STATISTICA E ANALISI DEI DATI

Distribuzione Geometrica

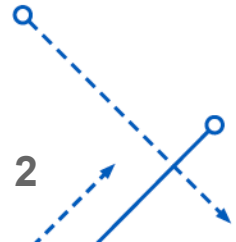
Distribuzione Geometrica

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$ dove ogni prova ha una probabilità di successo pari a p (e quindi di fallimento pari a 1−p)

$$F_r = \{\text{il numero di fallimenti che precedono il primo successo è } r\} \quad (r = 0, 1, \ldots)$$

La distribuzione geometrica è utilizzata, ad esempio, per modellare:

- numero di ritrasmissioni di un messaggio in un sistema informatico;
- numero di fallimenti di uno sportivo per riuscire a completare un percorso senza alcun incidente;
- numero di fallimenti ad una prova di esame prima di superarla;
- numero di tentativi falliti con le chiavi da un ubriaco prima che riesca ad aprire la porta di casa.



Distribuzione Geometrica

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

  F_r = {il numero di fallimenti che precedono il primo successo è r} $\quad (r = 0, 1, \ldots)$

  F_r = {il primo successo avviene dopo r fallimenti} $\quad (r = 0, 1, \ldots)$

- La probabilità di avere un successo all'r-esimo tentativo è data da:

$$P(F_r) = (1-p)^r p$$

che si ottiene dall'ipotesi di **indipendenza delle prove**

Distribuzione Geometrica

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

  F_r = {il numero di fallimenti che precedono il primo successo è r} $(r = 0, 1, \ldots)$

  F_r = {il primo successo avviene dopo r fallimenti} $(r = 0, 1, \ldots)$

- La probabilità di avere un successo all'r-esimo tentativo è data da:

$$P(F_r) = (1-p)^r p$$

che si ottiene dall'ipotesi di **indipendenza delle prove**

Questo valore rappresenta la probabilità che i primi r tentativi siano tutti fallimenti, ognuno con probabilità di fallimento pari a $1 - p$

Questo valore rappresenta la probabilità che il primo successo avvenga immediatamente dopo r fallimenti

Distribuzione Geometrica

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

$$F_r = \{\text{il numero di fallimenti che precedono il primo successo è } r\} \quad (r = 0,1,\dots)$$

- La probabilità di avere un successo all'r-esimo tentativo è data da:

$$P(F_r) = (1-p)^r p$$

che si ottiene dall'ipotesi di **indipendenza delle prove**

- Sia Y la variabile aleatoria che descrive il numero di fallimenti che precedono il primo successo

Funzione di probabilità: $F_Y(y) = P(Y = x) = \begin{cases} (1-p)^y p & y = 0,1,\dots \\ 0 & altrimenti \end{cases}$

con $0 < p < 1$ si dice avere distribuzione geometrica di parametro p

Distribuzione Geometrica

- La funzione di distribuzione della variabile aleatoria geometrica Y è la seguente:
- **Funzione Distribuzione**:

$$F_Y(r) = P(Y \leq r) = \sum_{k=0}^{r} P(Y = k) = \sum_{k=0}^{r} (1-p)^k p = p \cdot \frac{1-(1-p)^{r+1}}{1-(1-p)} = 1-(1-p)^{r+1}, \quad r = 0,1,2,\dots$$

- Da cui:

$$F_Y(y) = P(Y \leq y) = \begin{cases} 1-(1-p)^{k+1} & k \leq y \leq k+1 \quad (k = 0,1,\dots) \\ 0 & y < 0 \end{cases}$$

- **Valore atteso**: $E[Y] = \frac{1-p}{p}$ **Varianza**: $Var(Y) = \frac{1-p}{p^2}$

Esempio - Industria

- Un'azienda produce componenti elettronici con una probabilità $p = 0.9$ che un componente sia funzionante
- Si vuole calcolare il numero di pezzi difettosi (R) che vengono prodotti prima di ottenere il primo componente funzionante
  - La variabile geometrica R ha la seguente **funzione di probabilità**:

$$P(R = r) = (1 - p)^r p$$

  - La probabilità che $R = 0$, cioè che ci sia subito un componente:

$$P(R = 0) = (1 - p)^0 p = 0.1^0 * 0.9 = 0.9$$

  - La probabilità che $R = 1$, cioè che il primo componente sia difettoso e il secondo sia funzionante:

$$P(R = 1) = (1 - p)^1 p = 0.1^1 * 0.9 = 0.99$$

Esempio - Industria

- La probabilità che $R = 2$, cioè che i primi 2 componenti siano difettosi e il terzo sia funzionante:

$$P(R = 2) = (1 - p)^2 p = 0.1^2 * 0.9 = 0.009$$

- La probabilità che $R = 3$, cioè che i primi 3 componenti siano difettosi e il quarto sia funzionante:

$$P(R = 3) = (1 - p)^3 p = 0.1^3 * 0.9 = 0.0009$$

- La probabilità che ci vogliano **al massimo 3 tentativi** per ottenere un componente funzionante:

$$P(R \leq 3) = P(R = 0) + P(R = 1) + P(R = 2) + P(R = 3) =$$

$$= 0.9 + (0.1 * 0.9) + (0.1^2 * 0.9) + (0.1^3 * 0.9) = 0.9 + 0.09 + 0.009 + 0.0009 = 0.999$$

Esempio - Industria

- La probabilità che $R = 2$, cioè che i primi 2 componenti siano difettosi e il terzo sia funzionante:

$$P(R = 2) = (1 - p)^2 p = 0.1^2 * 0.9 = 0.009$$

- La probabilità che $R = 3$, cioè che i primi 3 componenti siano difettosi e il quarto sia funzionante:

$$P(R = 3) = (1 - p)^3 p = 0.1^3 * 0.9 = 0.0009$$

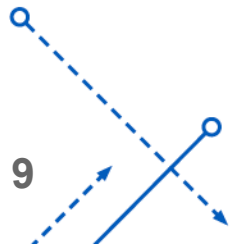
- La probabilità che ci vogliano **al massimo 3 tentativi** per ottenere un componente funzionante:

$$P(R \leq 3) = P(R = 0) + P(R = 1) + P(R = 2) + P(R = 3) =$$

$$= 0.9 + (0.1 * 0.9) + (0.1^2 * 0.9) + (0.1^3 * 0.9) = 0.9 + 0.09 + 0.009 + 0.0009 = 0.999$$

- **Valore atteso** $E[R] = \frac{1-p}{p} = \frac{0.1}{0.9} = 0.11$ → Significa che, in media, ci aspettiamo di produrre 0.11 pezzi difettosi prima di ottenere il primo funzionante

- **Varianza** $Var(R) = \frac{1-p}{p^2} = \frac{0.1}{0.81} = 0.123$ → La varianza rappresenta la variabilità attesa nel numero di pezzi difettosi.



Esempio – Software Testing

- Un team di sviluppatori sta eseguendo dei test su un software per rilevare difetti. Si stima che la probabilità di individuare un difetto in un singolo test sia $p = 0.3$
  - Quindi $1 - p = 0.7$ è la probabilità che un test non individui difetti
- La variabile R rappresenta il numero di test senza difetti prima di individuare il primo difetto
  - La variabile geometrica R ha la seguente **funzione di probabilità**:

$$P(R = r) = (1 - p)^r p$$

  - Probabilità di individuare il primo difetto al quarto test ($R = 3$):

$$P(R = 3) = (1 - p)^3 p = 0.7^3 * 0.3 = 0.343 * 0.3 = 0.1029$$

Significa che c'è circa il 10.29% di probabilità che il primo difetto **venga rilevato al quarto test**

Esempio – Software Testing

- Probabilità di individuare almeno un difetto nei primi 3 test ($R \leq 3$):

$$P(R \leq 2) = P(R = 0) + P(R = 1) + P(R = 2) =$$

$$= 0.3 + (0.7 * 0.3) + (0.7^2 * 0.3) = 0.3 + 0.21 + 0.147 = 0.657$$

  - C'è quindi una probabilità del 65.7% di rilevare almeno un difetto nei primi 3 test

- **Valore atteso** $E[R] = \frac{1-p}{p} = \frac{0.7}{0.3} = 2.33$ → Significa che, in media, ci aspettiamo di effettuare circa 2.33 test senza trovare difetti prima di rilevarne uno

- **Varianza** $Var(R) = \frac{1-p}{p^2} = \frac{0.7}{0.09} = 7.78$ → La varianza è moderatamente alta, indicando una certa variabilità nel numero di test necessari per trovare il primo difetto

Distribuzione Geometrica (R)

- Per il calcolo in R delle probabilità di una variabile aleatoria geometrica si utilizza la funzione:

dgeom(x, prob)

dove

  - **x** è il valore assunto (o i valori assunti) dalla variabile aleatoria geometrica considerata;
  - **prob** è la probabilità di successo in ciascuna prova

