Corso integrato di informatica, statistica e analisi dei dati sperimentali
Esercitazione VI

Test del χ^2 (il file Excel **test_chi.xls** mostra possibili sviluppi degli esercizi proposti**)**
Esercizio 1a)
un costruttore di lampadine afferma che la probabilità che una delle sue lampadina si fulmini prima di 100 ore di funzionamento è il 12% (p = 0.12). Per verificare questa affermazione prendiamo un gruppo di N=90 lampadine e le lasciamo accese per 100 ore. Scaduto il tempo verifichiamo che O_f = 18 lampadine sono fulminate e O_a = 72 sono ancora accese. Verificare l'ipotesi che il costruttore dica il vero.

Prepare un foglio elettronico per il calcolo del test χ^2.

	fulminate	accese	totale		p
exp.:	18	72	90		0.12
th.	10.8	79.2	90		
ΔN	7.20	-7.20	χ oss.	ν	P(χ oss.)
ΔN^2/N^th	4.80	0.65	5.45	1	0.019517
			χ Lim.	α	
			3.84	0.05	

$$\frac{(O_i - A_i)^2}{A_i}$$

Inserire i dati (**1**), inserire la probabilità (**2**) e calcolare le frequenze teoriche (**3**).

Calcolare i termini $\frac{(O_i - A_i)^2}{A_i}$ e quindi il χ^2 (**4**). Definire i gradi di libertà (**5**) e calcolare la probabilità associata al valore osservato di χ^2 per una distribuzione con ν gradi di libertà usando la funzione Excel DISTIB.CHI(χ,ν) (**6**). In base alla probabilità ottenuta si può affermare che …..?

Confrontare il valore limite sulle tabelle o calcolare il valore limite per a =0.05 utilizzando la funzione Excel INV.CHI(α,ν) (**7**). Trarre le conclusioni del caso.

Cambiando il valore della probabilità determinare i limiti di confidenza entro i quali la probabilità associata al χ^2 osservato è > di 0.05%. In mancanza di un programma di adeguato si cambiano a mano i valori di p e si osserva che per valori di p compresi tra [0.135 ; 0.29] non posso rigettare l'ipotesi con un livello di confidenza migliore del 95%. In base ai dati il valore di p è compreso tra 0.135 e 0.29

	fulminate	accese	totale		p
exp.:	18	72	90		0.29
th.	26.1	63.9	90		
ΔN	-8.10	8.10	χ oss.	ν	P(χ oss.)
ΔN^2/N^th	2.51	1.03	3.54	1	0.059885
			χ Lim.	α	
			3.84	0.05	

	fulminate	accese	totale		p
exp.:	18	72	90		0.135
th.	12.15	77.85	90		
ΔN	5.85	-5.85	χ oss.	ν	P(χ oss.)
ΔN^2/N^th	2.82	0.44	3.26	1	0.071151
			χ Lim.	α	
			3.84	0.05	

Esercizio 1b) Lanciando un dado 38 volte osservo 10 volte testa. Posso affermare con un limite di confidenza del 5% che il dato è truccato?

Dado

	6	<6	totale			p
exp.:	10	28	38			0.17
th.	6.33	31.67	38			
ΔN	3.67	-3.67	χ oss.	ν	P(χ oss.)	
ΔN^2/N^th	2.12	0.42				
			χ Lim.	α		
			3.84	0.05		

Esercizio 2)

Per verificare la sensibilitá di una specie di piante ad un virus di espongono al contagio M gruppi (M=90) di $n = 20$ individui ciascuno. Dopo un tempo certo tempo si contano il numero di contagi k per ogni gruppo e si riportano in tabella le frequenze N_k con cui sono osservati k contagi per gruppo. Si vuole verificare l'ipotesi che la probabilitá di contagio sia $p = 0.3$ e che quindi il numero di contagi per gruppo (k) segua una distribuzione binomiale $f(n, p)$.

I dati mostrati in tabella sono riportati nel file **Contagio.dat**
Importare i dati in Excel e costruire la tabella mostrata in figura.

