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

Un breve richiamo sul test t-Student

Siano A^{exp} (a_1, a_2 ….a_n) e B^{exp} (b_1, b_2 ….b_m) due set di dati i cui valori medi sono:

$$\bar{A} = \frac{1}{n}\sum_i a_i \quad e \quad \bar{B} = \frac{1}{m}\sum_i b_i$$

e le varianze campionarie:

$$\sigma_a^2 = \frac{1}{n-1}\sum_i (a_i - \bar{A})^2 \quad e \quad \sigma_b^2 = \frac{1}{m-1}\sum_i (b_i - \bar{B})^2$$

Il test t- di Student serve per verificare quanto siano significative le differenze osservate tra due gruppi di dati.

Il rapporto:

$$t_{oss} = \frac{\text{differenza tra le medie campionarie}}{\text{errore standard sulla differenza tra le medie campionarie}}$$

cioè:

$$t_{oss} = \frac{\bar{A} - \bar{B}}{\sigma_{\bar{A}-\bar{B}}}$$

è una variabile aleatoria. Nell'ipotesi (H_o) che i due campioni provengano dalla stessa popolazione, cioè nell'ipotesi che il valore atteso e la varianza vera siano gli stessi per ambedue i campione e che le differenze osservate siano dovute al caso, la t_{oss} segue la distribuzione di probabilità di una variabile t-Student: t_ν con $\nu = n + m - 2$ gradi di libertà.

Il test si può applicare in due modi:

a) si calcola la probabilità di osservare un valore t_ν maggiore o eguale a t_{oss}. Questa rappresenta la probabilità che, pur essendo valida l'ipotesi H_o (distribuzioni eguali), abbiamo osservato per caso una certa differenza tra i valori medi. Questa probabilità può essere calcolata utilizzando la funzione Excel

   DISTRIB.T(t_{oss}, ν, *code*)

   con *code* = 1 calcola la probabilità per una coda della distribuzione:

$$P(t_\nu > t_{oss}) = \int_{t_{oss}}^{\infty} f(t_\nu)dt_\nu$$

   Con *code* = 2 calcola la probabilità per una coda della distribuzione:

$$P(|t_\nu| > |t_{oss}|) = \int_{-\infty}^{-t_{oss}} f(t_\nu)dt_\nu + \int_{t_{oss}}^{\infty} f(t_\nu)dt_\nu$$

   Si usa *code* = 1 quando, per qualche motivo, si conosce il segno della differenza: $\bar{A} - \bar{B}$

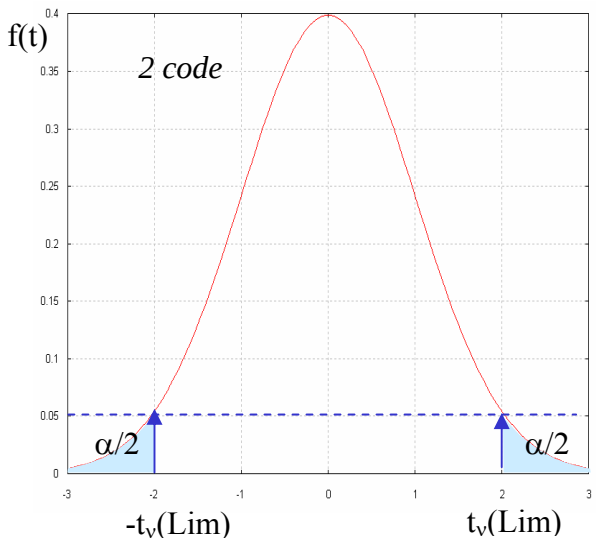
b) si definisce un limite di rischio α (confidenza 1-α) alla probabilità di sbagliare affermando che le distribuzioni sono diverse e si calcola il valore limite **t_ν(Lim)** tale che:

