Distribuzioni di probabilità

Saverio Vicario – CNR-ITB

Distribuzioni di probabilità note

- Per diversi casi particolari di esperimenti e osservazioni gli statistici hanno definito le probabilità che gli elementi della variabile aleatoria tratta dal campione prendano i diversi possibili valori
- Andremo a vedere come si possono interrogare queste distribuzioni di probabilità note attraverso le funzioni di ripartizione e le funzioni di densità di probabilità che si applicano alle variabili quantitative

Eventi su Variabili aleatorie

- Useremo lettere maiuscole X per indicare variabili aleatorie (v.a.) e lettere minuscole x per indicare i valori che una v.a. può assumere.
- Si può definire un Evento di interesse tutte le volte che dei valori di X siano più grandi di x_a e più piccoli di $x_{b.}$ Questo può essere descritto con la seguente notazione $\{x_a<X<x_b\}$
- Nel caso che l'Evento di interesse sia l'identità con x_a la notazione e' la seguente $\{X=x_a\}$

Probabilità di Eventi su V.a.

- Dato un evento vorremmo stabilire una probabilità connessa. Per esempio $P\{X \leq x_a\}$ per la probabilità di osservare nella variabile aleatoria X l'Evento minore uguale al valore x_a.
- Data la variabile X la suddetta probabilità e' una funzione del valore di x_a.

- La funzione che da il valore della probabilità per ogni x e' la funzione di ripartizione (o funzione di distribuzione cumulativa o funzione di distribuzione)

Funzione di ripartizione

- La funzione di ripartizione (c.d.f.) si definisce come:
  $F(x)=P\{X\leq x\}$
  Con le seguenti proprietà
  1. $0 \leq F(x) \leq 1$, *per ogni* $x \in R$
  2. *F(x) è una funzione non decrescente di x*
  3. *F(x)tende a 0 per* $x \rightarrow -\infty$ *e tende a 1 per* $x \rightarrow +\infty$
  4. $P\{x1<X \leq x2\} = F(x2)-F(x1)$.

  Es. $X=\{2,2,3,4,7,9\}$ *se* $x1=2.4$ *e* $x2=5.3$ allora
  $F(x1)=2/6=0.333$ e $F(x2)=4/6=0.666$
  $P\{x1<X \leq x2\}=0.666-0.333=0.333$ *e effettivamente*
  Elementi 3,4 sono tra x1 e x2 e dunque $P\{x1<X \leq x2\}=2/6$

Funzione di densità di probabilità

- La *funzione densità di probabilità (p.d.f.) f(x) della v.a. X è la derivata di F(x):*

$$f(x) = \frac{dF(x)}{dx}$$

- *f(x) dunque rappresenta il tasso di aumento della probabilità cumulativa nel punto x*

Funzione di densità di probabilità

- Con le seguenti proprietà

$$f(x) \geq 0, x \in R$$

$$F(x) = \int_{-\infty}^{x} f(\alpha) d\alpha$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$P\{x_1 \leq X \leq x_2\} = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x) dx$$

