PDF → Per una distribuzione continua, quanto è densa la probabilità in ogni punto (probability density function)

CDF → (CUMULATIVE DISTRIBUTION FUNCTION) ti dice la probabilità di $\leq x$, quindi sommi tutte le probabilità fino a quel punto. Per una distribuzione continua, la CDF è l'area sotto la PDF fino a quel punto.

5.3 Distribuzioni continue

Come abbiamo già detto all'inizio del paragrafo precedente, quando una variabile casuale ha il supporto coincidente con un intervallo di numeri reali, $S_x = [a, b]$, è detta variabile aleatoria continua.

Esempi di variabili aleatorie continue sono l'altezza o il peso di una persona, la durata di un evento in tempo...

In questo paragrafo considereremo quindi modelli che coinvolgono variabili aleatorie continue.

5.3.1 Funzioni di variabili casuali continue: PDF e CDF

Se X è una variabile aleatoria continua, si definisce *funzione di densità di probabilità (PDF)* f_X associata a X la funzione

$$f_X : S_X \to R$$

PDF ≡

↓

area sottografico

tale che

$$P(X \in A) = \int_A f_X(x)dx.$$

A come A=[a,b]

Alcune considerazioni.

- L'insieme A è a sua volta un intervallo $[c, d] \subset [a, b]$, oppure un'unione di intervalli contenuti in $[a, b]$. Quindi:

$$P(c \leq X \leq d) = \int_c^d f_X(x)dx$$

è l'area sottesa dalla la funzione f_X nell'intervallo $[c, d]$.
- In particolare, $P(X = c) = 0,\ \forall c$, poiché l'area sottesa da una retta è nulla.

Valgono per la PDF proprietà analoghe a quelle viste per la PMF di una variabile aleatoria discreta:

1. $f_X(x) \geq 0 \ \forall x \in S_X$
2. $\int_{x \in S_X} f_X(x) = 1$
3. $\cancel{P(X \in A) = \int_{x \in A} f_X(x) \text{ per ogni evento } A \subseteq S_X}$.

Se X è una variabile aleatoria continua, si definisce anche la *funzione di ripartizione o di distribuzione di X (Cumulative Distribution Function CDF)* $F_X(t)$:

$F_X : R \to [0,1]$

$$F_X(t) = P(X \leq t), \quad -\infty < t < \infty.$$

Anche per la CDF valgono proprietà analoghe a quelle viste per la CDF di variabile aleatorie discrete:

1. $F_X(t)$ è non decrescente,
2. $F_X(t)$ è continua (in questo caso non solo a destra),
3. vale: $lim_{t\to-\infty} F_X(t) = 0$ e $lim_{t\to+\infty} F_X(t) = 1$.

Esiste anche una relazione fra PDF e CDF continue:

$$F'_X(t) = \frac{d}{dt} F_X(t) = \frac{d}{dt} \int_{-\infty}^{t} f_X(x) dx.$$

Si definiscono anche nel caso continuo i seguenti valori particolari per le distribuzioni:

- **media**

$$\mu = \mathbf{E}X = \int_{x \in S_X} x f_X(x),$$

(supponendo che la serie infinita $\sum |x| f_X(x)$ sia convergente)

- **varianza**

$$\sigma^2 = \int_{x \in S_X} (x - \mu)^2 f_X(x),$$

- **deviazione standard** $\sigma = \sqrt{\sigma^2}$.

Si definisce anche per le variabili aleatorie continue la funzione generatrice dei momenti:

Definizione. Data una variabile aleatoria X continua con PDF $f_X(x)$, la sua *funzione generatrice dei momenti* (MGF) è definita da:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x),$$

supposto che l'integrale esista e sia finito in un intorno di $t = 0$.

