immediato provare che si ha $\lim_{n\to\infty} G_{Y_n}(t) = \mathrm{e}^{\frac{t^2}{2}}$.

L'argomentazione riportata però non è del tutto rigorosa: infatti per ipotesi le variabili X_i hanno solo momento primo e secondo e per fare uno

sviluppo di Taylor della loro funzione generatrice occorre invece che abbiano tutti i momenti. Inoltre non esiste una corrispondenza precisa tra convergenza in distribuzione e convergenza delle funzioni generatrici: occorre usare un altro strumento, le *trasformate di Fourier* (ed è questo lo strumento che ha usato Paul Lévy), ma arrivare ad una dimostrazione del tutto rigorosa è un po' troppo avanzato per questo corso introduttivo.

Capitolo 5

Prime nozioni di Inferenza statistica.

5.1 Campioni statistici e statistiche campionarie

Lo scopo dell'*Inferenza statistica* è partire dall'analisi di un *campione* per avere informazioni su una intera *popolazione*: ad esempio un sondaggio sulle intenzioni di voto parte da un certo numero di telefonate per ricostruire gli orientamenti elettorali della intera popolazione (in questo caso l'intero corpo elettorale).

Una ipotesi di base (in generale del tutto ragionevole) è che ci sia una implicita *distribuzione di probabilità* nell'intera popolazione, e che le osservazioni del campione siano le realizzazioni di n variabili aleatorie, che è ragionevole supporre indipendenti, aventi questa distribuzione di probabilità.

Questo suggerisce la seguente definizione:

Definizione 5.1.1 (**Campione statistico).** Si chiama *Campione statistico* (o anche *campione aleatorio*) una famiglia $X_1, \ldots, X_n$ di n variabili aleatorie indipendenti, tutte aventi funzione di ripartizione $F(.)$.

Le variabili sono dunque *equidistribuite*, la numerosità n è chiamata la *taglia* del campione e la probabilità associata alla c.d.f. $F(.)$ è chiamata la *legge di probabilità del campione*. Questa legge di probabilità è, secondo i casi, del tutto *sconosciuta* o solo *parzialmente conosciuta*.

È doveroso insistere sull'importanza di usare *lettere minuscole* $x_1, x_2, ...$ per indicare dati numerici (ad esempio risultati di alcune misurazioni) e *lettere maiuscole* $X_1, X_2, ..$ per indicare *variabili aleatorie*: le v.a. esistono *nel modello*, i dati numerici nella realtà.

Ad esempio, se i numeri $x_1, \ldots, x_n$ sono gli esiti di n misurazioni, questi possono essere *interpretati* come gli esiti $X_1(\omega), \ldots, X_n(\omega)$ di n variabili aleatorie.

Proposizione 5.1.2. *Sia $X_1, \ldots, X_n$ un campione statistico, supponiamo che le variabili possiedano momento secondo e siano $\mu = \mathbf{E}[X_i]$ e $\sigma^2 = Var(X_i)$: indicando $\overline{X} = \frac{X_1+\cdots+X_n}{n}$, si ha*

$$\mathbf{E}\big[\,\overline{X}\,\big] = \mu \qquad\qquad \mathbf{E}\Big[\frac{\sum_{i=1}^n \big(X_i - \overline{X}\big)^2}{n-1}\Big] = \sigma^2$$

Dimostrazione. Partiamo dall'eguaglianza (già incontrata nel primo capitolo) $\sum_{i=1}^n \big(X_i - \overline{X}\big)^2 = \sum_{i=1}^n X_i^2 - n\,\overline{X}^2$.

Per ogni X_i si ha $\mathbf{E}\big[\,X_i^2\big] = Var(X_i) + \mathbf{E}[X_i]^2 = \sigma^2 + \mu^2$; viceversa abbiamo già visto che $\mathbf{E}\big[\,\overline{X}\,\big] = \mu$ e $Var(\overline{X}) = \sigma^2/n$.

Ne segue che $\mathbf{E}\big[\sum_{i=1}^n \big(X_i - \overline{X}\big)^2\big] = \sum_{i=1}^n \mathbf{E}\big[X_i^2\big] - n\mathbf{E}\big[\,\overline{X}^2\,\big] =$
$= n(\sigma^2+\mu^2) - n(\sigma^2/n + \mu^2) = (n-1)\,\sigma^2$. □

Una funzione $g(X_1, \ldots, X_n)$ di un campione statistico è chiamata **statistica campionaria**: tali sono ad esempio la **media campionaria** $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ e la **varianza campionaria** $S^2 = \frac{\sum_{i=1}^n \big(X_i-\overline{X}\big)^2}{n-1}$.

Il risultato della Proposizione 5.1.2 si può interpretare dicendo che *la media campionaria e la varianza campionaria sono stime corrette rispettivamente del valore atteso e della varianza.*

Una stima è detta *corretta* o *non distorta* ("unbiased") se il suo valore atteso coincide con la quantità che si vuole stimare.

5.2 Statistiche campionarie di variabili Gaussiane

Quando il campione statistico è formato da Variabili Gaussiane, si può dire molto di più sulla *distribuzione congiunta* si $\overline{X}$ e S^2: i risultati esposti in questo paragrafo saranno fondamentali nel capitolo successivo. Cominciamo con variabili gaussiane standard.

Teorema 5.2.1. *Siano $(X_1, \ldots, X_n)$ indipendenti con densità $N(0,1)$. Valgono i seguenti risultati:*

a) le variabili $\overline{X}$ e $\sum_{i\le n} \big(X_i - \overline{X}\big)^2$ sono indipendenti;

b) $\overline{X}$ ha densità $N(0, \frac{1}{n})$ e $\sum_{i\le n} \big(X_i - \overline{X}\big)^2$ ha densità $\chi^2(n-1)$;

c) la variabile $T = \sqrt{n}\,\dfrac{\overline{X}}{S}$ *ha densità di Student* $T(n-1)$ *.*

La dimostrazione di questo risultato è piuttosto complessa e lasciata all'appendice. Notiamo però che in realtà dobbiamo dimostrare solo due cose: l'indipendenza di $\overline{X}$ e $\sum_{i\leq n}\left(X_i-\overline{X}\right)^2$ ed il fatto che $\sum_{i\leq n}\left(X_i-\overline{X}\right)^2$ ha densità chi-quadro(n-1).

Infatti già sappiamo che $\overline{X}$ ha densità $N(0,\frac{1}{n})$, notiamo che $\sqrt{n}\,.\,\overline{X}$ è gaussiana standard e infine

$$\sqrt{n}\,\frac{\overline{X}}{S}=\frac{\sqrt{n}\,\overline{X}}{\sqrt{\frac{\sum_{i\leq n}\left(X_i-\overline{X}\right)^2}{n-1}}}$$

Il risultato 5.2.1 è in realtà preparatorio del risultato generale, che è il seguente.

Teorema 5.2.2. *Siano* $(X_1,\ldots,X_n)$ *indipendenti con densità* $N(m,\sigma^2)$ *. Valgono i seguenti risultati:*

a) le variabili $\overline{X}$ *e* $\sum_{i\leq n}\left(X_i-\overline{X}\right)^2$ *sono indipendenti;*

b) $\overline{X}$ *ha densità* $N(m,\frac{\sigma^2}{n})$ *e* $\sum_{i\leq n}\frac{\left(X_i-\overline{X}\right)^2}{\sigma^2}$ *ha densità* $\chi^2(n-1)$ *;*

c) la variabile $T=\sqrt{n}\,\dfrac{\left(\overline{X}-m\right)}{S}$ *ha densità di Student* $T(n-1)$ *.*

Come sempre con le variabili Gaussiane, ci si riporta alle variabili Gaussiane standard: indicando $Y_i=\frac{X_i-m}{\sigma}$, queste ultime sono indipendenti $N(0,1)$ e ad esse si applica 5.2.1. Notiamo in particolare che $\sum_{i\leq n}\left(X_i-\overline{X}\right)^2/\sigma^2=\sum_{i\leq n}\left(Y_i-\overline{Y}\right)^2$, e che $\sqrt{n}\,.\,\dfrac{\left(\overline{X}-m\right)}{\sigma}$ è gaussiana standard.

