UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Relazione di un'indagine Statistica

Corso di Foundations of Probability and Statistics

Ad opera di:

Davide **Banfi**, matricola **806539**

Francesco Martino **Carugati**, matricola **795116**

In questa relazione viene proposta una breve indagine statistica svolta su dati raccolti dal *National Center for Health Statistics*, l'agenzia principale di statistica degli Stati Uniti.
Il dataset rappresenta il numero di nascite avvenute negli U.S nel 2018 ed è disponibile al link https://www.kaggle.com/des137/us-births-2018.
In particolare, il lavoro è stato progettato sull'impronta di uno studio del 2006 compiuto dai ricercatori del Royal Devon and Exeter Hospital (UK), il quale sancisce che…

… "l'altezza si eredita dal padre, il peso dalla madre" [1] [2]

Ci siamo posti il problema di andare a verificare se quella tesi trovasse una evidenza, anche parziale, nei nostri dati. Tuttavia, non disponendo né delle altezze del neonato né di quelle paterne, si propone principalmente uno studio su una eventuale correlazione tra il peso del bambino e quello materno.

Dapprima le *3.801.534* unità statistiche vengono separate in due classi distinte in base al sesso, evidenziando distribuzioni pressoché uguali.

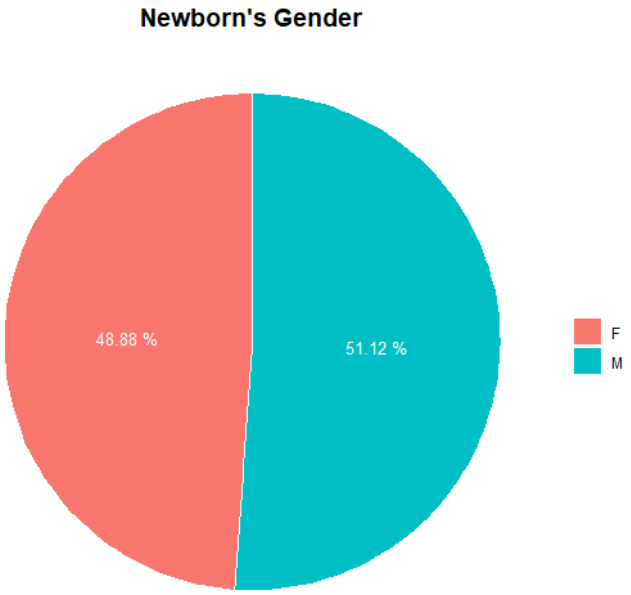


Grafico 1.1

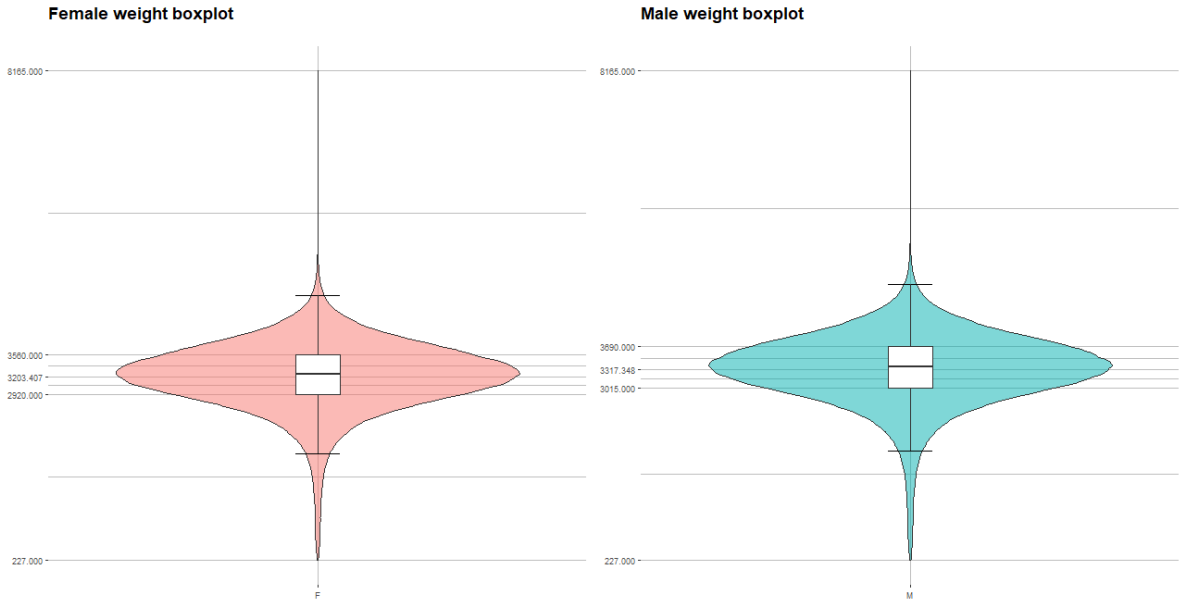
Successivamente, previa eliminazione delle unità statistiche che presentano un peso alla nascita sconosciuto[3], vengono calcolate le seguenti misure di tendenza centrale della variabile *X* "*peso*" espresso in grammi.