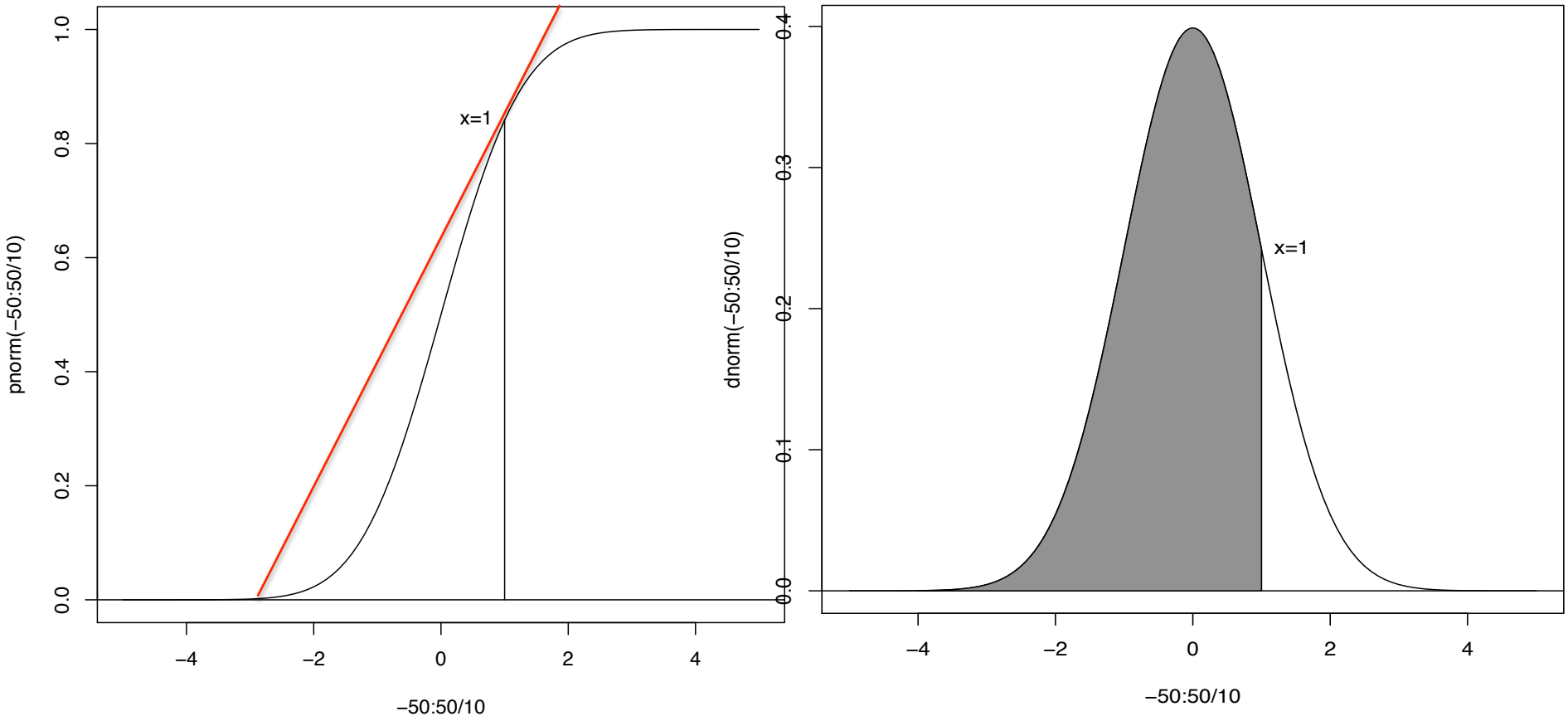
Interpretazione grafica

- la funzione dnorm e' la p.d.f della distribuzione nota gaussiana o normale mentre pnorm e' il c.d.f

```
>plot(-50:50/10,dnorm(-50:50/10), type='l')
> abline(h=0)
>polygon(c(-80/20,-80:20/20,1),c(0,dnorm(-80:20/20),0),col=gray(0.5))
> text(1,dnorm(1),label='x=1', pos=4)
> plot(-50:50/10,pnorm(-50:50/10), type='l')
> abline(h=0)
>segments(1,0,1,pnorm(1))
>text(1,pnorm(1),label='x=1', pos=2)
```

Interpretazione grafica

L'altezza del F(x) e' uguale a l'area in grigio del f(x)
Mentre il coefficiente angolare nel punto x di F(x) e' uguale a f(x)