5.3 Stima parametrica

Se non abbiamo ulteriori informazioni sulla comune c.d.f. $F(.)$ delle variabili $X_1,\ldots,X_n$ che formano il campione statistico, non si può dire molto di più di quello che è stato fatto; in verità è possibile dare risultati ulteriori ma ci si addentra nella *statistica non parametrica* che è al di là degli obiettivi di questo corso.

A volte però la comune distribuzione di probabilità è *parzialmente specificata* nel senso che appartiene ad una famiglia di c.d.f. dipendenti da un opportuno *parametro* usualmente indicato θ *.*

Ad esempio è ragionevole supporre che l'altezza di una popolazione (ad esempio gli abitanti di una zona) sia rappresentata da una variabile Gaussiana, della quale però non si conosce né media né varianza.

Oppure la densità esponenziale è usata per modellizzare i tempi di decadimento radioattivo o la durata di vita di certe apparecchiature (*apparecchiature "che non invecchiano"*), però non è specificato il parametro: questo parametro deve essere valutato a partire dalle osservazioni.

Cercare di ricostruire il parametro (o i parametri) a partire dalle osservazioni è l'oggetto della **stima parametrica**. Esponiamo due metodi: il *metodo della massima verosimiglianza* ed il *metodo dei momenti*.

Supponiamo dunque di avere un *campione statistico* la cui legge di probabilità dipende da un parametro $\theta \in \Theta$, nel quale le variabili possono o essere discrete con funzione di massa $p_\theta(x)$, oppure con *densità* $f_\theta(x)$.

Definizione 5.3.1. Si chiama **verosimiglianza** del campione $X_1, \dots, X_n$ la funzione $L\big(\theta; \dots\big)$ definita, nel caso delle *variabili discrete*, da

$$L\big(\theta\,;\, x_1, \dots, x_n\big) = \prod_{i=1}^{n} p_\theta(x_i)$$

e nel caso delle *variabili con densità* da

$$L\big(\theta\,;\, x_1, \dots, x_n\big) = \prod_{i=1}^{n} f_\theta(x_i)$$

Si può notare che, nel caso di variabili con densità, la verosimiglianza è la *densità congiunta* di $(X_1, \dots, X_n)$ pensata come variabile aleatoria n-dimensionale (o meglio come *vettore aleatorio*). Analoga è l'interpretazione nel caso delle variabili discrete.

Definizione 5.3.2. Si chiama **stima di massima verosimiglianza** (se esiste) una statistica campionaria (usualmente indicata $\widehat{\theta} = \widehat{\theta}(x_1, \dots, x_n)$) tale che valga l'eguaglianza

$$L\big(\widehat{\theta}\,;\, x_1, \dots, x_n\big) = \max_{\theta \in \Theta} L\big(\theta\,;\, x_1, \dots, x_n\big)$$

Un po' diverso è il *metodo dei momenti*: l'idea è di eguagliare i momenti teorici con i momenti empirici. Più precisamente calcoliamo i momenti teorici in funzione del o dei parametri $\theta_1, \dots, \theta_h$

$$m_k(\theta_1, \dots, \theta_h) = \mathbf{E}_{\theta_1, \dots, \theta_h}\big[\, X^k\big]$$

e confrontiamoli con i *momenti empirici*

$$\sum_{i=1}^{n} \frac{x_i^k}{n}$$

Definizione 5.3.3. Se esiste una scelta del parametro (dei parametri) che permette di eguagliare i momenti teorici con i momenti empirici, cioè di scrivere (per uno o più interi k)

$$\mathbf{E}_{\theta_1,\dots,\theta_h}\big[\, X^k \big] = \sum_{i=1}^{n} \frac{x_i^k}{n}$$

questa si chiama **stima col metodo dei momenti**

Non esiste una notazione consolidata per la stima col metodo dei momenti, in queste note usiamo la notazione $\widetilde{\theta}$.

Precisiamo subito che queste stime non sempre esistono, a volte coincidono e a volte sono sostanzialmente diverse. Per quanto riguarda la stima col metodo dei momenti, supponendo ad esempio che ci sia un solo parametro, a volte è sufficiente il momento primo, a volte bisogna passare a un momento di ordine superiore.

I metodi sopra definiti si possono capire solo affrontando degli esempi concreti, osserviamo preliminarmente che *la media e la varianza campionaria sono rispettivamente la stima della media e della varianza col metodo dei momenti.*

Esempio 5.3.4 (Campione con densità esponenziale). Supponiamo che $X_1, \dots, X_n$ abbiano densità esponenziale di parametro θ con $0 < \theta < +\infty$.

Innanzi tutto noi adatteremo questo modello ad un campione di osservazioni $x_1, \dots, x_n$ formato da numeri *positivi*: poiché $\mathbf{E}_\theta[X_i] = 1/\theta$, la stima col metodo dei momenti si ottiene imponendo $\dfrac{1}{\theta} = \displaystyle\sum_i \frac{x_i}{n}$, cioè $\widetilde{\theta} = \dfrac{1}{\overline{x}}$.

Allo stesso modo (sempre se gli x_i sono positivi) al verosimiglianza coincide con $\theta^n e^{-\theta(\sum_i x_i)}$: questa funzione tende a 0 per $\theta \to 0$ e per $\theta \to \infty$ (quindi ha massimo) e annullando la derivata si ottiene di nuove $\widehat{\theta} = \dfrac{n}{\sum_i x_i} = \dfrac{1}{\overline{x}}$.

Se si considera un campione di *variabili di Poisson* di parametro θ con $0 < \theta < +\infty$, con conti altrettanto semplici si trovano entrambe le stime, e sono coincidenti.

Invece è più complicato l'esempio che segue .

Esempio 5.3.5 (Densità uniformi su un intervallo variabile). Supponiamo dunque che, per $0 < \theta < +\infty$, la densità sia

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & \text{altrove} \end{cases}$$

Partiamo dunque dalle osservazioni $x_1, \dots, x_n$ che di nuovo supponiamo positive: poiché $\mathbf{E}_\theta[X_i] = \theta/2$, anche questa volta la stima col metodo dei momenti è facile e si ottiene $\widetilde{\theta} = 2\left(\sum_i \frac{x_i}{n}\right) = 2\,\overline{x}$.

È invece più complicata la stima di massima verosimiglianza: notiamo che la verosimiglianza è

$$L\big(\theta\,;\, x_1, \dots, x_n\big) = \begin{cases} \theta^{-n} & x_i \le \theta \quad \forall i \\ 0 & \text{altrimenti} \end{cases}$$

cioè la verosimiglianza è diversa da 0 solo se $\theta \ge \max(x_1, \dots, x_n)$, dopodiché è una funzione decrescente. Ne segue che si ha $\widehat{\theta}(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$.

5.4 Appendice: alcune dimostrazioni

Come avevamo detto, per provare il risultato fondamentale 5.2.1 bisogna in realtà provare soltanto due cose: l'indipendenza di $\overline{X}$ e $\sum_{i=1}^n \big(X_i - \overline{X}\big)^2$ ed il fatto che $\sum_{i=1}^n \big(X_i - \overline{X}\big)^2$ ha densità chi-quadro(n-1).