[1] Cfr. https://www.bionews.org.uk/page_90220

[2] Mentre l'altezza si eredita dal padre per via *genetica*, il peso del neonato sarebbe determinato dalle condizioni ambientali dell'utero materno. In particolare, il peso del bambino sarebbe collegato al tasso di zuccheri nel sangue della madre e quindi, nei casi estremi, all'obesità.

[3] Si è ritenuto opportuno eliminare tali righe per fare in modo che dati mancanti non influenzassero l'output dell'indagine. Il peso del neonato è raccolto nella colonna DBWT, mentre i rispettivi valori mancanti sono rappresentati dal valore 9999. Le unità che presentano un peso definito sono *3.798.574*.

Sesso	Media	Std deviation	Varianza	n°
Femmina	3203.407	573.5281	328934.5	1856902
Maschio	3317.348	600.9876	361186.1	1941672



Il peso medio complessivo è dato da

$$E[X] = \frac{3202.407 * 1856902 + 3317.348 * 1941672}{1941672 + 1856902} = 3261.649$$

con varianza pari a $Var[X] = 348664.1$ e deviazione standard $\sqrt{Var[X]} = 590.4779$. La moda è 8165 mentre la mediana è 3300.
Studi statistici affermano che il 90% dei bambini alla nascita sia normopeso, ovvero abbia un peso compreso tra i 2,5 e i 4,5 kg. Per poter calcolare la percentuale di bambini normopeso nel modello americano è necessario prima definire come la variabile casuale X si distribuisce.