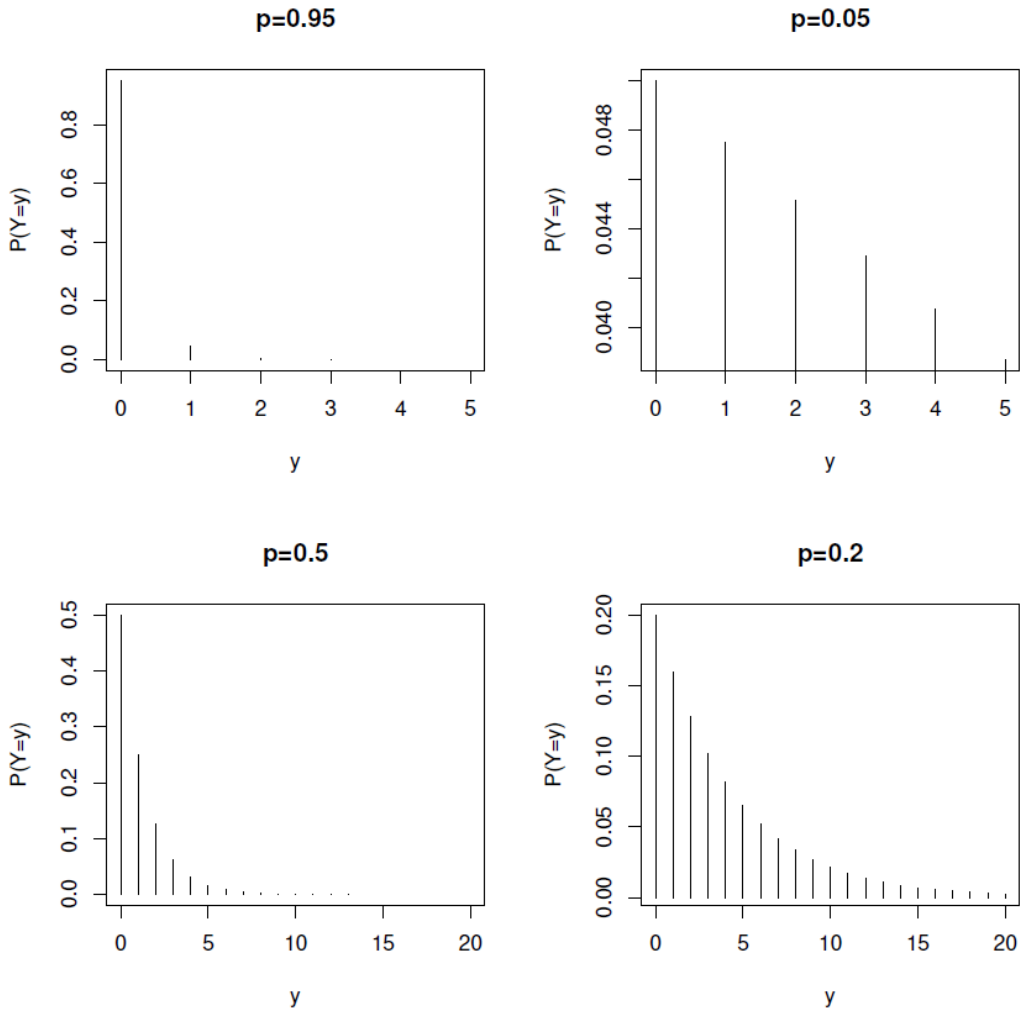
- Ad esempio, se $p = 0.95$ le probabilità di una variabile aleatoria geometrica si calcolano:

```
> x<-0:5
> dgeom(x,prob=0.95)
[1] 9.50000e-01 4.75000e-02 2.37500e-03 1.18750e-04 5.93750e-06
[6] 2.96875e-07
```

Distribuzione Geometrica (R)

- Variando il parametro *p* (probabilità) nella funzione in R possiamo visualizzare le **funzioni di probabilità** di una variabile aleatoria geometrica

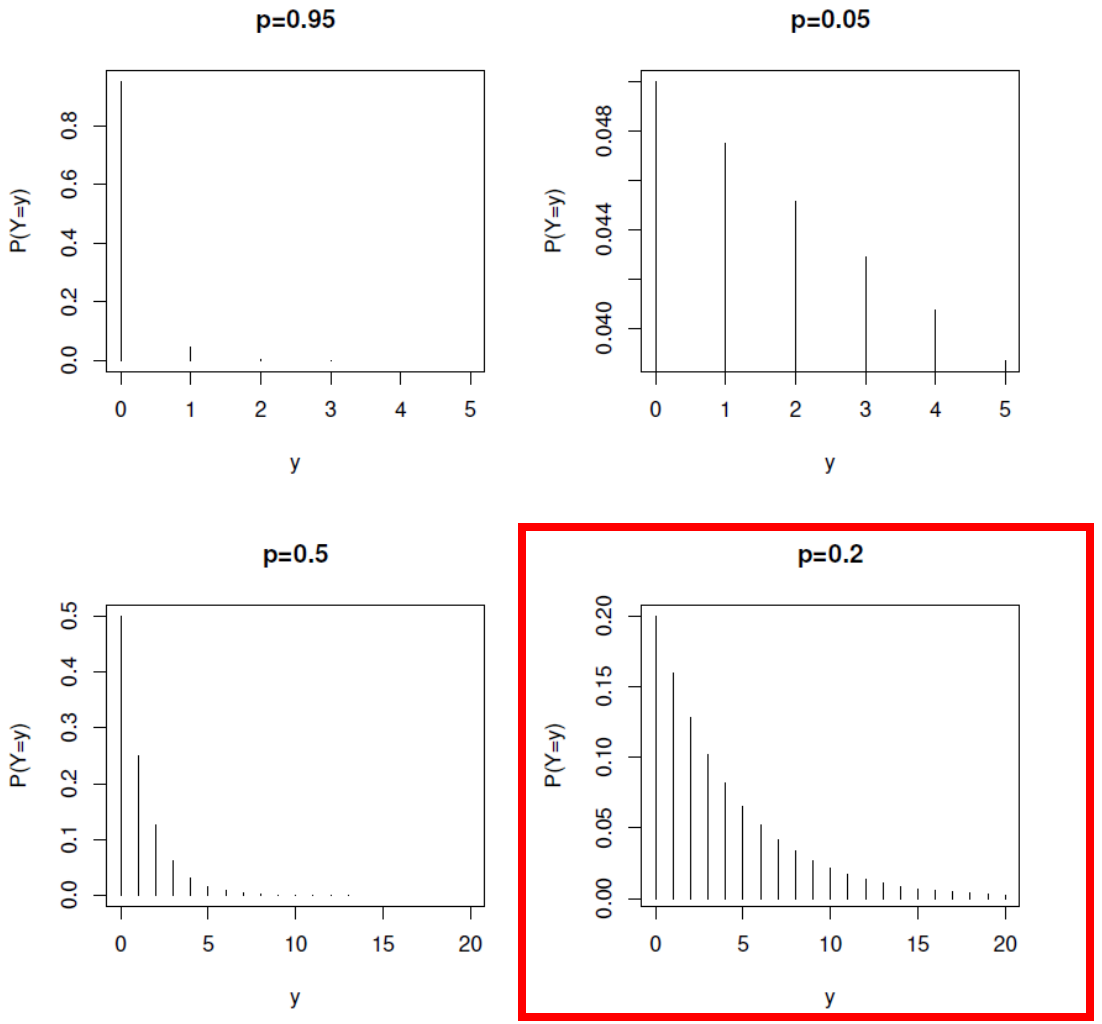
```
> par(mfrow=c(2,2))
> y<-0:5
> plot(y,dgeom(y,prob=0.95),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.95")
>
> y<-0:10
> plot(y,dgeom(y,prob=0.05),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.05")
>
> y<-0:20
> plot(y,dgeom(y,prob=0.5),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.5")
>
> y<-0:20
> plot(y,dgeom(y,prob=0.2),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.2")
```



Distribuzione Geometrica (R)

- Variando il parametro *p* (probabilità) nella funzione in R possiamo visualizzare le **funzioni di probabilità** di una variabile aleatoria geometrica

```
> par(mfrow=c(2,2))
> y<-0:5
> plot(y,dgeom(y,prob=0.95),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.95")
>
> y<-0:10
> plot(y,dgeom(y,prob=0.05),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.05")
>
> y<-0:20
> plot(y,dgeom(y,prob=0.5),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.5")
>
> y<-0:20
> plot(y,dgeom(y,prob=0.2),
+ xlab="y",ylab="P(Y=y)",type="h",
+ main="p=0.2")
```



Consideriamo $p = 0.2$: $\mathrm{E}(Y) = \frac{(1-p)}{p} = 4.0$ $Var(Y) = \frac{(1-p)}{p^2} = 20$ $CV(Y) = \frac{\sqrt{Var(Y)}}{E(Y)} = \frac{\sqrt{20}}{4} = 1.118034$

Coefficiente di Variazione

- Il **coefficiente di variazione (CV)** è una misura statistica che esprime la **dispersione relativa** di una distribuzione rispetto al suo valore medio
  - È particolarmente utile per confrontare la variabilità di due o più dataset con unità di misura diverse o valori medi molto differenti

$$CV(Y) = \frac{\sqrt{Var(Y)}}{E(\mathrm{Y})}$$

  - Il coefficiente di variazione indica **quanto è grande la deviazione standard rispetto alla media**
  - **CV basso**: i dati sono poco dispersi rispetto alla media (la distribuzione è più concentrata)
  - **CV alto**: i dati sono molto dispersi rispetto alla media (la distribuzione è più varia)
- Ad esempio:
  - Un $CV = 20\%$ indica che la deviazione standard è il 20% della media
  - Un $CV = 100\%$ indica che la deviazione standard è uguale alla media, segnalando alta variabilità

Coefficiente di Variazione nella Geometrica

- Nel caso di una variabile aleatoria geometrica con parametro p:
  - Poiché
    - $E[R] = \frac{1-p}{p}$
    - $Var(R) = \frac{1-p}{p^2}$

$$CV(Y) = \frac{\sqrt{Var(Y)}}{E(Y)} = \frac{\sqrt{\frac{1-p}{p^2}}}{\frac{1-p}{p}} = \frac{\frac{\sqrt{1-p}}{p}}{\frac{1-p}{p}} = \frac{\sqrt{1-p}}{p} * \frac{p}{1-p} = \frac{\sqrt{1-p}}{1-p}$$

$$= \frac{\sqrt{1-p}}{(\sqrt{1-p})^2} = \frac{1}{\sqrt{1-p}}$$

- Esempio con $p = 0.3$:

$$CV(Y) = \frac{1}{\sqrt{1-0.3}} = \frac{1}{0.836} = 1.20$$

Significa che la deviazione standard è circa 1.23 volte il valore medio, segnalando una distribuzione altamente dispersa

Distribuzione Geometrica (R)

- Per il calcolo in R della funzione di distribuzione geometrica si utilizza la funzione:

pgeom(x, prob, lower.tail = TRUE)

dove

- **x** è il valore assunto (o i valori assunti) dalla variabile aleatoria geometrica considerata;
- **prob** è la probabilità di successo in ciascuna prova
- **lower.tail** se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$