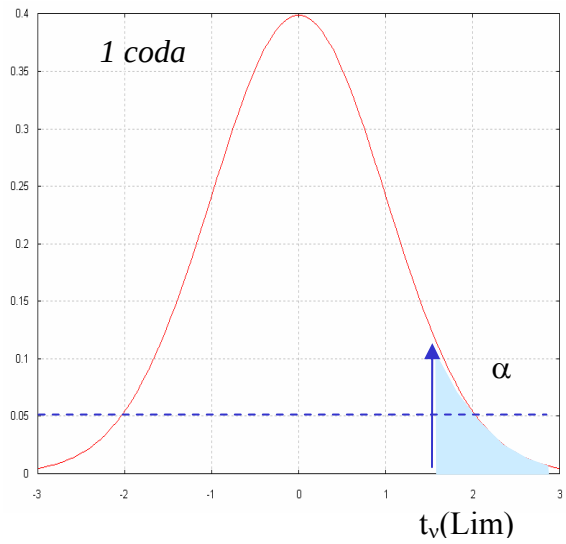
$$\alpha = \int_{t_\nu(Lim)}^{\infty} f(t_\nu)dt_\nu$$

   si confronta quindi il valore osservato con il valore limite, se $t_{oss} > t_\nu$(Lim) posso rigettare l'ipotesi con un livello di confidenza maggiore di 1-α. I valori limite si possono trovare sulle tavole (dispense, internet, etc…) o si possono calcolare utilizzando la funzione Excel:

   INV.T(α,ν)

Attenzione: INV.T(α,ν) calcola la probabilità associata alla variabile t_ν a due code (quando non conosco il segno della differenza $\bar{A} - \bar{B}$). Per un test ad una coda utilizzare: INV.T(2α,ν)





Problema: come calcolare $\sigma_{\bar{A}-\bar{B}}$:

a) se i due campioni hanno la stessa numerosità, cioè n = m:

$$\sigma_{\bar{A}-\bar{B}} = \sqrt{\sigma_{\bar{A}}^2 + \sigma_{\bar{B}}^2} = \sqrt{\frac{\sigma_a^2}{n} + \frac{\sigma_b^2}{n}} = \sqrt{\frac{\sigma_a^2 + \sigma_b^2}{n}}$$

dove σ_a^2 è la varianza dei dati A e $\sigma_{\bar{A}}$ l'errore standard sulla media $\bar{A}$ (lo stesso vale per i dati B), abbiamo usato il fatto che l'errore standard della media è:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

b) se i due campioni non hanno la stessa numerosità n ≠ m:

$$\sigma_{\bar{A}-\bar{B}} = \sqrt{\frac{(n-1)\sigma_a^2 + (m-1)\sigma_b^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)}$$

Che si riconduce alla precedente per m=n.

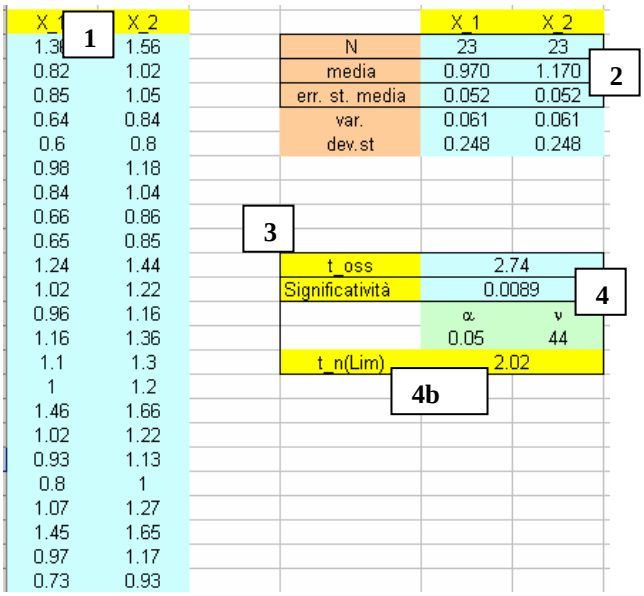
Il t-Test può essere usato per stabilire se, dato un set di dati X^{exp} (x_1, x_2 ….x_n) il valor medio $\bar{X}$ è significativamente diverso da un valore determinato μ. In questo caso la t_{oss}:

$$t_{oss} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

Deve essere confrontata con la statistica di una variabile t_ν con ν = n-1.

Esercizio 1) (il file Excel: **student.xls** mostra possibili soluzioni)
a) Il file **T_test1.dat** riporta i risultati di due set di misure ($n_1 = n_2$). Stabilire con un livello di confidenza 1-α=95% se le differenze osservate sono significative

1: importare i dati in un foglio Excel,
2: calcolare valori medi, varianza, dev.st ed errore standard della media
3: si calcola il valore di t_{oss}
4: usando la funzione DISTRIB.T si calcola il valore di significatività (due code): se i dati provenissero da una stessa distribuzione la probabilità di osservare un tale valore di t per caso è 0.89%, molto minore del limite del 5%, quindi le differenze sono significative con un livello di confidenza maggiore del 95% (maggiore del 99% in questo caso)
4b: usando la funzione INV.T si calcola il valore limite per una probabilità α = 0.05 (α=5%) per una variabile t_ν. Il valore t_{oss} = 2.74 è maggiore del valore limite t_ν(Lim)=2.02, quindi le differenze sono significative con un livello di confidenza maggiore del 95%