Dobbiamo prima provare questo risultato: *sia* $\mathbf{X} = (X_1, \dots, X_n)$ *un vettore aleatorio formato da n v.a. indipendenti con densità* $N(0,1)$, *sia A una matrice* $n \times n$ ortogonale *(cioè la matrice di un cambio di base) e sia* $\mathbf{Y} = A\,\mathbf{X}$. *Anche le componenti* $(Y_1, \dots, Y_n)$ *sono indipendenti con densità* $N(0,1)$.

La densità del vettore aleatorio $\mathbf{X}$ (scritta con notazione vettoriale) è $f(\mathbf{x}) = \big(2\pi\big)^{-\frac{n}{2}} \exp\big(-\frac{\|\mathbf{x}\|^2}{2}\big)$: la formula 3.2.6 ha un analogo vettoriale, dove al valore assoluto della derivata della funzione inversa h^{-1} si sostituisce il valore assoluto del *determinante dello Jacobiano* associato alla funzione inversa.

In questo caso la trasformazione $\mathbf{y} = A\,\mathbf{x}$ è un biunivoca, con inversa $\mathbf{x} = A^{-1}\,\mathbf{y}$: osservando che $\|A^{-1}\,\mathbf{y}\|^2 = \|\mathbf{y}\|^2$ (poiché A è una matrice ortogonale) e che il determinante dello Jacobiano associato ad A^{-1} è eguale ad 1, si conclude facilmente che $\mathbf{Y}$ ha densità eguale a quella di $\mathbf{X}$.

Sia ora $\mathbf{e}_1$ il vettore $\mathbf{e}_1 = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$ e sia E_1 il sottospazio vettoriale di $\mathbb{R}^n$ generato da $\mathbf{e}_1$; sia poi E_2 l'ortogonale di E_1 e sia $\mathbf{e}_2, \ldots, \mathbf{e}_n$ una base ortonormale di E_2. Sia infine A la matrice (ortogonale) di passaggio dalla base canonica di $\mathbb{R}^n$ alla base $\mathbf{e}_1, \ldots, \mathbf{e}_n$.

Indichiamo con $\mathbf{X}$ il vettore aleatorio $(X_1, \ldots, X_n)$ e sia $\mathbf{Y} = A\,\mathbf{X}$: in base al risultato appena annunciato, le componenti $Y_1, \ldots, Y_n$ sono ancora indipendenti con densità $N(0,1)$. Quindi Y_1 è indipendente da $(Y_2^2 + \cdots + Y_n^2)$ che ha densità $\chi^2(n-1)$.

Notiamo che $Y_1 = \sqrt{n}\ \overline{X}$, inoltre $Y_2^2 + \cdots + Y_n^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 = \sum_{i=1}^n \left(X_i - \overline{X}\right)^2$.

Segue dunque immediatamente l'indipendenza di $\overline{X}$ e di $\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$, ed il fatto che $\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ ha densità $\chi^2(n-1)$.

Capitolo 6

Intervalli di fiducia e verifica delle ipotesi.

6.1 Intervalli di fiducia (o di confidenza)

Gli intervalli di fiducia in italiano sono chiamati spesso *intervalli di confidenza*: si tratta in realtà di una cattiva traduzione dall'inglese (poiché la parola *"confidence"* si traduce in italiano con *"fiducia"*). Tuttavia il termine "intervallo di confidenza" è ormai entrato nell'uso comune e quindi useremo indifferentemente l'uno o l'altro termine.

Gli intervalli di fiducia spesso sono trattati insieme alla stima (alcuni testi parlano di *"stima intervallare"*), tuttavia come vedremo questi intervalli possono essere visti *"come un caso particolare di test"*.

Nella vita di tutti i giorni in realtà si è abituati (senza rendersene conto) agli intervalli di fiducia: ad esempio durante le proiezioni di voto al termine di una elezione dopo un'ora di spoglio il partito XY viene dato al $26.2 \pm 2.3\,\%$, dopo due ore e mezza al $25.9 \pm 1.2\,\%$ e infine dopo quattro ore al $26.1 \pm 0.4\,\%$. Man mano che i dati aumentano la misurazione diventa più precisa, viene cioè proposto un intervallo sempre più stretto.

Si parla di intervallo di fiducia in presenza di un campione statistico con distribuzione dipendente da un parametro $\theta \in \Theta$, dove $\Theta \subseteq \mathbb{R}$. Vediamo adesso la definizione precisa.

Definizione 6.1.1. Sia $0 < \alpha < 1$ ed I un intervallo con $I \subseteq \Theta$: si dice che I è un **intervallo di fiducia (o di confidenza)** per θ al livello $(1 - \alpha)$ se, per ogni θ

$$\mathbf{P}_\theta\big\{\theta \in I\big\} \geq (1 - \alpha)$$

Tipicamente α è un numero piccolo (ad esempio 0.05 o 0.02) in modo che il *livello di fiducia* sia vicino ad 1. Chiaramente l'ideale è avere intervalli di fiducia più piccoli possibile (però conservando il livello di significatività).

Non si cercano risultati generali: bisogna guardare specificamente esempi particolari, e lo faremo in alcuni dei casi più frequenti (con particolare attenzione ai campioni Gaussiani).

Tratteremo per primo il caso di intervalli di fiducia per la media *con varianza nota*: questo non è il caso realistico, ma si tratta più facilmente e ci guiderà poi per analogia a trattare il caso più realistico della *varianza sconosciuta*.

6.1.1 Intervalli di fiducia per la media di un campione Gaussiano

Supponiamo di avere un campione $X_1, \dots, X_n$ con varianza σ^2 fissa (**varianza nota**): vogliamo trovare un intervallo di fiducia per la media.

Poiché $\overline{X}(\omega)$ è la stima della media, è naturale prendere un intervallo della forma $I(\omega) = \left[\overline{X}(\omega) - d\,,\, \overline{X}(\omega) + d\right]$ con d da determinare. In modo sintetico, si usa anche scrivere $I(\omega) = \overline{X}(\omega) \pm d\,.$

Notiamo che $m \in I$ equivale a dire $\left|\overline{X} - m\right| \leq d$, quindi d deve essere tale che $\mathbf{P}_m\big\{\, \left|\overline{X} - m\right| \leq d \,\big\} \geq (1-\alpha)\,,$ con d *più piccolo possibile*: questo significa imporre $\mathbf{P}_m\big\{\left|\overline{X} - m\right| \leq d\,\big\} \approx (1-\alpha)\,.$

Ricordando che $\sqrt{n}\,\dfrac{\left(\overline{X} - m\right)}{\sigma}$ è gaussiana standard, si ha

$$\mathbf{P}_m\big\{\left|\overline{X} - m\right| \leq d\,\big\} = \mathbf{P}_m\Big\{\frac{\sqrt{n}}{\sigma}\left|\overline{X} - m\right| \leq \frac{d\sqrt{n}}{\sigma}\Big\} \approx (1-\alpha)$$

e di conseguenza si prende $\dfrac{d\sqrt{n}}{\sigma} = q_{1-\frac{\alpha}{2}}\,,$ cioè l'intervallo di fiducia risulta della forma

$$\overline{X}(\omega) \pm \frac{\sigma}{\sqrt{n}}\, q_{1-\frac{\alpha}{2}}$$

Il numero $\dfrac{\sigma}{\sqrt{n}}\, q_{1-\frac{\alpha}{2}}$ è chiamato **precisione della stima** e $\dfrac{\frac{\sigma}{\sqrt{n}}\, q_{1-\frac{\alpha}{2}}}{\overline{X}(\omega)}$ è la **precisione relativa**.