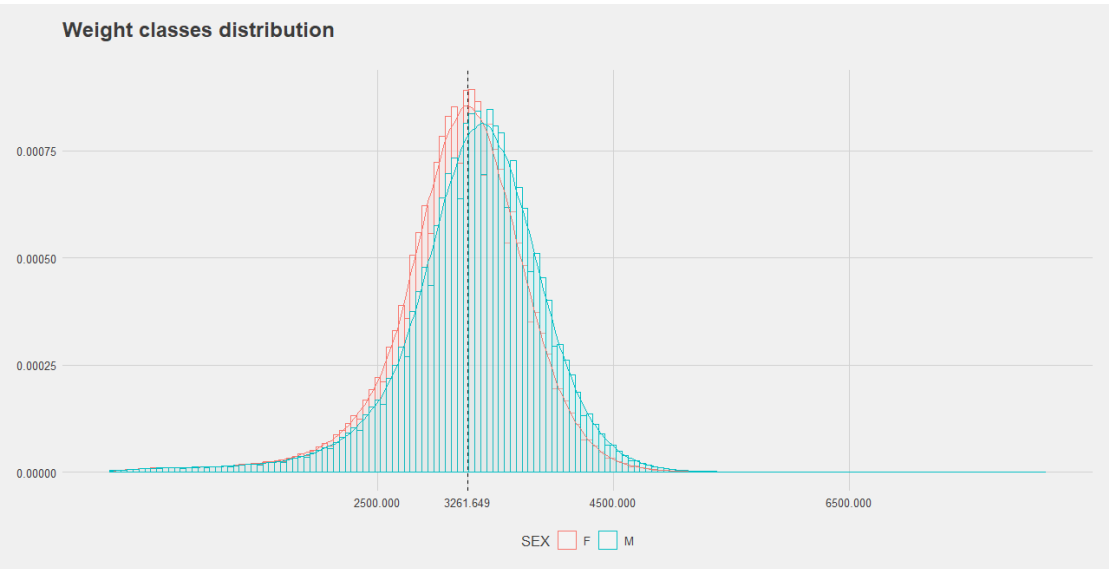


Grafico 1.2

Il grafico sovrastante è molto simile a quello di una distribuzione Gaussiana, tuttavia in una distribuzione normale la media tende a coincidere con la mediana e con la moda.
Abbiamo visto precedentemente come questi tre valori non siano uguali, perlopiù…

- …l'indice di asimmetria negativo $I = \frac{E[X-\mu]^3}{\sigma^3} = -0.86035$
- …il fatto che $Media < Mediana < Moda$
- …l'indice di Curtosi $K = \frac{E[X-\mu]^4}{\sigma^4} = 5.72$ (per una normale succede che $K = 3$)

...confermano la presenza di una curva asimmetrica negativa e leptocurtica.
Si può quindi affermare con certezza che la distribuzione non è approssimabile a una Gaussiana, pertanto la percentuale di bambini normopeso è calcolabile come:

$$P[2500 \leq X \leq 4500] = \frac{Casi\ favorevoli}{Casi\ possibili}$$

Per entrambi i sessi si ottiene:

$$P_{Maschio}[2500 \leq X \leq 4500] = \frac{1767605}{1943273} = 0.9096 = 90.96\%$$

$$P_{Femmina}[2500 \leq X \leq 4500] = \frac{1677778}{1858261} = 0.90287 = 90.29\%$$

Tali valori coincidono con l'area sottesa dalle due curve nell'intervallo $[2500\,;4500]$ del grafico 1.2 Effettivamente si nota come le probabilità siano prossime al 90% per entrambi i sessi, quindi anche in questo caso si conclude che i bambini sono prevalentemente normopeso.
Un'inchiesta dell'UNICEF[4] datata maggio 2019 rivela che nel 2015 la percentuale di bambini nati sottopeso negli U.S. è compresa tra il 7.9% e l'8.1%, calcolato con un intervallo di confidenza del 95% e senza distinzione di sesso.
Dai nostri dati si ricava che la percentuale di bambini nati sottopeso nel 2019 è pari a:

$$P_{sottopeso}[X \leq 2500] = \frac{317324}{3798574} = 0.08353 = 8.35\%$$

Rispetto al 2015, il tasso di bambini sottopeso è dunque aumentato del 4.3%.

Si passa allo studio della correlazione tra il peso materno e quello del bambino per evidenziare se tra i due vi sia un legame lineare. A tal proposito sono stati fatti alcuni accorgimenti:

1. causa l'enorme volume di dati e la limitazione della potenza di calcolo, si è estratto un campione totalmente casuale di 2000 individui[5].

[4] Cfr. https://www.unicef.it/Allegati/UNICEF-WHO_Low_Birthweight_Estimates_2019.pdf , tabella pag 26.

[5] Il campione è stato generato in modo completamente casuale e si aggiorna ogni volta che si esegue lo script, confrontare con lo script nella sezione finale.

2. come indice del peso della madre si è tenuto in considerazione il *Body Mass Index*[6], il quale è più opportuno perché tiene in conto anche della statura (una persona, adulto o adolescente che sia, può essere considerata sovrappeso anche se pesa 70 kg ma è alta 1,2 metri).

3. sono state escluse tutte le unità statistiche che presentavano valori sconosciuti nella colonna del peso del neonato (DBWT) o nella colonna del BMI della madre (BMI).

Il diagramma a dispersione riportato sotto fornisce una prima visualizzazione della relazione peso ~ BMI.