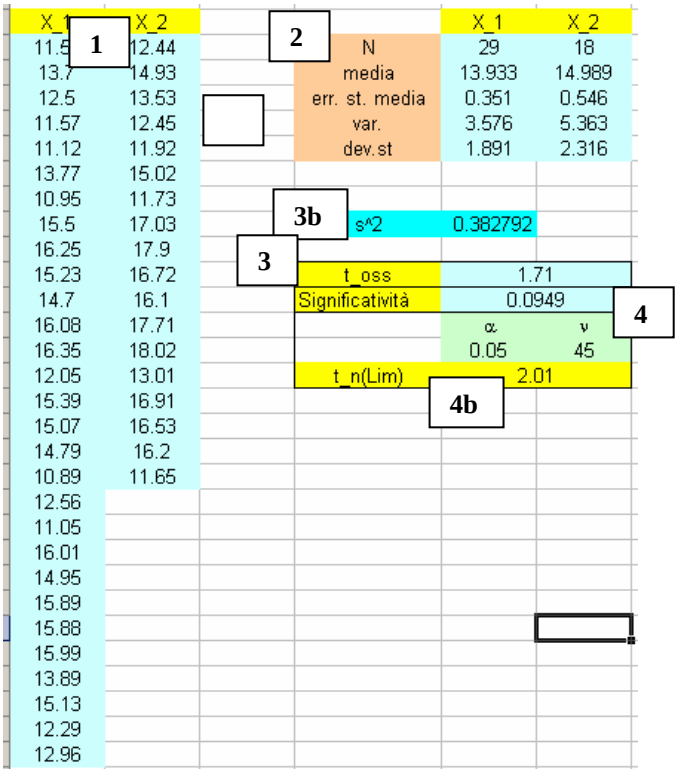


b) Il file **T_test2.dat** riporta i risultati di due set di misure (diversa numerosità). Stabilire con un livello di confidenza 1-α=95% se le differenze osservate sono significative

1: importare i dati in un foglio Excel,
2: calcolare valori medi, varianza, dev.st ed errore standard della media
3: si calcola il valore di t_{oss}.
3b: per non complicare troppo le formule è utile calcolare separatamente il valore:

$$\sigma_{\bar{A}-\bar{B}} = \sqrt{\frac{(n-1)\sigma_a^2 + (m-1)\sigma_b^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}$$

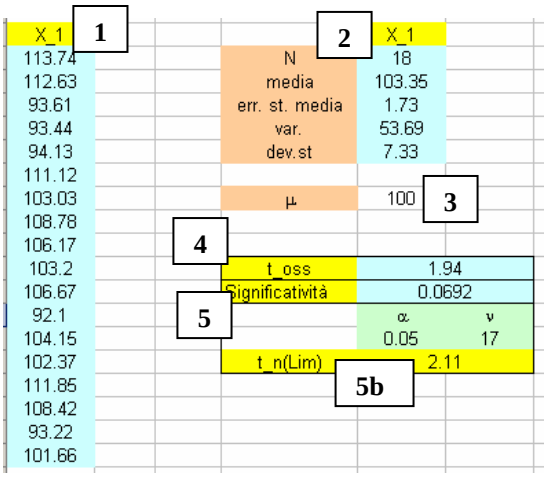
4: usando la funzione DISTRIB.T si calcola il valore di significatività (due code): se i dati provenissero da una stessa distribuzione la probabilità di osservare un tale valore di t per caso è 9.5%, maggiore del limite del 5%, quindi le differenze non sono significative con un livello di confidenza maggiore del 95%.
4b: usando la funzione INV.T si calcola il valore limite per una probabilità α = 0.05 (α=5%) per una variabile t_ν. Il valore osservato t_{oss} = 1.71 è minore del valore t_ν(Lim) quindi le differenze non sono significative con un livello di confidenza maggiore del 95%.



Nota: le differenze sono significative con un livello di confidenza del: 1-0.0949 = **90.5 %**

c) Il file **T_test3.dat** riporta i risultati di un set di misure. Stabilire con un livello di confidenza 1-α=95% se il valore medio è significativamente diverso da 100.