- Ad esempio, se p=0.95 le probabilità di una variabile aleatoria geometrica si calcolano:

```
> x<-0:5
> pgeom(x,prob=0.95)
[1] 0.9500000 0.9975000 0.9998750 0.9999938 0.9999997 1.0000000
```

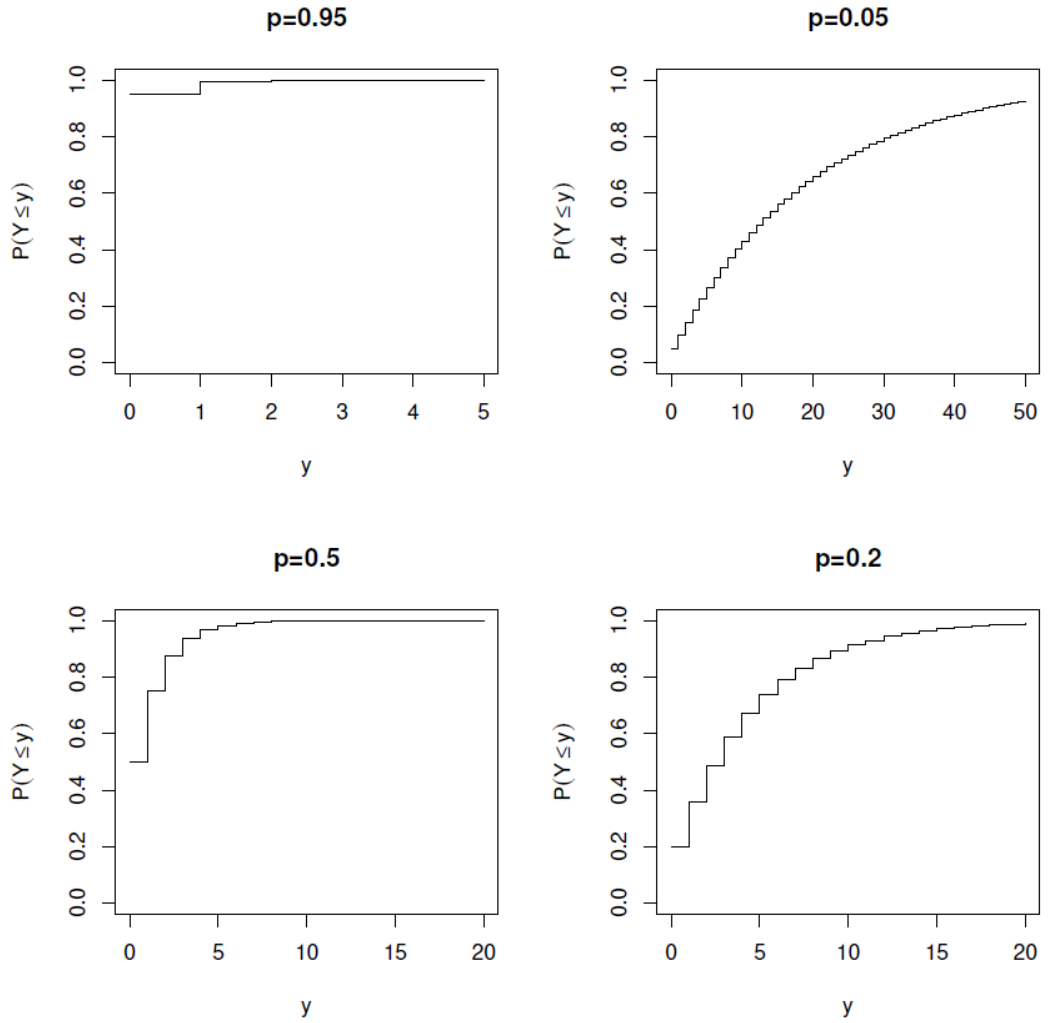
- I cui risultati sono le probabilità:

$$P(X \leq x) = \sum_{n=0}^{x} P(Y = x) \qquad x = 0,1, \dots, 5$$

Distribuzione Geometrica (R)

- Variando il parametro p (probabilità) nella funzione in R possiamo visualizzare le **funzioni di distribuzione** di una variabile aleatoria geometrica

```
> par(mfrow=c(2,2))
> y<-0:5
> plot(y,pgeom(y,prob=0.95),
+ xlab="y",ylab=expression(P(Y<=y)),ylim=c(0,1),type="s",
+ main="p=0.95")
>
> y<-0:50
> plot(y,pgeom(y,prob=0.05),
+ xlab="y",ylab=expression(P(Y<=y)),ylim=c(0,1),type="s",
+ main="p=0.05")
>
> y<-0:20
> plot(y,pgeom(y,prob=0.5),
+ xlab="y",ylab=expression(P(Y<=y)),ylim=c(0,1),type="s",
+ main="p=0.5")
>
> y<-0:20
> plot(y,pgeom(y,prob=0.2),
+ xlab="y",ylab=expression(P(Y<=y)),ylim=c(0,1),type="s",
+ main="p=0.2")
```



Consideriamo $p = 0.2$: $\mathrm{E}(Y) = \frac{(1-p)}{p} = 4.0$ $Var(Y) = \frac{(1-p)}{p^2} = 20$ $CV(Y) = \frac{\sqrt{20}}{4} = 1.118034$

Altri Valori in R

- Per calcolare i **Quantili** della distribuzione geometrica si utilizza la funzione:

  qgeom(z, prob)

  dove

  - **z** è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$-esimo;
  - **prob** è la probabilità di successo in ciascuna prova

- Per una distribuzione geometrica il percentile $z \cdot 100$-esimo è il più piccolo numero intero k assunto dalla variabile aleatoria binomiale Y tale che:

$$P(Y \leq k) = 1 - (1-p)^{k+1} \geq z \qquad k = 0,1,\dots,n$$

Quantili

- Per una distribuzione geometrica il percentile $z \cdot 100$-esimo è il più piccolo numero intero k assunto dalla variabile aleatoria binomiale Y tale che:

$$P(Y \leq k) = 1 - (1-p)^{k+1} \geq z \qquad k = 0,1, \ldots$$

Da cui segue

$$1 - (1-p)^{k+1} \leq 1 - z \qquad k = 0,1, \ldots$$

ossia:

$$(k+1)\log(1-p) \leq \log(1-z) \qquad k = 0,1, \ldots$$

per una distribuzione geometrica il percentile (quantile) $z \cdot 100$-esimo è il più piccolo intero k tale che:

$$k \geq \frac{\log(1-z)}{\log(1-p)} - 1 \qquad k = 0,1, \ldots$$

Quantili

- Esempio: consideriamo $p = 0.2$ avremo:

$$z = 0 \rightarrow k \geq \frac{\log(1-0)}{\log(1-0.2)} - 1 = -1 \rightarrow Q_0 = 0$$

$$z = 0.25 \rightarrow k \geq \frac{\log(1-0.25)}{\log(1-0.2)} - 1 = 0.2892 \rightarrow Q_1 = 1$$

$$z = 0.5 \rightarrow k \geq \frac{\log(1-0.5)}{\log(1-0.2)} - 1 = 2.1063 \rightarrow Q_2 = 3$$

$$z = 0.75 \rightarrow k \geq \frac{\log(1-0.75)}{\log(1-0.2)} - 1 = 5.2126 \rightarrow Q_3 = 6$$

$$z = 1 \rightarrow k \geq \frac{\log(1-1)}{\log(1-0.2)} - 1 = +\infty \rightarrow Q_4 = +\infty$$

Quantili (R)

- Per calcolare i **Quantili** della distribuzione geometrica si utilizza la funzione:

  qgeom(z, prob)

  dove

  - **z** è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$-esimo;
  - **prob** è la probabilità di successo in ciascuna prova

- Esempio: Riprendendo l'esempio precedente, se $p = 0.2$ le seguenti linee di codice forniscono i quartili Q_0, Q_1, Q_2, Q_3, Q_4:

```
>z<-c(0,0.25,0.5,0.75,1)
> qgeom(z,prob=0.2)
[1]   0   1   3   6 Inf
```

Popolazione Pseudocasuale

- È possibile simulare una variabile aleatoria geometrica in R utilizzando una sequenza di numeri pseudocasuali
- La funzione che genera numeri casuali secondo una distribuzione geometrica è:

rgeom(n, prob)

dove

  - **n** il numero di campioni (cioè, il numero di variabili aleatorie geometriche da generare;
  - **prob** è la probabilità di successo in ciascuna prova

- Esempio: Supponiamo di voler simulare una variabile aleatoria geometrica con parametro $p = 0.3$ e di voler generare 1000 campioni

Popolazione Pseudocasuale

- Esempio: Supponiamo di voler simulare una variabile aleatoria geometrica con parametro $p = 0.3$ e di voler generare 1000 campioni