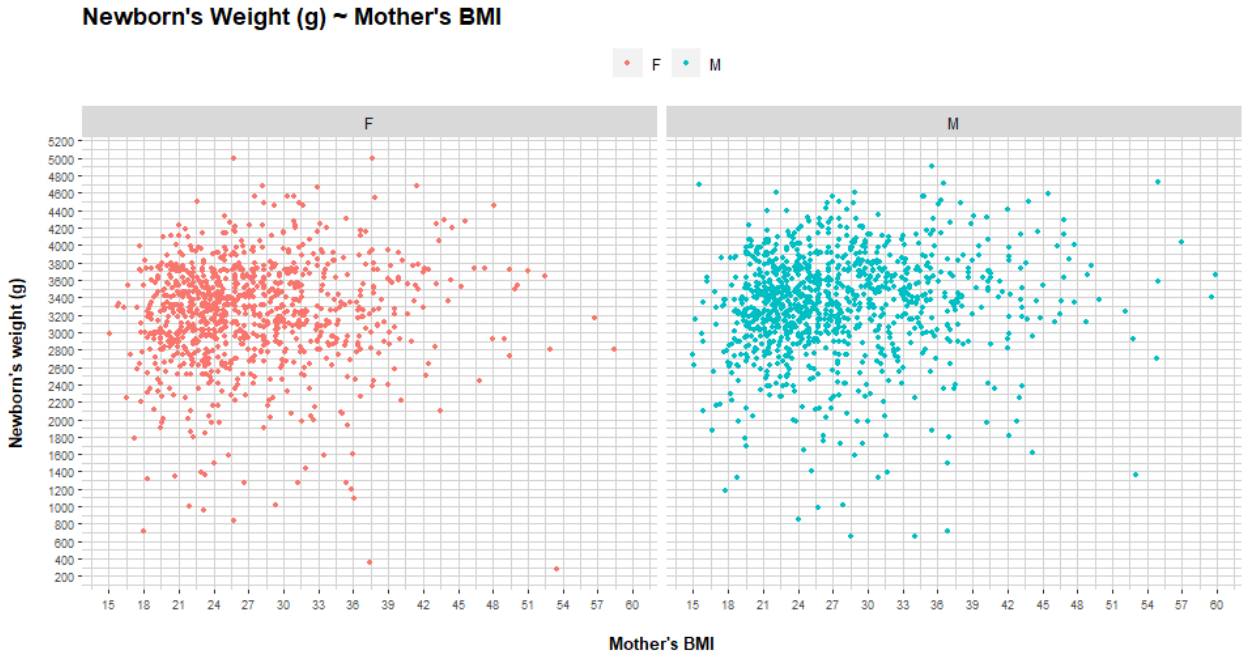


Grafico 1.3

Il campione riporta i seguenti valori.

Sesso	Media BMI madre	Media peso neonato	Std peso neonato	Std BMI madre	Covarianza BMI-peso	n°
Femmina	27.333	3211.251	595.562	6.75	271.247	942
Maschio	27.406	3309.257	589.684	7.1	406.097	1058

Tabella 1

Utilizzando la formula di Pearson si ricava il coefficiente di correlazione lineare.

$$\rho_{\text{Maschio}} = \frac{406.097}{7.1 * 589.684} = 0.097 \qquad \rho_{\text{Femmina}} = \frac{271.247}{6.75 * 595.562} = 0.067$$

[6] Il peso della madre è espresso in pounds anziché nello stesso ordine di grandezza del peso del neonato. Poiché teniamo in conto il rapporto BMI e non la grandezza scalare "peso", questo non distorce l'ordine di grandezza finale. Infatti, il BMI è stato calcolato sul peso in pounds per poi essere moltiplicato per un coefficiente di scala pari a 703.

La concordanza è positiva, ossia variazioni positive del BMI materno producono variazioni positive del peso del neonato indipendentemente dal sesso, tuttavia, il coefficiente è così basso da escludere un legame lineare influente (non si può dire nulla sull'esistenza di legami non lineari). In sintesi, possiamo concludere che non c'è una vera e propria correlazione forte a sufficienza affinché il peso della madre incida su quello del neonato.
Essendo l'indice prossimo a 0, la retta di regressione sarà quanto più "appiattita".

