Metodi Matematici e Informatici per la Biologia
a.a. 2023-2024
(Agliari - Panati - Presilla - Simonella)
12 luglio 2024
Prova di Verifica II

Nome: **Cognome:** **Matricola:**

Esercizio 1 (2+2+2=6 punti)

L'emofilia è una malattia di origine genetica che causa un difetto nella coagulazione del sangue. La patologia viene trasmessa mediante il cromosoma X, pertanto i maschi (XY) hanno una probabilità maggiore di essere emofilici rispetto alle femmine (XX) che risultano affette dalla malattia solo se hanno entrambi i cromosomi X difettosi mentre ne sono portatrici sane nel caso abbiano un solo cromosoma X difettoso. Una donna portatrice sana di emofilia ha quattro figlie femmine da un uomo non emofilico.

1.1 Stabilire quale distribuzione di probabilità regola le condizioni emofiliche di queste figlie.

Sia Z la variabile aleatoria che conta il numero di figlie portatrici sane (si noti che queste non possono essere emofiliche essendo il padre non emofilico). Ogni figlia ha il 50% di probabilità di essere portatrice sana e la condizione di ogni figlia è indipendente dalle altre. Pertanto Z segue una distribuzione binomiale

$$P(Z=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Nel nostro caso abbiamo parametri $n = 4$ e $p = 1/2$.

1.2 Determinare la probabilità che una sola delle quattro figlie sia portatrice sana di emofilia.

$$P(Z=1) = \binom{4}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = 4\frac{1}{2^4} = \frac{1}{4} = 0.25,$$

1.3 Determinare la probabilità che almeno una figlia sia portatrice sana di emofilia.

$$P(Z \geq 1) = 1 - P(Z=0) = 1 - \binom{4}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = 1 - \frac{1}{2^4} = \frac{15}{16} = 0.9375,$$

avendo utilizzato

$$\sum_{k=0}^{n} P(Z=k) = 1.$$

Esercizio 2 (2+2+2+2=8 punti)

Si consideri il seguente insieme di numeri interi:

$$\{3, 5, 1, 1, 0, 2, 8, 2, 3, x\}.$$

Si determini il valore dell'intero x affinché l'insieme abbia:

2.1 media 3,

La media è data dalla formula

$$\frac{1}{n}\sum_{i=1}^{n} x_i.$$

Nel nostro caso $n = 10$, gli x_i per $i = 1, \dots, 9$ sono dati, $x_{10} = x$ è incognito. Quindi imponendo che la media valga 3 otteniamo

$$\frac{1}{10}\left(\sum_{i=1}^{9} x_i + x\right) = 3 \qquad \Rightarrow \qquad 25 + x = 30 \qquad \Rightarrow \qquad x = 5.$$

2.2 moda 2,

Costruiamo la tabella delle frequenze per i 9 dati assegnati

dato	frequenza
0	1
1	2
2	2
3	2
5	1
8	1

Per avere moda uguale a 2 occorre aumentare la frequenza del dato uguale a 2 quindi $x = 2$.

2.3 mediana 2.5.

Per determinare la mediana riscriviamo i primi 9 dati assegnati in ordine crescente

$$\{0, 1, 1, 2, 2, 3, 3, 5, 8\}.$$

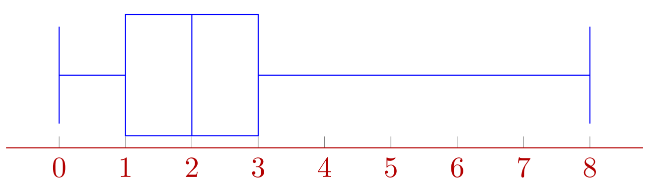
La mediana di questi 9 dati è 2. Se aggiungendo un decimo dato intero x la mediana diventa 2.5 non può che essere $x \geq 3$.