$S =]0,1[$

Esempio 5.28 *Supponiamo che X abbia la PDF $f(x) = 3x^2$, $0 < x < 1$.*

1. *Calcolare la probabilità che X sia nell'intervallo $[0.2, 0.65]$.*
2. *Calcolare inoltre la media della distribuzione f_X.*

Soluzione.

$$P(0,2 \le x \le 0,65) = \int_c^d f_X(x)\,dx = \int_{0,2}^{0,65} 3x^2\,dx = \left[\frac{\cancel{3}x^3}{\cancel{3}}\right]_{0,2}^{0,65} = 0,65^3 - 0,2^3$$

1. In R:

```
> f <- function(x) 3 * x ^ 2
> integrate(f, lower = 0.2, upper = 0.65)

## 0.266625 with absolute error < 3e-15
```

$$\mu = \int_{x \in S_X} x f_X(x) = \int_0^1 x\, 3x^2 \to 3\int_0^1 x^3 \to 3\left[\frac{x^4}{4}\right]_0^1 = \frac{3}{4}$$

$$\sigma^2 = \int_{x \in S_x} (x-\mu)^2 f_x(x) = \int_0^1 \left(x-\frac{3}{4}\right)^2 3x^2 dx = \int_0^1 \left(x^2+\frac{9}{16}-\frac{3}{2}x\right) 3x^2 dx$$

$$= \int_0^1 \frac{9}{16} \cdot 3x^2 + \int_0^1 3x^4 + \int_0^1 -\frac{3}{2} 3x^3$$

2.
```
> fm <- function(x) 3 * x ^ 3
> integrate(fm, lower = 0, upper = 1)

## 0.75 with absolute error < 8.3e-15
```

Consideriamo ora le più comuni distribuzioni continue.

5.3.2 Distribuzione continua uniforme

$S_x = [a, b]$

Una variabile aleatoria casuale ha *distribuzione continua uniforme sull'intervallo* $[a, b]$ se ha una PDF della forma:

$$f_X(x) = \frac{1}{b-a}, \quad a < x < b.$$

PDF

Si dice che $X \sim$ `unif(min=a,max=b)`.

Per questa particolare PDF si può esplicitamente scrivere la CDF associata:

$$F_X(t) = \begin{cases} 0, & t < 0, \\ \dfrac{t-a}{b-a}, & a \le t \le b, \\ 1, & t \ge b. \end{cases}$$

La distribuzione continua uniforme descrive, nel continuo, eventi che seguono il modello ELM, per i quali, quindi, tutti i risultati hanno la stessa probabilità di successo.

La media si calcola come:

$$\begin{aligned} \mu &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_a^b x \frac{1}{b-a} dx \\ &= \frac{1}{b-a} \left[\frac{x^2}{2}\right]_a^b \\ &= \frac{b+a}{2}. \end{aligned}$$

In R, la funzione `dunif(min=a,max=b)` è associata alla distribuzione continua uniforme. Anche per le funzioni relative alle distribuzioni continue, ci sono 4 diverse funzioni che cominciano con le lettere **d**, **p**, **q**, **r** e hanno lo stesso significato visto per le distribuzioni discrete.

Esempio 5.29 *Calcolare* 1 *numero casuale nell'intervallo* $[0, 1]$.

Soluzione.

```
> runif(n = 1, min = 0, max = 1)

## [1] 0.9914289
```

Se dobbiamo calcolare più numeri casuali, basta modificare il primo parametro di input della funzione `runif`.

5.3.3 Distribuzione normale

Una variabile aleatoria casuale ha *distribuzione normale* se ha una PDF della forma:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty.$$

Si dice che $X \sim$ `norm(mean=mu, sd=sigma)`.

La distribuzione normale viene quindi identificata da due parametri che coincidono con la sua media μ e la sua deviazione standard σ. La distribuzione normale, dalla caratteristica forma a campana, viene anche detta *distribuzione gaussiana* ed è sicuramente la distribuzione più diffusa nella realtà. In particolare, il caso in cui la media $\mu = 0$ e la deviazione standard $\sigma^2 = 1$ è detta *distribuzione normale standard*. In questo caso la PDF è:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} exp(-x^2/2).$$