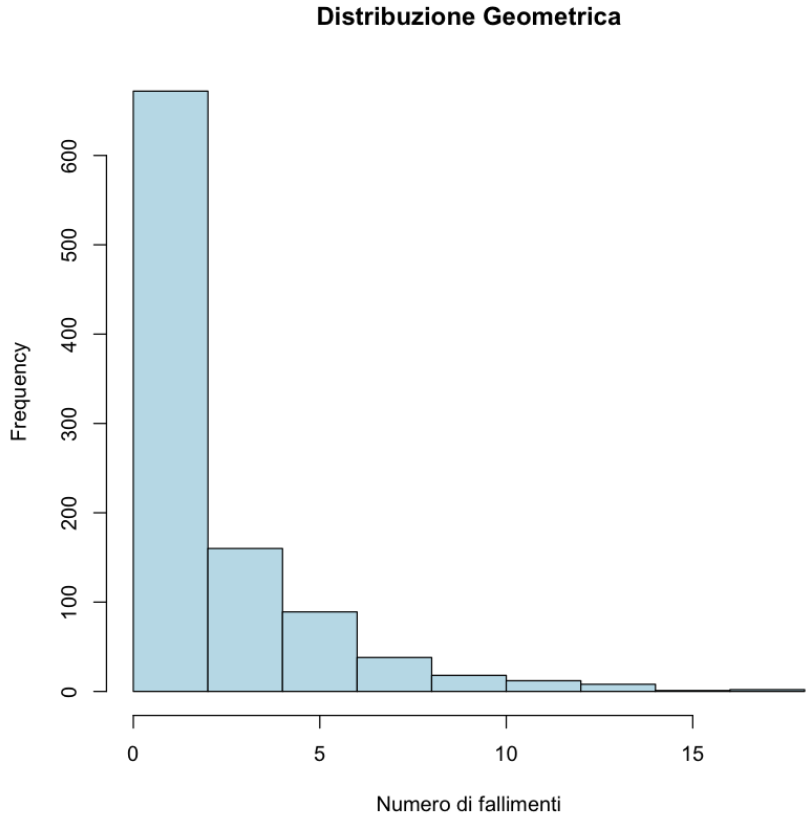
```
sim <- rgeom(1000, prob=0.3)
sim
```

```
3 · 2 · 4 · 2 · 2 · 0 · 9 · 3 · 0 · 2 · 8 · 11 · 0 · 1 · 0 · 2 · 0 · 2 · 0 · 2 · 1 · 5 · 4 · 2 · 0 · 6 · 18 · 3 · 12 · 2 · 0 · (
4 · 0 · 0 · 1 · 0 · 3 · 1 · 5 · 0 · 1 · 0 · 1 · 3 · 2 · 7 · 4 · 0 · 0 · 4 · 0 · 0 · 0 · 5 · 1 · 1 · 4 · 0 · 0 · 3 · 0 · 17 · 0 · '
0 · 0 · 5 · 4 · 1 · 6 · 4 · 1 · 3 · 0 · 2 · 1 · 1 · 1 · 0 · 2 · 1 · 0 · 1 · 0 · 1 · 3 · 5 · 1 · 0 · 2 · 2 · 7 · 12 · 1 · 0 · 7 · 0
0 · 0 · 1 · 0 · 1 · 0 · 0 · 9 · 1 · 0 · 0 · 1 · 4 · 1 · 1 · 3 · 2 · 4 · 0 · 2 · 1 · 1 · 5 · 3 · 2 · 3 · 0 · 0 · 2 · 0 · 0 · 1 · 8 ·
5 · 15 · 1 · 2 · 0 · 0 · 9 · 1 · 5 · 0 · 0 · 1 · 0 · 3 · 2 · 0 · 5 · 3 · 0 · 7 · 0 · 5 · 1 · 0 · 6 · 6 · 0 · 0 · 10 · 3 · 0 · 3
0 · 9 · 3 · 0 · 6 · 3 · 2 · 1 · 0 · 0 · 1 · 4 · 1 · 0 · 1 · 0 · 3 · 0 · 8 · 3 · 1 · 0 · 2 · 0 · 2 · 1 · 1 · 0 · 1 · 4 · 2 · 3 · 1 ·
2 · 1 · 1 · 14 · 0 · 0 · 4 · 3 · 2 · 3 · 0 · 0 · 0 · 2 · 0 · 3 · 1 · 8 · 0 · 3 · 0 · 0 · 2 · 2 · 2 · 0 · 0 · 1 · 1 · 1 · 1 · 0 · 1
1 · 6 · 6 · 0 · 0 · 11 · 0 · 0 · 5 · 0 · 7 · 2
```

```
table(sim)
```

```
sim
  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  17  18
307 242 123  98  62  53  36  22  16  14   4   6   6   4   4   1   1   1
```

```
hist(sim, main="Distribuzione Geometrica", xlab="Numero di fallimenti",
     col="lightblue", border="black")
```



- Ogni valore della sequenza rappresenta il numero di fallimenti prima del primo successo
- La distribuzione geometrica ha una lunga coda, quindi i valori maggiori sono relativamente rari, ma possibili

STATISTICA E ANALISI DEI DATI

Distribuzione Geometrica Modificata

Distribuzione Geometrica Modificata

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

$$E_r = \{\text{il primo successo si verifica alla prova } r\text{-esima}\} \quad (r = 1,2,\dots)$$

- La distribuzione geometrica modificata è utilizzata, ad esempio, per modellare:
  - numero di analisi da effettuare in un laboratorio prima di ottenere una risposta positiva;
  - numero di farmaci da sperimentare prima di trovarne uno efficace

Distribuzione Geometrica Modificata

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

$$E_r = \{\text{il primo successo si verifica alla prova } r\text{-esima}\} \quad (r = 1,2,\dots)$$

- La distribuzione geometrica modificata è utilizzata, ad esempio, per modellare:
  - numero di analisi da effettuare in un laboratorio prima di ottenere una risposta positiva;
  - numero di farmaci da sperimentare prima di trovarne uno efficace
- Alcune situazioni descrivibili con una distribuzione geometrica sono le seguenti:
  - Un motore di ricerca passa in rassegna un elenco di siti alla ricerca di una determinata frase chiave
    - Supponiamo che la ricerca termini non appena viene trovata la frase chiave
    - **Il numero di siti visitati** è descrivibile con una distribuzione geometrica modificata
  - Un responsabile delle assunzioni intervista i candidati, uno per uno, per coprire un posto vacante.
    - Il numero di candidati intervistati fino a quando un candidato riceve un'offerta è descrivibile con una distribuzione geometrica modificata.

Distribuzione Geometrica Modificata

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

$$E_r = \{\text{il primo successo si verifica alla prova } r\text{-esima}\} \quad (r = 1, 2, \ldots)$$

- La probabilità di avere un successo all'r-esimo tentativo è data da:

$$P(E_r) = (1-p)^{r-1}p$$

che si ottiene dall'ipotesi di indipendenza delle prove

- Sia X la variabile aleatoria che descrive il numero di fallimenti che precedono il primo successo

Funzione di probabilità: $F_X(x) = P(X = x) = \begin{cases} (1-p)^{x-1}p & x = 1, 2, \ldots \\ 0 & altrimenti \end{cases}$

con $0 < p < 1$ si dice avere distribuzione geometrica di parametro p

Relazione tra Geometrica e Geometrica Modificata

- Dalla funzione di probabilità: $F_X(x) = P(X = x) = \begin{cases} (1-p)^{x-1}p & x = 1,2,\dots \\ 0 & altrimenti \end{cases}$

  si ha che la funzione di probabilità è strettamente decrescente in $x = 1,2,\dots$

- Sia Y una variabile aleatoria geometrica, se consideriamo che $X = Y + 1$
  - Si ha che
    - Probabilità: $p_X(x) = P(X = x) = P(Y + 1 = x) = P(Y = x - 1) = p_Y(x - 1)$
    - Distribuzione: $F_X(x) = P(X \leq x) = P(Y + 1 \leq x) = P(Y \leq x - 1) = F_Y(x - 1)$
  - Poiché:

$$\sum_{r=1}^{k} p_X(x) = \sum_{r=1}^{k} (1-p)^{r-1}p = p\sum_{s=0}^{k-1} (1-p)^{s} = p\frac{1-(1-p)^k}{1-(1-p)} = 1-(1-p)^k$$

- la **funzione di distribuzione** di X è la seguente:

$$F_X(x) = \begin{cases} 1-(1-p)^k & k \leq x < k+1 \\ 0 & x < 1 \end{cases}$$

Relazione tra Geometrica e Geometrica Modificata

- Dalla funzione di probabilità: $F_X(x) = P(X = x) = \begin{cases} (1-p)^{x-1}p & x = 1,2,\dots \\ 0 & altrimenti \end{cases}$

  si ha che la funzione di probabilità è strettamente decresc<ente in $x = 1,2,\dots$

- Sia Y una variabile aleatoria geometrica, se consideriamo che $X = Y + 1$
  - Si ha che
    - Probabilità: $p_X(x) = P(X = x) = P(Y + 1 = x) = P(Y = x - 1) = p_Y(x - 1)$
    - Distribuzione: $F_X(x) = P(X \leq x) = P(Y + 1 \leq x) = P(Y \leq x - 1) = F_Y(x - 1)$
  - Poiché:

$$\sum_{r=1}^{k} p_X(x) = \sum_{r=1}^{k} (1-p)^{r-1}p = p\sum_{s=0}^{k-1}(1-p)^s = p\frac{1-(1-p)^k}{1-(1-p)} = 1-(1-p)^k$$

p

- la **funzione di distribuzione** di X è la seguente:

$$F_X(x) = \begin{cases} 1-(1-p)^k & k \leq x \leq k+1 \\ 0 & x < 1 \end{cases}$$

Distribuzione Geometrica Modificata

- Il valore atteso e la varianza della distribuzione geometrica modificata sono:

Valore atteso: $E(X) = E(Y + 1) = \frac{1}{p}$

Varianza:$Var(X) = Var(Y + 1) = \frac{1-p}{p^2}$

Distribuzione Geometrica Modificata

- Il valore atteso e la varianza della distribuzione geometrica modificata sono:

Valore atteso: $E(X) = E(Y + 1) = \frac{1}{p}$ **Varianza**:$Var(X) = Var(Y + 1) = \frac{1-p}{p^2}$

- Per il calcolo in R della funzione di probabilità e della funzione di distribuzione si utilizzano le funzioni:

dgeom(x-1, prob)

pgeom(x-1, prob, lower.tail = TRUE)

dove

- **x** è il valore assunto (o i valori assunti) dalla variabile aleatoria geometrica considerata;
- **prob** è la probabilità di successo in ciascuna prova
- **lower.tail** se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$

Proprietà Assenza di Memoria

- Una variabile aleatoria discreta Y a valori interi non negativi gode della proprietà di **assenza di memoria** se per ogni n ed m interi non negativi:

$$P(Y > n + m \mid Y > n) = P(Y > m)$$

  - La probabilità che il primo successo si verifichi dopo altri m tentativi, dato che non si è ancora verificato nei primi n tentativi, è uguale alla probabilità che il primo successo richieda più di m tentativi in assoluto
  - In altre parole, il comportamento futuro della distribuzione **non dipende** da ciò che è successo prima: la "storia" non influisce sulle probabilità future

Proprietà Assenza di Memoria

- Proprietà assenza di memoria:
  - Una variabile aleatoria discreta Y a valori interi non negativi gode della proprietà di mancanza di memoria se per ogni n ed m interi non negativi:

$$P(Y > n + m \mid Y > n) = P(Y > m)$$

Dimostrazione:

- Sia T l'istante di primo successo in una successione di prove ripetute di Bernoulli e indipendenti, in ciascuna delle quali la probabilità del successo è costante, ed è uguale a $p \in (0,1)$
  - Sappiamo che T ha distribuzione geometrica modificata di parametro p, ovvero $X = T - 1$ ha distribuzione **geometrica** di parametro p
- Per la distribuzione geometrica modificata, la funzione di probabilità cumulativa (CDF) di T (numero totale di tentativi necessari per il primo successo) è:

$$P(T \leq n) = 1 - (1-p)^n \;\; e\; quindi \;\; P(T > n) = (1-p)^n$$

- Ebbene, T gode della proprietà di mancanza di memoria e si ha:

$$P(T > n + m \mid T > n) = \frac{P(T > n+m, T > n)}{P(T > n)} = \frac{P(T > n+m)}{P(T > n)} = \frac{(1-p)^{n+m}}{(1-p)^n} = (1-p)^m = P(T > m)$$

Esempio Proprietà Assenza di Memoria

- Esempio:
  - **Caso 1**:
    - Lanciamo un dado e vogliamo che esca 6.
    - La variabile aleatoria T rappresenta il numero di lanci necessari per ottenere il primo 6.
    - Probabilità di successo: $p = \frac{1}{6}$
    - Probabilità di insuccesso: $1 - p = \frac{5}{6}$
      - Qual è la probabilità che ci servano più di 3 lanci per vedere il primo 6?

$$P(T > 3) = (1 - p)^3 = \left(\frac{5}{6}\right)^3 = \frac{125}{216} \approx 0{,}579$$

      - C'è circa il 58% di probabilità che nei primi 3 lanci NON esca mai il 6

Esempio Proprietà Assenza di Memoria

- **Caso 2**:
  - Abbiamo già lanciato il dado 2 volte e non è uscito 6.
  - *Qual è la probabilità che ci servano altri più di 3 lanci (quindi in totale più di 5 lanci) per vedere il primo 6?*

$$P(T > 2 + 3 \mid T > 2) = \frac{P(T > 5)}{P(T > 2)} = \frac{\left(\frac{5}{6}\right)^5}{\left(\frac{5}{6}\right)^2} = \left(\frac{5}{6}\right)^3 = \frac{125}{216} \approx 0{,}579$$

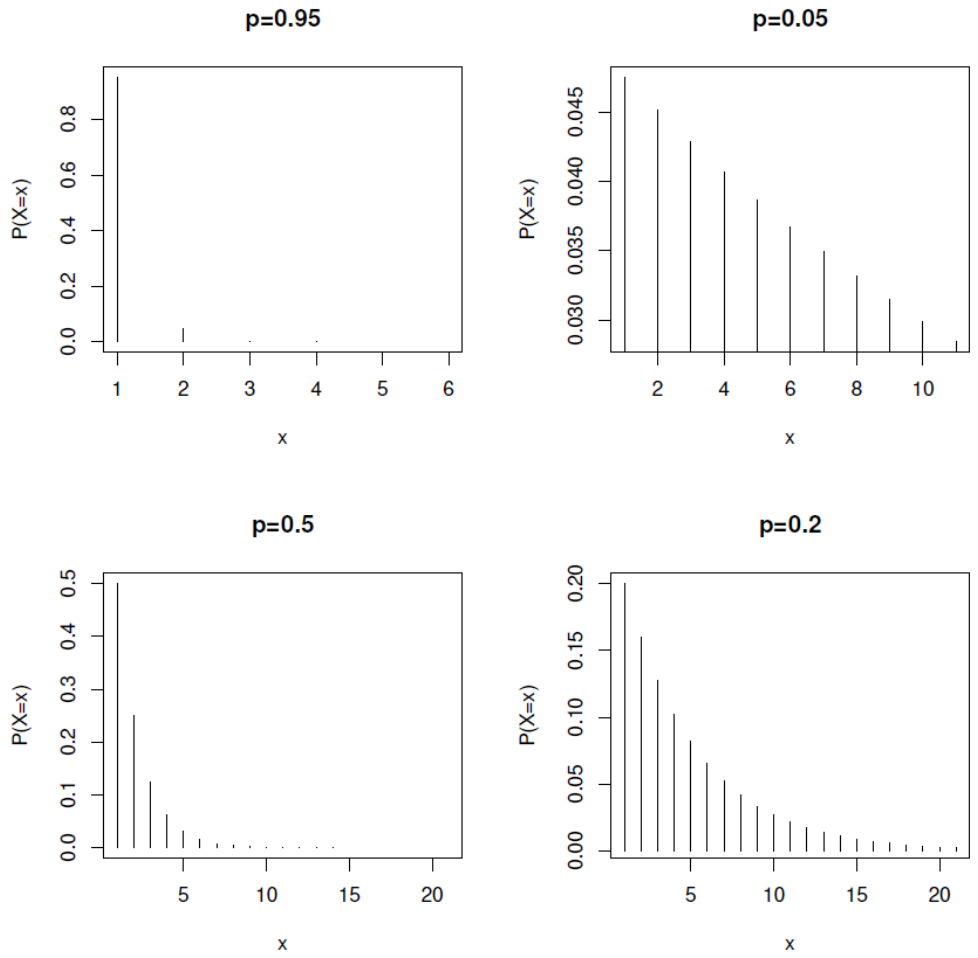
- Conclusione:
  - Il dado non ricorda i lanci precedenti!
  - Anche dopo 2 fallimenti, la probabilità di dover attendere altri 3 lanci è Esattamente la stessa che se iniziassi ora da zero

Distribuzione Geometrica Modificata (R)

- Per visualizzare le funzioni di probabilità di una variabile aleatoria geometrica modificata $X = Y + 1$ con le probabilità usate nell'esempio delle variabili aleatorie geometriche:

```
> x<-0:5
> dgeom(x,prob=0.95)
[1] 9.50000e-01 4.75000e-02 2.37500e-03 1.18750e-04 5.93750e-06
[6] 2.96875e-07
```

```
> par(mfrow=c(2,2))
> x<-1:6
> plot(x,dgeom(x-1,prob=0.95),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.95")
>
> x<-1:11
> plot(x,dgeom(x-1,prob=0.05),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.05")
>
> x<-1:21
> plot(x,dgeom(x-1,prob=0.5),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.5")
>
> y<-1:21
> plot(x,dgeom(x-1,prob=0.2),
+ xlab="x",ylab="P(X=x)",type="h",
+ main="p=0.2")
```



Esempio Sistema di Sicurezza Informatica

- Un hacker tenta di violare un sistema informatico.
- Ogni tentativo ha una probabilità $p = 0.2$ (20%) di successo
- Per proteggersi, il sistema ha un meccanismo di blocco che interrompe i tentativi dopo $n = 5$ fallimenti consecutivi. Ci interessa calcolare:
  1. La probabilità che il primo successo avvenga esattamente al r-esimo tentativo, per $r = 1,2,3,4,5$
  2. La probabilità cumulativa che il successo avvenga entro $r = 3$

Esempio Sistema di Sicurezza Informatica

- Un hacker tenta di violare un sistema informatico.
- Ogni tentativo ha una probabilità $p = 0.2$ (20%) di successo
- Per proteggersi, il sistema ha un meccanismo di blocco che interrompe i tentativi dopo $n = 5$ fallimenti consecutivi. Ci interessa calcolare:
  1. La probabilità che il primo successo avvenga esattamente al r-esimo tentativo, per $r = 1,2,3,4,5$

$$P(X = 1) = (1-p)^{x-1}p = (1-0.2)^{1-1}0.2 = 0.2$$

$$P(X = 2) = (1-p)^{x-1}p = (1-0.2)^{2-1}0.2 = 0.16$$

$$P(X = 3) = (1-p)^{x-1}p = (1-0.2)^{3-1}0.2 = 0.128$$

$$P(X = 4) = (1-p)^{x-1}p = (1-0.2)^{4-1}0.2 = 0.1024$$

$$P(X = 5) = (1-p)^{x-1}p = (1-0.2)^{5-1}0.2 = 0.08192$$

Esempio Sistema di Sicurezza Informatica

- Un hacker tenta di violare un sistema informatico.
- Ogni tentativo ha una probabilità $p = 0.2$ (20%) di successo
- Per proteggersi, il sistema ha un meccanismo di blocco che interrompe i tentativi dopo $n = 5$ fallimenti consecutivi. Ci interessa calcolare:

2. La probabilità cumulativa che il successo avvenga entro $r = 3$

$$P(X = 1) = 0.2$$

$$P(X = 2) = 0.16$$

$$P(X = 3) = 0.128$$

$$F(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.2 + 0.16 + 0.128 = 0.488$$

- Quindi, c'è circa il 48,8% di probabilità che l'attacco riesca entro i primi 3 tentativi.