2.4 Scelto poi $x = 1$, determinare i numeri di sintesi necessari a graficarne il boxplot relativo.

Riscriviamo i 10 dati assegnati in ordine crescente

$$\{0, 1, 1, 1, 2, 2, 3, 3, 5, 8\}.$$

I relativi numeri di sintesi e il corrispondente boxplot sono:
minimo 0, massimo 8, mediana 2, primo quartile 1, terzo quartile 3.



Esercizio 3 (2+2+2=6 punti)

Una trivella di perforazione in azione a Piazza Venezia per X ore raggiunge quota Y metri (rispetto a una opportuna quota di riferimento) come monitorato dalle seguenti 7 misurazioni:

X(ore)	3.30	3.80	4.99	5.74	7.03	8.16	9.34
Y(metri)	-15.10	-17.67	-22.85	-26.05	-28.47	-34.77	-48.48

3.1 Determinare l'equazione della corrispondente retta di regressione lineare e dell'associato coefficiente di correlazione r.

Per il campione specificato con $n = 7$ dati abbiamo i seguenti valori medi e deviazioni standard

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = 6.05 \text{ ore}, \qquad S_{X\,n} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} = 2.24 \text{ ore},$$

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = -27.63 \text{ metri}, \qquad S_{Y\,n} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2} = 11.31 \text{ metri}.$$

La retta di regressione lineare ha equazione $Y = aX + b$ con

$$a = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} = -4.88 \text{ metri/ore},$$

$$b = \overline{Y}_n - a\overline{X}_n = 1.93 \text{ metri}.$$

Il coefficiente di correlazione è il numero

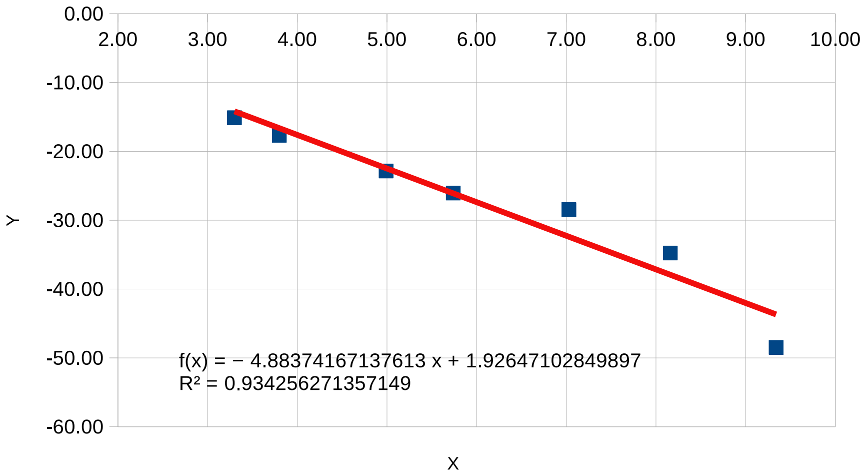
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}_n}{S_{X\,n}}\right)\left(\frac{Y_i - \overline{Y}_n}{S_{Y\,n}}\right) = -0.967.$$

3.2 Prendendo una quota di riferimento diversa come cambia il valore di r? Perché?

Prendendo una quota di riferimento diversa i valori Y_i vengono traslati di una quantità costante ΔY. Pertanto le differenze $Y_i - \overline{Y}_n$ rimangono invariate e così pure il valore di r.

3.3 In quanto tempo la trivella dovrebbe raggiungere quota $Y = -70$ metri?

$$X = (Y - b)/a = (-70 - 1.93) \text{ metri}/(-4.88 \text{ metri/ore}) = 14.74 \text{ ore}.$$



Esercizio 4 (4+2=6 punti)

Durante l'anno scolastico mio figlio ha fatto assenze per un totale di 30 giorni così ripartite dal lunedì al venerdì:

lunedì	martedì	mercoledì	giovedì	venerdì
10	4	3	4	9

Mi dice che le sue assenze non si sono verificate strategicamente nei giorni che allungano il week end, come a me viene da pensare, ma in modo del tutto casuale.

