Esercizi di Statistica per gli studenti di Scienze Politiche, Università di Firenze

Esercizi svolti da una selezione di compiti degli Esami scritti di Statistica del 1999 e del 2000

VERSIONE PROVVISORIA APRILE 2001

A cura di
L. Matrone
F.Mealli
L.Mencarini
A.Petrucci

A. ESERCIZI DI STATISTICA DESCRITTIVA

Esercizio 1.

Si consideri la seguente distribuzione delle industrie tessili secondo il fatturato annuo in milioni:

Fatturato	***[300,800]***	***]800,1500]***	***]1500,3000]***	***]3000,5000]***
Aziende	*50*	*80*	*40*	*30*

a) Determinare la distribuzione di frequenze relative.
Le frequenze relative si ottengono dividendo ciascuna frequenza assoluta per il totale delle frequenze n=200:

Classi di modalità	Frequenze assolute (n_i)	Frequenze relative (n_i/n)	Ampiezza intervallo (a_i)	Densità di frequenza (d_i)	Valori centrali di classe (c_i)
[300,800]	50	50/200=0,25	500	50/500 =0,10	550
]800,1500]	80	80/200=0,40	700	80/700=0,11	1150
]1500,3000]	40	40/200=0,20	1500	40/1500=0,03	2250
]3000,5000]	30	30/200=0,15	2000	30/2000=0,01	4000
Totale	200	1			

b) Qual è la percentuale di industrie con fatturato annuo superiore a 800 milioni e non superiore a 3 miliardi?
Il numero di industrie con tali caratteristiche risulta dalla somma delle frequenze assolute delle classi]800,1500] e]1500,3000]. La percentuale richiesta è quindi (120/200)*100= 60%.

c) Calcolare il fatturato modale.
È la classe con la densità di frequenza più elevata, che risulta essere la classe]800, 1500].

d) Calcolare il fatturato medio
Essendo le modalità raggruppate in classi è necessario fare qualche ipotesi sulla distribuzione dell'età all'interno di ciascuna classe. Si può ipotizzare, ad esempio, che le frequenze siano concentrate sul valore centrale, $c_i=(x_i+x_{i-1})/2$, di ogni classe, oppure che l'età media in ogni classe sia pari al valore centrale. Entrambe queste ipotesi conducono al calcolo della età media come:

$$\mu_x=\frac{\sum_{i=1}^{k} c_i n_i}{\sum_{i=1}^{k} n_i} = (550*50 + 1150*80 +2250*40 + 4000*30)/200 = 329500/200 = 1647,5.$$

Esercizio 2.

I tentativi di suicidio nel 1995 secondo l'età sono descritti dalla seguente distribuzione di frequenza:

Età	***N° tentativi***
[14,18[	*133*
[18,25[	*499*
[25,45[	*1515*
[45,65[	*770*
[65,75]	*409*

Si sa inoltre che la somma delle età di coloro che hanno tentato il suicidio è 141233 anni.

a) Calcolare l'età media

Il totale dei soggetti che hanno tentato il suicidio è $n=\Sigma_i\ n_i=3326$. Inoltre, poiché è nota la somma delle età di coloro che hanno tentato il suicidio, ovvero

$$\sum_{i=1}^{3326} x_i = 141233,$$

la media aritmetica risulta pari a:

$\mu_x=141233/3326= 42{,}46$. (Non è necessaria alcuna ipotesi semplificatrice per il calcolo come invece è stato necessario nell'esercizio 1).

b) Calcolare la percentuale di minorenni che ha tentato il suicidio

Essendo i minorenni coloro che hanno età nella classe [14,18[, tale percentuale risulta pari a (133/3326)*100 =3,99%.

c) Calcolare la percentuale di coloro che hanno tentato il suicidio di età non inferiore a 18 anni e minore di 65 anni

Il numero di persone che soddisfa la condizione richiesta è dato dalla somma delle frequenze assolute delle tre classi di età [18,25[, [25,45[e [45,65[. Dunque la percentuale è pari a ((499+1515+770)/3326)*100=83,7%.

d) Calcolare la classe modale

Essendo le classi di ampiezza diversa, è necessario individuare la classe a cui corrisponde la densità di frequenza più elevata:

Classi di età	Frequenze assolute (n_i)	Ampiezza dell'intervallo (a_i)	Densità di frequenza $\frac{n_i}{a_i}$
[14, 18[	133	4	33,25
[18, 25[	499	7	71,29
[25, 45[	1515	20	75,75
[45, 65[	770	20	38,50
[65, 75]	409	10	40,90
Totale	3326		

La classe modale è dunque la classe [25, 45[.

Esercizio 3.

Sia data la variabile ***X*** *= reddito mensile in milioni, rilevata su un collettivo di famiglie come segue:*

Reddito	***N° di famiglie***
1	*1*
2	*0*
3	*5*
4	*4*

a) Trovare la moda del reddito

La moda è la modalità che si presenta più frequentemente (ovvero che presenta frequenza assoluta più elevata); il reddito modale è dunque pari a 3 milioni.

b) Trovare lo scarto quadratico medio del reddito

Lo scarto quadratico medio, o deviazione standard, è la media quadratica degli scarti dalla media μ:

$$\sigma = \sqrt{\frac{1}{N}\left[\sum_{i=1}^{N}(x_i-\mu)^2 n_i\right]}.$$

In questo esempio risulta

μ=(1*1 + 2*0 + 3*5 + 4*4) /10=(1+15+16)/10=32/10=3,2

e quindi
σ^2= [(1-3,2)^2*1 + (3-3,2)^2 *5 + (4-3,2)^2 *4]/10= (4,84 + 0,04*5 + 0,64*4)/10=
=(4,84+0,2+2,56)/10=7,6/10=0,76
da cui
σ=0,8717.

Dato che risulta $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i)^2 n_i - \mu^2}$, lo stesso risultato può essere ottenuto come radice quadrata della differenza tra la media quadratica al quadrato e la media aritmetica al quadrato, infatti:
σ^2=(1+4*0+9*5+16*4)/10 -(3,2)^2=(1+45+64)/10-10,24=110/10-10,24=11-10,24=0,76.

c) Trovare lo scarto quadratico medio del reddito nell'ipotesi che ad ogni famiglia venga dato un aumento di stipendio di 500 mila lire

Lo scarto quadratico medio, così come la varianza, è invariante per traslazione, ovvero se viene aggiunta una costante a ciascuna determinazione del carattere lo scarto quadratico medio non cambia: σ(X+a)=σ(X)=0.8717.
Si ricordi, più in generale, che $\sigma^2(aX+b) = a^2\sigma^2X$ e σ(aX+b) =|a|σ(X).

d) Trovare il rapporto di concentrazione per il reddito

Il rapporto di concentrazione ottenuto con la formula di Gini, $R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$, può essere utilizzato solo se i dati vengono riorganizzati in forma di successione (che in questo caso risulta 1 3 3 3 3 3 4 4 4 4).

Non volendo, o non potendo, organizzare l'informazione in questo modo è possibile utilizzare la formula del rapporto di concentrazione ottenuta dividendo la differenza media semplice (Δ) per il valore che tale indice di variabilità assume nel caso di massima concentrazione (2μ): $R=\Delta/2\mu$.
La differenza semplice media è ottenibile utilizzando la seguente formula per distribuzioni di frequenza $(x_i, n_i)_{i=1\ldots k}$:

$$\Delta = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}|x_i - x_j|n_i n_j}{n(n-1)}$$

Per calcolare $|x_i-x_j|$ possiamo fare riferimento al seguente prospetto di calcolo:

	1	3	4
1	0	2	3
3	2	0	1
4	3	1	0

Per calcolare $n_{i*}n_j$ utilizziamo un prospetto di calcolo analogo:

	1	*5*	*4*
1	-	5	4
5	5	-	20
4	4	20	-

Dunque risulta:
$\Delta=[(2*5)+(3*4)+(1*20)+ (2*5)+(3*4)+(1*20)]/(10*9)=$
$2[(2*5)+(3*4)+(1*20)]/(10*9)=84/90=0,933$
$\mu=3,2$
$R=\Delta/2\mu=0,933/6,4=0,146$

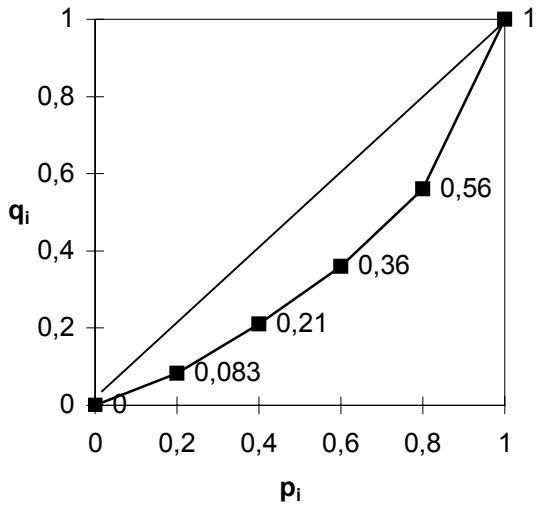
Esercizio 4.

Nell'a.a. 1988-89, gli iscritti all'Università in Italia per Facoltà risultano:

Facoltà	***Studenti in corso*** *(in migliaia)*
Scientifiche	*146*
Mediche	*100*
Ingegneria	*193*
Economiche-Giuridiche-Sociali	*520*
Letterarie	*239*

a) Disegnare il diagramma di Lorenz del numero di studenti.
Il diagramma di Lorenz è un grafico che permette di evidenziare la concentrazione di un carattere trasferibile. Per costruire il grafico è necessario ordinare le modalità del carattere "numero di studenti in corso" in senso non decrescente. Si procede poi calcolando la cumulata dell'intensità assoluta (c_i), la cumulata dell'intensità relativa (q_i) e la cumulata di frequenza relativa (p_i), come risulta nella seguente tabella:

x_i	Intensità cumulate c_i	Cumulate intensità relative q_i	p_i
100	100	0,083	0,2
146	246	0,21	0,4
193	439	0,36	0,6
239	678	0,56	0,8
520	1198	1	1



b) Calcolare il rapporto di concentrazione.
Il rapporto di concentrazione può essere calcolato utilizzando la formula di Gini:

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$$

Dalla precedente tabella risulta:

Cumulate intensità relative q_i	p_i	$p_i - q_i$
0,083	0,2	0,117
0,21	0,4	0,19
0,36	0,6	0,24
0,56	0,8	0,24
$\sum_{i=1}^{n-1}$	2	0,787

E dunque R= 0,787/2= 0,3935

Esercizio 5.

Il capitale (in miliardi) di una Società è suddiviso tra i soci nel seguente modo:

Socio	*1*	*2*	*3*	*4*	*5*
Capitale	*3*	*1*	*0,5*	*10*	*5*

a) Calcolare la variabilità del capitale mediante la differenza media semplice.

La differenza semplice media è $\Delta = \frac{\sum_{i \neq j=1}^{n} |x_i - x_j|}{n(n-1)}$

(È l'approccio più elementare per misurare la mutua variabilità e consiste nell'esaminare tutte le differenze tra le modalità a due a due, facendone una sintesi tramite un'opportuna media. Si considera il valore assoluto delle differenze per evitare che ogni confronto di annulli con il confronto opposto. I possibili confronti tra le n unità statistiche sono n(n-1) escludendo i confronti tra una unità e se stessa. È una media aritmetica delle differenze prese in valore assoluto.)