Vediamo ora il caso più realistico della **varianza sconosciuta**. L'idea che ha avuto Student è stato di sostituire a σ^2 (che non è nota) la *varianza campionaria* $S^2 = \dfrac{\sum_{i=1}^n \left(X_i^2 - \overline{X}\right)}{n-1}$: si procede quindi per analogia con

il caso precedente. Questa volta però la variabile $\sqrt{n}\,\dfrac{\big(\overline{X}-m\big)}{S}$ non è più gaussiana ma con densità di Student $T(n-1)$.

L'intervallo di fiducia risulta allora della forma

$$\overline{X}(\omega) \pm \frac{S(\omega)}{\sqrt{n}}\,\tau_{(1-\frac{\alpha}{2}\,,\,n-1)}$$

essendo τ il quantile della variabile di Student. Quando n è grande ($n \geq 60$) si approssima il quantile della variabile di Student con quello della Gaussiana standard.

Gli intervalli di fiducia appena descritti sono detti **bilateri**, a volte può essere interessante considerare un intervallo **unilatero** (ad esempio può interessare sapere che la media non è troppo alta): si considera allora un intervallo **unilatero sinistro** della forma $I(\omega) = \big]-\infty\,,\,\overline{X}(\omega)+d\,\big]$.

I passaggi sono simili a quelli svolti precedentemente: $m \in I$ se $m \leq \overline{X}+d$ cioè $\overline{X}-m \geq -d$. Imponendo un livello di fiducia $(1-\alpha)$ si richiede

$$\mathbf{P}_m\big\{\overline{X}-m \geq -d\,\big\} = \mathbf{P}_m\Big\{\frac{\sqrt{n}}{\sigma}\big(\overline{X}-m\big) \geq -\frac{d\sqrt{n}}{\sigma}\,\Big\} \approx (1-\alpha)$$

e si ricava $-\dfrac{d\sqrt{n}}{\sigma} = q_\alpha$ ossia $\dfrac{d\sqrt{n}}{\sigma} = q_{1-\alpha}$ (ricordiamo infatti che si ha $q_\alpha = -q_{1-\alpha}$).

Si ha quindi un intervallo (in realtà una semiretta)

$$\Big]-\infty\,,\,\overline{X}(\omega)+\frac{\sigma}{\sqrt{n}}\,q_{1-\alpha}\Big]$$

Invece un intervallo di fiducia **unilatero destro**, sempre al livello $(1-\alpha)$, sarà

$$\Big[\overline{X}(\omega)+\frac{\sigma}{\sqrt{n}}\,q_\alpha\,,\,+\infty\Big[$$

o, equivalentemente

$$\Big[\overline{X}(\omega)-\frac{\sigma}{\sqrt{n}}\,q_{1-\alpha}\,,\,+\infty\Big[$$

Gli intervallo di fiducia unilateri con varianza sconosciuta sono perfettamente analoghi, sostituendo a σ la variabile S e ai quantili della variabile Gaussiana standard quelli della variabile di Student.

6.1.2 Intervalli di fiducia approssimato per la media di un campione di Bernoulli

Consideriamo un campione $X_1, \ldots, X_n$ di variabili di Bernoulli di parametro p, $0 < p < 1$: il parametro p appare come una *proporzione* (ad esempio percentuale di pezzi difettosi in una produzione). I conti sono più agevoli se n è grande perché si può utilizzare l'approssimazione data dal Teorema Limite Centrale 4.2.2, tuttavia affiché questa approssimazione sia significativa la numerosità n deve assere elevata (almeno 80).

I passaggi che hanno portato a determinare gli intervalli di fiducia della media in un campione gaussiano con varianza nota erano basati sul fatti che, in quel caso, $\sqrt{n}\,\frac{\left(\overline{X}-m\right)}{\sigma}$ è gaussiana standard; in un campione con legge di Bernoulli di parametro p (ed n grande) la variabile $\dfrac{X_1+\cdots+X_n-n\,p}{\sqrt{n\,p(1-p)}} = \sqrt{n}\,\dfrac{\overline{X}-p}{\sqrt{p(1-p)}}$ è approssimativamente gaussiana standard.

Tuttavia, mentre in quel caso la varianza σ^2 era nota, questa volta il parametro p non è noto e lo sostituiamo con la sua *stima di massima verosimiglianza* $\widehat{p} = \overline{X}$.

Si arriva di conseguenza ad un intervallo di fiducia **bilatero** per p al livello $(1-\alpha)$ della forma

$$\overline{X} \pm \frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}\, q_{1-\frac{\alpha}{2}} \;=\; \widehat{p} \pm \frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}\, q_{1-\frac{\alpha}{2}}$$

La *precisione della stima* (approssimata) è pertanto $\frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}\, q_{1-\frac{\alpha}{2}}$; in un *controllo di qualità* $\widehat{p}$ è la percentuale di pezzi difettosi riscontrata, in un sondaggio la percentuale di risposte positive. Per evitare equivoci, qui con *percentuale* si intende un numero compreso tra 0 e 1, ad esempio il 25 % qui è inteso 0.25.

Esempio 6.1.2. Si vuole condurre un sondaggio telefonico per determinare la percentuale di gradimento del governo: quante telefonate almeno bisogna effettuare per avere una precisione della stima inferiore all'uno % con fiducia al 95 % ?

Anche qui 95 % va inteso $0.95 = (1-\alpha)$, la condizione sulla precisione diventa $\dfrac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}\, q_{0.975}$ e troviamo che $q_{0.975} \approx 1.96$: si arriva pertanto alla maggiorazione $\sqrt{n} \geq \dfrac{\sqrt{\widehat{p}(1-\widehat{p})} \times 1.96}{0.01}$, ma $\widehat{p}$ non è noto. Si considera

pertanto il valore massimo possibile, e poiché $\max_{0<p<1} p(1-p) = \frac{1}{4}$, si ha di conseguenza $n \geq \left(\frac{1.96}{0.02}\right)^2 = 9604$.

6.1.3 Intervalli di fiducia per la varianza di un campione Gaussiano

Se vogliamo trovare intervalli di fiducia per la *varianza* di un campione gaussiano, non c'è sostanziale differenza tra il caso *media nota* ed il caso *media sconosciuta* e quindi ci mettiamo direttamente in questo secondo caso.

Sostanzialmente tutto si appoggia su questo risultato (già visto in 5.2.2): se $X_1, \ldots, X_n$ è un campione di variabili $N(m, \sigma^2)$, la variabile $\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{\sigma^2} = (n-1)\,\frac{S^2}{\sigma^2}$ ha densità chi-quadro(n-1).

Il calcolo di intervalli di fiducia *bilateri* in questo caso non è agevole, ci concentriamo su intervalli *unilateri* che del resto sono quelli di maggiore interesse: sapere ad esempio che, nella misurazione di prodotti industriali, la varianza è inferiore a una certa soglia equivale a dire che la produzione è di buona qualità.

Cerchiamo un intervallo di fiducia **unilatero sinistro** per la varianza, e lo cerchiamo della forma $I = \left]0\,,\, \frac{\sum_i (X_i - \overline{X})^2}{a}\right]$, dove a è una costante da determinare (vedremo svolgendo i conti che è più opportuno considerare una costante *moltiplicativa* anziché *additiva*).