1: importare i dati in un foglio Excel,
2: calcolare valori medi, varianza, dev.st ed errore standard della media
3: definire il valore per il quale testare la compatibilità μ=100 in questo caso
4: si calcola il valore di t_{oss}.
5: usando la funzione DISTRIB.T si calcola il valore di significatività (due code): se i dati provenissero da una stessa distribuzione la probabilità di osservare un tale valore di t per caso è 6.9%, maggiore del limite del 5%, quindi le differenze **non sono significative** con un livello di confidenza maggiore del 95%.



5b: usando la funzione INV.T si calcola il valore limite per una probabilità α = 0.05 (α=5%) per una variabile t_v. Il valore osservato t_{oss} = 1.94 è minore del valore t_v(Lim) = 2.11 quindi le differenze **non sono significative** con un livello di confidenza maggiore del 95%.

Nota: le differenze sono significative con un livello di confidenza del (1-0.0692) = **93%.**

Regressione lineare

Una retta di regressione permette di stimare i parametri che caratterizzano una relazione lineare tra due variabili: una variabile indipendente (X) e una variabile dipendente (Y).
Siano X^{exp} (x_1, x_2 ….x_n) e Y^{exp} (y_1, y_2 ….y_n) un set di misure sperimentali tra le quali si ipotizza esista una relazione lineare del tipo:

$$Y^{th} = aX + b$$

Nell'ipotesi che

i) **l'errore sulle ascisse** (X) sia **nullo**,
ii) **l'errore sulle variabili dipendenti sia costante** e indipendente da X

si può utilizzare il metodo dei minimi quadrati: i parametri **a** e **b** sono quelli che minimizzano la somma dei quadrati degli scarti tra valori sperimentali (Y^{exp}) e valori misurati (Y^{th}):

$$\mathrm{Min}\sum_i (y_i^{exp} - y_i^{th})^2 = \mathrm{Min}\sum_i (y_i^{exp} - ax_i^{exp} - b)^2$$

Si dimostra che il termine noto (o intercetta) è:

$$b = \frac{(\sum_i y_i)(\sum_i x_i^2) - (\sum_i x_i)(\sum_i x_i y_i)}{n(\sum_i x_i^2) - (\sum_i x_i)^2}$$

e il coefficiente angolare è:

$$a = \frac{n(\sum_i x_i y_i) - (\sum_i x_i)(\sum_i y_i)}{n(\sum_i x_i^2) - (\sum_i x_i)^2}$$

I parametri a e b possono essere calcolati in modo esplicito (calcolando i vari termini) ma può essere complesso e fonte di errori di battitura. Più semplice è utilizzare le funzioni Excel:

PENDENZA(dati_Y;Dati_X): restituisce il coefficiente angolare della retta: **a**
INTERCETTA(dati_Y;Dati_X): restituisce l'intercetta della retta: **b**

Il parametro $\mathbf{s_{xy}}$ quantifica la variabilità dei dati attorno alla retta di regressione:

$$s_{xy} = \sqrt{\frac{\sum (y_i^{exp} - (ax_i + b))^2}{n-2}}$$

dove **n-2** sono i gradi di libertà per la stima della varianza. Ho dovuto ridurre di 2 i gradi di liberà poiché ho utilizzato i dati sperimentali per stimare il valore di due parametri: **a** e **b**. Per calcolare il parametro $\mathbf{s_{xy}}$ si può usare la funzione Excel:

ERR.STD.YX(dati_Y;Dati_X)

Attenzione all'ordine con cui vengono inseriti i dati: prima la colonna per la variabile dipendente Y e poi la colonna dei dati per la variabile indipendente X.

Il parametro $\mathbf{s_{xy}}$ serve per calcolare gli errori standard sui parametri **a** e **b** della retta di regressione:

$$s_a = \frac{s_{xy}}{\sqrt{\sum_i (x_i - \bar{x})^2}} \quad \text{e} \quad s_b = s_{xy}\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

Per il calcolo si noti che: $\sum_i (x_i - \bar{x})^2$ = **(n-1) var(x)**
Per stabilire se **a** o **b** sono significativamente diverso da 0 (o da un valore dato) utilizziamo il test t-

Student: la variabile t è definita:

$$t_\nu = \frac{\text{stima del parametro} - \text{valore da testare}}{\text{err. standard sulla stima del parametro}}$$

con ν=numero di gradi di libertà = numero di punti sperimentali – numero di parametri stimati, nel caso di una retta di regressione i parametri stimati sono 2: a e b, quindi ν=n-2. Si usano i valori s_a e s_b come errori standard sulle stime dei parametri.
Per applicare il test di Student si segue quanto fatto nell'esercizio precedente: o si usano le tabelle, o si usano le funzioni Excel.