Tabellina delle differenze in valore assoluto

	3	1	0,5	10	5
3	0	2	2,5	7	2
1	2	0	0,5	9	4
0,5	2,5	0,5	0	9,5	4,5
10	7	9	9,5	0	5
5	2	4	4,5	5	0

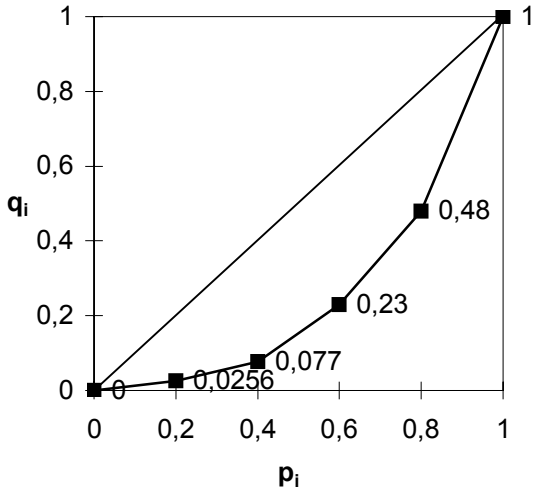
Δ=2(2+2,5+7+2+0,5+9+4+9,5+4,5+5)/(5*2)=46/10=4,6

b) Rappresentare la concentrazione del capitale mediante la spezzata di Lorenz.
Si tratta di una rappresentazione grafica della concentrazione per caratteri trasferibili

Si ordinano le modalità del carattere "capitale" in senso non decrescente, poi si calcola la cumulata dell'intensità assoluta e la cumulata dell'intensità relativa (si veda punto a) dell'esercizio precedente).

x_i	Intensità cumulate c_i	Cumulate intensità relative q_i	p_i
0,5	0,5	0,0256	0,2
1	1,5	0,0769	0,4
3	4,5	0,2307	0,6
5	9,5	0,0487	0,8
10	19,5	1	1

Si costruisce quindi il diagramma di Lorenz di ascissa p_i e ordinata q_i:



c) Determinare il rapporto di concentrazione.

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$$ ma anche $R=\Delta/2\mu$

Quindi R= 4,6/(2*3,9)= 4,6/7,8=0,5897

Esercizio 6.

Su un collettivo formato da 120 maschi e 80 femmine è stata rilevata l'età in anni ottenendo la seguente distribuzione percentuale per sesso:

Età	***% Maschi***	***% Femmine***
0 - 19	*10*	*20*
20 - 29	*10*	*20*
30 - 49	*30*	*30*
50 - 89	*50*	*30*
Totale	*100*	*100*

a) Trovare il numero di unità statistiche nel collettivo di età minore di 20 anni
Sono il 10% del totale dei 120 maschi, cioè 12 maschi + il 20% delle 80 femmine, cioè 16 femmine = 28 unità

b) Trovare la percentuale di unità statistiche nel collettivo di età maggiore o uguale a 50 anni
Sono 60 maschi + 24 femmine per un totale di 84 su 200 = 42%

c) Trovare il numero di maschi di età maggiore o uguale a 30 anni
Sono l'80% di 120= 96 maschi

d) Trovare le classi modali di età per i maschi e le femmine
È necessario individuare la classe di modalità cui corrisponde la massima densità di frequenza.

	Frequenze assolute Maschi n_{i1}	Frequenze assolute Femmine n_{i2}	Numero di modalità della classe a_i	Densità di frequenza $\frac{n_{i1}}{a_i}$	Densità di frequenza $\frac{n_{i2}}{a_i}$
0-19	12	16	20	0,6	0,8
20-29	12	16	10	1,2	1,6
30-49	36	24	20	1,8	1,2
50-89	60	24	40	1,5	0,6
TOT	120	80			

Per i maschi la classe modale è 30-49 anni, per le femmine 20-29 anni.

Esercizio 7.

Un collettivo di 200 studenti di cui 30 sono lavoratori è stato rilevato il voto ad un certo esame ottenendo la seguente distribuzione percentuale del voto per condizione dello studente:

Voto	***% Studenti non lavoratori***	***% Studenti lavoratori***
18- 22	*10*	*20*
23 – 25	*10*	*40*
26 – 28	*30*	*20*
29 – 30	*50*	*20*
Totale	*100*	*100*

L'ESERCIZIO È SIMILE AL PRECEDENTE, QUINDI VENGONO FORNITI SOLO I RISULTATI

a) Trovare il numero di unità statistiche nel collettivo con voto minore di 23
23 unità

b) Trovare la percentuale di unità statistiche nel collettivo con voto maggiore o uguale a 29
45,5%

c) Trovare il numero di studenti lavoratori con voto maggiore o uguale a 26
12

d) Trovare le classi modali del voto per gli studenti e gli studenti lavoratori
La classe modale per gli studenti è quella di voto 29-30, per gli studenti lavoratori invece è quella 23-25.

Esercizio 8.

Le abitazioni di una città vengono distinte in quelle abitate dai proprietari e in quelle abitate da affittuari. La distribuzione di frequenza relativa delle abitazioni per numero di vani delle due classificazioni vengono riportate nella tabella che segue; si sa inoltre che il numero di abitazioni abitate dai proprietari è 4000 e quello delle abitazioni in affitto è 6000.

Numero di vani	***1***	***2***	***3***	***4***	***5***	***6***	***Totale***
Abitate da proprietari	*0,05*	*0,10*	*0,15*	*0,16*	*0,23*	*0,31*	*1*
Abitate da affittuari	*0,17*	*0,21*	*0,22*	*0,18*	*0,13*	*0,09*	*1*

a) Calcolare il numero totale di abitazioni con un numero di vani non inferiore a 5

Si sa che le abitazioni di proprietari sono 4000, mentre quelle in affitto 6000

Quelle con un numero di vani ≥ 5 sono il 23%+il 31% di 4000 (proprietari), cioè il 54%, cioè 2160,e il 13% + il 9% di 6000 (affittuari), cioè il 22%, 1320.

2160+1320= 3480

b) Calcolare il numero medio di vani per il complesso delle abitazioni

Si può fare in due modi

1) Sviluppando la seguente tabella delle frequenze assolute

N° vani	1	2	3	4	5	6	Tot
Proprietari	200	400	600	640	920	1240	4000
Affittuari	1020	1260	1320	1080	780	540	6000
Tot	1220	1660	1920	1720	1700	1780	10000

Si dispone della distribuzione di frequenza delle k modalità distinte e delle corrispondenti frequenze assolute, allora la formula della media aritmetica da applicare è

$$\mu_x = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$$

μ_x= (1220*1 + 1660*2 + 1920*3 + 1720*4 + 1700*5 + 1780*6)/10000=

= (1220+3320+5760+6880+8500+10680)/10000= 3,636

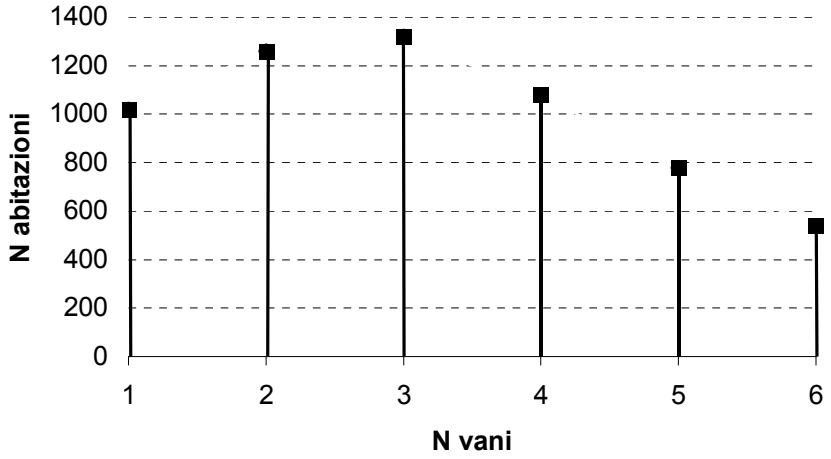
2) Oppure si può calcolare la media con le frequenze relative (metodo in questo caso più veloce perché il testo dell'esercizio fornisce proprio le frequenze relative)

$$\mu_p = \Sigma x_i f_i^{(p)} = (1*0,05)+(2*0,10)+(3*0,15)+(4*0,16)+(5*0,23)+(6*0,31)=4,35$$

$$\mu_a = \Sigma x_i f_i^{(a)} = (1*0,17)+(2*021)+(3*0,22)+(4*0,18)+(5*0,13)+(6*0,09)=3,16$$

$$\mu = \frac{4000\mu_p + 6000\mu_a}{10000} = (17400+18960)/65520=3,636$$

c) Rappresentare graficamente le abitazioni per numero di vani abitate dai affittuari



Esercizio 9.

Data la seguente distribuzione doppia di frequenza riferita alla quantità di colesterolo in milligrammi per 100 millilitri di sangue ed al sesso in un collettivo di pazienti:

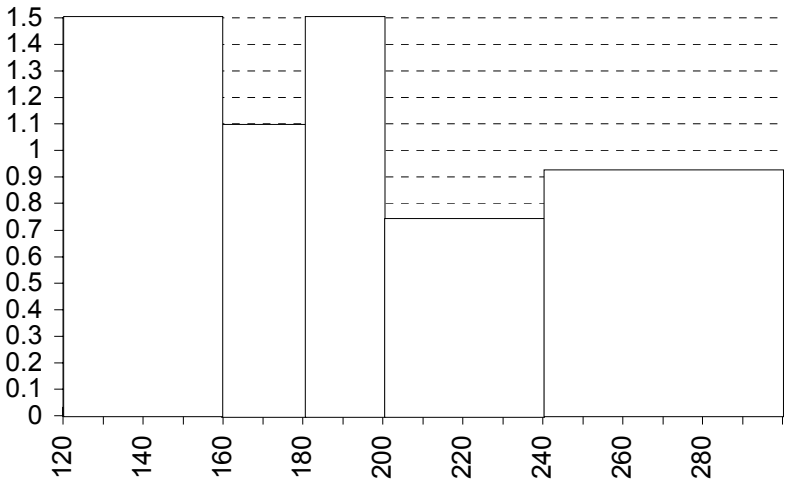
Colesterolo	***Maschio***	***Femmina***
[120,160]	*40*	*20*
]160,180]	*10*	*12*
]180,200]	*20*	*10*
]200,240]	*10*	*20*
]240,300]	*45*	*10*

a) Rappresentare graficamente la distribuzione del colesterolo

Il carattere "quantità di colesterolo" è di tipo quantitativo continuo, suddiviso in classi, pertanto la rappresentazione grafica opportuna è l'istogramma. Per fare ciò bisogna calcolare l'ampiezza e la densità delle classi.

Colesterolo $[x_i, x_{i+1}]$	Frequenze assolute n_i	Ampiezza dell'intervallo a_i	Densità di frequenza $\frac{n_i}{a_i}$	Valori centrali $c_i=(x_i+x_{i+1})/2$
[120, 160]	60	40	1,5	140
]160, 180]	22	20	1,1	170
]180, 200]	30	20	1,5	190
]200, 240]	30	40	0,75	220
]240, 300]	55	60	0,91	270

La rappresentazione per istogrammi avviene costruendo tanti rettangoli quante sono le classi, le cui basii hanno lunghezza uguale all'ampiezza di classe, con gli estremi negli estremi di classe, e le cui altezze sono pari alla densità di classe: la base in ascissa è la classe di modalità, l'altezza in ordinata è la densità di frequenza della classe i-esima



b) Calcolare la media del colesterolo per ciascuno dei due sessi

I dati sono raggruppati in classi (variabili continue), allora si sostituisce alla classe di modalità $]x_{i-1}; x_i]$ il corrispondente valore centrale $c_i=(x_i+x_{i-1})/2$ con i=1….k
quindi si può determinare solo una media approssimata nell'ipotesi che tutte le unità statistiche della classe assumano una intensità pari al valore centrale di classe. Di conseguenza per determinare la media aritmetica approddimata di utilizza l'espressione:

$$\mu_x = \frac{\sum_{i=1}^{k} c_i n_i}{\sum_{i=1}^{k} n_i}$$

μ_M = (140*40 + 170*10 +190*20 + 220*10 +270*45)/125 =
(5600+1700+3800+2200+12150)/125 = 25450/125 = 203,6

μ_F= (140*20 + 170*12 +190*10 + 220*20 +270*10)/72=
(2800+2040+1900+4400+2700)/72= 13840/72 = 192,22

c) Calcolare la classe modale del colesterolo per i maschi

La classe (o le classi modali) sono quelle con densità di frequenza più elevate

Colesterolo	Frequenze assolute (maschi) n_i	Ampiezza di classe a_i	Densità di frequenza $\frac{n_i}{a_i}$
[120, 160]	40	40	1
]160, 180]	10	20	0,5
]180, 200]	20	20	1
]200, 240]	10	40	0,25
]240, 300]	45	60	0,75

Ci sono due classi modali: [120, 160] e]180, 200], come si evidenzia anche nell'istogramma.