Partiamo dall'eguaglianza insiemistica $\left\{\sigma^2 \in I\right\} = \left\{\sigma^2 \leq \frac{\sum_i (X_i - \overline{X})^2}{a}\right\}$: imporre la condizione sul livello di fiducia equivale a chiedere, per ogni σ

$$\mathbf{P}_{\sigma^2}\left\{\sigma^2 \leq \frac{\sum_i (X_i - \overline{X})^2}{a}\right\} = \mathbf{P}_{\sigma^2}\left\{a \leq \frac{\sum_i (X_i - \overline{X})^2}{\sigma^2}\right\} \geq (1-\alpha)$$

Notiamo che la probabilità $\mathbf{P}_{\sigma^2}\left\{a \leq \frac{\sum_i (X_i - \overline{X})^2}{\sigma^2}\right\}$ non dipende dal parametro σ e per ottenere un intervallo più piccolo possibile imponiamo $\mathbf{P}_{\sigma^2}\left\{a \leq \frac{\sum_i (X_i - \overline{X})^2}{\sigma^2}\right\} \approx (1-\alpha)$: questo equivale a prendere $a = \chi^2_{(\alpha,n-1)}$ (cioè lo α-quantile della variabile chi-quadro).

In conclusione l'intervallo di fiducia *unilatero sinistro* al livello $(1-\alpha)$ risulta essere

$$\left]\,0\,,\, \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\chi^2_{(\alpha,n-1)}}\right] = \left]\,0\,,\, \frac{(n-1)\,S^2}{\chi^2_{(\alpha,n-1)}}\right]$$

Analogo è l'intervallo di fiducia **unilatero destro**, scriviamo direttamente la formula

$$\left[\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\chi^2_{(1-\alpha,n-1)}} \, , \, +\infty \right[= \left[\frac{(n-1)\, S^2}{\chi^2_{(1-\alpha,n-1)}} \, , \, +\infty \right[$$

6.2 Verifica delle ipotesi

Pianificare un test significa per prima cosa **formulare una ipotesi** e poi *pianificare un esperimento* per decidere se accettare o respingere l'ipotesi: come tutte le conclusioni della statistica, la risposta del test non è *una verità* ma viene fornita con un opportuno *grado di fiducia*.

Per formalizzare l'ipotesi, si divide l'insieme dei parametri Θ in due sottinsiemi Θ_0 (insieme dei parametri dell'ipotesi) ed il suo complementare Θ_1 (parametri dell'alternativa).

Ad esempio, in un *controllo di qualità* nel quale si desidera valutare la percentuale di pezzi difettosi in una produzione, l'insieme dei parametri è $\Theta =]0, 1[$, l'ipotesi *"la percentuale di pezzi difettosi non supera il 2 %"* equivale a considerare $\Theta_0 =]0\,,\, 0.02]$ e $\Theta_1 =]0.02\,,\, 1[$.

Nella pratica si preferisce usare questo linguaggio: l'ipotesi (chiamata anche "ipotesi nulla") è $\mathcal{H}_0\big)\ \theta \leq 0.02$ contro l'alternativa (o "ipotesi alternativa") $\mathcal{H}_1\big)\ \theta > 0.02$.

Fissata l'ipotesi, si determina un insieme di risultati che portano a rifiutarla: questo si identifica con un sottinsieme C dello spazio campionario Ω chiamato **regione critica** o anche **regione di rifiuto**, il suo complementare $A = C^c$ è chiamato invece **regione di accettazione**. Tornando all'esempio del controllo di qualità, si rifiuterà l'ipotesi se la percentuale di pezzi difettosi riscontrata è superiore a un certo d (da scegliere secondo opportune regole): scriveremo sinteticamente $C = \left\{\omega \in \Omega \;\middle|\; \overline{X}(\omega) > d\right\} = \left\{\overline{X} > d\right\}$.

Vi sono due tipi di errore: l'**errore di prima specie** consiste nel rifiutare l'ipotesi quando è soddisfatta, e l'**errore di seconda specie** consiste nell'accettare l'ipotesi quando non è soddisfatta.

Definizione 6.2.1 (**Livello del test).** Fissato $0 < \alpha < 1$, si dice che il test è di livello α se per ogni $\theta \in \Theta_0$ si ha $\mathbf{P}_\theta(C) \leq \alpha$.

Fissare un livello significa pertanto fissare *un limite superiore per la probabilità dell'errore di prima specie*, i valori tipici del livello sono piccoli, ad esempio 0.05 o 0.02.

Definizione 6.2.2 (**Potenza del test).** Si chiama potenza del test la funzione, definita sui parametri dell'alternativa Θ_1 , $\theta \to \mathbf{P}_\theta(C)$.

La potenza rappresenta in un certo senso la *"capacità di accorgersi che l'ipotesi non è soddisfatta"* e naturalmente l'ideale sarebbe avere livello basso e potenza alta, ma occorre bilanciare queste due esigenze che sono in contrapposizione.

Definizione 6.2.3 (**Curva Operativa).** La curva operativa (chiamata anche Curva Operativa Caratteristica e indicata sinteticamente O.C.) è la funzione, definita sull'intero insieme dei parametri Θ, $\beta(\theta) = \mathbf{P}_\theta(A)$ (cioè la probabilità, in funzione del parametro θ, di accettare l'ipotesi).

È evidente che dalla *curva operativa* si possono ricavare *livello* e *potenza*, in quanto $\mathbf{P}_\theta(C) = 1 - \mathbf{P}_\theta(A)$: il linguaggio *livello e potenza* è più tipico dell'inferenza statistica, mentre il linguaggio della *Curva Operativa* è più tipico della teoria dell'affidabilità.

A rigore, perché si possa parlare di *curva*, Θ dovrebbe essere un intervallo (in senso lato) di $\mathbb{R}$.

Naturalmente se il livello del test diminuisce, deve diminuire la *regione critica* e di conseguenza *diventa più facile accettare l'ipotesi*; potremmo trovarci quindi a dover verificare il test con livelli diversi e concludere ad esempio *"l'ipotesi è rifiutata la livello 0.05 ma è accettata al livello 0.02 .."*

Questo suggerisce il concetto di *"p-value"*: questo termine anglosassone non ha una traduzione italiana consolidata, alcuni lo chiamano *"valore-p"*, altri il *"p-dei-dati"* ... io ritengo preferibile usare il termine anglosassone.

Definizione 6.2.4 (**p-value).** Si chiama *p-value* il numero $\overline{\alpha}$ tale che, se $\alpha < \overline{\alpha}$, l'ipotesi viene accettata al livello α, se invece $\alpha > \overline{\alpha}$, l'ipotesi viene rifiutata al livello α.

Il calcolo del *p-value* è molto importante perché sintetizza in un solo numero la plausibilità di una ipotesi: se il p-value è basso (poniamo inferiore a 0.1) questo vuol dire che l'ipotesi è *poco plausibile* mentre se è più alto (poniamo superiore a 0.3) vuol dire che l'ipotesi è *molto plausibile*.

La definizione di *p-value* che usualmente viene data è forse più intuitiva ma meno precisa, la riporto qua sotto.

Definizione 6.2.5 (**p-value: definizione alternativa).** Il *p-value* è la probabilità che l'eventuale rifiuto dell'ipotesi sia dovuta al caso (e non strutturale).

Non è possibile a questo stadio dimostrare l'equivalenza delle due definizioni; quando esamineremo in dettaglio come si calcola il p-value nel *test sulla media di un campione gaussiano con varianza nota*, apparirà con chiarezza che queste due definizioni apparentemente diverse in realtà coincidono.

È importante sottolineare che mentre la regione critica (che dipende dal livello scelto) viene decisa *prima di raccogliere i dati* cioè prima di fare l'esperimento, il p-value *dipende dai dati che sono stati raccolti* attuando l'esperimento.