Quantili (R)

- Per calcolare i quantili oppure per simulare una variabile aleatoria geometrica modificata X, basta ricordare che $X = Y + 1$ e aggiungere 1 ai quantili oppure alla simulazione della variabile geometrica
- Esempio: Generiamo una sequenza di 20 numeri pseudocasuali simulando una variabile aleatoria geometrica modificata con $p = 0.2$ si ha che:

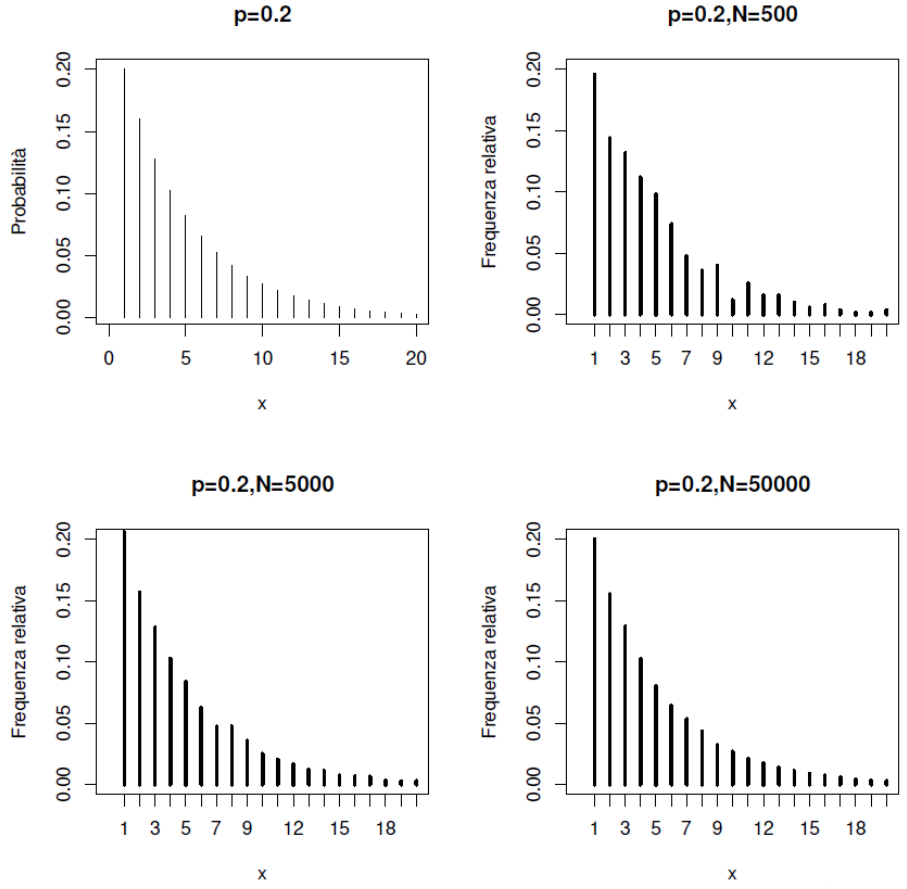
```
> sim<-rgeom(20,prob=0.2)+1
> sim
 [1] 28  1  2  1  9  2  2  1  5 12  9  1  2  7  2  4  3  1 21  2
> table(sim)
sim
 1  2  3  4  5  7  9 12 21 28
 5  6  1  1  1  1  2  1  1  1
> table(sim)/length(sim)
sim
   1    2    3    4    5    7    9   12   21   28
0.25 0.30 0.05 0.05 0.05 0.05 0.10 0.05 0.05 0.05
```

dove $\frac{table(sim)}{length(sim)}$ fornisce le frequenze relative con cui i numeri $0,1,\dots,20$

Generazione di Numeri Casuali (R)

- Esempio: Il codice seguente permette di confrontare la funzione di probabilità geometrica modificata con quella simulata all'aumentare della lunghezza $N = 500, 5000, 50000$ della sequenza generata

```
> par(mfrow=c(2,2))
> x<-1:21
> plot(x,dgeom(x-1,prob=0.2),xlab="x",ylab="Probabilita' ",
+ type="h",main="p=0.2",xlim=c(0,20))
>
> sim1<-rgeom(500,prob=0.2)+1
> plot(table(sim1)/length(sim1),xlab="x",type="h",
+ ylab="Frequenza relativa",xlim=c(0,20),ylim=c(0,0.20),
+ main="p=0.2,N=500")
>
> sim2<-rgeom(5000,prob=0.2)+1
> plot(table(sim2)/length(sim2),xlab="x",type="h",
+ ylab="Frequenza relativa",xlim=c(0,20),ylim=c(0,0.20),
+ main="p=0.2,N=5000")
>
> sim3<-rgeom(50000,prob=0.2)+1
> plot(table(sim3)/length(sim3),xlab="x",type="h",
+ ylab="Frequenza relativa",xlim=c(0,20),ylim=c(0,0.20),
+ main="p=0.2,N=50000")
```

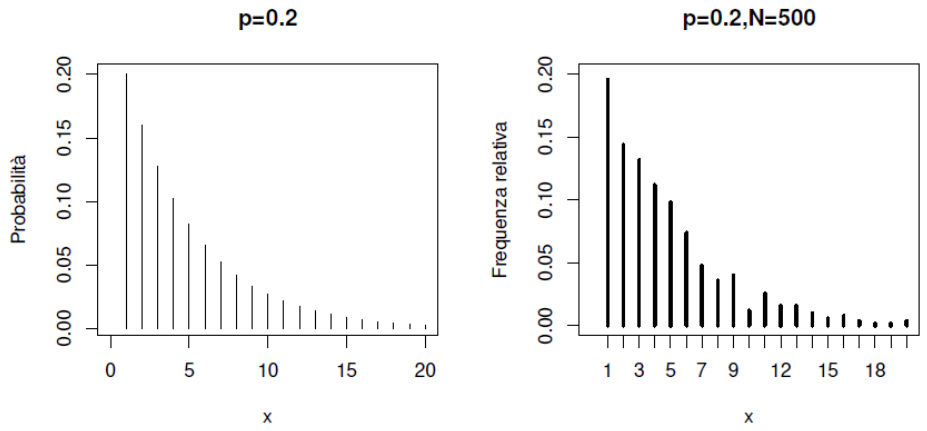


Nota: all'aumentare della lunghezza della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità geometrica modificata

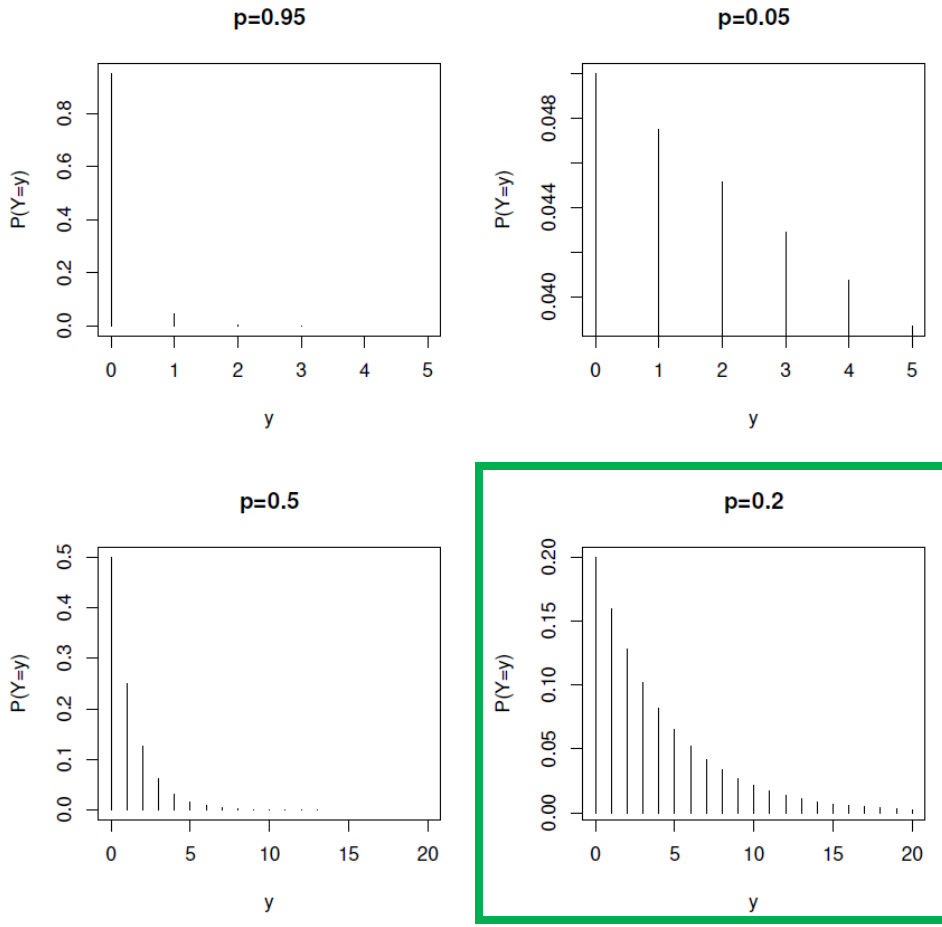
Generazione di Numeri Casuali (R)

- All'aumentare della lunghezza della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità geometrica modificata

Geometrica Mod.



Geometrica



STATISTICA E ANALISI DEI DATI

Distribuzione binomiale negativa

Distribuzione Binomiale Negativa

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

F_r = {il numero di fallimenti che precedono il successo n-esimo in una sequenza di prove di Bernoulli è r}

$$(r = 0,1,2,\dots)$$

- La distribuzione binomiale negativa è utile, ad esempio, per modellare:
  - numero di ritrasmissioni di un messaggio costituito da n blocchi in un sistema informatico;
  - numero di bit ricevuti senza errori in un collegamento con rumore prima dell' n-esimo bit di errore

Distribuzione Binomiale Negativa

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

F_r = {il numero di fallimenti che precedono il successo n-esimo in una sequenza di prove di Bernoulli è r}

$$(r = 0,1,2,\dots)$$

- Sia Y la variabile aleatoria che descrive il numero di fallimenti che precedono successo *n*-esimo

Funzione di probabilità: $p_Y(y) = P(Y = y) = \begin{cases} \binom{n+y-1}{y} p^n (1-p)^y & y = 1,2,\dots \\ 0 & altrimenti \end{cases}$

con $0 < p < 1$ si dice avere distribuzione binomiale negativa di parametri n e p

Distribuzione Binomiale Negativa

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

F_r = {il numero di fallimenti che precedono il successo n-esimo in una sequenza di prove di Bernoulli è r}

$$(r = 0,1,2,\dots)$$

- Sia Y la variabile aleatoria che descrive il numero di fallimenti che precedono successo *n*-esimo

Funzione di probabilità: $p_Y(y) = P(Y = y) = \begin{cases} \binom{n+y-1}{y} p^n(1-p)^y & y = 1,2,\dots \\ 0 & altrimenti \end{cases}$

con $0 < p < 1$ si dice avere distribuzione binomiale negativa di parametri n e p

y è il numero di fallimenti

È il coefficiente binomiale, che conta il numero di modi di distribuire y fallimenti e n successi in una sequenza di y + n prove