Esercizio 10.

In un collettivo di studenti è stato rilevato il voto riportato all'esame di Statistica e quello riportato all'esame di Storia Contemporanea:

Studente	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Voto a Statistica (X)	*28*	*22*	*18*	*18*	*20*	*30*	*20*	*23*	*23*	*27*
Voto a Storia Contemporanea (Y)	*30*	*28*	*27*	*18*	*28*	*28*	*28*	*27*	*27*	*18*

a) Costruire la distribuzione doppia di frequenze (*X,Y)*
Si tratta di una tabella a doppia entrata, che registra quante volte (cioè la frequenza assoluta) una coppia di modalità (x_i, y_j) si presenta contemporaneamente per le unità statistiche
È una distribuzione bivariata

Voto a statistica (X)	Voto a storia contemporanea (Y)				
	18	27	28	30	*Tot di riga*
18	1	1	0	0	2
20	0	0	2	0	2
22	0	0	1	0	1
23	0	2	0	0	2
27	1	0	0	0	1
28	0	0	0	1	1
30	0	0	1	0	1
Tot di colonna	2	3	4	1	10

I totali per riga e per colonna sono le frequenze corrispondenti alle variabili X e Y e sono definite frequenze assolute marginali (cioè distribuzioni di frequenza marginali univariate di X eY)

b) Calcolare il voto mediano dell'esame di Statistica
Per prima cosa occorre ordinare i voti riportati
18, 18, 20, 20, 22, 23, 23, 27, 28, 30

Il numero di unità statistiche è di N=10 quindi pari; allora si devono considerare i voti riportati dalle unità statistiche che occupano le posizioni N/2=5° e (N/2)+1=6° e fare la media
Me = (22+23)/2 = 22,5

c) Stabilire se vi è indipendenza in media di *X* ***da*** *Y*
Vi è indipendenza in media di X da Y, se al variare di Y le medie condizionate di X rimangono costanti (tuttavia si noti che l'indipendenza in media di Y da X non implica l'indipendenza in media di X da Y, ovvero l'indipendenza in media non è una relazione simmetrica)
Per calcolarlo occorrono:
1) Media aritmetica $\bar{x} = \mu_x = (18*2 + 20*2 + 22 + 23*2 + 27 + 28 + 30)/10 = 22,9$
2) Le medie condizionate: cioè media della variabile X condizionata al valore y_j assunto dalla variabile Y , sono gli $\bar{x}_i$ e dato che y_j=18, 27, 28, 30
$\mu_{X|Y=18} = (18+27)/2 = 22,5$

$\mu_{X|Y=27}$ = (18+23*2)/ 3= 21,33
$\mu_{X|Y=28}$ = (20*2+22+30)/ 4= 23
$\mu_{X|Y=30}$ = (28)/1= 28
Le medie condizionate non sono costanti quindi si può dire che non c'è indipendenza in media

Esercizio 11.

In un collettivo di giovani si è osservato l'atteggiamento verso il fumo per classi di età ottenendo la seguente distribuzione di frequenze:

	Classi di età			
	[16 , 18]	***]18 , 22]***	***[22 , 25]***	***[25 , 30]***
Fuma	*7*	*8*	*21*	*30*
Non fuma	*16*	*18*	*9*	*10*

a) Calcolare la classe modale per l'età di chi fuma e di chi non fuma.

	Frequenze assolute Fumatori n_{i1}	Frequenze assolute Non fumatori n_{i2}	Ampiezza della classe a_i	Densità di frequenza Fumatori $\frac{n_{i1}}{a_i}$	Densità di frequenza Non fumatori $\frac{n_{i2}}{a_i}$	**TOT**
[16, 18]	7	16	2	3,5	8	23
]18, 22]	8	18	4	2	4,5	26
]22, 25]	21	9	3	7	3	30
]25, 30]	30	10	5	6	2	40
TOT	66	53				119

Per chi fuma la classe modale è]22, 25] e per chi non fuma]16, 18]

b) Calcolare il rapporto di correlazione dell'età dall'atteggiamento verso il fumo.

$$\eta^2_{X|y} = \text{Dev (B)/Dev(X)} = \frac{\sum_{j=1}^{h}(\bar{x}_j - \bar{x})^2 n_{.j}}{\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_{i.}}$$

I calcoli occorrenti per la scomposizione della devianza sono i seguenti
1) media aritmetica dell'età
Si parte da questa tabella

Valori centrali di classe di età	tot
17	23
20	26
23,5	30

27,5	40

$\mu_{età}$= (17*23+ 20*26 +23,5*30 + 27,5*40)/119 = 22,8235

2) le medie dell'età condizionate per i fumatori e non fumatori)
$\mu_{fumatori}$= (17*7+ 20*8 +23,5*21 + 27,5*30)/66 = 24,2045
$\mu_{non\ fumatori}$= (17*16+ 20*18 +23,5*9 + 27,5*10)/53 = 21,1037

3) devianza tra i gruppi
Dev (B) = $[(24,2 - 22,8)^2 *66]$+ $[(21,1 - 22,8)^2 *53]$= 282,5

4) devianza totale
Dev (X) = $[(17 -22,8)^2 *23]$ + $[(20 - 22,8)^2 *26]$ + $[(23,5 - 22,8)^2 *30]$ + $[(27,5 - 22,8)^2 *40]$ = 1875,87
$\eta^2_{X|y}$= Dev (B)/Dev(X) = 282,5/1875,8 = 0,1506 allora η= 0,388

Esercizio 12.

In un collettivo di 420 volontari si è osservato la frequenza di attività di volontariato per classi di età ottenendo la seguente distribuzione di frequenze relative percentuali:

	Classi di età			
	[14 , 20]	***]20 , 35]***	***]35 , 55]***	***]55 , 60]***
Almeno una volta la settimana	*10*	*15*	*10*	*5*
Una o più volte al mese	*10*	*20*	*20*	*10*

a) Quanti sono i volontari con età superiore a 20 anni e non superiore a 55 anni.
Sono il (15+10+20+20)%=65% del totale di 420, cioè 273

b) Quanti sono i volontari che prestano la loro attività almeno una volta la settimana e che hanno un'età superiore a 55 anni e non superiore a 60 anni.
Sono il 5% dei 420 in totale, cioè 21

c) Determinare il rapporto di correlazione dell'età dalla regolarità del servizio di volontariato.

$$\eta^2_{X|y}= \text{Dev (B)/Dev(X)} = \frac{\sum_{j=1}^{h}(\bar{x}_j - \bar{x})^2 n_{.j}}{\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 ni_{.}}$$

Tenuto conto che $\mu_x = \frac{\Sigma x_i n_i}{N} = \Sigma x_i \frac{n_i}{N}\frac{100}{100} = \frac{1}{100}\Sigma x_i p_i$ dove p_i è la percentuale di unità statistiche corrispondenti alla modalità $x_{i,}$, i calcoli occorrenti per la scomposizione della devianza sono i seguenti, partendo da questa tabella

Valori centrali di classe di età	Frequenza %
17,0	20
27,5	35
45,0	30

57,5	15
Tot	

$\mu_{età}$= (17*20+ 27,5*35 +45*30 + 57,5*15)/100 = (340+962,5+1350+862,5)/100 = 3515/100=35,15

Occorre poi determinare le medie condizionate per i due tipi di volontari:

$\mu_{1 \text{ volta a settimana}}$= (17*10+ 27,5*15 +45*10 + 57,5*5)/40 = (170+412,5+450+287,5)/40 =1320/40=33

$\mu_{1 \text{ volta o + al mese}}$= (17*10+ 27,5*20 +45*20 + 57,5*10)/60 = (170+550+900+575)/60 =2195/60=36,58

Dev (B) = $[(33 - 35,15)^2 *40]$+ $[(36,58 - 35,15)^2 60]$ = (4,6225*40)+(2,0449*60) = 184,9+122,694 = 307,594

Dev (X) = $[(17 -35,15)^2 *20]$ + $[(27,5 - 35,15)^2 *35]$ + $[(45 - 35,15)^2 *30]$ + $[(57,5 - 35,15)^2 *15]$ = (329,4225*20) + (58,5225*35) + (97,0225*30) +(499,5225*15) = 6588,45+2048,28+2910,67+ 7492,83 = 19040,23

$\eta^2_{X|y}$= Dev (B)/Dev(X) = 307,59/19040,23 = 0,01615

$\eta_{X|y}$ = 0,127

Esercizio 13.

Su un collettivo di individui sono stati rilevati i caratteri ***X*** *Peso(in kg) e* ***Y*** *Altezza (in cm) otteniamo la seguente distribuzione di frequenza congiunta:*

	Y		
X	*165*	*170*	*175*
60	*2*	*0*	*0*
70	*0*	*1*	*0*
80	*1*	*0*	*1*

a) Ricostruire la successione dell'altezza

165 165 165 170 175

b) Calcolare la media e la mediana dell'altezza

Essendo N=5 la mediana è la modalità che occupa il terzo posto nella successione ordinata:
Me=165

Per calcolare la media è necessario ottenere per ciascuna modalità di Y la frequenza marginale

Y	n_i
165	3
170	1
175	1

Dunque:
$\mu = (165*3+170+175)/5=840/5=168$

c) Calcolare il peso medio per gli individui che hanno un'altezza di 165 cm

$\mu_{X|Y=165}$=(60*2+80)/3=66,66

d) Calcolare il coefficiente di correlazione lineare tra peso e altezza

Il coefficiente di correlazione lineare è:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Per il suo calcolo è necessario quindi calcolare la varianza di X (peso):

$$\sigma_X^2 = \frac{1}{n}\left[\sum_{i=1}^{3}(x_i - \mu_x)^2 n_i\right].$$

La media aritmetica del peso risulta:

μ_X=(60*2 + 70*1 +80*2)/5=(120+70+160)/5=350/5=70

da cui:

σ^2_x= 1/5[(60-70)^2 *2 + (70-70)^2*1 + (80-70)^2 *2]=1/5(200+200)=80

La varianza di Y (altezza) è:

$$\sigma_Y^2 = \frac{1}{n}\left[\sum_{i=1}^{3}(y_i - \mu_y)^2 n_i\right]$$

σ^2_Y=1/5 [(165-168)^2*3 + (170-168)^2*1 + (175-168)^2*1]= 1/5(27+4+49)=80/5=16

La covarianza risulta dalla formula:

Cov (X, Y)=$\frac{1}{n}\sum_{i=1}^{3}\sum_{j}^{3}(x_i - \mu_x)(y_j - \mu_y)n_{ij}$

Quindi Cov (X, Y)= [(60-70)(165-168)*2 + (80-70)(165-168)*1 + (70-70)(170-168)*1 + (80-70)(175-168)*1] /5= (60-30+0+70)/5=100/5=20

Il coefficiente di correlazione lineare risulta

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = 20/\sqrt{80*16} = 20/\sqrt{1280} = 0{,}56$$

B. ESERCIZI DI PROBABILITÀ E VARIABILI CASUALI

Premessa

Al fine di poter effettuare una estrazione casuale di una unità statistica del collettivo si può pensare di associare a ciascuna di essa una pallina, di diametro costante e di un dato materiale in determinate condizioni fisico-chimiche, sulla quale annotare genere e stato occupazionale. Le 400 palline così costruite vengono inserite in una scatola e mescolate accuratamente. La prova consiste nell'estrarre una sola pallina dalla scatola. In queste condizioni ciascuna pallina ha la stessa probabilità di essere estratta. Si è così costruito uno spazio di eventi (le 400 palline) necessari (una pallina verrà estratta), incompatibili (una sola pallina verrà estratta) ed equiprobabili (ciascuna pallina ha la stessa probabilità di essere estratta): ciascuna pallina ha probabilità 1/400 .