6.2.1 Test sulla media di un campione Gaussiano con varianza nota, o Z-test

Convenzionalmente in questo paragrafo indichiamo con Z una variabile Gaussiana standard, e (come per la determinazione dell'intervallo di fiducia) in presenza di un campione gaussiano con varianza nota i calcoli sono basati sul fatto che la variabile $\sqrt{n}\,\dfrac{(\overline{X}-m)}{\sigma} = Z$ è gaussiana standard.

Occupiamoci del **test bilatero** dell'ipotesi

$$\mathcal{H}_0)\ m = m_0 \qquad \text{contro} \qquad \mathcal{H}_1)\ m \neq m_0$$

Poiché $\overline{X}$ è la stima della media m, l'intuizione ci porta a rifiutare l'ipotesi se $\overline{X}$ si scosta troppo da m_0, cioè a scegliere una regone critica delle forma $C = \left\{\, \left|\overline{X} - m_0\right| > d \,\right\}$ dove il numero d deve essere determinato in funzione del livello α scelto.

Si deve cioè avere $\mathbf{P}_{m_0}\left\{\, \left|\overline{X} - m_0\right| > d \,\right\} \leq \alpha$ e, per ottenere una regione critica più grande possibile (allo scopo di aumentare la *potenza* del test) cambiamo la diseguaglianza ponendo $\mathbf{P}_{m_0}\left\{\, \left|\overline{X} - m_0\right| > d \,\right\} = \alpha$.

Abbiamo dunque

$$\mathbf{P}_{m_0}\left\{\left|\overline{X} - m_0\right| > d\right\} = \mathbf{P}_{m_0}\left\{\frac{\sqrt{n}}{\sigma}\left|\overline{X} - m_0\right| > \frac{d\sqrt{n}}{\sigma}\right\} = \mathbf{P}\left\{|Z| > \frac{d\sqrt{n}}{\sigma}\right\} = \alpha$$

e questo ci porta a porre $\frac{d\sqrt{n}}{\sigma} = q_{1-\frac{\alpha}{2}}$: la regione critica al livello α diventa pertanto $C = \left\{\, \left|\overline{X} - m_0\right| > \dfrac{\sigma}{\sqrt{n}}\ q_{1-\frac{\alpha}{2}} \,\right\}$.

Osservazione 6.2.6 (**Legame tra intervalli di fiducia e test).** Si nota facilmente che *l'ipotesi* $\mathcal{H}_0)\ m = m_0$ *è accettata al livello* α se e solo se m_0 *appartiene all'intervallo di fiducia per la media* con livello di fiducia $(1-\alpha)$. Questa è una proprietà generale: si può dimostrare che è equivalente identificare un *intervallo di fiducia* al livello $(1-\alpha)$ oppure un test *nel quale l'ipotesi è semplice*, cioè ridotta a un solo parametro.

Tornando al test, in pratica di fronte a dati concreti $x_1, \ldots, x_n$, si calcola la media empirica $\overline{x}$ e si rifiuta l'ipotesi se $|\overline{x} - m_0| > \frac{\sigma}{\sqrt{n}}\ q_{1-\frac{\alpha}{2}}$.

Occupiamoci ora del **calcolo del p-value**: abbiamo visto che l'ipotesi è rifiutata al livello α se $\frac{\sqrt{n}}{\sigma}\,|\overline{x}-m_0| > \ q_{1-\frac{\alpha}{2}}$, cioè (ricordando che, se $m = m_0$, la variabile $\sqrt{n}\,\frac{(\overline{X}-m_0)}{\sigma} = Z$ è gaussiana standard) se

$$\mathbf{P}_{m_0}\Big\{\,\sqrt{n}\,\frac{\big|\overline{X}-m_0\big|}{\sigma} > \frac{\sqrt{n}}{\sigma}\,|\overline{x}-m_0|\,\Big\} < \mathbf{P}_{m_0}\Big\{\,\sqrt{n}\,\frac{\big|\overline{X}-m_0\big|}{\sigma} > q_{1-\frac{\alpha}{2}}\Big\} = \alpha$$

mentre l'ipotesi viene *accettata* se vale la diseguaglianza opposta: l'elemento discriminante tra accettazione e rifiuto si ha pertanto quando vale l'eguaglianza, cioè

$$\mathbf{P}_{m_0}\Big\{\,\sqrt{n}\,\frac{\big|\overline{X}-m_0\big|}{\sigma} > \frac{\sqrt{n}}{\sigma}\,|\overline{x}-m_0|\,\Big\} = \overline{\alpha}$$

In formule, il *p-value* è pertanto

$$\mathbf{P}\big\{\,|Z| > \frac{\sqrt{n}}{\sigma}\,|\overline{x}-m_0|\big\} = 2\,\Big[1-\Phi\big(\frac{\sqrt{n}}{\sigma}\,|\overline{x}-m_0|\big)\,\Big]$$

Ma torniamo all'eguaglianza precedente: se l'ipotesi è soddisfatta, la variabile $\sqrt{n}\,\frac{(\overline{X}-m_0)}{\sigma}$ è gaussiana standard, e quindi il p-value diventa effettivamente la probabilità che, se l'ipotesi è soddisfatta, questa venga erroneamente rifiutata in seguito ai dati dell'esperimento: si ha cioè l'equivalenza tra le definizioni 6.2.4 e 6.2.5.

Calcoliamo ora in dettaglio la **curva operativa** al livello α : si ha

$$\beta(m) = \mathbf{P}_m\Big\{\,\frac{\sqrt{n}}{\sigma}\,\big|\overline{X}-m_0\big| \le q_{1-\frac{\alpha}{2}}\Big\} = \mathbf{P}_m\Big\{-q_{1-\frac{\alpha}{2}} \le \frac{\sqrt{n}}{\sigma}\,\big(\overline{X}-m_0\big) \le q_{1-\frac{\alpha}{2}}\Big\} =$$

$$= \mathbf{P}_m\Big\{\,\sqrt{n}\frac{(m_0-m)}{\sigma} - q_{1-\frac{\alpha}{2}} \le \frac{\sqrt{n}}{\sigma}\,\big(\overline{X}-m\big) \le \sqrt{n}\frac{(m_0-m)}{\sigma} + q_{1-\frac{\alpha}{2}}\,\Big\} =$$

$$= \Phi\Big(\sqrt{n}\frac{(m_0-m)}{\sigma} + q_{1-\frac{\alpha}{2}}\Big) - \Phi\Big(\sqrt{n}\frac{(m_0-m)}{\sigma} - q_{1-\frac{\alpha}{2}}\Big)$$

Occupiamoci ora, questa volta senza ripetere tutti i dettagli, del **test unilatero** dell'ipotesi

$$\mathcal{H}_0)\ m \le m_0 \qquad \text{contro} \qquad \mathcal{H}_1)\ m > m_0$$

L'intuizione ci spinge a rifiutare l'ipotesi se $(\overline{X}-m_0)$ è troppo grande, cioè a considerare una regione critica della forma $C = \big\{\,\big(\overline{X}-m_0\big) > d\,\big\}$, e la condizione sul livello diventa

$$\forall\ m \le m_0\,, \quad \mathbf{P}_m\big\{\,\big(\overline{X}-m_0\big) > d\,\big\} \le \alpha$$

È intuitivo (e anche facile da mostrare) il fatto che la probabilità sopra scritta cresce al crescere di m e si arriva (procedendo come nel caso precedente) a

$$\mathbf{P}_{m_0}\big\{\big(\overline{X}-m_0\big)>d\big\}=\mathbf{P}_{m_0}\Big\{\frac{\sqrt{n}}{\sigma}\big(\overline{X}-m_0\big)>\frac{d\sqrt{n}}{\sigma}\Big\}=\alpha$$

da cui segue $\frac{d\sqrt{n}}{\sigma}=q_{1-\alpha}$: la *regione critica* al livello α diventa pertanto $C=\Big\{\big(\overline{X}-m_0\big)>\dfrac{\sigma}{\sqrt{n}}\,q_{1-\alpha}\Big\}$.