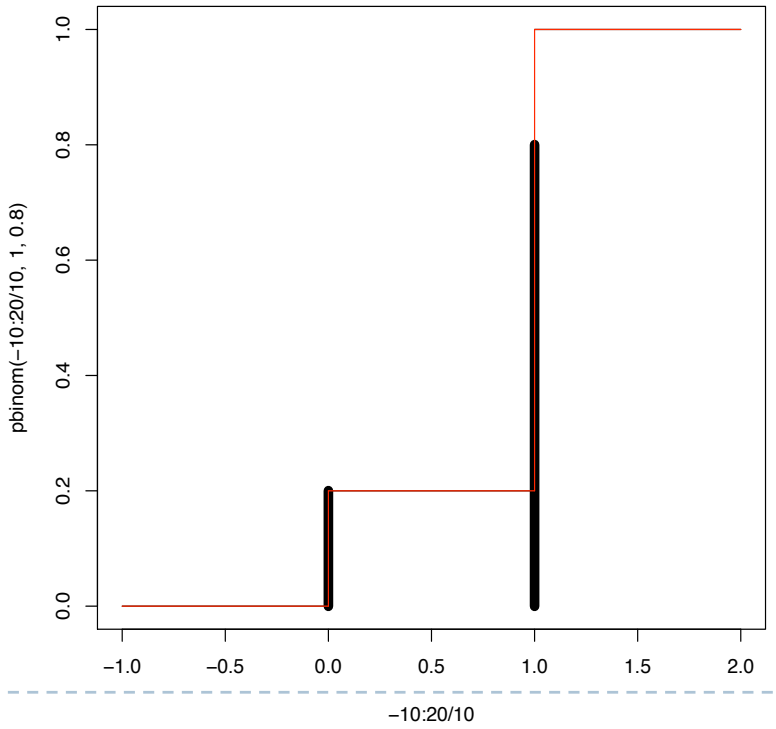
Applicazioni per V.A. discrete

- Nel caso di una v.a. *X discreta, la f(x)è una somma di impulsi,* e *i valori sono direttamente probabilità non densità di probabilità*
- *mentre la F(x)è una funzione a scala (costante a tratti)*
- Nel caso di un campione da una distribuzione non nota si può calcolare la funzione di ripartizione empirica che tratta il campione come fosse tratto da una distribuzione discreta.

```
>plot(ecdf(InsectSprays[,1]))
```

Distribuzione di Bernoulli

- Una v.a. *X è distribuita secondo la legge di Bernoulli B (1,p) quando essa assume due possibili valori: 1 con probabilitàpe 0 con probabilità1-p, con 0 ≤p ≤1.*
- *Quindi: P{X=1}=p e P{X=0}=1-p*



In rosso la funzione di ripartizione o c.d.f.
In nero la funzione di densità di probabilità.

Distribuzione Binomiale

- In un esperimento l'evento *A si verifica con probabilità p:* $P(A) = p$ e $P(\bar{A}) = q$, $p + q = 1$.
- *Vogliamo determinare la probabilità* $p_n(k)$ *che in n ripetizioni indipendenti dell'esperimento di tipo Bernoulli, l'evento A si verifichi k volte in qualsiasi ordine.*

Problema dell'ordine degli eventi

- n ripetizioni di un esperimento possono essere rappresentate con una stringa binaria di lunghezza n (es. 00110010) con un 1 nella i-esima posizione se A si è verificato nella i-esima ripetizione e 0 altrimenti. Dobbiamo quindi contare il numero di stringhe di lunghezza n con k '1'.
- A tale scopo, estraiamo a caso k posizioni della stringa ed a queste assegniamo 1, alle altre 0.
- La 1° posso estrarla in n modi diversi;
- la 2° posso estrarla in n-1 modi diversi;
- …
- la k° posso estrarla in n-k+1 modi diversi. Quindi n(n-1)(n-2) · · · (n-k+1) è il numero di possibili modi di estrarre k posizioni dalle n disponibili.

Problema dell'ordine degli eventi

- Poiché non siamo interessati all'ordine con cui estraiamo le posizioni, allora questo numero deve essere diviso per il numero di modi diversi in cui possiamo disporre i *k oggetti estratti in k posizioni*