Il coefficiente di correlazione lineare (coefficiente di correlazione di Pearson) permette di quantificare quanto la variabile dipendente dipende linearmente dalla variabile indipendente:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = (n-1)\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

dove s_x e s_y sono le deviazioni standard stimate sulla variabile X e sulla variabile Y.

La funzione Excel:

CORRELAZIONE(matriceY,matriceX)

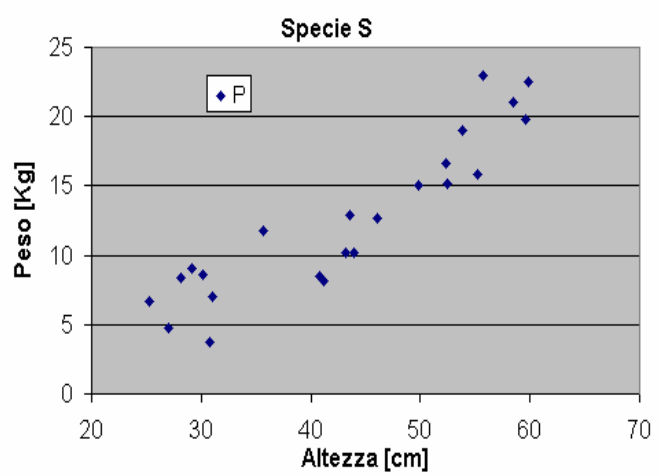
calcola il coefficiente di correlazione di Pearson.

Il coefficiente di determinazione R^2 rappresenta la frazione di variabilità dei dati Y^{exp} che può essere spiegata tramite la retta di correlazione. Nel caso di regressione lineare semplice è uguale al quadrato del coefficiente di correlazione: $R^2 = r^2$

Esercizio 2) il file **regressione.xls** contiene suggerimenti utili per la soluzione.
Sono stati misurati l'altezza (H) e il peso (P) per un gruppo di n=23 animali della specie **S** e i dati sono riportati nel file **Specie1.dat**.

0. importare i dati in un foglio Excel
1. graficare i dati Peso in funzione dell'altezza,
2. calcolare i parametri che definiscono la retta di regressione con i loro errori standard (riportare i dati con il numero corretto di cifre significative)
3. riportare su grafico la retta di regressione y = ax + b
4. stabilire se **a** e **b** sono diversi da 0 con un livello di confidenza del 95 % ($\alpha = 0.05$)
5. calcolare il coefficiente di correlazione e il parametro R^2

1.: si usi l'opzione grafica di Excel: **dispersione (xy)** per graficare i dati riportando sugli assi l'unità di misura (cm per l'altezza e kg per il peso)

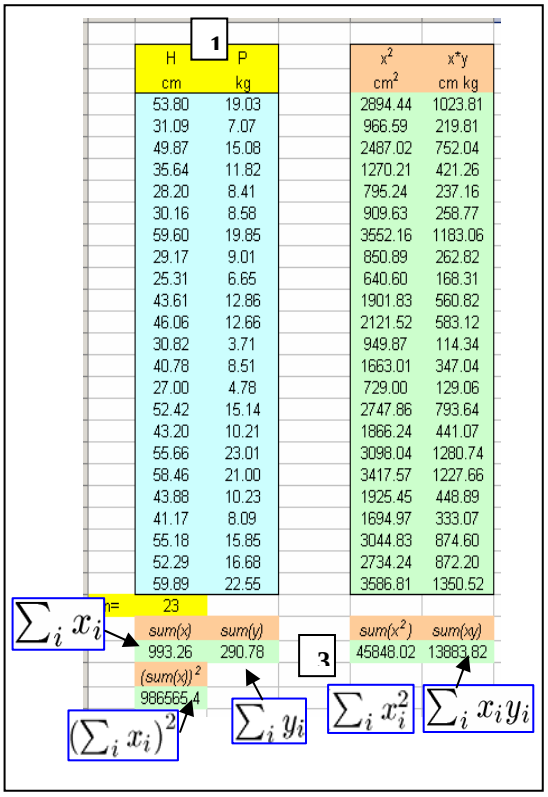


2/3.: per calcolare i parametri della retta di regressione possiamo seguire tre metodi. In modo complicato: si calcolano i termini x^2 e xy per ogni coppia (2) di dati e quindi le somme (3):