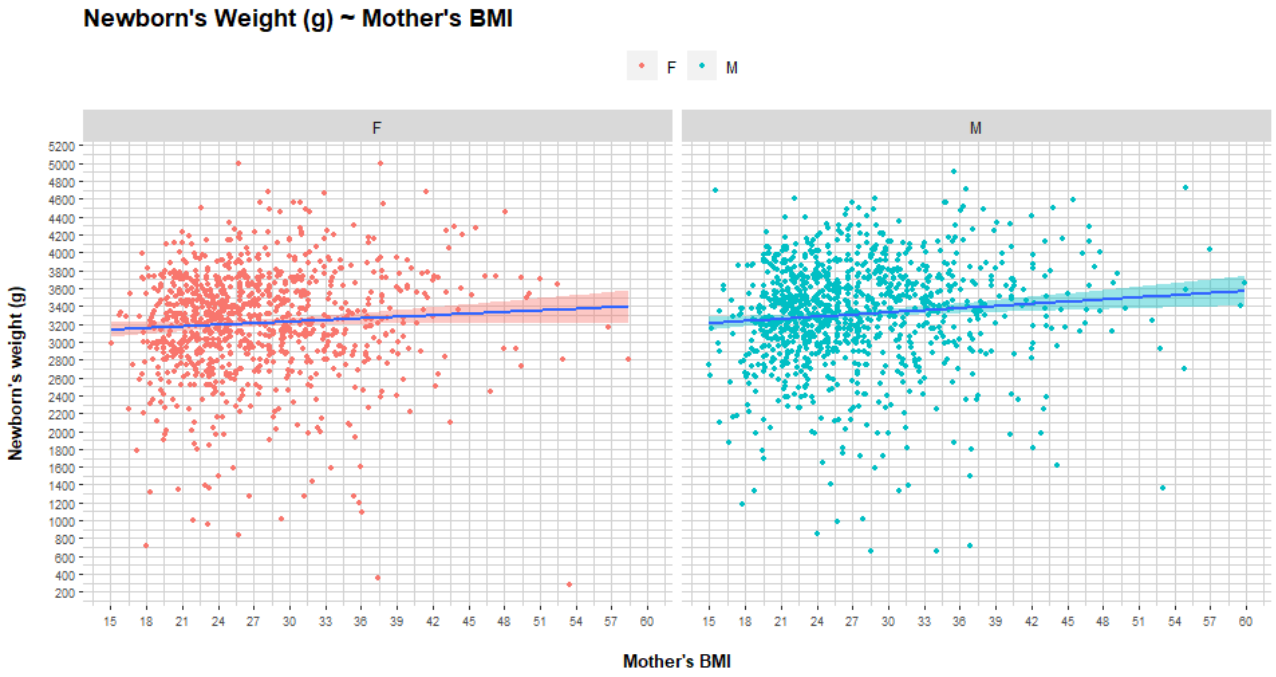


Grafico 1.4: la retta di regressione assume forma piatta

Si è ritenuto opportuno non calcolare il modello di regressione lineare per via dell'assenza di un legame forte. L'interpretazione di una retta appiattita ci dice che a una variazione unitaria del BMI materno corrisponderebbe una variazione di pochissimi grammi del peso natale, tuttavia, questo è insignificante anche per via del fatto che l'approssimazione a un modello lineare, quando può non essere tale, genererebbe errori di distorsione del modello stesso.

Appurato che tra il peso del bambino e il BMI della madre non via sia correlazione, è lecito domandarsi se il peso possa dipendere da altri fattori.
Come sostiene Joel Ray[7], l'etnia paterna potrebbe essere una di questi: bambini che pesano poco e che appartengono ad alcune etnie, specialmente quelle orientali, possono essere ritenuti normopeso sebbene venissero classificati come sottopeso in culture occidentali.

La seconda task è stata studiare la connessione tra il peso del neonato e l'etnia paterna.

Poiché la relazione è tra un carattere quantitativo e uno qualitativo viene preferito l'indice Chi-quadro normalizzato, essendo interpretabile tra i valori 0 e 1. L'esperimento ha avuto come esito $\chi 2 = 0.00068845$.

