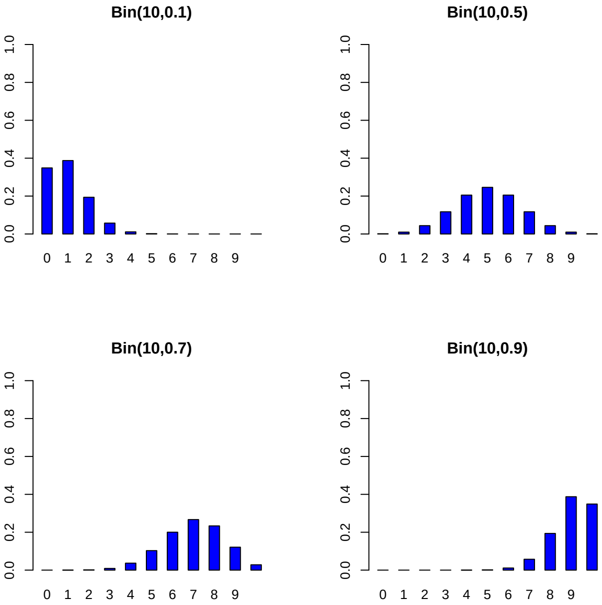
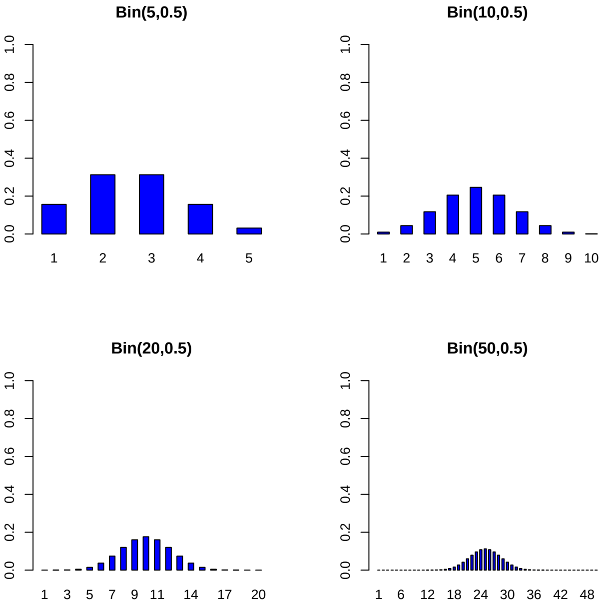
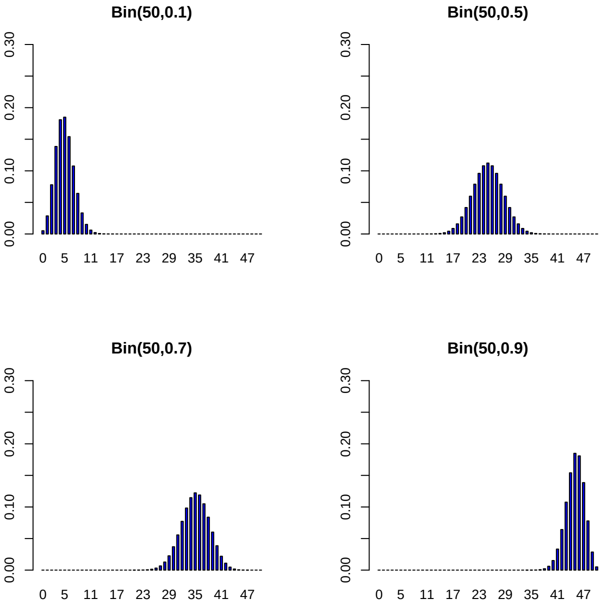
Il modello di riferimento

- Come per le altre grandezze considerate anche per conteggi e proporzioni possiamo definire una distribuzione di probabilità che svolga la funzione di modello teorico di riferimento per le nostre procedure.
- La gaussiana, la t di Student e la F di Fisher sono distribuzioni che descrivono variabili continue, mentre i conteggi sono variabili discrete. Quindi occorre una distribuzione differente:
- **Distribuzione Binomiale**: $Bin(n, p)$.
- E' governata da due parametri: n numero totale di *osservazioni* e p probabilità di "successo", ovvero di osservare un certo evento.
- Media e varianza possono essere calcolate come $\mu = np$ e $\sigma^2 = np(1 - p)$


Bin(10,0.1)
Bin(10,0.5)
Bin(10,0.7)
Bin(10,0.9)
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3 4 5 6 7 8 9


Bin(5,0.5)
Bin(10,0.5)
Bin(20,0.5)
Bin(50,0.5)
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5
1 2 3 4 5 6 7 8 9 10
1 3 5 7 9 11 14 17 20
1 6 12 18 24 30 36 42 48


Bin(50,0.1)
Bin(50,0.5)
Bin(50,0.7)
Bin(50,0.9)
0.00
0.10
0.20
0.30
0 5 11 17 23 29 35 41 47

- Al crescere della numerosità delle osservazioni e per $p \approx 0.5$ la distribuzione binomiale può essere approssimata con la distribuzione normale di media $\mu = np$ e varianza $\sigma^2 = np(1-p)$.
- Quando e come usare l'approssimazione gaussiana:
    a) Sia $\mu = np$ che $\sigma^2 = np(1-p)$ sono maggiori di 5.
    b) Per $n > 5$ l'approssimazione normale è valida se:

$$|(1/\sqrt{n})(\sqrt{(1-p)/p} - \sqrt{p/(1-p)})| < 0.3$$

    c) In ogni caso è opportuno applicare una *correzione di continuità* ovvero se voglio calcolare $P(X < 8)$ per una variabile binomiale usando l'approssimazione normale, calcolerò $P(X < 8.5)$, dove l'aggiunta di 0.5 è appunto tale correzione.