	Classi k	frequenze N_k
	2	5
	3	2
	4	11
	5	14
	6	18
	7	12
	8	14
	9	6
	10	4
N. di classi osservate	11	3
	12	1
	out	0
totali	11	90

La distribuzione binomiale calcola la probabilitá di ottenere k di successi (nel nostro caso contagi) su n prove indipendenti (nel nostro caso n=20):

$$f(k,n) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

In base al modello Binomiale le frequenze teoriche sono:$N_k^{th} = M f(k, n)$ dove M é il numero delle osservazioni (M=90) e sono riportate in tabella.

Parametri esperimento · calcolo del termine fattoriale: n! / k! (n-k)! · calcolo della distribuzione binomiale · calcolo N_T *f · calcolo (N-N(th)^2/N_th

	Classi k	frequenze N_k	n 20	fatt	binom	N_k(th)	ΔN^2/N
	2	5		190	0.027845873	2.506	2.482
	3	2	p	1140	0.071603672	6.444	3.065
	4	11	0.3	4845	0.130420974	11.738	
	5	14	p da testare	15504	0.178863051	16.098	
	6	18		38760	0.191638983	17.248	
	7	12		77520	0.164261985	14.784	[illegible]
	8	14		125970	0.11439674	10.296	1.333
	9	6		167960	0.065369566	5.883	0.002
	10	4		184756	0.030817081	2.774	0.542
N. di classi osservate	11	3		167960	0.012006655	1.081	3.409
	12	1		125970	0.003859282	0.347	1.226
	out	0			out	0.802	0.802
totali	11	90					
				normalizz.	0.991083861	89.198	regione lontana

$$\frac{(O_i - A_i)^2}{A_i}$$

Riportare in una cella opportuna il numero di prove per esperimento (n=20) e la probabilità p che si vuole sottoporre a test. Inserire lunghe formule in Excel può essere complicato: i calcoli si possono dividere ad esempio calcolando prima il termine fattoriale: n!/k!/(n-k)! (FATTORIALE(N) è la funzione Excel per il calcolo di N!) e quindi la distribuzione f(k,20).

Si possono calcolare le frequenze attese M*f e i $\frac{(O_i - A_i)^2}{A_i}$ termini

e quindi il $$\chi^2_o = \sum_{i=1}^{N} \frac{(O_i - A_i)^2}{A_i}$$

χ^2	
12.94	*Senza regione esterna*
13.74	*Inclusa la regione esterna*

facendo attenzione a includere il termine che tiene conto della regione “lontana” dove non abbiamo avuto osservazioni.
Per sottoporre a test l’ipotesi che i dati seguono l’andamento teorico calcolato in base all’ipotesi di una distribuzione binomiale si può:

a) verificare utilizzando le tabelle che il valore osservato del χ^2 è minore del valore limite (α=0.05) per un χ^2 con 10 (o 11 se si considera la regione esterna) gradi di libertà.

b) calcolare i valori limite del χ^2 con 10 (o 11) gradi di libertà e α=0.05 usando la funzione Excel= INV.CHI(α,ν)