$$\frac{n(n-1)(n-2)\ldots(n-k+1)}{k(k-1)\ldots 1} =$$

$$\frac{n!}{k!(n-k)!} = \binom{n}{k} = {}_nC_k$$

Probabilità di eventi Binomiali

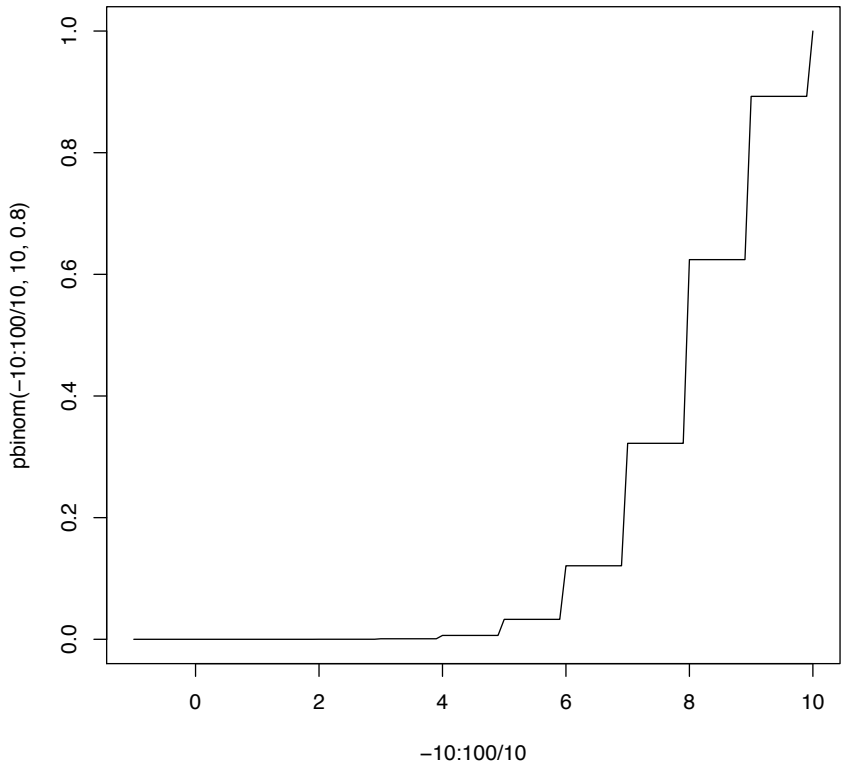
- Il coefficiente binomiale ${}_nC_k$ *conta il numero di modi diversi in cui l'evento A può verificarsi in n ripetizioni di un esperimento. Poiché le prove sono indipendenti allora:*
- *P{A si verifica k volte in un ordine specifico}=* $p^k q^{n-k}$
- *Inoltre l'evento { A si verifica k volte in qualunque ordine} è l'unione di* ${}_nC_k$ *eventi { A si verifica k volte in un ordine specifico} e questi sono mutuamente esclusivi.*
- *Allora:*