PDF della distribuzione normale standard $x=0$ $\sigma^2=1$

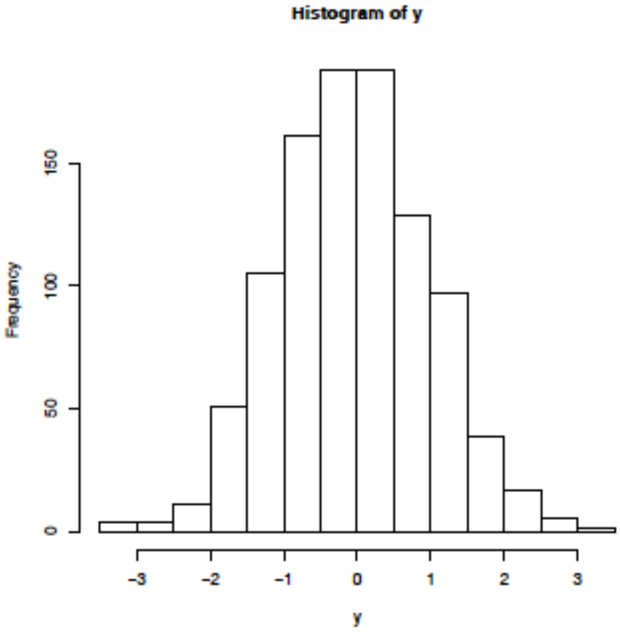
Invece la CDF è:

$$F_X(x) = \int_{-\infty}^{t} f_X(x)dx, \quad -\infty < x < \infty.$$

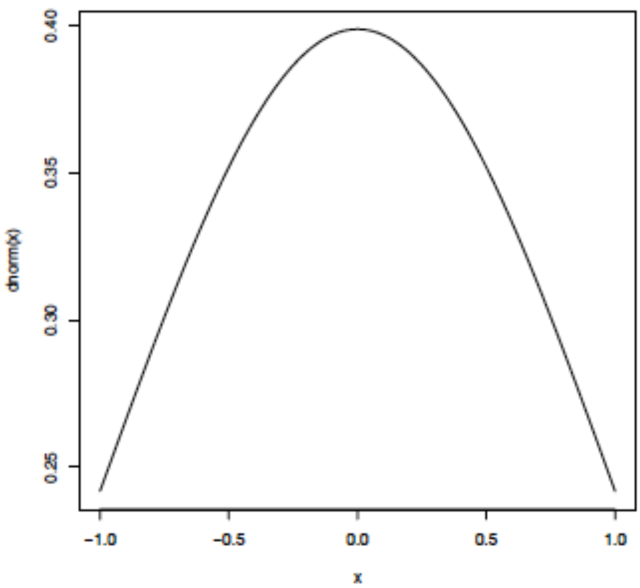
In particolare, si ha che se X `norm(mean=mu, sd=sigma)` allora la variabile aleatoria $Z = \frac{X-\mu}{\sigma}$ ha distribuzione normale standard.

In R, le funzioni associate alla distribuzione normale sono: `dnorm`, `pnorm`, `qnorm`, `rnorm`.

```
> y <- rnorm(1000, mean = 0, sd = 1)
> hist(y, breaks = 20)
```




```
> curve(dnorm(x), from = -1, to = 1)
```



$X \sim N(2, 2.5)$

Calcolo la PDF poi calcolo la CDF

PDF $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ lascio in funzione di x

CDF $P(1{,}5 \le x \le 3) = \int_{1{,}5}^{3} f_X(x) = \int_{1{,}5}^{3} \frac{1}{\sqrt{2{,}5 \cdot 2\pi}} \exp\left(\frac{-(x-2)^2}{2 \cdot 2{,}5}\right)$

Esempio 5.30 *Consideriamo X variabile aleatoria che rappresenta il rumore di registrazione di un'immagine digitale, con distribuzione normale con media 2 e varianza 2.5. Qual è la probabilità che $1.5 \le X \le 3$?*

$\mu = 2$

$\sigma^2 = 2{,}5$

Soluzione.

```
> pnorm(3, mean = 5, sd = 2.5) - pnorm(1.5, mean = 2,
+                                       sd = 2.5)

## [1] 0.2346815
```

5.3.4 Distribuzioni di variabili aleatorie di misura di un tempo di attesa

(raro)

Se la variabile aleatoria rappresenta un *tempo prima che accada un evento*, tale variabile è continua, poiché il suo supporto è un intervallo di numeri reali. In generale, l'evento è un *arrivo*, e quindi si misura il tempo prima dell'arrivo (cliente, fallimento di una procedura, ecc.).

Vediamo allora due distribuzioni di probabilità relative a variabili aleatorie di questo tipo.