Il **p-value** diventa

$$\mathbf{P}_{m_0}\Big\{\sqrt{n}\,\frac{\big(\overline{X}-m_0\big)}{\sigma}>\frac{\sqrt{n}}{\sigma}\,(\overline{x}-m_0)\Big\}=1-\Phi\Big(\frac{\sqrt{n}}{\sigma}\,(\overline{x}-m_0)\Big)$$

e la **curva operativa** è la funzione

$$\beta(m)=\mathbf{P}_m\Big\{\frac{\sqrt{n}}{\sigma}\,\big(\overline{X}-m_0\big)\le q_{1-\alpha}\Big\}=\Phi\Big(\sqrt{n}\,\frac{(m_0-m)}{\sigma}+q_{1-\alpha}\Big)$$

Se invece l'ipotesi è della forma $\mathcal{H}_0\big)\ m\ge m_0$, si ribaltano tutte le diseguaglianze e si sostituisce $q_{1-\alpha}$ con q_α.

6.2.2 Test sulla media di un campione Gaussiano con varianza sconosciuta, o T-test

I passaggi che abbiamo svolto in dettagli per i test sulla media di un campione Gaussiano, con varianza nota, ci guidano a esaminare il caso *più realistico* della varianza sconosciuta: questa volta il punto di partenza è il fatto che, se la media è m, la variabile $T=\sqrt{n}\,\dfrac{\big(\overline{X}-m\big)}{S}$ ha densità di Student $T(n-1)$.

Ad esempio il test **bilatero** dell'ipotesi

$$\mathcal{H}_0\big)\ m=m_0\,,\ \sigma\text{ qualsiasi}\qquad\text{contro}\qquad\mathcal{H}_1\big)\ m\ne m_0\,,\ \sigma\text{ qualsiasi}$$

al livello α, ha come regione critica $C=\Big\{\big|\overline{X}-m_0\big|>\dfrac{S}{\sqrt{n}}\,\tau_{(1-\frac{\alpha}{2},n-1)}\Big\}$ mentre la formula per il *p-value* diventa

$$\mathbf{P}_{m_0}\Big\{\sqrt{n}\,\frac{\big|\overline{X}-m_0\big|}{S}>\frac{\sqrt{n}}{s}\,|\overline{x}-m_0|\Big\}=2\Big[1-F_{n-1}\Big(\frac{\sqrt{n}}{s}\,|\overline{x}-m_0|\Big)\Big]$$

(dove F_n è la c.d.f. della variabile $T(n)$ e $\tau_{(\alpha,n)}$ il relativo α-quantile).

Di fronte ai dati concreti $x_1,\dots,x_n$, si hanno le eguaglianze $\overline{x}=\sum_i\frac{x_i}{n}$ e $s^2=\sum_i\frac{(x_i-\overline{x})^2}{n-1}$ (cioè s^2 è la varianza campionaria dei dati).

Nel caso di una variabile Gaussiana, sia la funzione di ripartizione che il quantile si ricavano dalla tavola della variabile $N(0,1)$ ma nel caso della variabile di Student le cose sono diverse: in passato si usavano delle opportune tavole ma queste sono ora superate e si ricorre a un qualsiasi software statistico.

Sostanzialmente tutte le formule si riportano con queste modifiche con un'unica eccezione: la *curva operativa*. Per essere più precisi in questo caso *la curva operativa non ha senso* ed il motivo è il seguente: se la media delle variabili è m_0, qualunque sia la varianza la variabile $T = \sqrt{n}\,\dfrac{(\overline{X} - m_0)}{S}$ ha densità di Student $T(n-1)$, ma se la media è diversa da m_0 la densità di T non dipende solo da m ma anche da σ.

6.2.3 Test approssimato su un campione di Bernoulli

Esattamente come abbiamo fatto per gli intervalli di fiducia approssimati, in presenza di un campione $X_1, \ldots, X_n$ di variabili di Bernoulli di parametro p con $0 < p < 1$, si utilizza il fatto che la variabile $\sqrt{n}\,\frac{\overline{X}-p}{\sqrt{p(1-p)}}$ è approssimativamente gaussiana standard.

Consideriamo ad esempio il test **bilatero** dell'ipotesi

$$\mathcal{H}_0)\ p = p_0 \qquad \text{contro} \qquad \mathcal{H}_1)\ p \neq p_0$$

al livello α: questo ha **regione critica** $C = \left\{ \dfrac{\sqrt{n}\,\left|\overline{X} - p_0\right|}{\sqrt{p_0(1-p_0)}} > q_{1-\frac{\alpha}{2}} \right\}$. In questo caso $\overline{X} = \widehat{p}$, cioè la *percentuale* osservata.

In modo simile si ottiene il **p-value** e precisamente

$$\mathbf{P}_{p_0}\Big\{ \frac{\sqrt{n}\,\left|\overline{X} - p_0\right|}{\sqrt{p_0(1-p_0)}} > \frac{\sqrt{n}\,\left|\widehat{p} - p_0\right|}{\sqrt{p_0(1-p_0)}} \Big\} \approx 2\Big[1 - \Phi\Big(\frac{\sqrt{n}\,\left|\widehat{p} - p_0\right|}{\sqrt{p_0(1-p_0)}}\Big)\Big]$$

È un po' diverso il discorso per quanto riguarda la Curva Operativa: a differenza del caso varianza sconosciuta (nel quale la curva operativa *non aveva senso*) questa volta può essere calcolata (approssimativamente). Tuttavia i calcoli sono più lunghi di quelli fatti per il caso del test sulla media di un campione gaussiano con varianza nota poiché quando cambia p cambia anche la varianza ... in sostanza non è interessante riportare i conti.

Anche per quanto riguarda i test *unilateri* si procede in analogia con il caso Gaussiano, non riporto i dettagli.

6.2.4 Test sulla varianza di un campione Gaussiano

Come per gli intervalli di fiducia, ci occupiamo solo del caso *unilatero*, ricordando che, se il campione $X_1, \ldots, X_n$ è formato da variabili gaussiane con varianza σ^2, la variabile $\sum_{i=1}^n \frac{\left(X_i - \overline{X}\right)^2}{\sigma^2}$ ha densità $\chi^2(n-1)$.

Guardiamo allora il **test unilatero** dell'ipotesi

$$\mathcal{H}_0)\ \sigma^2 \leq \sigma_0^2 \qquad \text{contro} \qquad \mathcal{H}_1)\ \sigma^2 > \sigma_0^2$$

al livello α: è naturale considerare una regione critica della forma $C = \left\{ \sum_i \left(X_i - \overline{X}\right)^2 > d \right\}$ con d tale che si abbia, per $\sigma \leq \sigma_0$

$$\mathbf{P}_\sigma \Big\{ \sum_{i=1}^n \left(X_i - \overline{X}\right)^2 > d \Big\} \leq \alpha$$

È facile provare che la probabilità sopra scritta cresce con σ e, cercando una regione critica più piccola possibile (purché di livello α) si ottiene

$$\mathbf{P}_{\sigma_0} \Big\{ \sum_{i=1}^n \left(X_i - \overline{X}\right)^2 > d \Big\} = \mathbf{P}_{\sigma_0} \Big\{ \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma_0^2} > \frac{d}{\sigma_0^2} \Big\} = \alpha$$

e questo ci porta a prendere $\frac{d}{\sigma_0^2} = \chi^2_{(1-\alpha,n-1)}$. Si ottiene quindi una regione critica della forma $C = \left\{ \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma_0^2} > \chi^2_{(1-\alpha,n-1)} \right\}$, mentre il **p-value** è eguale a

$$\mathbf{P}_{\sigma_0} \Big\{ \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma_0^2} > \frac{\sum_{i=1}^n \left(x_i - \overline{x}\right)^2}{\sigma_0^2} \Big\} = 1 - G_{n-1}\Big(\frac{\sum_{i=1}^n \left(x_i - \overline{x}\right)^2}{\sigma_0^2} \Big)$$

dove con G_n si indica la c.d.f. della variabile chi-quadro(n).