[7] Cfr. https://www.sciencedaily.com/releases/2014/06/140630094846.htm

L’indice, approssimabile a 0, indica assenza di connessione nonché indipendenza delle due variabili. A conferma della non connessione ci si aspetta un indice di indipendenza eta quadro molto vicino a 0:

$$\eta 2 = 0.018262$$

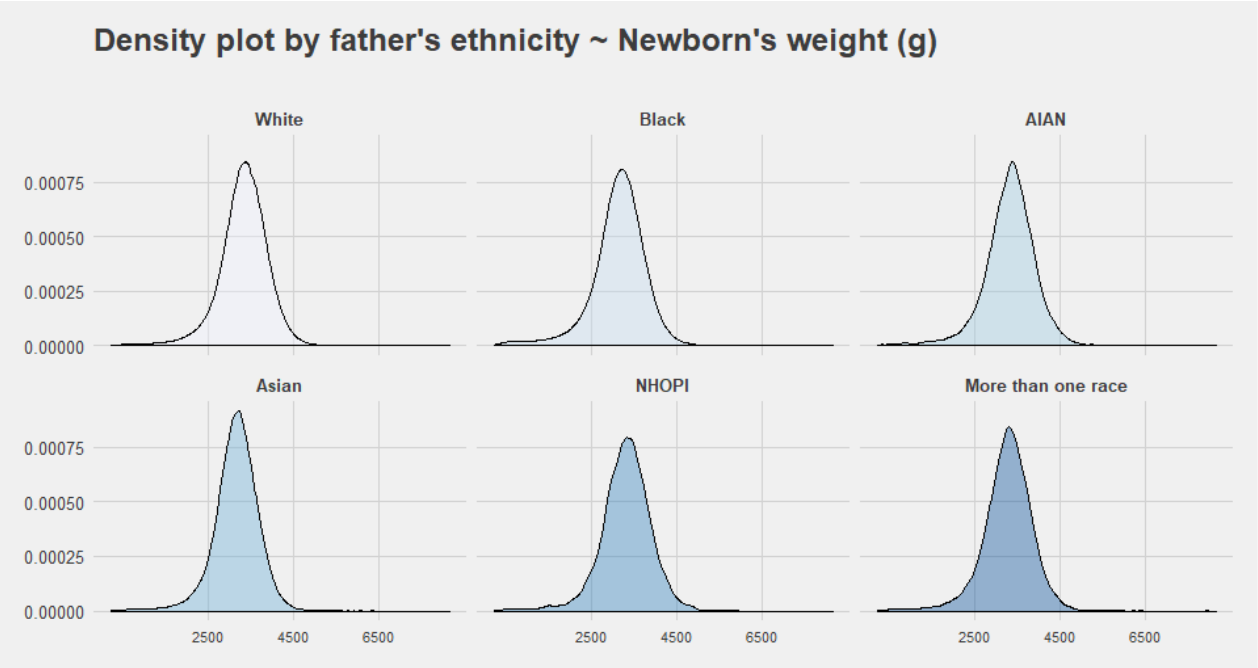


Grafico 1.5: grafico che rappresenta la densità del peso natale per ciascuna etnia paterna

Il grafico 1.5 mostra come la quasi completa assenza di connessione e di dipendenza si traduca a livello visivo in curve di densità quasi identiche.

Viene riportato un altro articolo[8] in cui si sostiene che gli elementi che possono condizionare il peso di un nascituro siano il fumo, il basso incremento di peso durante la gravidanza o un’età della madre minore di 17 anni o maggiore di 35
Essendo i caratteri quantitativi, la correlazione lineare è data ancora dall’indice di Pearson. Previa esclusione dei bambini con un peso maggiore di 2500 g e…

1. …delle unità statistiche per il quale non si ha un numero di sigarette fumate definito, si ottiene $\rho_{\mathrm{CIG_0,DBWT}} = 0.02644$.
2. …delle unità statistiche per il quale il peso guadagnato in gravidanza è sconosciuto, si ottiene $\rho_{\mathrm{WTGAIN,DBWT}} = 0.16743$.
3. …delle unità statistiche per le quali l’età materna è compresa tra 17 e 35, si ottiene $\rho_{\mathrm{MAGER,DBWT}} = 0.028781$.