Per effettuare un test d'ipotesi procederemo in modo analogo a quanto già visto.

- Impostiamo il problema di ipotesi defininendo H_0 ed H_1
- Da questo scegliamo la statistica test, ad esempio la proporzione $\hat{p}$ calcolata sulle osservazioni
- Calcoliamo il p-value ovvero la probabilità del valore osservato sotto la distribuzione definita dall'ipotesi nulla.

Esempio delle balene: la storia

I pescherecci in alcune zone dell'oceano Atlantico hanno occasionalmente problemi con le balene che, con la loro presenza, spaventano e fanno scappare i pesci. Le balene di solito non si fermano a lungo, forse disturbate dal rumore delle barche, abbandonano la zona. Gli operatori sonar hanno osservato che circa il 40% delle balene si allontana quasi subito all'arrivo dei pescherecci.

Nel tentativo di allontanare le balene viene condotto un esperimento, nel quale una barca trasmette sott'acqua i suoni prodotti dalle orche. Si osserva che di 30 balene presenti la metà abbandona immediatamente la zona. Poiché si osserva una proporzione superiore al 40% ci chiediamo se possiamo concludere che l'apparecchio spaventa balene è efficace.

Esempio delle balene: test binomiale

Impostiamo il problema di ipotesi:

H_0 : $p = 0.4$ l'apparecchio spaventa balene non ha effetto

H_1 : $p > 0.4$ l'apparecchio spaventa balene ha effetto

Abbiamo osservato 15 balene che si allontanano quindi il p-value si ottiene calcolando $P(X > 15)$ da una distribuzione Binomiale(30,0.4).

Poiché la distribuzione è discreta otteniamo questo valore come

$$P(X = 15) + P(X = 16) + \ldots + P(X = 30) = 0.1754$$

fissando $\alpha = 0.05$ accettiamo H_0.

Esempio delle balene: test usando la distribuzione normale

Lo stesso problema di ipotesi potrebbe essere affrontato utilizzando l'approssimazione normale. Vediamo però se questa da una buona approssimazione della binomiale nelle condizioni dell'esperimento. A tal fine calcoliamo le condizioni prima introdotte:

a) $\mu = 30 * 0.4 = 12$, $\sigma^2 = np(1-p) = 30 * 0.4 * 0.6 = 7.2$ verificata

b) $|(1/\sqrt{n})(\sqrt{(1-p)/p} - \sqrt{p/(1-p)}) =$
$|(1/\sqrt{30})(\sqrt{(1-0.4)/0.4} - \sqrt{0.4/(1-0.4)})| = 0.0745356 < 0.3$

Quindi posso calcolare $P(X > 15)$ usando la distribuzione gaussiana di media 12 e varianza 7.2 (sd=2.68), $P(X > 15) = 0.13177$ e di nuovo accetto l'ipotesi nulla.

Quando effettuiamo una osservazione basata su di un campione ci interessa capire di quanto ci possiamo sbagliare, ovvero data una certa grandezza valutata su base campionaria, quale sia l'intervallo di valori plausibile nell'intera popolazione. Come già visto a questa domanda rispondiamo con gli intervalli di confidenza. Per le proporzioni possiamo costruirli basandoci sull'approssimazione normale.

In questa situazione faremo riferimento ad una distribuzione normale con media $\mu = \hat{p}$ e varianza $\sigma^2 = \frac{p(1-p)}{n}$. Quindi l'intervallo di confidenza sarà costruito come:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Esempio dei moscerini del mango: la storia

Il moscerino della frutta *Bactrocera tryoni* che vive nel Queensland può depositare le sue uova sui frutti del mango. Pertanto, i manghi australiani, vanno trattati prima di poter essere esportati.

A tal fine è stato condotto un esperimento per verificare l'efficacia di un trattamento dei frutti basato sul riscaldamento degli stessi. I risultati sono riportati nella tavola seguente, il riscaldamento applicato era di 43 gradi Celsius.

Adulti Sopravvissuti	637
Uova eliminate	5266
Totale uova presenti	5903

Esempio dei moscerini del mango: l'IC al 95%

Fissiamo un livello di confidenza (di fiducia nelle nostre stime) del 95%, quindi ammettiamo un errore massimo di $\alpha = 0.05$. In questo esempio $\hat{p} = 637/5903 = 0.1079$ con una deviazione standard di $\sqrt{\frac{p(1-p)}{n}} = 0.0069$. Inoltre $z_{\alpha/2} = 1.96$:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 0.1079 \pm 0.01359 = [0.0943, 0.1215]$$

Possiamo allora dire che nel 95% dei casi il trattamento permette la sopravvivenza tra il 9.43% e il 12% delle uova.