È agevole calcolare la *curva operativa*: si ha infatti

$$\beta(\sigma) = \mathbf{P}_\sigma \Big\{ \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma_0^2} \leq \chi^2_{(1-\alpha,n-1)} \Big\} =$$

$$= \mathbf{P}_\sigma \Big\{ \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma^2} \leq \frac{\sigma_0^2}{\sigma^2}\ \chi^2_{(1-\alpha,n-1)} \Big\} = G_{n-1}\Big(\frac{\sigma_0^2}{\sigma^2}\ \chi^2_{(1-\alpha,n-1)} \Big)$$

Naturalmente, se si considera un test unilatero nell'altra direzione, cioè dell'ipotesi $\mathcal{H}_0)\ \sigma^2 \geq \sigma_0^2$, si ribaltano tutte le diseguaglianze ed il quantile $\chi^2_{(1-\alpha,n-1)}$ è cambiato in $\chi^2_{(\alpha,n-1)}$.

6.3 Confronto tra due (o più) campioni statistici

Introduciamo l'argomento di questo paragrafo con due esercizi, apparentemente simili.

Esercizio 6.3.1. Viene propagandata una cura dimagrante che promette *7 chili in 7 giorni*: vengono misurati *"prima e dopo"* 7 pazienti ottenendo i valori che seguono

a) 72 89 94 77 86 91 83

b) 68 83 90 71 80 88 74

Si può accettare l'affermazione dell'istituto che offre la cura oppure si deve concludere che *hanno esagerato*? Quale sarebbe stata la conclusione se invece di promettere 7 chili in 7 giorni avessero promesso solo 6 chili in 7 giorni?

Esercizio 6.3.2. Vengono misurate le lunghezze delle tibie di uomini adulti provenienti da reperti di tombe etrusche: dal sito di Cerveteri vengono effettuate 13 misurazioni ottenendo un valore medio di 47.2 ed una varianza campionaria di 7.98 , mentre dal sito di Ladispoli si ottengono 8 misurazioni con un valore medio di 44.9 ed una varianza campionaria di 8.85.

Si può affermare che la differenza sia una semplice fluttuazione statistica oppure si deve concludere che gli abitanti di Cerveteri erano veramente più alti?

Fermo restando che questi campioni sono troppo poco numerosi per trarre delle conclusioni veramente significative, si tratta di due situazioni *radicalmente diverse*: nel primo caso ci troviamo due campioni *accoppiati* e nel secondo *indipendenti*.

Nel caso dei campioni *accoppiati* non c'è sostanzialmente nulla di nuovo da aggiungere dal punto di vista teorico: abbiamo il campione $X_1, \dots, X_n$ con densità $N(m_1, \sigma_1^2)$, ed il secondo campione $Y_1, \dots, Y_n$ con densità $N(m_2, \sigma_2^2)$, prendendo le differenze $Z_i = X_i - Y_i$ queste formeranno un campione con densità $N(m_1 - m_2, \sigma_1^2 + \sigma_2^2)$. Per risolvere un problema come il 6.3.1, sostanzalmente un confronto tra le medie, ci si riporta al test di Student.

Quando invece i due campioni sono *indipendenti* (tra l'altro in genere di numerosità diverse), la situazione cambia: i campioni possono essere anche più di due, e si apre tutto un capitolo dell'inferenza statistica chiamato *analisi della varianza*. Qui ci limitiamo a illustrare un primo passo.

Il risultato teorico che è alla base del confronto tra le medie di due campioni è il risultato seguente, che si dimostra sostanzialmente come l'ultimo punto del Teorema 5.2.2.

Teorema 6.3.3. *Siano $X_1, \dots, X_n$ un campione con densità $N(m_1, \sigma^2)$ e $Y_1, \dots, Y_k$ un campione indipendente dal primo con densità $N(m_2, \sigma^2)$ (si suppone cioè che le varianze siano sconosciute ma eguali): la variabile*

$$T_{n,k} = \frac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{\sqrt{\sum_i (X_i - \overline{X})^2 + \sum_j (Y_j - \overline{Y})^2}} \; \frac{\sqrt{n+k-2}}{\sqrt{\frac{1}{n} + \frac{1}{k}}}$$

ha densità di Student $T_{(n+k-2)}$ (cioè a $(n+k-2)$ gradi di libertà).

Naturalmente rimane aperto il problema di decidere se è verosimile che i due campioni abbiano la stessa varianza; a questo scopo esiste il test F (di Fisher) che però non illustriamo.
Ora siamo in grado di risolvere i due esercizi proposti.

Soluzione esercizio 6.3.1 Se chiamiamo $Z_i = X_i - Y_i$, queste variabili formano un campione di legge $N(m_1 - m_2, \sigma^2)$ (cioè varianza sconosciuta) al quale si può applicare il T-test $\mathcal{H}_0) \, m \geq 7$ contro $\mathcal{H}_1) \, m < 7$.

Prendendo la variabile $T = \sqrt{7} \, \frac{(\overline{Z} - 7)}{S(Z)}$, il *p-value* del test è $\mathbf{P}\{T \leq t\} = F_6(t)$, dove t è il valore che risulta dai dati.

Con dei conti elementari, $t = -2.09$ e $F_6(t) = 0.04$, cioè l'ipotesi è molto poco credibile.

Se si considera l'ipotesi $\mathcal{H}_0) \, m \geq 6$, risulta $t = -0.76$ e $F_6(t) = 0.23$ e questa volta l'ipotesi è molto più verosimile.

Soluzione esercizio 6.3.2 Possiamo supporre che le varianze dei due campio siano eguali (le varianze campionarie sono abbastanza simili); considerando che in questo caso $n = 13$ e $k = 8$, e ponendo $m = m_1 - m_2$, la variabile

$$T = \frac{\overline{X} - \overline{Y} - m}{\sqrt{\sum_i (X_i - \overline{X})^2 + \sum_j (Y_j - \overline{Y})^2}} \; \frac{\sqrt{19}}{\sqrt{\frac{1}{13} + \frac{1}{8}}}$$

ha densità di Student $T(19)$.

Consideriamo il test dell'ipotesi $\mathcal{H}_0) \, m = 0$ contro $\mathcal{H}_1) \, m > 0$ e calcoliamo il valore che risulta dai dati: $t = \frac{47.2 - 44.9}{\sqrt{7.98 \times 12 + 8.85 \times 7}} \; \frac{\sqrt{19}}{\sqrt{\frac{1}{13} + \frac{1}{8}}} = 1.776$.

Il *p-value* che risulta è $\mathbf{P}\{T > t\} = 1 - F_{19}(1.776) = 0.045$: si tratta di un valore decisamente basso e l'ipotesi è da scartare. Si conclude che gli abitanti di Cerveteri erano effettivamente più alti.