Nessuna delle variabili considerate influenza in modo considerevole il peso del neonato.

[8] Cfr. https://www.stanfordchildrens.org/en/topic/default?id=low-birthweight-90-P02382

Si sposta ora l'attenzione sul campione di 2000 neonati estratto precedentemente per stimare alcuni valori.

La probabilità totale che un neonato estratto casualmente da una popolazione di 3.798.574 unità statistiche sia sottopeso è calcolabile rapportando i casi favorevoli ai casi possibili.

$$P_{X\leq 2500} = \frac{casi\ favorevoli}{casi\ possibili} = \frac{317324}{3798574} = 0.0835377$$

Osservando la tabella 1, la probabilità che **nel campione** un bambino sia sottopeso è compresa tra il 6.9% e il 9.79% adottando un grado di fiducia del 98%.

$$I.C. = \left[0.08354 \pm 2.32635 * \frac{\sqrt{0.08354 * (1 - 0.08354)}}{\sqrt{2000}}\right] = [0.06914\ ;\ 0.09793]$$

Analogamente è possibile calcolare un intervallo entro il quale la media del peso del neonato è compresa all'interno del campione.

$$I.C. = \left[3261.649 \pm 2.32635 * \frac{\sqrt{348664.1}}{\sqrt{2000}}\right] = [3230.933\ ;\ 3292.365]$$

I due valori ottenuti indicano che, estraendo tutti i campioni possibili di 2000 individui dalle 3.798.574 unità statistiche, si riscontrano i seguenti aspetti con un'attendibilità del 98%.

- Il valore medio reale del peso alla nascita è compreso tra 3230.933 e 3292.365 grammi.

- La percentuale di bambini sottopeso, cioè con un peso inferiore a 2500 grammi, varia tra 6.91% e 9.79%.
  Questo è equivalente a dire che i bambini sottopeso in qualunque campione sono quasi sempre compresi tra 138 e 196.

Le misure appena calcolate variano al variare dell'ampiezza del campione oppure scegliendo un grado di confidenza diverso. Ad esempio, ponendo una fiducia pari a 95% si produce un intervallo più ampio sia per la probabilità di essere sottopeso sia per la media reale.

Per verificare questo risultato ci viene in aiuto un codice progettato ad hoc per generare campioni casuali di 2000 unità ogni qualvolta lo si esegua.
Nel nostro caso, questo è stato eseguito 80 volte. La tabella sottostante riporta le statistiche di ciascun campione casuale con la relativa percentuale di bambini sottopeso.

Esperimento: eseguendo per *n* volte il codice si registrano i seguenti valori.

n	media peso	n° neonati sottopeso	% neonati sottopeso		n	media peso	n° neonati sottopeso	% neonati sottopeso
1	3247.92	185	0.0925		41	3251.08	174	0.087
2	3258.74	165	0.0825		42	3261.39	182	0.091
3	3277.51	171	0.0855		43	3264.77	167	0.084
4	3277.04	144	0.072		44	3253.89	160	0.08
5	3244.25	161	0.0805		45	3253	171	0.086
6	3248.49	175	0.0874		46	3250.24	159	0.0795
7	3257.51	158	0.079		47	3255.54	154	0.077
8	3269.61	154	0.077		48	3246.7	194	0.097
9	3277.93	162	0.081		49	3267.69	162	0.081
10	3279.4	153	0.0765		50	3266.83	169	0.0845
11	3236.96	181	0.09		51	3253.24	168	0.084
12	3246.89	159	0.0795		52	3260.29	167	0.0835
13	3261.46	162	0.081		53	3274.84	159	0.0795
14	3276.79	146	0.073		54	3278.22	159	0.0795
15	3261.28	163	0.0815		55	3252.26	178	0.089
16	3258.99	162	0.081		56	3254.58	168	0.084
17	3252.46	166	0.083		57	3261.18	166	0.083
18	3273.73	157	0.0785		58	3240.94	165	0.0825
19	3262.22	202	0.101		59	3276.05	152	0.076
20	3275.06	166	0.083		60	3260.62	174	0.087
21	3263.18	162	0.081		61	3259.87	159	0.0795
22	3260.35	160	0.08		62	3298.35	154	0.077
23	3276.82	176	0.088		63	3244.39	162	0.081
24	3254.46	162	0.081		64	3270.22	162	0.081
25	3269.97	147	0.0735		65	3293.46	154	0.077
26	3259.19	185	0.925		66	3277.28	167	0.0835
27	3262.33	161	0.081		67	3295.05	141	0.071
28	3267.97	164	0.082		68	3270.47	172	0.086
29	3279.15	147	0.0735		69	3265.24	159	0.0795
30	3266.98	153	0.077		70	3244.24	176	0.088
31	3243.02	173	0.087		71	3248.82	184	0.092
32	3251.08	164	0.082		72	3281.69	150	0.075
33	3255.55	168	0.084		73	3274.92	163	0.082
34	3248.06	172	0.086		74	3263.64	159	0.0795
35	3262.21	163	0.082		75	3266.89	157	0.0785
36	3248.24	170	0.085		76	3278.91	161	0.0805
37	3261.8	174	0.087		77	3242.38	163	0.0815
38	3248.02	176	0.088		78	3272.42	147	0.0735
39	3258.28	178	0.089		79	3273.15	163	0.0815
40	3247.85	177	0.089		80	3240.05	170	0.085