Il problema di calcolare la probabilità di estrarre una pallina con una particolare annotazione, per esempio femmina, si risolve considerando tale annotazione (femmina) come un evento composto dalla disgiunzione di un numero k (le 250 palline con femmina) di eventi incompatibili ed equiprobabili e quindi la sua probabilità sarà data dalla somma delle probabilità di questi k eventi equiprobabili cioé $\frac{k}{400}$ (la probabilità di femmina sarà 250 /400), ovvero dal rapporto fra il numero di casi favorevoli (le 250 palline con femmina) e il numero di casi possibili (le 400 palline).

Dalle considerazioni esposte si può concludere che la frequenza relativa di una modalità di un carattere può essere vista come la probabilità di un evento: quello individuato dalla modalità fissata.

B1. Calcolo delle probabilità

Esercizio 1.

In una popolazione di 400 laureati in Scienze Politiche la distribuzione secondo il sesso e lo stato lavorativo a due anni dalla laurea è la seguente:

	Occupato	***Disoccupato***
Maschio	*100*	*50*
Femmina	*150*	*100*

Si estrae a caso un laureato.

a) Quale è la probabilità che sia disoccupato?

	O	*D*	
M	100	50	150
F	150	100	250
	250	150	400

Eventi:
A={essere disoccupato}
B={essere maschio}

Dalla premessa risulta
Pr(A)=150/400=0,375

b) Quale è la probabilità che sia disoccupato e maschio?
Pr(A∩B)= 50/400=0,125

c) Quale è la probabilità che sia disoccupato dato che è stato estratto un maschio?
Pr(dis|maschio)=Pr(A|B)=$\frac{\Pr(A \cap B)}{\Pr(B)} = \frac{50/400}{150/400}$ =50/150=0,125

Esercizio 2.

Un collettivo di 200 studenti è stato classificato secondo il voto riportato ad un esame e se è o meno il primo esame come segue:

	Primo esame	
Voto	***si***	***no***
= 24	*40*	*15*
= 25	*45*	*100*

Si estrae a caso dal collettivo uno studente.

*Sia **A** l'evento «voto = 24» e **B** l'evento «è il primo esame».*

a) Calcolare Pr(A)

	SI 1° esame	*NO*	
≤24	40	15	55
≥25	45	100	145
	85	115	200

Pr(A)= 55/200=0,275

b) Calcolare Pr(B)
Pr(B)= 85/200=0,425

c) Calcolare Pr($A \cup B$)
Pr ($A \cup B$)= Pr(A)+Pr(B) - Pr($A \cap B$)=55/200+85/200 – 40/200=100/200=0,5

d) Calcolare Pr($B \mid A$)
Pr(B|A)=)= $\frac{\Pr(A \cap B)}{\Pr(A)} = \frac{40/200}{55/200}$ = 40/55=0,727

Esercizio 3.

Un collettivo di 200 donne è stato classificato secondo lo stato civile e l'età come segue:

	Stato civile	
Età	***Nubile***	***Coniugata***
fino a 25	*40*	*15*
pi□ di 25	*45*	*100*

Si estrae dal collettivo casualmente una donna.
*Sia **A** l'evento «avere una età fino a 25» e **B** l'evento «essere coniugata».*

a) Calcolare Pr(A)

	N	*C*	
Fino 25	40	15	55
+ di 25	45	100	145
	85	115	200

Eventi:
A={avere un'età fino a 25 anni}
B={essere coniugata}
Pr(A)= 55/200=0,275

b) Calcolare Pr(B)
Pr(B)= 115/200=0,575

c) Calcolare Pr(A∪B)
Pr (A∪B)= Pr(A)+Pr(B) - Pr(A∩B)= 55/200+115/200-15/200=155/200=0,775

d) Calcolare Pr(B|A)
Pr(B|A)=)= $\frac{\Pr(A \cap B)}{\Pr(A)} = \frac{15/200}{55/200}$ 15/55=0,273

e) Calcolare Pr(A∩B)
Pr(A∩B)= 15/200= 0,075
(vale anche Pr(A∩B)=Pr(A) Pr(B|A)= 0,275* 0,273=0,075)

f) A e B sono eventi indipendenti?
Due eventi si dicono stocasticamente indipendenti se
Pr(A|B)=Pr(A) o Pr(B|A)=Pr(B) o Pr(A∩B)=Pr(A)Pr(B)

Essendo Pr(A|B)=15/115=0,130 ≠ Pr(A) e Pr(B|A)=15/55=0,27 ≠ Pr(B)
i due eventi sono dipendenti

Esercizio 4.

Un collettivo di 200 giovani è stato classificato secondo lo stato civile e la condizione lavorativa come segue:

	Stato civile	
Condizione lavorativa	***Celibe***	***Coniugato***
lavora	*50*	*60*
non lavora	*70*	*20*

Si estrae dal collettivo casualmente una uomo.
*Sia **A** l'evento «non lavora» e **B** l'evento «essere celibe».*

a) Calcolare Pr(A)

	Celibe	*Coniugato*	
L	50	60	110
NL	70	20	90
	120	80	200

Pr(A)= 110/200=0,55

b) Calcolare Pr(A∩B)
Pr(A∩B)= 60/200= 0,3 e infatti anche Pr(A∩B)=Pr(A) Pr(B|A)=0,55*0,55=0,3
Perché Pr(B|A)=60/110=0,55

c) A e B sono eventi indipendenti?
Due eventi si dicono stocasticamente indipendenti se
Pr(A|B)=Pr(A) o Pr(B|A)=Pr(B) o Pr(A∩B)=Pr(A)Pr(B)

Dato che risulta
P(A|B)=60/80=0,75 ≠ P(A)=0,55 e P(B|A)=60/110=0,55 ≠ P(B)=80/200=0,4 i due eventi sono dipendenti

d) Calcolare Pr(A∪B)
Pr (A∪B)= Pr(A)+Pr(B) - Pr(A∩B)= 110/200 + 80/200 – 60/200=120/200=0,6

Esercizio 5.

Delle auto prodotte da una certa casa automobilistica si sa che 1 su 100 presenta difetti di carrozzeria e che 4 su 180 presentano difetti meccanici, inoltre fra le auto con difetti di carrozzeria la probabilità di trovarne una con difetti meccanici è pari a 0.002.

Calcolare la probabilità di produrre un'auto con difetti di un tipo o dell'altro.

Eventi:
A={difetti di carrozzeria}
B={difetti meccanici}

Sappiamo che:
Pr(A)= 1/100=0,01
Pr(B)= 4/180=0,022
Pr(B|A)=0,002

Da cui risulta:
Dato che $Pr(A \cap B)= Pr(A)\ Pr(B|A)=0,01*0,002=0,00002$
e dunque $Pr(A \cup B)= Pr(A)+Pr(B)- Pr(A \cap B)=0,01+0,022-0,00002=0,0319$

Esercizio 6.

In un collettivo di 600 studenti dell'Università di Firenze consideriamo i seguenti eventi:
***A** = {ha superato l'esame di Economia}* ***B** = {frequenta il corso di Statistica}.*
Sapendo che 400 studenti hanno superato l'esame di Economia, 300 studenti frequentano e che 200 sono gli studenti che hanno superato l'esame di Economia e frequentano il corso di Statistica

a) Calcolare Pr(A)
Pr(A) = 400/600 = 0,66

b) Calcolare $Pr(A \cap B)$
$Pr(A \cap B)$ = 200/600 = 0,33

c) Calcolare $Pr(A \cup B)$
$Pr(A \cup B) = Pr(A)+Pr(B)- Pr(A \cap B)= 400/600+300/600-200/600=500/600=0,83$
dato che appunto Pr(B) = 300/600

Esercizio 7.

Per un paziente con certi sintomi si considerino i seguenti eventi:

***A** := { ha l'influenza }* ***B** := { ha la polmonite }* ***C** := { ha la febbre a 40}*

sapendo che:

$A \cap B = \phi$ $A \cup B = I$ $P(A) = 0.7$ $P(C|A)=0.3$ $P(C|B)=0.8$

a) Qual è la probabilità che il paziente abbia la polmonite?
Essendo *I* l'evento certo risulta che:
Pr(B)=1-Pr(A)=0,3

b) Qual è la probabilità che abbia l'influenza se ha la febbre a 40 ?
Pr(A|C)= Pr(A∩C)/P(C)
Allora, dal teorema delle probabilità totali, Pr(C) = Pr(C|A)Pr(A) + Pr(C|B)Pr(B) =
0,3*0,7+0,8*0,3=0,21+0,24=0,45 e Pr(A∩C)=Pr(C|A)Pr(A)=0,21
Quindi Pr(A|C)=0,21/0,45=0,47 (teorema di Bayes)

Esercizio 8.

Uno studente al primo anno di università vuole conoscere le sue possibilità di laurearsi entro 4 anni. Gli vengono fornite le seguenti informazioni:

1) *il 15% degli iscritti si laurea entro 4 anni;*
2) *su 10 laureati entro 4 anni 6 hanno riportato il massimo dei voti all'esame di diploma di scuola media superiore;*
3) *su 100 laureati con tempi superiori ai 4 anni 10 hanno riportato il massimo dei voti all'esame di diploma di scuola media superiore.*

Sapendo che lo studente in questione ha riportato il massimo dei voti all'esame di diploma di scuola media superiore, qual è al probabilità che si laurei entro 4 anni?

Eventi:
A={laurea entro 4 anni}
B={voto massimo}
Sappiamo che
Pr(A)= 15/100=0,15
Pr(B|A)=0,6
Pr(B|$\overline{A}$)=0,1
Devo quindi ricavare Pr(A|B)=Pr(A∩B)/Pr(B)
Pr(A∩B)=Pr(A)Pr(B|A)=0,15*0,6=0,09.

Pr(B)=Pr(B∩$\overline{A}$)+Pr(A∩B)= Pr(B|A)Pr(A)+ Pr(B|$\overline{A}$)Pr($\overline{A}$) (teorema delle Probabilità totali)
da cui risulta:
0,09+0,1*(1-0,15)= 0,09+0,085=0,175

Dunque Pr(A|B)= Pr(A∩B)/Pr(B)=0,09/0,175=0,514

Esercizio 9.

Un giovane deve decidere se iscriversi all'Università per conseguire una laurea o mettersi sul mercato del lavoro. Egli sa che tra i giovani lavoratori il 30% ha la laurea mentre tra i disoccupati il 20% è laureato. Inoltre, data la situazione economica, la probabilità di trovare lavoro è 0.8.

Consiglierebbe al giovane di iscriversi all'Università per conseguire una laurea?

Eventi:
A={lavora}
B={ha la laurea}
Sappiamo che:
Pr(B|A)=0,3
Pr(B|$\overline{A}$)=0,2
Pr(A)=0,8
Pr ($\overline{A}$)=0,25

Devo confrontare Pr(A|B) con la Pr(A|$\overline{B}$)
E quindi ricavare Pr(A|B)=Pr(A∩B)/Pr(B)
Pr(A∩B)=Pr(A)Pr(B|A)=0,8*0,3=0,24

Pr(B)=Pr(B∩$\overline{A}$)+Pr(A∩B)= Pr(B|A)Pr(A)+ Pr(B|$\overline{A}$)Pr($\overline{A}$) (teorema delle Probabilità totali)
da cui risulta:
0,3*0,8+0,2*0,25)= 0,24+0,05=0,29

allora Pr(A|B)= Pr(A∩B)/Pr(B)=0,24/0,29=0,83

l'83% dei laureati è occupato, che risulta superiore alla % di occupati tra i non laureati e quindi conviene laurearsi
Infatti Pr(A|$\overline{B}$)= Pr(A∩$\overline{B}$)/Pr($\overline{B}$)=[Pr(A)-Pr(A∩B)]/[1-Pr(B)]=(0,8-0,24)/(1-0,29)=0,56/0,71=0,78

Esercizio 10.