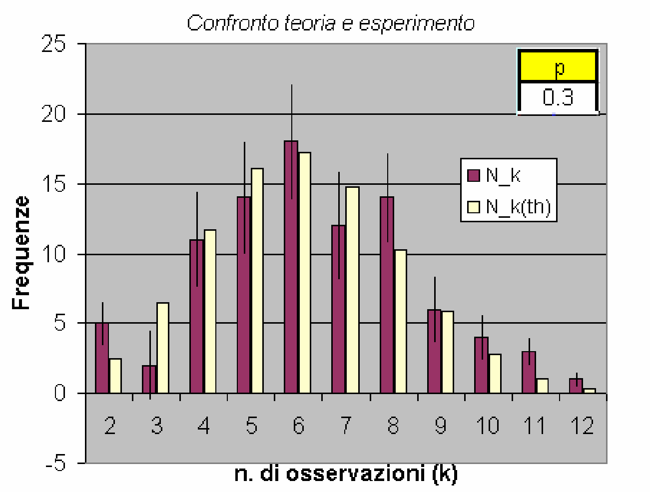
ν	10
α_L	0.05

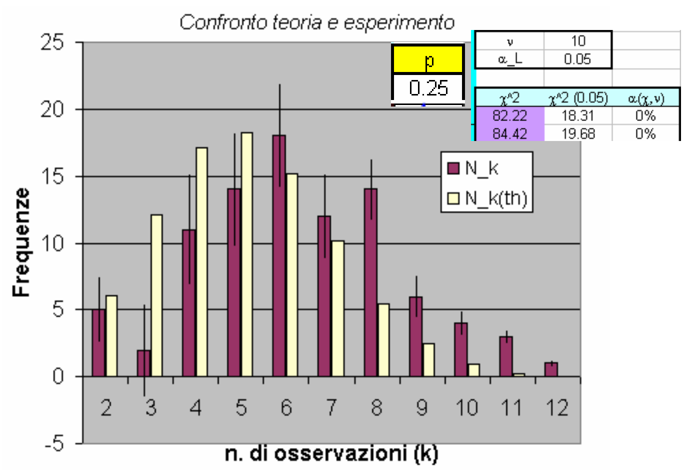
χ^2	χ^2 (0.05)	α(χ,ν)
12.94	18.31	23%
13.74	19.68	25%
	b	c

c) calcolare la probabilità associata ad un χ^2 con 10 (o 11) gradi di libertà eguale al valore osservato

Il test dice che….

Un istogramma che mette a confronto i dati sperimentali con le frequenze teoriche fa vedere che c’e’ un buon accordo tra teoria ed esperimento. Se di cambia il valore della probabilità utilizzato per il test si vede che per p<0.28 o p>0.36 l’accordo peggiora e il test del χ^2 permette di rigettare le ipotesi (le differenze tra dato e teoria sono significative)





Esercizio 3a) Distribuzione di Poisson

Per studiare la distribuzione di piante in una regione si eseguono M=33 campionamenti in zone diverse di eguale superficie **S**. L'istogramma in tabella mostra le frequenze con cui sono state osservate k piante per campionamento. (frequenze di osservazione). I dati sono riportati nel file **Piante.dat**.

In base au dati sperimentali si può assumere che la distribuzione delle piante sia uniforme con densità d = <k>/S (<k> numero medio di piante in un'area S) con una confidenza del 95%? Ovvero: le osservazioni seguano una distribuzione di Poisson con valore atteso <k> e la probabilità associata al chi^2 è maggiore di 0.05%?

	classi	N_k	somma	scarti^2
	20	4	80	1600
	21	2	42	882
	22	1	22	484
	23	2	46	1058
	24	0	0	0
	25	4	100	2500
	26	1	26	676
	27	3	81	2187
	28	2	56	1568
	29	2	58	1682
	30	2	60	1800
	31	1	31	961
	32	0	0	0
	33	1	33	1089
	34	3	102	3468
	35	1	35	1225
	36	1	36	1296
	37	1	37	1369
	38	0	0	0
	39	0	0	0
	40	2	80	3200
N_oss	21			
	totale	**33**		
		media	**28.0303**	
			varianza	819.55

In base ai dati calcoliamo prima il numero medio di piante per area di campionamento:

$$\bar{k} = \frac{\sum_i N_k k}{\sum_i N_k} = 28.0$$

e la varianza:

$$\sigma_k^2 = \frac{\sum_i N_k (k - \bar{k})^2}{\sum_i N_k} = 33.8$$

(attenzione al calcolo delle medie pesate!)