4.1 Con quale test statistico potrei verificare l'ipotesi che mio figlio dice la verità?

L'ipotesi nulla H_0 da verificare è che le 30 assenze siano ripartite in modo equiprobabile negli $n = 5$ giorni, cioè che le frequenze attese siano

$$E_1 = 6, \qquad E_2 = 6, \qquad E_3 = 6, \qquad E_4 = 6, \qquad E_5 = 6.$$

D'altro canto le frequenze osservate sono

$$F_1 = 10, \qquad F_2 = 4, \qquad F_3 = 3, \qquad F_4 = 4, \qquad F_5 = 9.$$

Confrontiamo l'adattamento tra frequenze osservate e attese mediante il test del χ^2. Il valore della variabile aleatoria χ^2 corrispondente al test con dati $\{E_i\}_{i=1,\dots,n}$ e $\{F_i\}_{i=1,\dots,n}$ risulta

$$\begin{aligned}\chi^2_{\text{test}} = \sum_{i=1}^{n} \frac{(F_i - E_i)^2}{E_i} &= \frac{(10-6)^2}{6} + \frac{(4-6)^2}{6} + \frac{(3-6)^2}{6} + \frac{(4-6)^2}{6} + \frac{(9-6)^2}{6} \\ &= \frac{16+4+9+4+9}{6} = \frac{42}{6} = 7.\end{aligned}$$

4.2 Qual'è il livello di significatività con cui questa ipotesi può essere accettata?

Possiamo accettare l'ipotesi H_0 con un livello di significatività $\alpha = P(\chi^2 \geq \chi^2_{\text{test}})$. Abbiamo $n - 1 = 4$ gradi di libertà ($n = 5$ categorie con vincolo $\sum_{i=1}^{n} E_i = 30$) e dalla tabella allegata troviamo che $P(\chi^2 \geq 7) > 0.1$ (ovvero che, essendo $0.1 = P(\chi^2 \geq 7.779)$, risulta $\chi^2_{\text{test}} < 7.779$). Possiamo quindi accettare l'ipotesi nulla con un livello di significatività superiore (di poco) al 10%. Il calcolo numerico, CHIDIST(7;4)=0.135, mostra che il livello esatto di significatività è 13.5%.

Esercizio 5 (2+2+3=7 punti)

I seguenti $n = 64$ valori costituiscono un campione di una variabile aleatoria

22.4068	21.4640	21.6448	20.8153	22.3582	21.8522	19.8134	20.3051
17.5286	21.5596	20.8131	24.1789	20.8267	17.8681	21.9928	16.5818
20.2513	20.6278	17.4116	20.5234	21.9711	17.2719	20.9209	22.7997
18.3895	22.8883	21.9153	20.4422	21.8812	19.3265	18.9496	21.8754
19.1839	21.6838	20.4167	19.9607	20.9080	21.7677	19.1629	20.4361
17.8960	19.2969	22.8496	21.7890	14.7062	16.5618	21.6812	19.9248
18.2798	17.4298	19.3612	21.3175	20.3548	18.0676	21.5738	18.2420
16.7587	20.5813	22.3904	21.1385	20.0346	20.0549	19.5419	19.8395

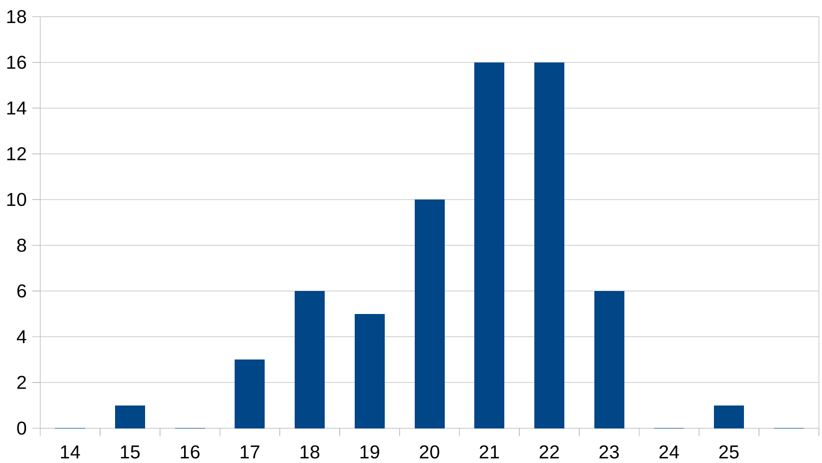
5.1 Determinare valore minimo, valore massimo, valore medio e deviazione standard del campione.