In un ufficio le pratiche relative ad una certa procedura amministrativa vengono affidate ***casualmente*** *a tre impiegati che indicheremo con A,B,C. La probabilità che una pratica venga completata entro una settimana per ciascun impiegato è indicata nella tabella che segue:*

Impiegato	*A*	*B*	*C*
Probabilità	*0.4*	*0.8*	*0.3*

Avendo ricevuto una pratica espletata entro una settimana qual è, secondo voi, l'impiegato al quale era stata affidata?

Sappiamo che
Pr(1 settimana|A)=0,4
Pr(1 settimana|B)=0,8
Pr(1 settimana|C)=0,3
Pr(A)=Pr(B)=Pr(C)=1/3=0,33
Bisogna confrontare
Pr(A|1 settimana) con Pr(B|1 settimana) con Pr(C|1 settimana)
Dal teorema di Bayes
Pr(A|1 set)=Pr(A)Pr(1 set|A)/[Pr(A)Pr(1 set|A)+Pr(B)Pr(1 set|B)+ Pr(C)Pr(1 set|C)]=
0,33*0,4/(0,33+0,4+0,33*0,8+0,33*0,3)=0,132/(0,132+0,264+0,099)=0,132/0,495=0,266
Pr(B|1 set)= Pr(B)Pr(1 set|B)/[Pr(A)Pr(1 set|A)+Pr(B)Pr(1 set|B)+ Pr(C)Pr(1 set|C)]=
0,264/0,495=0,533
Pr(C|1 set)= Pr(C)Pr(1 set|C)/[Pr(A)Pr(1 set|A)+Pr(B)Pr(1 set|B)+ Pr(C)Pr(1 set|C)]=
0,099/0,495=0,2

B risulta l'impiegato più probabile

Esercizio 11.

Si consideri un mazzo di 40 carte costituito da 10 carte per ciascun seme (♥ ♦ ♣ ♠) e per ciascun seme le carte sono numerate da 1 a 10. Si estraggano da tale mazzo due carte senza reintroduzione.

a) Calcolare la probabilità che entrambe siano ♥

Eventi:
C_1={esce una carta di cuori alla prima estrazione}
C_2={ esce una carta di cuori alla seconda estrazione}

Pr(entrambe cuori)=Pr($C_1 \cap C_2$)
Ω={($C_1 \cap C_2$)∪ ($C_1 \cap \overline{C}_2$))∪ ($\overline{C}_1 \cap \overline{C}_2$))∪ ($\overline{C}_1 \cap C_2$)}
Pr($C_1 \cap C_2$)=Pr(C_1)Pr($C_2|C_1$)=(10/40)*(9/39)=0,25*0,23=0,0575

b) Calcolare la probabilità che la seconda sia ♠ (P_2) dato che la prima è un 2 (2_1)

Pr($P_2|\ 2_1$)=(9/39)*(1/4)+(10/39)*(3/4)=9/156+30/156=39/156=1/4=0,25
oppure si può considerare Pr($P_2|\ 2_1$)= Pr($P_2 \cap 2_1$)/ Pr(2_1)=(1/40)/(4/40)=1/4
dato che Pr (2_1)=4/40 e Pr($P_2 \cap 2_1$)=1/40

c) Calcolare la probabilità che la seconda sia ♦

Pr(Q_2)=Pr(($Q_1 \cap Q_2$)∪ ($\overline{Q}_1 \cap Q_2$))= =Pr(($Q_1 \cap Q_2$) + ($\overline{Q}_1 \cap Q_2$))= Pr(Q_1)Pr($Q_2|Q_1$) + Pr($\overline{Q}_1$)Pr($Q_2|\overline{Q}_1$)
=(10/40*9/39)+(30/40*10/39)=0,0575+(0,75*0,256)=0,192

Esercizio 12.

Vengono estratte, senza reintroduzione, tre carte da un mazzo di 52 contenente 13 carte di ciascun seme (fiori, quadri, picche, cuori), per ciascun seme le carte sono contrassegnate dai numeri da 2 a 10, da fante, regina, re, asso.

a) Trovare la probabilità che abbiano tutte lo stesso contrassegno

(Pr 3 carte stesso contrassegno)=13 (0,076*0,058*0,04)=0,0023
dato che infatti ad esempio
Pr (3 assi)=Pr($A_1 \cap A_2 \cap A_3$)=4/52 * 3/51 *2/50 = 0,076*0,058*0,04

b) Trovare la probabilità che nessuna sia asso

Pr(nessuna sia asso)=Pr($\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3$)= Pr($\overline{A}_1$) Pr($\overline{A}_2|\overline{A}_1$) Pr($\overline{A}_3|\overline{A}_1,\ \overline{A}_2$) =
=(1-Pr(A_1))* (1- Pr($\overline{A}_2|\overline{A}_1$))* (1-Pr($\overline{A}_3|\overline{A}_1,\ \overline{A}_2$)) =
(1-4/52)*(1- 3/51) *(1-2/50)=0,924*0,942*0,96=0,835

B2.1. Variabili casuali discrete

Esercizio 1.

Vi propongono di giocare al seguente gioco: si lanciano due monete, se si verificano due teste si vince 1 euro, se si verificano due croci si vince 0.5 euro, in tutti gli altri casi non si vince nulla. Per partecipare al gioco si paga 0.5 euro.

Conviene giocare a questo gioco? (calcolare la vincita media)

Lo spazio Ω dei possibili risultati è costituito da:
E_1={TT, CC, CT, TC }

A ciascuno dei risultati è attribuibile una probabilità pari ad ¼.

La variabile casuale della quale dobbiamo calcolare il valore atteso assume valore 1 se si verifica {TT} (con probabilità ¼), valore 0,5 se si verifica {CC} (con probabilità ¼), e valore 0 se si verifica {CT}∪{TC} (con probabilità 2/4).

La vincita media è quindi 1*0,25 + 0,5*0,25=0,375 euro
Non conviene giocare al gioco dato che la vincita media è inferiore al costo di partecipazione al gioco.

Esercizio 2.

Un'urna contiene palline bianche e nere con probabilità rispettivamente uguale 0.3 e 0.7. La prova consiste nell'estrarre ripetutamente una pallina dall'urna rimettendo la pallina nell'urna dopo ogni estrazione.

a) Calcolare la probabilità di ottenere la prima pallina bianca alla decima estrazione.
Eventi:
PN={esce una pallina nera}
PB={ esce una pallina bianca}

Pr(PB)=0,3
Pr(PN)=0,7
Pr (B alla decima estrazione) =
Pr(PN)+Pr(PN)+Pr(PN)+Pr(PN)+Pr(PN)+Pr(PN)+Pr(PN)+Pr(PN)+Pr(PN)+Pr(PB)= $(0,7)^9$ *0,3

b) Calcolare la probabilità di ottenere la prima pallina bianca fra la settima e la nona estrazione.
L'evento "la prima PB fra la 7 e la 9 estrazione" è l'unione dei seguenti eventi incompatibili:
E_1={NNNNNNB}, $Pr(E_1)=(0,7)^6$*0,3
E_2={NNNNNNNB}, $Pr(E_2)=(0,7)^7$*0,3
E_3={NNNNNNNNB}, $Pr(E_3)=(0,7)^8$*0,3

Allora $Pr(E_1 \cup E_2 \cup E_3) = (0,7)^6 * 0,3 + (0,7)^7 * 0,3 + (0,7)^8 * 0,3 = [((0,7)^6 * 0,3][(1 + 0,7 + (0,7)^2]$.

Esercizio 3.

Un'urna contiene 7 palline gialle e 3 rosse.

a) Calcolare la probabilità che, estraendo dall'urna due palline senza reintroduzione, alla seconda estrazione si verifichi pallina gialla

7 gialle e 3 rosse, estrazioni senza ripetizione
$Pr(G_2)=Pr((R_1 \cap G_2) \cup (G_1 \cap G_2))= Pr((R_1 \cap G_2) + (G_1 \cap G_2))=7/10*6/9 + 3/10*7/9=0,7$

b) Calcolare la probabilità che, estraendo dall'urna due palline senza reintroduzione, si verifichi pallina rossa alla prima estrazione e gialla alla seconda
$Pr (R_1 \cap G_2)$?= $Pr(R_1)Pr(G_2| R_1)=3/10*7/9=21/90=0,23$

Esercizio 4.

La proporzione di studenti di una certa Facoltà che hanno superato un determinato esame è 0.4 e si ipotizza di estrarre un campione casuale di 50 studenti della stessa Facoltà.

Stabilire la probabilità di ottenere una proporzione campionaria di studenti che hanno superato quell'esame pari a 0.4.

Dobbiamo calcolare la probabilità di ottenere 50, 04 successi in 50 prove di Bernoulli indipendenti, ciascuna con probabilità di successo θ pari a 0,4.
Utilizziamo la funzione di probabilità di una variabile casuale binomiale:

$Pr(X=x)=\binom{n}{x}\theta^x(1-\theta)^{n-x}$ per x=0, 1, 2, ..., n

In questo caso
θ=0,4
n=50
X/n=0,4
quindi x=0,4*50=20

$Pr(X=20)=\binom{50}{20}0,4^{20}(1-0,4)^{50-20}=\binom{50}{20}0,4^{20}0,6^{30}=\frac{50!}{20!(50-20)!}0,4^{20}0,6^{30}$

B2.2. Variabili casuali continue

Esercizio 1.

Il tempo di percorrenza di un tratto autostradale è descritto da una variabile casuale con la seguente distribuzione di probabilità:

***Tempo** (minuti)*	***Probabilità***
[15,20]	*0.15*
]20,23]	*0.25*
]23,27]	*0.40*
]27,31]	*0.20*

a) Calcolare la probabilità di percorrere il tratto autostradale in non più di 23 minuti

Sappiamo che
Pr(A)=0,15
Pr(B)=0,25
Pr(C)=0,40
Pr(D)=0,20

P(A∪B)= Pr(A) + Pr (B)=0,15+0,25=0,40
dato che A∩B=ø perché gli eventi sono incompatibili (il tratto di autostrada o è percorso in un tempo o nell'altro, non si possono verificare contemporaneamente i 2 eventi).

b) Calcolare la probabilità di percorrere il tratto autostradale in un tempo T tale che 20<T<27
Pr(B∪C)=0,25+0,40=0,65

Esercizio 2.

*La quantità **P** in grammi di farina erogati in ogni confezione da una macchina si distribuisce normalmente con media 500 g. e scarto quadratico medio 10 g.*

a) Calcolare la probabilità che vengano erogati meno di 485 g.

Sappiamo che P ~ N(500, 100) quindi σ=10

$$\Pr(P\leq 485)= \Pr\left(\frac{P-\mu}{\sigma}\leq\frac{485-\mu}{\sigma}\right)= \Pr\left(z\leq\frac{485-\mu}{\sigma}\right)= \Pr\left(z\leq\frac{485-500}{10}\right)= \Pr(z\leq -1{,}5)=$$

$\Phi(-1{,}5)=1-\Phi(1{,}5)= 1-0{,}933=0{,}067$
cioè il 6,7% di probabilità che vengano erogati meno di 485 grammi

b) Calcolare la probabilità che la quantità erogata sia compresa fra 490 g. e 512 g.

$$\Pr(490<P\leq 512) = \Pr\left(\frac{490-\mu}{\sigma}< z\leq\frac{512-\mu}{\sigma}\right)= \Phi\left(\frac{512-\mu}{\sigma}\right)-\Phi\left(\frac{490-\mu}{\sigma}\right)=$$

$\Phi\left(\frac{512-500}{10}\right)-\Phi\left(\frac{490-500}{10}\right)$= Φ(1,2)-Φ(-1)= Φ(1,2)-(1-Φ(1))=0,885-1+0,841= 0,727

La probabilità è del 72,7%

c) Stabilire quel peso P_0 per il quale la probabilità che la macchina eroghi una quantità di farina maggiore di P_0 è pari a 0.14.