5.3.4.1 Distribuzione esponenziale

$S_X = (0, \infty)$

La variabile aleatoria continua X ha *distribuzione esponenziale* se ha una PDF della forma:

$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0,$$

λ = media eventi unità di tempo

e si dice che $X \sim$ `exp(rate=lambda)`.

La CDF è:

$$F_X(t) = 1 - e^{-\lambda t}, \quad t > 0.$$

La media è $\mu = 1/\lambda$ e la varianza $\sigma^2 = 1/\lambda^2$.

Per la distribuzione esponenziale vale la seguente relazione:

$$P(X > a + b | X > b) = P(X > a), \quad \forall a, b > 0,$$

che significa che la probabilità che il primo arrivo si verifichi dopo un tempo b quando è già trascorso il tempo a è la stessa che avrei *anche se il tempo a non fosse trascorso*. Cioè la distribuzione esponenziale *non ha memoria*.

Analizziamo ora la relazione fra la distribuzione esponenziale e la distribuzione discreta di Poisson. Consideriamo l'esempio, visto nel paragrafo precedente, delle auto che arrivano alla pompa di benzina, con una media di 25 auto all'ora. La variabile aleatoria X che conta *quante auto* arrivano in un intervallo di tempo t ha distribuzione $X \sim$ `pois(lambda= 25 t)`. Se consideriamo invece la variabile aleatoria Y che *misura l'intervallo di tempo prima che la prima auto arrivi*, allora $Y \sim$ `exp(rate=25)`. Infatti, se vogliamo scrivere la CDF di Y, possiamo considerare la relazione $P(Y \leq t) = 1 - P(Y \geq t)$, dove $P(Y > t)$ è la probabilità che il primo arrivo si abbia *dopo* il tempo t o, equivalentemente, la probabilità che *non ci siano arrivi in [0,t]*. Quindi:

$$P(Y > t) = P(Y = 0) = e^{-\lambda t},$$

infine:

$$P(Y \leq t) = 1 - e^{-\lambda t},$$

che è esattamente la CDF della distribuzione esponenziale.

5.3.4.2 Distribuzione gamma

Una generalizzazione della distribuzione esponenziale si può considerare la *distribuzione gamma.*

Una variabile aleatoria continua X ha una distribuzione gamma se ha una PDF della forma:

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0, \lambda > 0, \alpha > 0.$$

Si dice $X \sim$ `gamma(shape=alpha, rate=lambda)`. In particolare, se $\alpha = 1$ la distribuzione coincide con `exp(rate=lambda)`.

La media della distribuzione è $\mu = \alpha/\lambda$ e la varianza $\sigma^2 = \alpha/\lambda^2$.

Considerando la relazione con la distribuzione di Poisson, se la distribuzione esponenziale misura l'intervallo di tempo prima che accada il primo evento in una distribuzione di Poisson, la distribuzione gamma misura l'intervallo di tempo prima che accada l'evento α-esimo in una distribuzione di Poisson. Nell'esempio delle auto, se Y è la variabile aleatoria che misura l'intervallo di tempo prima che la quinta auto arrivi (sempre con una media di 25 auto all'ora), allora Y `gamma(shape=5, rate=25)`.

Le funzioni Rassociate sono: `dgamma`, `pgamma`, `qgamma`, `rgamma`.

5.3.5 Distribuzione Chi quadro

Una variabile aleatoria continua X ha una *distribuzione chi-quadro,* χ^2, *con p gradi di libertà* se ha una PDF della forma:

$$f_X(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} e^{-x/2}, \quad x > 0, \quad p > 0,$$