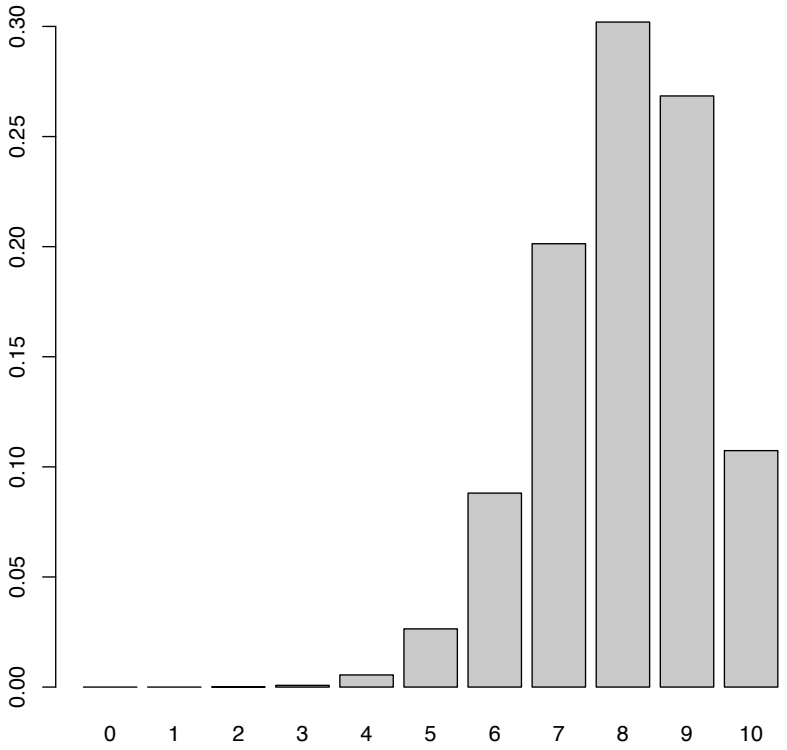
$$p_n(k) = \binom{n}{k} p^k q^{n-k}$$

- >*dbinom(k,n) # probability*
- >choose(n,k) # binomial coefficient

c.d.f e p.d.f per v.a. binomiale n=10, k=0,1,...10, p=0.8

```
>plot(-10:80/10, pbinom(-10:80/10,10,0.8), type ='l')
>barplot(dbinom(0:10,10,0.8), names.arg=0:10)
```





Esempio

- Esempio Il 30% di una certa popolazione è immune da una malattia. Se si estrae un campione di dimensione 10 da questa popolazione, qual è la probabilità che esso contenga esattamente 4 persone immuni?
- Soluzione: p=0.3, n=10, k=4

```
> dbinom(4,10,0.3)
[1] 0.2001209
```

- E la probabilità di avere almeno 4 persone immuni ?
- La funzione di ripartizione per k=3 identifica l'evento opposto dunque 1-F(3)
- >1-pbinom(3,10,0.3)
- > 0.3503893

Distribuzione di Poisson

- Una v.a. *X segue la legge di Poisson P(λ) con λ>0 quando essa assume tutti i valori interi k ∈ N con le seguenti probabilità:*

$$p_k = P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$$

- se *n è molto grande e p è molto piccola allora una legge Binomiale B(n, p) è ben approssimata dalla legge di Poisson con parametro λ=np.*
- La legge di Poisson è particolarmente adatta a descrivere v.a. che rappresentano conteggi e che possono assumere un numero illimitato di valori: numero di telefonate che arrivano ad un centralino in un dato periodo di tempo; numero di clienti che si presentano allo sportello di un ufficio durante una giornata, ecc.

Alcune caratteristiche

- media: λ
- varianza: λ
- simmetria: $\lambda^{-1/2}$
- curtosi: λ^{-1}

Che cosa rappresenta la distribuzione di poisson?

- Rappresenta un approssimazione della distribuzione binomiale quando e' difficile calcolare le probabilità degli eventi di successo perche troppo numerosi o indeterminati

Convergenza di Poisson al variabile di k e n

- > dbinom(3,10,0.3)
- [1] 0.2668279
- > dpois(3,10*0.3)
- [1] 0.2240418
- > dbinom(3,10,0.03)
- [1] 0.002617864
- > dpois(3,0.3)
- [1] 0.003333682
- > dbinom(3,1000,0.03)
- [1] 2.906214e-10
- > dpois(3,0.03*1000)
- [1] 4.21093e-10
- > dbinom(3,1000,0.003)
- [1] 0.2243786
- > dpois(3,0.003*1000)
- [1] 0.2240418

Esempio: shotgun sequencing

- Supponiamo di voler sequenziare un genoma *S costituito da g basi. Quindi possiamo pensare ad S come una stringa di lunghezza g di lettere dall'alfabeto {A,C,G,T}. Metodi di sequenziamento chimico possono essere applicati a filamenti relativamente corti di DNA, da 500 a 2000 basi. Shotgun sequencing è una strategia per sequenziare genomi di lunghezza g dell'ordine di 10^5, 10^6 basi. Numerose copie di S vengono rotte a caso in frammenti di lunghezza L. N di questi frammenti vengono scelti a caso e sequenziati, dove N è un parametro scelto dallo sperimentatore. Due frammenti che si sovrappongono contengono una sottosequenza di DNA in comune, e tale sovrapposizione è utilizzata per ricostruire S.*