Sappiamo che $Pr(P>P_0)=0,14$. Dobbiamo determinare P_0.

Si può partire dalla trasformazione $Pr(P>P_0)=1- Pr(P\leq P_0)=0,14$

$$1- Pr\left(\frac{P-\mu}{\sigma} \leq \frac{P_0-\mu}{\sigma}\right)= 0,14$$

$1- Pr(z \leq z_0)=1 - \Phi(z_0)=0,14$

$\Phi(z_0)=1- 0,14=0,86$

Dalle tavole ricavo $z_0=1,08$.

Essendo $z_0 = \frac{P_0-\mu}{\sigma}$, e quindi $1,08 = \frac{P_0-500}{10}$, dunque $P_0=500+10,8=510,8$

La probabilità che la macchina eroghi una quantità di farina maggiore di 510,8 grammi è pari a 0,14.

Esercizio 3.

I laureati di una certa facoltà hanno una votazione media di 100 con uno scarto quadratico medio di 4. Supponiamo che la distribuzione dei voti sia normale:

a) Calcolare la percentuale di laureati che ha ottenuto un voto compreso tra 96 e 104

Sappiamo che L ~ N(100, 16).

$\Pr(96<L\leq 104) = \Pr\left(\frac{96-\mu}{\sigma} < z \leq \frac{104-\mu}{\sigma}\right) = \Phi\left(\frac{104-\mu}{\sigma}\right) - \Phi\left(\frac{96-\mu}{\sigma}\right) =$

$\Phi\left(\frac{104-100}{4}\right) - \Phi\left(\frac{96-100}{4}\right) = \Phi(1)-\Phi(-1) = \Phi(1)-(1-\Phi(1)) = 2\,\Phi(1) - 1 = 2 * 0{,}841 - 1 = 0{,}682.$

Il 68,2% ha ottenuto un voto compreso tra 96 e 104.

b) Calcolare la percentuale di laureati che ha ottenuto un voto maggiore di 108

$\Pr(L>108) = 1 - \Pr(L\leq 108) = 1 - \Pr\left(\frac{L-\mu}{\sigma} \leq \frac{108-\mu}{\sigma}\right) = 1 - \Pr\left(z \leq \frac{108-\mu}{\sigma}\right) = \Pr\left(z \leq \frac{108-100}{4}\right) =$

$1-\Pr(z \leq 2) = 1 - \Phi(2) = 1 - 0{,}97725 = 0{,}02275$

Solo circa il 2,3% dei laureati ha ottenuto un voto maggiore di 108

c) Calcolare la differenza interquartile

Calcolo della differenza interquartile $D=Q_3-Q_1$

$\Pr(L \leq Q_3) = 0{,}75$

$\Pr\left(\frac{L-\mu}{\sigma} \leq \frac{Q_3-\mu}{\sigma}\right) = 0{,}75$

$\Pr(z \leq z_0) = 0{,}75$

$\Phi(z_0) = 0{,}75$ quindi dalle tavole $z_0 = 0{,}68$

$(Q_3 - 100)/4 = 0{,}68$ quindi $Q_3 = 100 + 0{,}68 * 4 = 100 + 2{,}72 = 102{,}72$

Q_1 è simmetrico rispetto a Q_3 rispetto alla media $\mu=100$. Quindi $Q_1 = 100 - 2{,}72 = 97{,}28$

$D = 102{,}72 - 97{,}28 = 5{,}44$

C. ESERCIZI DI INFERENZA STATISTICA

C1. Stima per intervalli

Esercizio 1.

Per analizzare la riuscita scolastica degli adolescenti si estrae un campione casuale semplice con reintroduzione di 600 studenti della prima classe superiore. In tale campione il numero di ragazzi bocciati è pari a 220.

Calcolare l'intervallo di confidenza al 90% per la percentuale dei bocciati nell'intera popolazione.

Indichiamo con X la v.c. che rappresenta l'esito della prima classe superiore:

$$X=\begin{cases}0 \ promosso\\ 1 \ bocciato\end{cases}$$

Con P(X=1)=p e P(X=0)=1-p. Se B è il numero di studenti bocciati in n prove Bernoullinae, la v.c. B/n, che è la proporzione di bocciati, segue una distribuzione binomiale relativa con parametro p: B/n~ Bin(n,p). Dato che la numerosità del campione è pari a 600, la variabile B/n può essere approssimata con una distribuzione normale $\frac{B}{n} \sim N\left(p;\ \frac{p(1-p)}{n}\right)$.

A questo punto possiamo quindi ricondurci al caso della stima di un intervallo di confidenza per la media di una normale con varianza incognita.

Il corrispondente intervallo di confidenza asintotico per il parametro P (percentuale di bocciati per l'intera popolazione) è dato da:

IC_α (P): $$\hat{P}_n - z_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}} \le P \le \hat{P}_n + z_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}} \text{ con } \hat{p}=\frac{1}{n}\Sigma x_i \text{ e } \hat{\sigma}=\hat{p}(1-\hat{p})$$

Ove $z_{\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ è il quantile (1- α/2) della v.c. Normale standardizzata

Quindi la stima dell'intervallo di confidenza è data da

IC_α (P): $$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \theta \le \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Dalle tavole della normale standardizzata

Se α=0,1 allora α/2=0,05

z=1,65

Tenendo conto di $\hat{p}$ =200/600=0,36, l'intervallo cercato è quindi

0,36±1,65*($\sqrt{0,23}$ / $\sqrt{600}$)= 0,36±0,79/24,49= 0,36±0,032=

0,328≤ p ≤0,392

Esercizio 2.

In una città ci sono 100000 persone di età compresa fra i 18 e i 25 anni; si estrae da questa popolazione un campione casuale semplice di 500 soggetti, 194 di questi risultano iscritti all'Università.

Determinare un intervallo di confidenza al 95% per la proporzione di persone con età compresa fra i 18 e i 25 anni che sono iscritte all' Università.

LA SOLUZIONE È ANALOGA AL PRECEDENTE ESERCIZIO

Indichiamo con X la v.c. che rappresenta la proporzione di iscritti e non iscritti all'università:

$$X=\begin{cases}0 \ \textit{non iscritto}\\ 1 \ \textit{iscritto}\end{cases}$$

Con P(X=1)=p e P(X=0)=1-p. Se B il numero di soggetti iscritti all'università, la v.c. B/n, che è la proporzione di iscritti, segue una distribuzione binomiale relativa con parametro p: B/n~ Bin(n,p). Dato che il campione è pari a 500 soggetti, la numerosità è tale che B/n può essere approssimata con una distribuzione normale

$$\frac{B}{n} \sim N\left(p;\ \frac{p(1-p)}{n}\right).$$

A questo punto possiamo quindi ricondurci al caso della stima di un intervallo di confidenza per la media di una normale con varianza incognita.

L'intervallo cercato è quindi $0{,}388 \pm 1{,}96 * (\sqrt{0{,}237}) / \sqrt{500} = 0{,}388 \pm 1{,}96 * 0{,}022 = 0{,}36 \pm 0{,}04312$

$0{,}345 \leq p \leq 0{,}431$

Esercizio 3.

Per studiare l'effetto della marijuana sulle capacità intellettuali di soggetti (senza esperienze precedenti) dei ricercatori hanno verificato su un campione di soggetti i cambiamenti nei punteggi ad opportuni test dopo aver fumato della marijuana. I risultati sono presentati nella seguente tabella:

Soggetto	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Differenza punteggio	*5*	*-17*	*-7*	*-3*	*-7*	*-9*	*-6*	*1*	*3*

Si determini l'intervallo di confidenza per la media della differenza dei punteggi al 99%.

Si tratta di un problema di stima di un intervallo di confidenza per la media di una popolazione con varianza incognita. In questo caso, poiché la numerosità campionaria è "piccola", non si può invocare il teorema del limite centrale, di conseguenza per risolvere il problema bisogna ipotizzare che la differenza dei punteggi si distribuisce normalmente e il corrispondente intervallo per il valore medio θ è dato da:

$IC_\alpha(\theta)$: $$\overline{X_n} - t_{\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}} \le \theta \le \overline{X_n} + t_{\frac{\alpha}{2}} \frac{S_n}{\sqrt{n}}$$

Ove $t_{\frac{\alpha}{2}}$ è il quantile di ordine 1-α/2 della v.c. t di Student con g=n-1 gradi di libertà

Dal campione osservato si determinano i valori $\overline{x_n}$ e s_n^2 (valori campionari rispettivamente di $\overline{X_n}$ e S_n^2), quindi l'intervallo di confidenza stimato sarà dato da

$IC_\alpha(\theta)$: $$\overline{x_n} - t_{\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}} \le \theta \le \overline{x_n} + t_{\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}$$

Dal campione abbiamo:
n=9
$\bar{x}$ =(5-17-7-3-7-9-6+1+3)/9=-4,4
$s_n=\sqrt{s_n^2}$ con $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ varianza campionaria corretta=
1/8 $[(5+4,4)^2 + (-17+4,4)^2 + (-7+4,4)^2 + (-3+4,4)^2 + (-7+4,4)^2 + (-9+4,4)^2 + (-6+4,4)^2 + (1+4,4)^2 + (3+4,4)^2]$ = 88,36+158,76+6,76+1,96+6,76+21,16+2,56+29,16+54,76= 370,24/8=46,28 Quindi s_n=6,803

L'intervallo cercato è $$-4,4 - t_{\frac{\alpha}{2}} \frac{6,8}{\sqrt{9}} \le \theta \le -4,4 + t_{\frac{\alpha}{2}} \frac{6,8}{\sqrt{9}}$$, cioè $4,4 \pm \frac{t_{\frac{\alpha}{2}} 6,8}{\sqrt{9}}$

Dato che $t_{0,005}$=3,355 (α=0,01 allora α/2=0,005 con n-1=8 gradi di libertà) l'intervallo stimato per θ sarà **12≤ θ ≤3,2**

C2. Verifica delle ipotesi

Si tratta di test sui parametri di una popolazione normale. In tutti gli esercizi proposti si ipotizza che la/le popolazione/i coinvolta dalla quale si effettua il campionamento sia normale.

Esercizio 1.

Una macchina per il riempimento delle buste di patatine ha uno scarto quadratico medio 6 grammi e una media incognita. La macchina è stata costruita per riempimento medio delle buste di patatine di 100 grammi. Per verificare la conformità di riempimento si estrae un campione di 100 buste ottenendo un contenuto medio di 99 grammi.

Effettuare un test delle ipotesi per stabilire se il riempimento medio di 100 grammi è accettabile al livello di significatività 0.05.