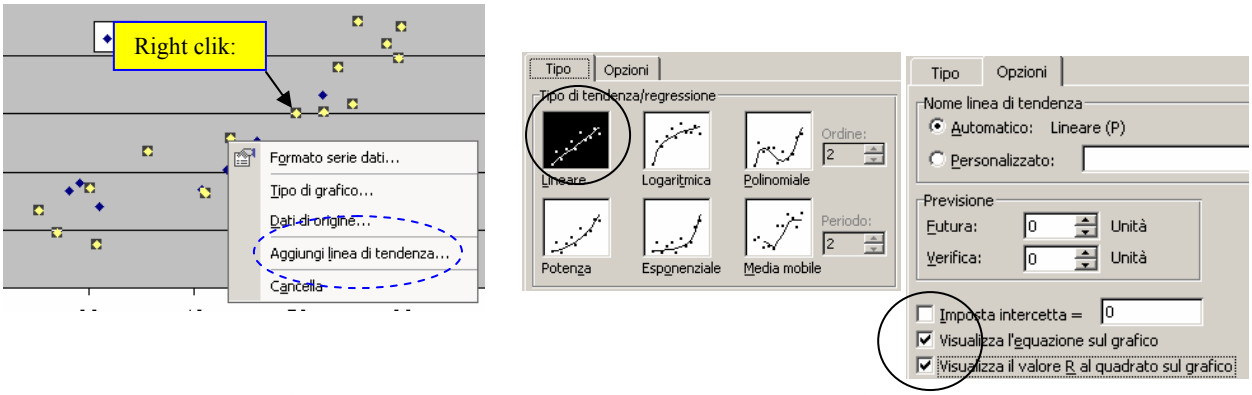
$$\sum_i x_i \quad \sum_i x_i^2 \quad \sum_i y_i$$
$$\left(\sum_i x_i\right)^2 \quad \sum_i x_i y_i$$

E, infine, si calcolano i termini a e:

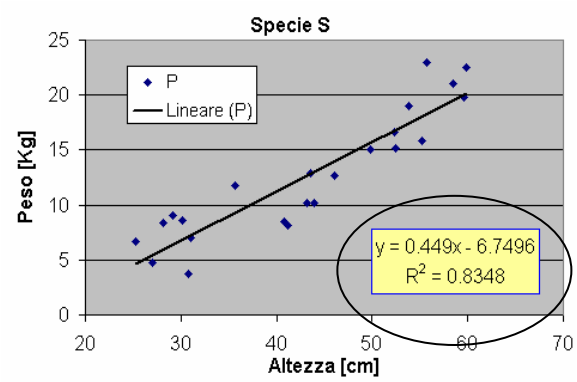
b=	-6.74956
a=	0.449047



Questo è un modo complesso. In modo più rapido si possono calcolare direttamente i parametri della retta di regressione dal grafico ciccando con il tasto destro del mouse su uno dei punti e scegliendo l'opzione: "aggiungi linea di tendenza". nella finestra "Opzioni" scegliere di visualizzare l'equazione e il valore di R^2 sul grafico:



Si ottiene:



La formula sul grafico mostra anche il parametro R^2, dal quale si ricava il coefficiente di correlazione:

$$r = (R^2)^{0.5}$$

Un terzo modo per calcolare i parametri della retta di regressione consiste nell'utilizzare le funzioni Excel:

PENDENZA(dati_Y;Dati_X): restituisce il coefficiente angolare della retta: **a**
INTERCETTA(dati_Y;Dati_X): restituisce l'intercetta della retta: **b**

Si dovrebbero ottenere gli stessi risultati mostrati in precedenza.

Per calcolare gli errori sui parametri bisogna calcolare $\mathbf{s_{xy}}$**:** si può usare la funzione Excel: **ERR.STD.YX(dati_Y;Dati_X)** e si ottiene:

$$\mathbf{s_{xy} = 2.37\ [kg]}$$

Il parametro $\mathbf{s_{xy}}$ serve per calcolare gli errori standard sui parametri **a** e **b** della retta di regressione:

Funzioni Excel			
b=	-6.749562	+/-	1.945896
a=	0.449047	+/-	0.043584
	s_{xy}	2.368748	