Distribuzione Binomiale Negativa

- Per la mancanza di memoria della distribuzione geometrica, si ha:

$$Y = Y_1 + Y_2 + \cdots + Y_n$$

dove $Y_1 + Y_2 + \cdots + Y_n$ sono variabili aleatorie indipendenti di tipo geometrico descriventi il numero di fallimenti al primo successo

Valore atteso: $E(Y) = \frac{n(1-p)}{p}$ **Varianza**: $Var(Y) = \frac{n(1-p)}{p^2}$

- Per il calcolo in R della funzione di probabilità si utilizza la funzione:

dnbinom(x, size, prob)

dove

- **x** è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale negativa
- **prob** è la probabilità di successo in ciascuna prova
- **size** è il numero richiesto di successi

Coefficiente di Variazione

- Nel caso di una variabile aleatoria binomiale negativa con parametro p:
  - Poiché
    - $E[R] = \frac{n(1-p)}{p}$
    - $Var(R) = \frac{n(1-p)}{p^2}$

$$CV(Y) = \frac{\sqrt{Var(Y)}}{E(Y)} = \frac{\sqrt{\frac{n(1-p)}{p^2}}}{\frac{n(1-p)}{p}} = \frac{\frac{\sqrt{n(1-p)}}{p}}{\frac{n(1-p)}{p}}$$

$$= \frac{\sqrt{n(1-p)}}{p} * \frac{p}{n(1-p)} = \frac{\sqrt{n(1-p)}}{n(1-p)} = \frac{\sqrt{n(1-p)}}{(\sqrt{n(1-p)})^2} = \frac{1}{\sqrt{n(1-p)}}$$

- Esempio con $p = 0.4$ e $n = 3$:

$$CV(Y) = \frac{1}{\sqrt{3(1-0.4)}} = \frac{1}{1.34} = 0.74$$

Significa che la deviazione standard è circa 0.74 volte il valore medio, segnalando una distribuzione non eccessivamente dispersa

Distribuzione Binomiale Negativa

- Per il calcolo in R della funzione di probabilità si utilizza la funzione:

  pnbinom(x, size, prob, lower.tail = TRUE)

  dove

  - **x** è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale negativa
  - **prob** è la probabilità di successo in ciascuna prova
  - **size** è il numero richiesto di successi
  - **lower.tail** se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$

- Per calcolare i quantili (percentili) della distribuzione binomiale negativa si utilizza la funzione

  qnbinom(z, size, prob)

  dove

  - **z** è il valore assunto (o i valori assunti) dalle probabilità relative al percentile z · 100–esimo;
  - **prob** è la probabilità di successo in ciascuna prova
  - **size** è il numero richiesto di successi

Esempio - Azienda

- Un'azienda vuole produrre 3 componenti funzionanti utilizzando una macchina che ha probabilità $p = 0.8$ di produrre un componente funzionante in un ciclo di produzione
  - Vogliamo calcolare quante volte, in media, la macchina produrrà componenti difettosi prima di ottenere i 3 componenti funzionanti richiesti
  - Il numero di componenti difettosi segue una distribuzione binomiale negativa con parametri $n = 3$ (successi richiesti) e $p = 0.8$ (probabilità di successo)

Valore atteso: $E(Y) = \frac{n(1-p)}{p} = \frac{3(1-0.8)}{0.8} = 0.75$ **Varianza**: $Var(Y) = \frac{n(1-p)}{p^2} = \frac{3(1-0.8)}{0{,}8^2} = 0.95$

  - Probabilità di produrre 2 componenti difettosi prima di ottenere 3 funzionanti:

$$p_Y(y) = P(Y = 2) = \binom{2+3-1}{3} 0.8^3(1-0.8)^2 = 0.08192. \qquad y = 1,2, \ldots$$

C'è circa l'8.1% di probabilità che ci siano esattamente 2 fallimenti

Esempio - Software

- Immaginiamo di lavorare su un **sistema distribuito** con molteplici microservizi
  - Durante il deployment di una nuova funzionalità, i test vengono eseguiti in sequenza sui diversi servizi per verificare che il codice funzioni correttamente
  - Ogni test ha una probabilità p di avere esito positivo e ogni fallimento richiede debug e fix del servizio prima di proseguire con il test successivo

Quanti servizi devono fallire prima che il team riesca a testare correttamente con successo un totale di n microservizi?

- Supponiamo che:
  - La probabilità di successo per un test su un microservizio sia $p = 0.6$
  - Il team vuole verificare che $n = 4$ microservizi siano correttamente deployati senza bug
  - Vogliamo sapere qual è la probabilità che ci siano esattamente $r = 3$ fallimenti prima di raggiungere i 4 successi

$$p_Y(y) = P(Y = 3) = \binom{4 + 3 - 1}{3} 0.6^4(1 - 0.6)^3 = 20 * 0.1296 * 0.064 = \mathbf{0.165888}$$

STATISTICA E ANALISI DEI DATI

Distribuzione binomiale negativa modificata

Distribuzione Binomiale Negativa Modificata

- Consideriamo l'esperimento consistente in n prove di Bernoulli indipendenti ed effettuate tutte in condizioni identiche con $p \in (0,1)$

E_r = {il successo n-esimo si verifica alla prova r-esima in una sequenza di prove di Bernoulli}

$(r = 0,1,2,\ldots)$

Qui **contiamo il numero totale di prove** (sia successi che fallimenti) fino al successo n-esimo.

- Perché ciò avvenga, deve succedere quanto segue:
  - $\boldsymbol{n-1}$ **successi** devono verificarsi nelle prime $r-1$ prove
  - La r-esima prova deve essere un **successo**
- La distribuzione binomiale negativa modificata è utile, ad esempio, per modellare:
  - Numero di analisi da effettuare in un laboratorio prima di ottenere l'n-esima risposta positiva;
  - Numero di farmaci da sperimentare prima di trovarne l' n-esimo efficace

Distribuzione Binomiale Negativa Modificata

- Sia X la variabile aleatoria che descrive il numero di fallimenti che precedono successo *n*-esimo

Funzione di probabilità: $p_X(x) = P(X = x) = \begin{cases} \binom{x-1}{n-1} p^n (1-p)^{x-n} & x = n, n+1, \dots \\ 0 & altrimenti \end{cases}$

con $0 < p < 1$ si dice avere distribuzione binomiale negativa modificata di parametri n e p

- dove:
  - $\binom{x-1}{n-1}$ è il numero di modi per scegliere $n-1$ successi nelle prime $x-1$ prove,
  - p^n è la probabilità di avere n successi totali
  - $(1-p)^{x-n}$ è la probabilità di avere $x-n$ insuccessi

Distribuzione Binomiale Negativa Modificata

- Sia X la variabile aleatoria che descrive il numero di fallimenti che precedono successo *n*-esimo la **Funzione di distribuzione** è:

$$F_X(x) = P(X \leq x) = \sum_{k=n}^{x} \binom{k-1}{n-1} p^n (1-p)^{k-n} \qquad x = n, n+1, \ldots$$

  - $F_X(x) = 0$ per $x < n$ (non si possono avere n successi prima della $n-$esima prova)
  - $F_X(x) \to 1$ quando $r \to \infty$ (la probabilità cumulativa tende a 1)

Distribuzione Binomiale Negativa Modificata

- Per la mancanza di memoria della distribuzione geometrica, si ha:

$$X = X_1 + X_2 + \cdots + X_n$$

dove $X_1, X_2, \ldots, X_n$ sono variabili aleatorie indipendenti di tipo geometrico modificato descriventi il numero di prove necessarie per ottenere il primo successo

- Poiché $X = Y + n$, con Y variabile binomiale negativa, si ha che:

Valore atteso: $E(X) = E(Y + n) = \frac{n}{p}$ **Varianza**: $Var(X) = Var(Y + n) = \frac{n(1-p)}{p^2}$

- Da cui ci si riconduce anche alla funzione di probabilità e di distribuzione con dove Y è una variabile binomiale negativa:
  - Probabilità: $p_X(x) = P(X = x) = P(Y + n = x) = P(Y = x - n) = p_Y(x - n)$
  - Distribuzione: $F_X(x) = P(X \leq x) = P(Y + n \leq x) = P(Y \leq x - n) = F_Y(x - n)$
- **Nota**: Il processo di derivazione delle funzioni di probabilità e distribuzione è simile a quello visto per le variabili aleatoria geometriche e geometriche modificate

Coefficiente di Variazione

- Nel caso di una variabile aleatoria binomiale negativa modificata con parametro p:
  - Poiché
    - $E[R] = \frac{n}{p}$
    - $Var(R) = \frac{n(1-p)}{p^2}$

$$CV(Y) = \frac{\sqrt{Var(Y)}}{E(Y)} = \frac{\sqrt{\frac{n(1-p)}{p^2}}}{\frac{n}{p}} = \frac{\frac{\sqrt{n(1-p)}}{p}}{\frac{n}{p}} = \frac{\sqrt{n(1-p)}}{p} * \frac{p}{n}$$

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

- Esempio con $p = 0.5$ e $n = 2$:

$$CV(Y) = \sqrt{\frac{0.5}{2}} = \sqrt{0.25} = 0.5$$

Significa che la deviazione standard è la metà del valore medio, segnalando una distribuzione **moderatamente** dispersa

Distribuzione Binomiale Negativa Modificata

- Per il calcolo in R delle funzioni di probabilità e distribuzione della binomiale negativa modificata si usano

```
dnbinom(x-n, size, prob)
pnbinom(x-n, size, prob, lower.tail = TRUE)
```

dove

- $x - n$ è il valore assunto (o i valori assunti) dalla variabile aleatoria binomiale negativa modificata;
- **prob** è la probabilità di successo in ciascuna prova
- **size** è il numero richiesto di successi
- **lower.tail** se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$