Il test da effettuare riguarda la media di una normale $X \sim N(\mu, \sigma^2)$ di cui è nota la varianza ($\sigma^2=6^2$) e si deve sottoporre a test l'ipotesi
H_0: $\mu=100$ (come da dichiarazione del produttore della macchina) contro H_1: $\mu\neq100$.
Il test è quindi bidirezionale, la statistica test può essere la media campionaria e in questo caso la regione critica sarà:

$$R_{co(\alpha)}:\begin{cases}\overline{X}\geq\mu_0+z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\\ oppure\\ \overline{X}\leq\mu_0-z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\end{cases}$$ con z~N(0,1) e $\Phi\left(z_{\frac{\alpha}{2}}\right)$=1- $\alpha/2$

Dato che il livello di significatività è α=0,05
$\Phi(z_{\alpha/2})$ =1- $\alpha/2$=1- (0,05/2)=0,975
$z_{\alpha/2}$ è quel valore della normale standardizzata determinato dalle tavole tenuto conto della condizione precedente
$z_{0,025}$=1,96
e tenuto conto inoltre che n=100 si ha

$$R_{co(\alpha)}:\begin{cases}\overline{X}\geq100+1{,}96\dfrac{6}{10}\\ oppure\\ \overline{X}\leq100-1{,}96\dfrac{6}{10}\end{cases}$$

e quindi la regione critica è

$$R_{co(0,05)}:\begin{cases}\overline{X}\geq101{,}76\\ oppure\\ \overline{X}\leq98{,}82\end{cases}$$

Di conseguenza la regione di non rifiuto $R_{co(0,05)}:\{98{,}82\leq\overline{X}\leq101{,}76\}$. Il campione estratto ci fornisce una stima della media campionaria $\overline{x}$=99 che non appartiene alla regione critica ma alla regione di non rifiuto, quindi, sulla base delle informazioni campionarie, al livello di significatività

del 5%, non può essere rifiutata l'ipotesi nulla e quindi la conformità del riempimento delle buste di patatine della macchina alla dichiarazione del produttore.

Esercizio 2.

Su un campione di giovani fra i 20 e 25 anni è stato rilevato X:="numero di libri letti in un anno" ottenendo i seguenti risultati campionari

X	*4*	*5*	*5*	*2*	*6*	*1*	*4*

Si può confutare l'ipotesi di un editore che il numero medio di libri letti è 2 al livello di significatività 0,05?

Si deve effettuare un sul valore medio di una popolazione la cui varianza è incognita.
Se ipotizziamo che X è distribuita normalmente con una media μ e una varianza σ^2, bisognerà effettuare un test delle ipotesi H_0: $\mu=2$ contro H_1: $\mu\neq 2$
Il test è bidirezionale e di conseguenza la regione critica è:

$$R_{co(\alpha)}:\begin{cases}\overline{X}\geq \mu_0 + t_{\left(\frac{\alpha}{2},g\right)}\dfrac{S_n}{\sqrt{n}} \\ \qquad\qquad oppure \\ \overline{X}\leq \mu_0 - t_{\left(\frac{\alpha}{2},g\right)}\dfrac{S_n}{\sqrt{n}}\end{cases}$$

con $t(\alpha/2, g)$ quantile di ordine $1-\alpha/2$ della v.c. t di Student con g=n-1 gradi di libertà.

Dalla rilevazione campionaria (n=7) possono essere calcolate la media e la varianza campionaria:
$\overline{x}$ =(4+5*2+2+6+1+4)/7=3,857
$s_n=\sqrt{s_n^2}$ con $s_n^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2$ =(1/6)[2*(4-3,857)^2+2*(5-3,857)^2+(2-3,857)^2+(6-3,857)^2+(1-3,857)^2=18,853/6=3,142
$s_n=\sqrt{3,142}$ =1,77

In questo caso α=0,05 e g=n-1=6 gradi di libertà e dalla tavola della t otteniamo $t_{0,025}$=2,447
Di conseguenza la regione di rifiuto è:

$$R_{co(0,05)}:\begin{cases}\overline{X}\geq 2+2{,}447\dfrac{1{,}77}{2{,}65}=3{,}64 \\ \overline{X}\leq 2-2{,}447\dfrac{1{,}77}{2{,}65}=0{,}36\end{cases}$$

$\overline{x}$ =3,857 appartiene alla regione di rifiuto, di conseguenza deve essere rifiutata l'ipotesi nulla che il numero medio di libri letto sia 2.

Esercizio 3.

In un campione di pazienti trattati con una terapia per l'abbassamento del colesterolo si sono osservati i seguenti valori di colesterolo in milligrammi per 100 millilitri di sangue:

130	*145*	*128*	*169*	*132*	*138*	*141*	*153*	*129*	*135*	*140*

Sapendo che in una popolazione di persone sane la quantità di colesterolo in media è pari a 130 cosa fareste per stabilire se la terapia adottata ha avuto effetto?

SOLUZIONE UGUALE AL PRECEDENTE ESERCIZIO

Le ipotesi da sottoporre a test sono H_0: μ=130 contro H_1: μ>130.
Assumendo α=0,05, la regione critica è data da:

$$R_{co(0,05)}: \left\{ \overline{X} \geq 130 + 1{,}812 \frac{12{,}22}{\sqrt{11}} = 136{,}67 \right.$$

$\overline{x} = 140$ è dentro la regione critica, quindi va rifiutata l'ipotesi nulla e confutata l'ipotesi che la terapia adottata abbia affetto (al livello di significatività scelto).

Esercizio 4.

Per un'indagine sul lavoro femminile sono state rilevate le ore lavorate giornalmente di un campione di 60 lavoratrici residenti in Toscana e di un campione di 45 lavoratrici residenti in Lombardia. I risultati sono i seguenti:

Regione	*Media Campionaria*	*Varianza campionaria*	*Numerosità campionaria*
Toscana	*5.5*	*4*	*60*
Lombardia	*6.5*	*9*	*45*

Verificare se le osservazioni campionarie possono suffragare l'ipotesi che in Toscana ci sia una tendenza maggiore all'uso del part-time (α=0.05).

Si tratta di un test di confronto tra 2 medie, cioè di un test sulla differenza dei valori medi con varianza non nota, ma che ipotizziamo uguale.
Si sottopone a verifica l'ipotesi H_0: $\mu_{Toscana} = \mu_{Lombardia}$ cioè $\mu_T - \mu_L = 0$ contro l'ipotesi H_1: $\mu_{Toscana} < \mu_{Lombardia}$ cioè $\mu_T - \mu_L < 0$
Le varianze delle popolazioni sono incognite, ma vengono ipotizzate uguali e quindi la regione critica: $RC(\alpha)$: $T_{n,m} \leq - t_{(\alpha,g)}$

La statistica test è $T_{n,m} = \dfrac{\overline{X}_n - \overline{Y}_m}{\sqrt{S_x^2(n-1) + S_y^2(m-1)}} \sqrt{\dfrac{nm(n+m-2)}{n+m}} = \dfrac{\overline{X}_n - \overline{Y}_m}{\tilde{S}^2_{n+m}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$

con $\tilde{S}^2_p = [(n-1)\,\tilde{S}^2_n + (m-1)\tilde{S}^2_m]/n+m-2$
T~ t di student con g=n+m-2 gradi di libertà

Poiché la numerosità campionaria è elevata i valori numerici dei quantili della T e della Z differiscono di poco, si può usare la tavola della Z, normale standardizzata.
E la regione critica in questo caso è $RC(\alpha)$: $T_{n,m} \leq - t_\alpha$
Dal campione osservato si hanno le seguenti informazioni:
$\bar{x}_T$=5,5; $\bar{x}_L$=6,5; s^2_T=4; s^2_L=9; n_T=60; n_L=45
Quindi la statistica test T= 5,5-6,5/$\sqrt{9/45 + 4/60}$ =-1/0,516=-1,93
Con α=0,05 allora Φ z_α=1- α=1-0,05=0,95 e di conseguenza, dalle tavole della z, t_α=1,65
La regione critica sarà quindi T<-1,65, e il valore T=–1,93<-1,65 quindi è compreso nella regione di rifiuto. L'ipotesi nulla va rifiutata e quindi si può suffragare l'ipotesi che in Toscana ci sia una maggiore tendenza al part-time.

Esercizio 5.

Si supponga di voler comparare la durata media delle lampadine prodotte da due fabbriche e di disporre delle seguenti informazioni campionarie

	Numerosità	*Durata media (ore)*	$\frac{S}{\sqrt{n}}$
Fabbrica A	*100*	*107*	*22*
Fabbrica B	*80*	*122*	*10*

Sottoporre a test l'ipotesi di uguaglianza fra le medie al livello di significatività 0.01

L'ESERCIZIO È SIMILE AL PRECEDENTE.

La regione critica è RC(α): $|T_{n,m}| \geq t_{\alpha/2}$
E quindi, in questo caso, con α=00,1 si ottiene Φ $t_{\alpha/2}$=1- α/2=1-0,01/1=1-0,005=0,995 dalle tavole della Z
$t_{\alpha/2}$=2,58 e la regione critica ottenuta è RC(0,0o1): $|T_{n,m}| \geq 2,58$
Calcolando la statistica-test T si ottiene T= 107-122/$\sqrt{220^2/100 + 89,44^2/80}$ =-15/24,166=-0,62207
Z sta dentro la regione di non rifiuto e quindi non si può rifiutare l'ipotesi nulla: cioè al livello di significatività dell'1% non si può rifiutare l'ipotesi che la durata media della lampadine prodotte sia uguale.

Esercizio 6.

Si è misurata la durata in ore delle pile prodotte da due diverse industrie su due campioni casuali estratti dalla produzione di pile delle due marche, i risultati campionari sono riportati nella tabella che segue:

Marca A	*1094*	*1137*	*1161*	*1092*	*1123*	*1084*
Marca B	*1159*	*1224*	*1153*	*1229*	*1260*	

Stabilire attraverso un test di ampiezza 0.05 se vi è differenza fra la durata delle pile delle due marche.

Si tratta di un test di confronto tra 2 medie: test sulla differenza dei valori medi con varianza non nota. Le due varianze vengono ipotizzate uguali

L'ipotesi da confrontare è H_0: $\mu_A-\mu_B=0$ contro H_1: $\mu_A-\mu_B\neq 0$

La regione critica RC(α): $\left|T_{n,m}\right| \geq t_{\frac{\alpha}{2}}$

la statistica test è $T_{n,m}= \dfrac{\overline{X}_n-\overline{Y}_m}{\sqrt{S_x^2(n-1)+S_y^2(m-n)}}\sqrt{\dfrac{nm(n+m-2)}{n+m}}=\dfrac{\overline{X}_n-\overline{Y}_m}{\tilde{S}^2_{n+m}\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}$

con $\tilde{S}^2_p$= [(n-1) S^2_n + (m-1)S^2_m]/n+m-2

T~ t di student con g=n+m-2 gradi di libertà

Dal campione osservato si ha:

n_A=5 e n_B=6

$\bar{x}_A$=(1159+1224+1153+1229+1269)/5=1205

$\bar{x}_B$=(1094+1137+1161+1092+1123+1084)/6=1115,16

$S_A^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$ =(1/4)[(1159-1205)^2+(1224-1205)^2+(1153-1205)^2+(1229-1205)^2+(1260-1205)^2=(2116+361+2704+576+3025)/4=8782/4=2195,5

$S_B^2=\dfrac{1}{m-1}\sum_{i=1}^{n}(x_i-\bar{x}_m)^2$ =(1/5)[(1094-1115,16)^2+(1137-1115,16)^2+(1161-1115,16)^2+(1092-1115,16)^2+(1123-1115,16)^2+(1084-1115,16)^2=
(447,75+476,98+2101,31+536,38+61,46+970,95)/5=4594,83/5=918,7

In questo caso di può calcolare quindi $\tilde{S}^2_p$= [(n-1) S^2_n + (m-1)S^2_m]/N+M-2 =
[(5-1)*2195,5+(6-1)*918,7]/(5+6-2) = (8782+4593,5)/9=13375,5/9=1486,16

$\tilde{S}_P=\sqrt{\tilde{S}_P^2}=\sqrt{1486,16}$=38,55

Di conseguenza la statistica-test T è pari a

T=(1205-115,16)/(38,55*$\sqrt{0,366}$)=89,84/23,34=3,849

Con α=0,05 e g= n+m-2=5+6-2=9 gradi di libertà, $t_{\alpha/2}= t_{0,005}$ = 3,25

E quindi la regione di rifiuto è RC(0,05): $\left|T_{a,b}\right| \geq 3{,}25$

La statistica-test T=3,849 è dentro la regione critica e quindi si deve rifiutare l'ipotesi nulla e confutare l'ipotesi che tra le due pile non vi sia differenza (a favore dell'ipotesi alternativa che tra la durata delle pile prodotte dalle due diverse marche sia diversa, al livello di significatività del 5%)

C3. Modello di regressione semplice

Esercizio 1.