Esempio: shotgun sequencing

- Supponiamo di aver accesso ad un numero elevato N di frammenti di lunghezza L(L<<g) del genoma tale che:
- NL=cg dove c è detto valore di copertura.
- Vogliamo la probabilità di leggere tutti i siti del genoma assumendo un certo coefficiente di copertura c.
- Se un certo frammento i contiene un certo sito x segue una legge di Bernoulli B(1,L/g)
- *Sulla somma di tutti gli N frammenti l'esperimento segue una legge binomiale B(N,* L/g)

Esempio: shotgun sequencing

- Vogliamo la probabilità di leggere tutti i siti del genoma assumendo un certo coefficiente di copertura c.

$$p = 1 - \left(\frac{N}{0}\right)\left(1 - \frac{L}{g}\right)^{N}$$

- Se L<<g e N grande allora puo essere approssimato da una distribuzione di poisson
- $P=1-e^{NL/g}=1-e^{c}$
- Se N=100000 L=500 e $g=10^7$
- >1-dbinom(0,100000,500/10^7)
- 0.9932629
- >1-dpois(0,5)
- 0.993262

V.A. continue

- Una v.a. *X si dice continua quando può assumere tutti i valori in un intervallo di numeri reali, o in tutto R.*
- *Osservazione. Se X è una v.a. continua con p.d.f. f(x), allora f(x) NON è la probabilità che X assuma valore x, in quanto P{X=x}=0. Inoltre f(x) può assumere valori maggiori di 1.*

Distribuzione uniforme

- E' molto semplice e ne parlo solo per ragioni didattiche
- La distribuzione uniforme descrive una V.A. che può assumere con uguale probabilità valori all'interno di un intervallo prefissato

```
>plot(-10:110/100, dunif(-10:110/100, max=0.8, min=0.2), type='l')
```

- lines(-10:110/100, dunif(-10:110/100, max=1, min=0), col=2)
- Da questo esempio risulta chiaro che dovendo la p.d.f avere un area totale uguale ad 1 se la base, l'intervallo, e' minore di 1, l'altezza, la densità di probabilità, dovrà essere maggiore di 1.
- La densità sarà alta $\frac{1}{b-a}$
- Con b=max
- a=min

Distribuzione Gaussiana o normale

- $N(\mu, \sigma^2)$
- μ coincide con la media della distribuzione
- σ^2 coincide con la varianza della distribuzione

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

- dnorm(x,mean= μ ,sd= σ)=f(x) c.d.f o funzione di densita di probabilita'
- pnorm(x,mean= μ ,sd= σ)=F(x) p.d.f. o funzione di ripartizione

Distribuzione Gaussiana o normale

1)f(x)>0;

2)f(x) è una curva a campana simmetrica intorno a $x=\mu$;

3)la larghezza della campana è regolata dal valore di σ;

4)f(x)ha due punti di flesso in $x=\mu\pm\sigma$;

5)F(X) ha una forma di S allungata con un flesso in $x=\mu$;

6)F(X) diventa sempre più ripida per $\sigma\rightarrow 0$.

Provate con

```
plot(-110:110/20, dnorm(-110:110/20),type='l')
plot(-110:110/20, pnorm(-110:110/20),type='l')
```

Variate mean (μ) e sd (σ) *nei grafici e osservate i cambiamenti*

Aree di probabilità nella distribuzione normale

$$P\{\mu - \sigma < X < \mu + \sigma\} = F(\mu + \sigma) - F(\mu + \sigma) = 0.68$$

$$P\{\mu - 2\sigma < X < \mu + 2\sigma\} = F(\mu + 2\sigma) - F(\mu + 2\sigma) = 0.95$$

$$P\{\mu - 3\sigma < X < \mu + 3\sigma\} = F(\mu + 3\sigma) - F(\mu + 3\sigma) = 0.997$$

```
>mean=3
>sd=4
>pnorm(mean+sd,mean=mean,sd=sd)-pnorm(mean-sd,mean=mean,sd=sd)
```

La distribuzione normale standard *N(0, 1)*

- Stadardizando la media e' la varianza di una distribuzione normale e' possibile ottenere una distribuzione normale standard. Questo consente di paragonare diverse distribuzione normali fra di loro