La funzione densità di probabilità nel modello di Poisson è:

$$f(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Le frequenze teoriche sono quindi:

$$N_k^{th} = M f(k, \bar{k})$$

	f_th	Nth	z
	2.47E-02	0.82	12.430
	3.30E-02	1.09	0.763
	4.20E-02	1.39	0.108
	5.12E-02	1.69	0.057
	5.98E-02	1.97	1.975
	6.71E-02	2.21	1.441
	7.23E-02	2.39	0.806
	7.51E-02	2.48	0.110
	7.52E-02	2.48	0.093
	7.27E-02	2.40	0.066
	6.79E-02	2.24	0.026
	6.14E-02	2.03	0.519
	5.38E-02	1.77	1.774
	4.57E-02	1.51	0.171
	3.77E-02	1.24	2.486
	3.02E-02	1.00	0.000
	2.35E-02	0.77	0.065
	1.78E-02	0.59	0.291
	1.31E-02	0.43	0.433
	9.43E-03	0.31	0.311
	6.61E-03	0.22	14.562
tot. in	9.40E-01	31.02	
resto (out)	5.99E-02	1.98	1.98
controllo	**1.0**	**33.0**	

Attenzione a considerare il contributo della regione "lontana" (out): in base al modello teorico avrei dovuto osservare 1.98 eventi per k<20 o k>40, regioni in cui non ho avuto osservazioni.

Si calcolano i temini: $\frac{(O_i - A_i)^2}{A_i}$

E quindi:

$$\chi_o^2 = \sum_{i=1}^{N} \frac{(O_i - A_i)^2}{A_i}$$

chi^2	38.486
nlib=	20
prob. chi	0.0077194

Attenzione al calcolo dei gradi di libertà: ho usato un grado di libertà per il calcolo di M e un grado di libertà per stimare la media, quindi i gradi di libertà sono:

M + 1(regione esterna) – 2 = 20

La probabilità associata al valore del chi^2 osservato con 20 gradi di libertà è: 0.77 %, ben inferiore a 0.05% quindi possiamo rigettare l'ipotesi che le piante siano distribuite in modo uniforme.

Esercizio 3b) Distribuzione uniforme:

Utilizzando i dati dell'esercizio precedente dire se i dati sono compatibili con una distribuzione uniforme del numero di piante osservate nell'area di misura.

Il numero minimo di piante osservato è 20, il numero massimo è 40. In tutto abbiamo 21 classi di osservazione. Per una distribuzione uniforme la distribuzione la probabilità di osservare k piante è costante: $f_u = 1/21 = 0.0476$. Il numero di piante per intervallo è $N^{th} = Mf_i = 1.57$. Quindi calcolo la tabella mostrata, il chi^2 associato e la probabilità associata al chi^2 osservato con $\nu = 21 - 3$ (ho utilizzato i dati per calcolare M e i limiti k_{min} e k_{max}, quindi ho utilizzato tre gradi di libertà).

Ottengo la tabella mostrata da cui risulta un chi^2 = 18.5.

La probabilità ad esso associata per n=18 gradi di libertà è alta (>> 0.05) quindi non posso rifiutare l'ipotesi che il numero di piantine nelle superfici di campionamento segua una distribuzione uniforme tra k_{min} = 20 e k_{max} = 40.

	Uniforme		
	f_th	Nth	z
	4.76E-02	1.57E+00	3.75E+00
	4.76E-02	1.57E+00	1.17E-01
	4.76E-02	1.57E+00	2.08E-01
	4.76E-02	1.57E+00	1.17E-01
	4.76E-02	1.57E+00	1.57E+00
	4.76E-02	1.57E+00	3.75E+00
	4.76E-02	1.57E+00	2.08E-01
	4.76E-02	1.57E+00	1.30E+00
	4.76E-02	1.57E+00	1.17E-01
	4.76E-02	1.57E+00	1.17E-01
	4.76E-02	1.57E+00	1.17E-01
	4.76E-02	1.57E+00	2.08E-01
	4.76E-02	1.57E+00	1.57E+00
	4.76E-02	1.57E+00	2.08E-01
	4.76E-02	1.57E+00	1.30E+00
	4.76E-02	1.57E+00	2.08E-01
	4.76E-02	1.57E+00	2.08E-01
	4.76E-02	1.57E+00	2.08E-01
	4.76E-02	1.57E+00	1.57E+00
	4.76E-02	1.57E+00	1.57E+00
	4.76E-02	1.57E+00	1.17E-01
tot. in	1.00E+00	33.00	
resto (out)	0.00E+00	0.00	0.00
controllo	**1.0**	**33.0**	
		chi^2	18.545
		nlib=	18
		prob. chi	0.420301