Esempio - Pubblicità

- Un venditore fa telefonate indipendenti a potenziali clienti con una probabilità di successo $p = 0.2$ (chiusura del contratto) per ciascuna telefonata
- Vogliamo calcolare:
  - La probabilità che il terzo contratto venga chiuso alla 8ª telefonata
  - La probabilità cumulativa che il terzo contratto venga chiuso entro la 10ª telefonata
    - $n = 3$: il terzo successo
    - $p = 0.2$: probabilità di successo per singola telefonata
    - $x = 8$ o $x = 10$: numero della telefonata
- Soluzione 1:

$$p_X(x) = P(X = 8) = \binom{8-1}{3-1} 0.2^3 (1-0.2)^{8-3} = 21 * 0.008 * 0.32768 = 0.055$$

Cioè c'è circa il 5,5% di probabilità che il terzo contratto venga chiuso alla 8ª telefonata

Esempio - Pubblicità

- Un venditore fa telefonate indipendenti a potenziali clienti con una probabilità di successo p = 0.2 (chiusura del contratto) per ciascuna telefonata
- Vogliamo calcolare:
  - La probabilità che il terzo contratto venga chiuso alla 8ª telefonata
  - La probabilità cumulativa che il terzo contratto venga chiuso entro la 10ª telefonata
    - $n = 3$: il terzo successo
    - $p = 0.2$: probabilità di successo per singola telefonata
    - $x = 8$ o $x = 10$: numero della telefonata
- Soluzione 2:

$$p_X(x) = P(X \leq 10) = \sum_{k=n}^{x} P(X = k) = \sum_{k=3}^{10} \binom{k-1}{n-1} p^n (1-p)^{k-n}$$

Esempio - Pubblicità

- Un venditore fa telefonate indipendenti a potenziali clienti con una probabilità di successo p = 0.2 (chiusura del contratto) per ciascuna telefonata
- Vogliamo calcolare:
  - La probabilità che il terzo contratto venga chiuso alla 8ª telefonata
  - La probabilità cumulativa che il terzo contratto venga chiuso entro la 10ª telefonata
    - $n = 3$: il terzo successo
    - $p = 0.2$: probabilità di successo per singola telefonata
    - $x = 8$ o $x = 10$: numero della telefonata
- Soluzione 2:

$$p_X(x) = P(X \leq 10) = \sum_{k=n}^{x} P(X = k) = \sum_{k=3}^{10} \binom{k-1}{n-1} p^n (1-p)^{k-n}$$

```r
n <- 3    # Numero di successi desiderati
p <- 0.2  # Probabilità di successo per prova
x <- 10   # Numero massimo di prove considerate

# Calcolo della probabilità cumulativa
F_X <- pnbinom(x - n, size = n, prob = p)

# Risultato
cat("La probabilità cumulativa è:", F_X*100, "%\n")
```

```
La probabilità cumulativa è: 32.22005 %
```

Esempio - Query

- In un progetto di data science, stai valutando l'efficacia di un motore di ricerca.
- Sai che la probabilità di ottenere un risultato pertinente da una query è $p = 0.3$
- Vogliamo calcolare:
  - La probabilità che il 5º risultato pertinente venga trovato esattamente alla **12ª query**.
  - La probabilità cumulativa che il 5º risultato pertinente venga trovato **entro le prime 15 query**.
    - $n = 5$: il terzo successo
    - $p = 0.4$: probabilità di successo per ogni query
    - $x = 12$ o $x = 15$: numero di query

Esempio - Query

- In un progetto di data science, stai valutando l'efficacia di un motore di ricerca.
- Sai che la probabilità di ottenere un risultato pertinente da una query è $p = 0.3$
- Vogliamo calcolare:
  - La probabilità che il 5º risultato pertinente venga trovato esattamente alla **12ª query**.
  - La probabilità cumulativa che il 5º risultato pertinente venga trovato **entro le prime 15 query**.
    - $n = 5$: il terzo successo
    - $p = 0.4$: probabilità di successo per ogni query
    - $x = 12$ o $x = 15$: numero di query
- Soluzione 1:

```r
n <- 5    # Numero di successi desiderati
p <- 0.3  # Probabilità di successo
x <- 12   # Numero della query

# Calcolo della probabilità puntuale
P_X <- dbinom(n - 1, x - 1, p)
cat("La probabilità che il 5º risultato venga trovato alla 12ª query è:", P_X*100,"%\n")
```

```
La probabilità che il 5º risultato venga trovato alla 12ª query è: 22.0133 %
```

Esempio - Query

- In un progetto di data science, stai valutando l'efficacia di un motore di ricerca.
- Sai che la probabilità di ottenere un risultato pertinente da una query è $p = 0.3$
- Vogliamo calcolare:
  - La probabilità che il 5º risultato pertinente venga trovato esattamente alla **12ª query**.
  - La probabilità cumulativa che il 5º risultato pertinente venga trovato **entro le prime 15 query**.
    - $n = 5$: il terzo successo
    - $p = 0.4$: probabilità di successo per ogni query
    - $x = 12$ o $x = 15$: numero di query
- Soluzione 2:

```r
n <- 5    # Numero di successi desiderati
p <- 0.3  # Probabilità di successo
x <- 15   # Numero massimo di query

# Calcolo della probabilità cumulativa
F_X <- pnbinom(x - n, size = n, prob = p)

# Risultato
cat("La probabilità cumulativa che il 5º risultato venga trovato entro le prime 15 query è:", F_X*100, "%\n")
```

```
La probabilità cumulativa che il 5º risultato venga trovato entro le prime 15 query è: 48.45089 %
```

Quantili e Generazione Randomica

- Per calcolare i quantili (percentili) della distribuzione binomiale negativa si utilizza la funzione basta aggiungere *n* ai quantili e alla simulazione della variabile binomiale negativa:

```
qnbinom(z, size,prob)+n
rnbinom(N, size, prob )+n
```

dove

- ***z*** è il valore assunto (o i valori assunti) dalle probabilità relative al percentile $z \cdot 100$–esimo;
- **prob** è la probabilità di successo in ciascuna prova
- **size** è il numero richiesto di successi

Quantili e Generazione Randomica

- Esempio:
  - Generiamo una sequenza di 10 numeri pseudocasuali che rappresentano il numero di prove che occorrono per ottenere il secondo successo (n) supponendo che $p = 0.2$

```
> sim<-rnbinom(10,size=2,prob=0.2)+2
> sim
 [1] 10  5 18 18  5 15 16  8  6  7
> table(sim)
sim
 5  6  7  8 10 15 16 18
 2  1  1  1  1  1  1  2
> table(sim)/length(sim)
sim
  5   6   7   8  10  15  16  18
0.2 0.1 0.1 0.1 0.1 0.1 0.1 0.2
```

Quantili e Generazione Randomica

- Esempio:
  - Generiamo una sequenza di 10 numeri pseudocasuali che rappresentano il numero di prove che occorrono per ottenere il secondo successo (n) supponendo che $p = 0.2$

```
> sim<-rnbinom(10,size=2,prob=0.2)+2
> sim
 [1] 10  5 18 18  5 15 16  8  6  7
> table(sim)
sim
 5  6  7  8 10 15 16 18
 2  1  1  1  1  1  1  2
> table(sim)/length(sim)
sim
  5   6   7   8  10  15  16  18
0.2 0.1 0.1 0.1 0.1 0.1 0.1 0.2
```

Target per numero di prove riuscite

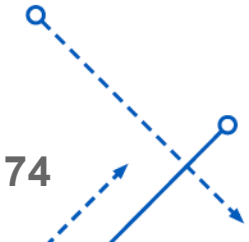
Numero di osservazioni

Frequenze assolute

Frequenze relative

Esempio - Pubblicità

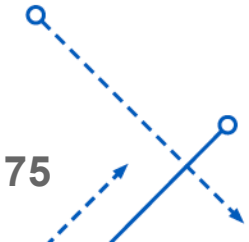
- Un'azienda vuole simulare il numero di clic necessari per **ottenere 10 conversioni** su una campagna pubblicitaria
  - Ogni clic ha una probabilità di successo (conversione) pari a $p = 0.05$
  - Usiamo **rnbinom** per simulare i risultati di 1000 campagne



Esempio - Pubblicità

- Un'azienda vuole simulare il numero di clic necessari per **ottenere 10 conversioni** su una campagna pubblicitaria

```r
# Parametri
n <- 10      # Numero di successi desiderati (conversioni)
p <- 0.05    # Probabilità di successo per clic
num_campaigns <- 1000  # Numero di campagne da simulare

# Simulazione
set.seed(123)  # Per riproducibilità
results <- rnbinom(num_campaigns, size = n, prob = p)

# Aggiungi i successi al numero di insuccessi generati
clicks_required <- results + n

# Risultati
cat("Numero di clic necessari (primi 10 risultati):", clicks_required[1:10], "\n")

# Analisi
mean_clicks <- mean(clicks_required)  # Media dei clic necessari
cat("Media dei clic necessari per ottenere 10 conversioni:", mean_clicks, "\n")

# Istogramma
hist(clicks_required, breaks = 20, main = "Distribuzione del numero di clic necessari",
     xlab = "Numero di clic", col = "blue", border = "white")
```



Esempio - Pubblicità

- Un'azienda vuole simulare il numero di clic necessari per **ottenere 10 conversioni** su una campagna pubblicitaria

```
Numero di clic necessari (primi 10 risultati): 173 106 312 105 275 216 175 241 134 162
Media dei clic necessari per ottenere 10 conversioni: 197.918
```

```r
# Parametri
n <- 10      # Numero di successi desiderati (conversioni)
p <- 0.05    # Probabilità di successo per clic
num_campaigns <- 1000  # Numero di campagne da simulare

# Simulazione
set.seed(123)  # Per riproducibilità
results <- rnbinom(num_campaigns, size = n, prob = p)

# Aggiungi i successi al numero di insuccessi generati
clicks_required <- results + n

# Risultati
cat("Numero di clic necessari (primi 10 risultati):", clicks_required[1:10], "\n")

# Analisi
mean_clicks <- mean(clicks_required)  # Media dei clic necessari
cat("Media dei clic necessari per ottenere 10 conversioni:", mean_clicks, "\n")

# Istogramma
hist(clicks_required, breaks = 20, main = "Distribuzione del numero di clic necessari",
     xlab = "Numero di clic", col = "blue", border = "white")
```