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

Campione standardizzato (dalla lezione3)

- E' un campione con $M_x=0$ e $s^2_x=1$
- Qualunque campione può essere standardizzato sottraendo la media e dividendo per la deviazione standard
- Più formalmente applicando la trasformazione lineare
- $y_i=ax_i+b$ *con* $a=1/s_x$ e $b=-M_x/s_x$

```
a=1/sd(Q)
> b=-mean(Q)/sd(Q)
> y=a*Q+b
> mean(y)
[1] -5.692927e-17    # molto prossimo a zero
> var(y)
[1] 1
```

Vantaggi della standardizzazione

- In caso di mancanza di un programma come R le probabilità possono essere lette da apposite tavole riferite alla distribuzione N(0,1) a cui una qualunque distribuzione può essere riportata con la procedura di standardizzazione

- Se Y e' di tipo N(μ , σ^2) allora la v.a. X=(Y- μ)/ σ
- Viceversa, se X e' di tipo N(0, 1), allora la v.a. Y = σ X + μ è di tipo è N(0, 1).

Esempio

- La statura di una certa popolazione di individui è approssimativamente normale, X: N(170cm, 100 cm^2). Qual è la probabilità che una persona estratta a caso da questa popolazione sia alta tra i 160 e 171 cm ?
- >mean=170;sd=100^0.5
- >pnorm(171,mean,sd)-pnorm(160,mean,sd)

Legge dei grandi numeri e Teorema del Limite Centrale

Saverio Vicario – CNR-ITB

Introduzione

- Per collegare quello che abbiamo visto sui i campioni e le v.a. di un singolo campione con le distribuzione teoriche appena illustrate c'e' bisogno di chiarire come valutare i parametri delle distribuzione teoriche sui campioni. O meglio come collegare i valori osservati in uno o più campioni con quelli delle popolazione oggetto dello studio ma non direttamente osservata nella sua totalità

Considerazioni

- Supponiamo di avere una popolazione costituita da individui di tipo A e di tipo B. Sia p la percentuale degli individui di tipo A e 1-p la percentuale degli individui di tipo B, e supponiamo che p sia incognito. Vogliamo determinare p.
- Se la popolazione è piccola allora basta contare il numero di individui di tipo A e dividerlo per il numero totale di individui.
- Se la popolazione è grande, o non interamente accessibile, allora possiamo solo ottenere una stima di p. A tale scopo estraiamo un campione casuale di dimensione n dall'intera popolazione e contiamo il numero N_A di individui di tipo A tra gli n estratti. Allora una stima $\overline{p}$ di p è: N_A/n
- e questa stima migliora al crescere di n.

Considerazioni

- Quale differenza tra $\bar{p}$ e p?
- p e' un numero fisso specifico della popolazione. Mentre la stima $\bar{p}$ e' una variabile aleatorio che varia a seconda del campione estratto

- E' chiaro che N_A e' una variabile aleatoria che segue una legge binomiale di tipo B(n,p) in cui n e' la dimensione del campione mentre ogni singola unità x che appartiene (1) o non appartiene (0) al tipo A segue una legge di Bernoulli B(1,p)
- La media dei valori di 0 e 1 dei singole unita e' dunque il nostro stimatore intuitivo

Considerazioni

- La media dei valori di 0 e 1 dei singole unita e' dunque il nostro stimatore intuitivo
- E' logico pensare che anche la stima della varianza del campione ci dia delle informazioni sulla varianza della popolazione
- Quanto sono realistiche queste stime
- Come varia la qualità della stima al variare di n

Teorema del Limite Centrale

- La media *di un campione casuale estratto da una qualsiasi distribuzione con varianza σ^2 finita* e *media* μ è *approssimativamente distribuita come* $N(\mu, \sigma^2/n)$ *con un accuratezza che cresce al crescere di n*

Dimostrazione numerica

- Ponendo n=10 e scegliendo un distribuzione di poisson con media 10 e varianza 10
- res=c()
- for(x in 1:100){res=c(res,mean(rpois(10,10)))}
- qqnorm(res)
- mean(res)
- var(res)
-