valore minimo $= 14.7062$, valore massimo $= 24.1789$,

valore medio $= 20.1976$, deviazione standard $= 1.8814$

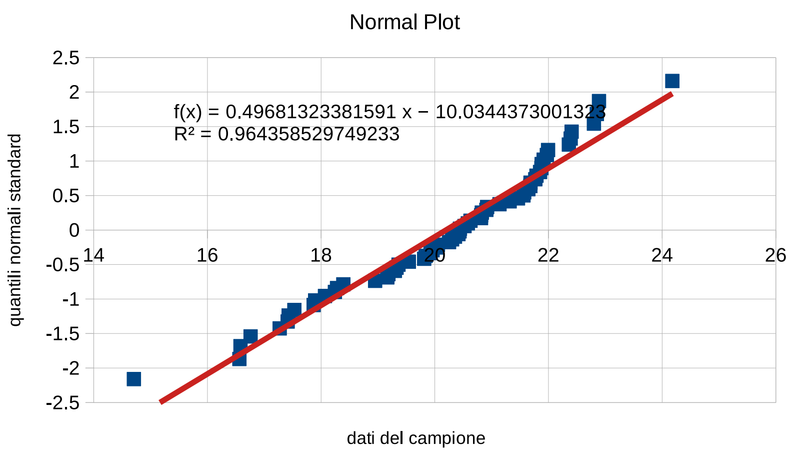
5.2 Calcolare le frequenze dei dati del campione usando classi di ampiezza 1 e quindi produrre il relativo istogramma.

classe	frequenza
(13,14]	0
(14,15]	1
(15,16]	0
(16,17]	3
(17,18]	6
(18,19]	5
(19,20]	10
(20,21]	16
(21,22]	16
(22,23]	6
(23,24]	0
(24,25]	1
(25,26]	0



5.3 Giudicare il carattere normale dei dati del campione realizzando un normal plot e riportando il coefficiente di determinazione della corrispondente retta di regressione lineare.

Il normal plot mostra che i dati hanno una distribuzione approssimativamente normale; il coefficiente di determinazione della retta di regressione lineare vale $r^2 = 0.964$.



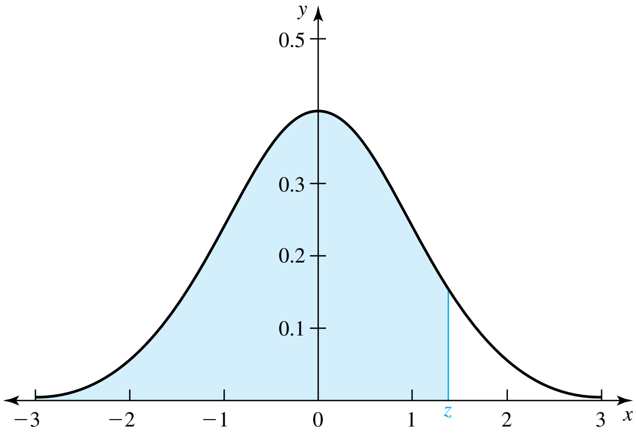


Figure B.1 Areas under the standard normal curve from $-\infty$ to z.

z	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5754
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7258	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7518	.7549
0.7	.7580	.7612	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7996	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986

Critical values of chi-square (right tail)

Degrees of freedom (*df*)	Significance level (α) .99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807