Per una certa compagnia aerea relativamente ad una certa tratta si sono osservati, per alcuni voli scelti casualmente, i prezzi praticati in Euro ed il numero di passeggeri:

n° passeggeri	*100*	*130*	*110*	*90*	*80*	*120*
Prezzo	*129*	*147*	*130*	*102*	*109*	*140*

Trovare l'equazione della retta di regressione del prezzo dal numero di passeggeri.

Il modello di regressione lineare semplice è $Y_i=\beta_0+\beta_1 x_i+\varepsilon_i$
Bisogna stimare i due parametri β_0 e β_1.

Gli stimatori dei parametri con il metodo dei minimi quadrati sono dati da:

$$\begin{cases} \beta_1 = \dfrac{S_{xy}}{S_x^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}$$

Dal campione osservato, se X è il numero di passeggeri e Y il prezzo:
$\bar{x}$ = (100+130+110+90+80+120)/6=630/6=105
$\bar{y}$ = (129+147+130+102+109+140)/6=757/6=126,16

$s_{xy}= \dfrac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n} =$ [(100-105)(129-126,16)+(130-105)(147-126,16)+(110-105)(130-126,16)+(90-105)(102-126,16)+(80-105)(109-126,16)+(120-105)(140-126,16)]/6=254,17

$s^2_x = \dfrac{\Sigma(x_i-\bar{x})^2}{n} =$ [(100-105)^2+(130-105)^2+(110-105)^2)+(90-105)^2+(80-105)^2+(120-105)^2]/6=291,67
E in definitiva le stime dei parametri sono date da:

$\hat{\beta}_1 = \dfrac{s_{xy}}{s_x^2}$ =254,17/291,67=0,87

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ =126,16-0,87*105=34,8
La retta di regressione stimata è quindi **y=34,8 + 0,87x**

Esercizio 2.

I sette studenti di un corso di specializzazione hanno riportato le seguenti votazioni indue prove successive di uno stesso corso.

	Studente						
	1	*2*	*3*	*4*	*5*	*6*	*7*
Voto alla 1 prova (X)	*4,1*	*2,2*	*2,7*	*6,0*	*8,5*	*4,1*	*9,0*
Voto alla 2 prova (Y)	*2,1*	*1,5*	*1,7*	*2,5*	*3,0*	*2,1*	*3,2*

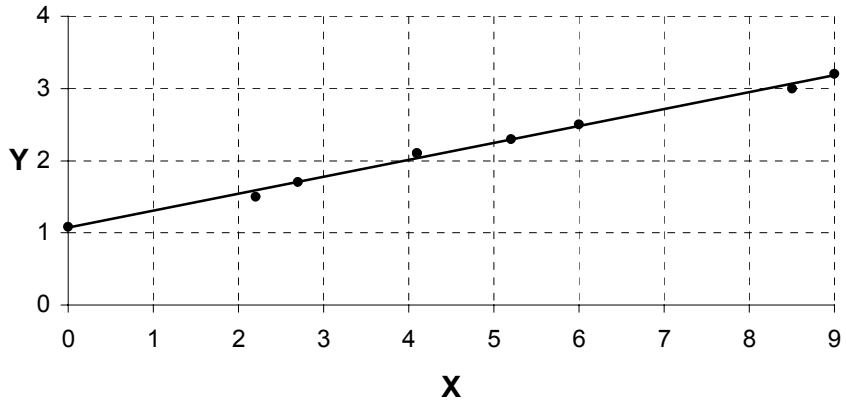
a) Stimare la retta di regressione di Y da X dato che s_{xy}=1,4614

Come l'esercizio precedente

La retta di regressione stimata è quindi **y=1,08 + 0,23x**

b) Costruire lo scatter delle osservazioni e rappresentare su di esso la retta stimata.

In un diagramma cartesiano si possono individuare tutti i punti. Per disegnare la retta stimata bastano comunque due soli punti: intercetta (0; 1,08) e punto di coordinate medie (5,2; 2,3).



c) Determinare il voto previsto alla seconda prova se alla prima si è riportato 3,5

Se alla prima prova si è riportato il voto 3,5, alla seconda prova sarà riportato il seguente voto:
Y(x=3,5)=1,08+0,23*3,5=1,885

Esercizio 3.

*Viene condotto uno studio su di un gruppo di studenti e si rileva che mediamente essi consumano in un mese 4 birre con una deviazione standard di 8 e 4 pizze con una deviazione standard di 4. Si osserva una certa associazione positiva fra il consumo di birra (**Y**) ed il consumo di pizze (**X**). La retta di regressione stimata ha la seguente equazione:*
$Y = 2 + \beta_1 X$

a) Il valore della pendenza β_1 è andato perduto, sapreste ricostruirlo?
Se $\bar{y}$ =4 e s_y=8; $\bar{x}$ = 4 e s_x=4 e la retta di regressione è y=2 + β_1x

Tenuto conto che la retta dei minimi quadrati passa per il punto $(\bar{x}, \bar{y})$ si potrà ricavare $\bar{y} = 2 + \hat{\beta}_1 \bar{x}$ e quindi 4=2+$\beta_1$4 da cui 4β_1=2 e β_1=0,5

b) Calcolare l'indice di determinazione lineare.
L'indice di determinazione lineare R^2 può essere espresso in funzione delle varianze campionarie e del β_1 stimato: $R^2=(\beta_1)^2 * (s^2_x/s^2_y)=(0,5)^2*16/64=0,25+0,25=0,062$.

Esercizio 4.

Su un campione di 50 studenti sono stati ottenuti i seguenti punteggi ad un test intermedio e uno finale:

	Punteggio medio	*Deviazione standard*
Test intermedio	*70*	*10*
Test finale	*55*	*20*

Si sa inoltre che il coefficiente di correlazione lineare fra i due punteggi è 0,8.

a) Trovare l'equazione della retta di regressione per prevedere il punteggio in un test finale sulla base di quello ottenuto al test intermedio.
Gli elementi conosciuti sono:
$\bar{x}$ = 70 e s_x=10; $\bar{y}$ =55 e s_y=20; n=50; r_{xy}=0,8
Il modello di regressione lineare semplice è $Y_i=\beta_0+\beta_1 x_i+\varepsilon_i$
Bisogna stimare i due parametri β_0 e β_1.
Tali parametri possono essere ricavati risolvendo il seguente sistema

$$\begin{cases} \beta_1 = \dfrac{S_{xy}}{S_x^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}$$

Lo stimatore β_1 si può anche esprimere come $\beta_1=r_{xy}\frac{s_y}{s_x}$ e quindi la sua stima è β_1=0,8*(20/10)=1,6

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ =55-1,6*70=55-112=-57
La retta di regressione è y=-57+1,6x

b) Costruire un intervallo di confidenza al 95% per il coefficiente angolare della retta di regressione.

Intervallo di confidenza al 95% per il coefficiente angolare della retta di regressione

$\beta_1 - t_{\left(\frac{\alpha}{2};n-2\right)} es(B_1) \le \beta_1 \le \beta_1 + t_{\left(\frac{\alpha}{2};n-2\right)} es(B_1)$

dove $es(B_1) = \sqrt{es^2(B_1)} = \sqrt{\frac{S^2}{nS^2x}}$

e tenuto conto che $s^2=(n/n-2)\ s^2_y[1-(r_{xy})^2]=(50/48)*400\ [1-(0,8)^2]=1,0416*400*0,36=149,99$

si ottiene la stima di $es(\hat{B}_1) = \sqrt{149,99} / \sqrt{50} * 10 = 12,24/70,71 = 0,173$

Essendo $t_{0,025}$ con n-2=48 gradi di libertà uguale a 2,009 l'intervallo cercato diviene

$1,6 - 2,009 * 0,173 \le \beta_1 \le 1,6 + 2,009 * 0,173$ cioè $1,253 \le \beta_1 \le 1,947$

Esercizio 5.

Su un campione di famiglie è stato rilevato X:=«numero di componenti la famiglia» e Y:=«reddito familiare mensile(in milioni)»:

X	*1*	*2*	*3*	*3*	*5*	*2*	*3*
Y	*4*	*5*	*4.5*	*2*	*6*	*1.5*	*4*

a) Stimare i coefficienti del modello di regressione

Il modello di regressione lineare semplice è $Y_i=\beta_0+\beta_1 x_i+\varepsilon_i$

Bisogna stimare i due parametri β_0 e β_1.

Tali parametri possono essere ricavati risolvendo il seguente sistema

$$\begin{cases} \beta_1 = \frac{S_{xy}}{S_x^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}$$

Dal campione osservato:

$\bar{x}$ = (1+2+3+3+5+2+3)/7=19/7=2,71

$\bar{y}$ = (4+5+4,5+2+6+1,5+4)/7=27/7=3,86

$s_{xy}= \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n} =$ (1-2,71)(4-3,86)+(2-2,71)(5-3,86)+(3-2,71)(4,5-3,86)+

+(3-2,71)(2-3,86)+(5-2,71)(6-3,86)+(2-2,71)(1,5-3,86)+(3-2,71)(4-,386) = 5,2142/7 =0,7448

$s^2_x = \frac{\Sigma(x_i-\bar{x})^2}{n} =$ (1-2,71)^2+(2-2,71)^2+(3-2,71)^2+(3-2,71)^2+(5-2,71)^2+(2-2,71)^2+(3-2,71)^2=2,9224+0,5041+0,0841+0,0841+5,244+0,5041+0,084=9,427/7=1,347

$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$=0,7448/1,347=0,55

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ =3,86-0,55*2,71=2,37

La retta di regressione è y=0,55 + 2,37x

b) Detto β_1 il coefficiente angolare del modello stabilire se, al livello di significatività 0,05, si più rifiutare l'ipotesi che esso sia uguale a 0.

Dobbiamo sottoporre a test l'ipotesi H_0: $\beta_1=0$ contro H_1: $\beta_1 \neq 0$

La statistica test sotto l'ipotesi H_0 è data da $T=\frac{B_1}{\sqrt{es(B_1)}} \sim t_{(n-2)}$

$$\hat{\beta}_1 - t_{\left(\frac{\alpha}{2};n-2\right)} es(B_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\left(\frac{\alpha}{2};n-2\right)} es(B_1)$$

con $es(B_1) = \sqrt{es^2(B_1)} = \sqrt{\frac{S^2}{nS^2x}}$

e con $s^2=(n/n-2)\ [s^2_y - (\hat{\beta}_1)^2 s^2_x]= (7/5)[\ 2,194-0,55^2*1,347]= 1,4*(2,194-0,407)=2,501$

dalle informazioni campionarie si ha

$$R_{co}: \begin{cases} T \leq -t_{\left(\frac{\alpha}{2};n-2\right)} = -2,571 \\ oppure \\ T \geq t_{\left(\frac{\alpha}{2};n-2\right)} = 2,571 \end{cases}$$

e $s^2_y=\frac{\Sigma(y_i - \bar{y})^2}{n} =$ $(4-3,86)^2+(5-3,86)^2+(4,5-3,86)^2+(2-3,86)^2+(6-3,86)^2+(1,5-3,86)^2+(4-3,86)^2==15,3572/7=2,194$

e quindi $es(\hat{B}_1) == \frac{\sqrt{2,501}}{\sqrt{7*1,347}} =1,58/3,07=0,515$

il valore della statistica test è

$T_c=\frac{0,55}{\sqrt{0,515}} =0,766$

che non cade nella regione critica in quanto $-2,571 \leq T_c \leq 2,571$ e di conseguenza non si può rifiutare l'ipotesi H_0.