A questo punto è necessario riportare i risultati con il numero corretto di cifre significative: utilizzando due cifre significative per l'errore e accordando le cifre significative dei parametri si ottiene:

$$\mathbf{a = 0.449 \pm 0.044\ [kg\ cm^{-1}]}$$
$$\mathbf{b = -6.7 \pm 1.9\ [kg]}$$

4.: Per stabilire se **a** è significativamente diverso da 0 (lo stesso vale per b) utilizziamo il test t-Student

come descritto sopra. La t_{oss}:

$$t_{oss}(a) = \frac{a}{s_a}$$

quindi, per **a**:

$$t_{oss}(a) = 0.449 / 0.044 = 10.3$$

confrontiamo il valore ottenuto con il valore della variabile t-Student con 23-2=21 gradi di libertà calcolate per un rischio $\alpha = 0.05$.

Per questo possiamo
- utilizzare le tabelle della variabile t-Student,
- utilizzare la funzione

  INV.T(α , ν)

  e otteniamo un valore limite per $t_{21}(Lim) = 2.08$. Dal momento che osserviamo un valore di t molto maggiore ($t_{oss}(a) = 10.3$) possiamo affermare che il valore trovato è diverso da 0 con un livello di confidenza maggiore del 95% (rischio minore del 5%).
- possiamo utilizzare la funzione

  DISTRIB.T($t_{oss}(a)$, ν , 2)

  per calcolare la probabilità associata al valore di t osservato (Si usa un tes a due code poiché siamo interessati a sapere se a è significamene diverso da 0, se volessimo sapere se a è significamene maggiore di 0 potremmo usare un test a una coda). Nel nostro caso abbiamo:

  $$DISTRIB.T(10.03, 21, 2) = 1.1 \cdot 10^{-9}$$

  cioè la probabilità di osservare per caso il valore di t osservato nell'ipotesi che il valore di **a** sia nullo è molto piccola, quindi il valore trovato è significativo

Un procedimento analogo si può seguire per il parametro b. Si ottiene la tabella seguente:

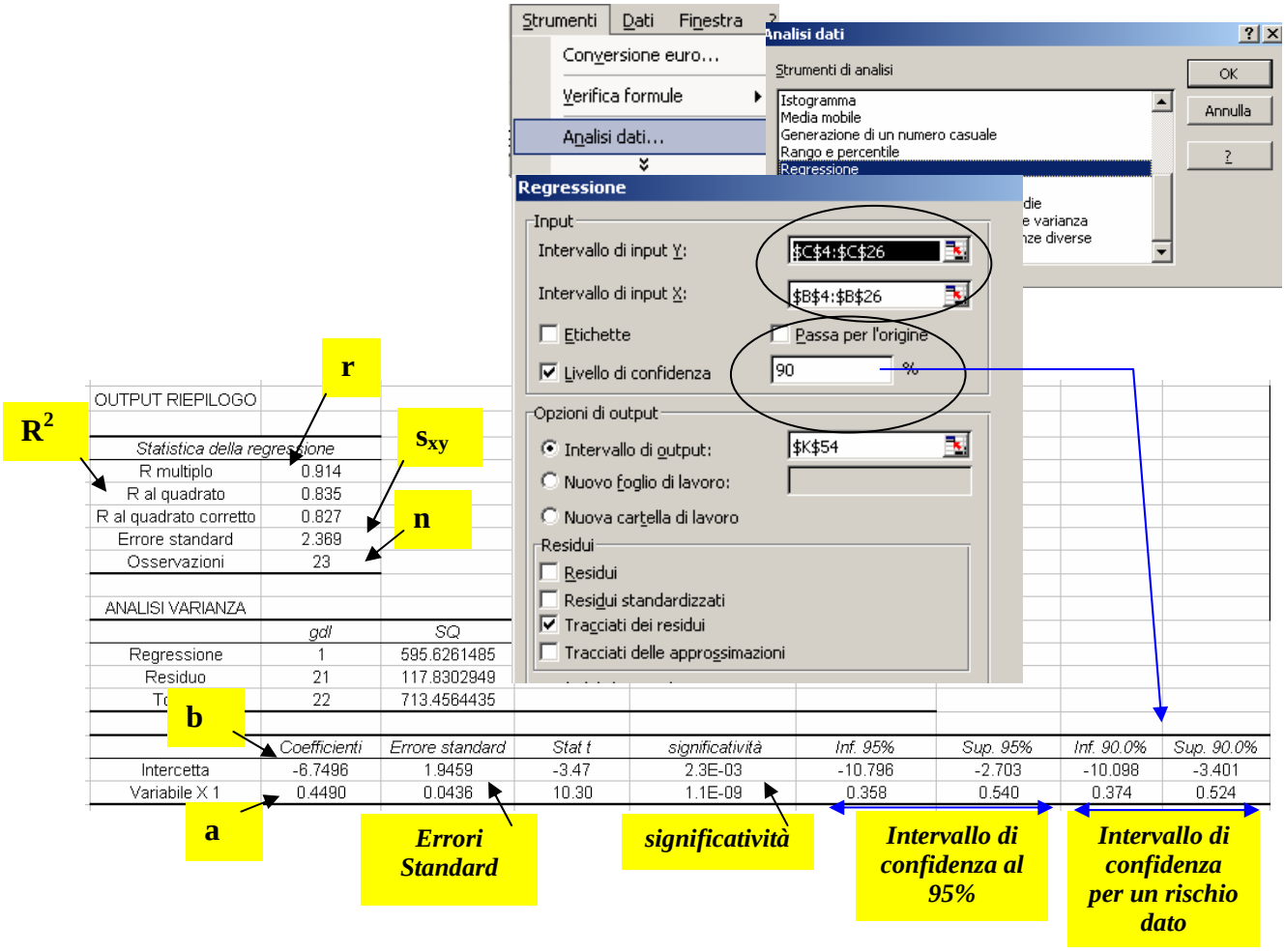
		t_limite
t(a)=	10.30	2.08
significatività	1.15E-09	
t(b)=	3.47	2.08
significatività	2.30E-03	