dove

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

Distribuzione chi quadro

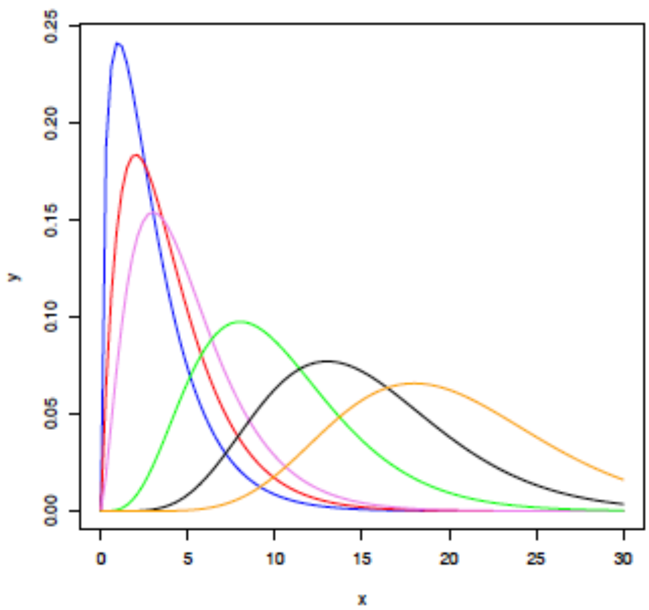
↳ è una distribuzione di probabilità continua utilizzata in statistica per verificare ipotesi, stimare varianze e confrontare le frequenze osservate con quelle stesse.

Si dice che $X \sim$ `chisq(df=p)`.

Le funzioni R associate sono: `dchisq`, `pchisq`, `qchisq`, `rchisq`.

Al variare di p la figura mostra le diverse distribuzioni χ^2.

```
> curve(dchisq(x, df = 3), from = 0, to = 30,
+       col = "blue", ylab="y")
> pind <- c(4, 5, 10, 15, 20)
> pcol <- c("red", "violet", "green", "black", "orange")
> j <- 1
> for (i in pind)
+ {
+   curve(dchisq(x, df=i), from = 0, to = 30,
+         add = TRUE, col = pcol[j])
+   j <- j + 1
+ }
```



χ^2 al variare di p

Distribuzione di t-Student

↳ e' una distribuzione di probabilita' continua che governa il rapporto tra due variabili aleatorie, la prima con distribuzione normale standard e la seconda al quadrato, segue una distribuzione Chi quadrato.

Alcune proprietà:

1. Se la variabile aleatoria X ha distribuzione normale standard (`norm(mean=0,sd01)`) allora vale: $X^2 \sim$ `chisq(df=1)`.
2. La distribuzione chi-quadro è una distribuzione con asimmetria a destra.

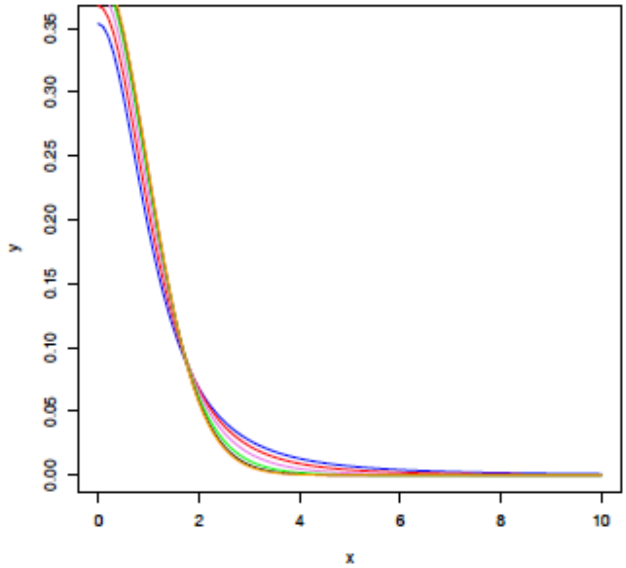
Una variabile aleatoria X ha *distribuzione t di Student con r gradi di libertà* se ha una PDF della forma:

$$f_X(x) = \frac{\Gamma[(r+1)/2]}{\sqrt{r\pi}\,\Gamma(r/2)} \left(1 + \frac{x^2}{r}\right)^{-(r+1)/2}, \quad x \in R.$$

Si dice che $X \sim$ t(df=r).

Le funzioni R associate sono dt, pt, qt, rt.

```
> curve(dt(x, df = 2), from = 0, to = 10,
+        col="blue", ylab="y")
> pind <- c(3, 5, 10, 20, 30)
> pcol <- c("red", "violet", "green", "black", "orange")
> j <- 1
> for (i in pind)
+ {
+   curve(dt(x, df = i), from = 0, to = 10,
+         add = TRUE, col = pcol[j])
+   j <- j + 1
+ }
```



t di student al variare di r