Tabella 2

Viene calcolato il margine di errore relativo alla media del peso:

$$ME = 2.32635 * \frac{\sqrt{348664.1}}{\sqrt{2000}} = 30.71593$$

Per garantire un errore di peso non troppo consistente (ad esempio non superiore a 50 g), posto un grado di fiducia del 98%, si ricava che il campione deve essere costituito da almeno 755 neonati.

$$ME < 50 \rightarrow Z_{\frac{a}{2}} * \frac{\sigma}{\sqrt{n}} < 50 \rightarrow n > \left(\frac{Z_{\frac{a}{2}} * \sigma}{MSE}\right)^2 \rightarrow n > \left(\frac{2.32635 * 590.4779}{50}\right)^2 = 755$$

Un aumento del numero di individui nel campione produca un margine di errore sempre più piccolo grazie alla *legge debole dei grandi numeri*.

Lo studio proposto fino a questo punto ha descritto come il carattere "peso" si manifesti senza tenere conto di eventuali distinzioni di sesso. In particolare, la tabella 1 evidenzia come i valori del peso natale si disperdano attorno alla media con un livello molto simile per le due classi di genere.
Alla luce dei dati raccolti si potrebbe ipotizzare che un campione casuale non dovrebbe mostrare nessuna differenza sostanziale per quanto riguarda il peso nei due sottogruppi.
Ad esempio, si potrebbe ipotizzare che tra le medie dei sottogruppi non ci sia una differenza maggiore di 50 g. Anche questa ipotesi viene verificata estraendo un nuovo campione di 2000 neonati in cui si rilevano le seguenti statistiche.

Sesso	Media peso	Std peso	Min	Max	n°
Femmina	3213.743	585.9182	390	5295	959
Maschio	3312.885	589.7494	312	5280	1041

Si stabilisce un grado di fiducia $1 - a$ del 98%: è ragionevole ipotizzare che i neonati maschi pesino mediamente di più delle femmine per un valore inferiore a 50 g?

$$\begin{cases} H_0: \mu_0 - \mu_1 \leq 50 \\ H_1: \mu_0 - \mu_1 > 50 \end{cases}$$

$$Z_{test} = \frac{(3312.885 - 3213.743) - 50}{\sqrt{\frac{585.92^2}{959} + \frac{589.75^2}{1041}}} = 1.8679$$

Per rifiutare l'ipotesi che le due medie si discostino di più di 50g deve accadere che Z_{test} sia maggiore di Z_a. Interessa conoscere il valore di z_a tale per cui $P[Z < z_a] = 0.98$, verificabile solo se z_a è uguale a 2.05375.

$$1.8679 < 2.05375$$

Lo Z_{test} non cade nella regione di rifiuto, pertanto si accetta l'ipotesi che le due medie si discostano per un valore inferiore a 50 g.