5.: per il calcolo del coefficiente di correlazione di Pearson usiamo la funzione Excel CORRELAZIONE(datiX,datiY): otteniamo la tabella:

Correlazione	0.913699
R^2	0.834846

Ci sono diversi modi per effettuare il calcolo in modo più veloce utilizzando Excel, ne vediamo uno di

questi che utilizza la macro "**regressione**" del menu: "**componenti aggiunitivi-Analisi dati**" :



La macro regressione fornisce molte informazioni statistiche sulla retta di regressione. Le principali sono evidenziate sopra. Oltre a quelle già viste (a, b, s_{xy}, errori su a e b) ci sono:

La significatività dei parametri: rappresenta la probabilità associata a un t-test sul parametro: rappresenta la probabilità di osservare per caso il valore di a o b trovato, nell'ipotesi che il valore vero di a o b sia nullo. Se questa probabilità è alta (es. > 5%) il valore trovato è poco significativo rispetto a 0.
In questo caso i valori trovati epr i parametri a e b sono diversi da 0 con un livello di confidenza molto maggiore del 95%.

L'intervallo di confidenza: intervallo di valori per il parametro (a o b) entro il quale è stimato trovarsi il valore vero con una probabilità del 95% Questo intervallo è dato per default. Si può definire un intervallo con un livello di confidenza definito.

Esercizio 3)
Sono stati misurati l'altezza (H) e il peso (P) per un gruppo di n=23 animali della specie **S** e i dati sono riportati nel file **Specie2.dat**. Gli animali sono stati trattati con una dieta di tipo diverso rispetto al caso precedente. Si vuole sapere se la dieta ha influenza sui parametri di crescita. Per questo si vuole stabilire se le rette di regressione Peso(altezza) sono diverse per i due gruppi di animali.

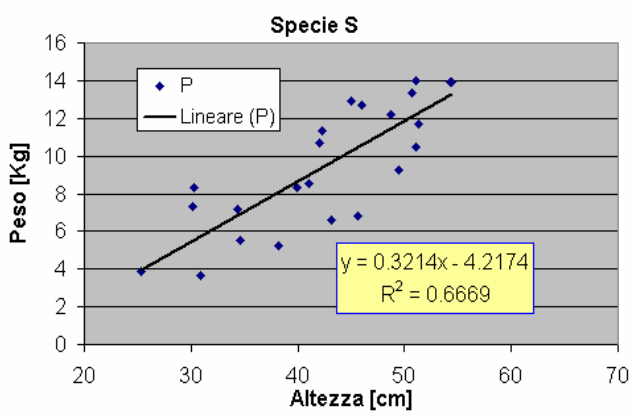
1.: calcolare la retta di regressione per gli animali del secondo gruppo.
2.: utilizzare un test t per stabilire se i coefficienti delle retta di regressione per i due gruppi sono diverse con un livello di significatività migliore del 5%.

Per il punto 1 possiamo utilizzare una copia del foglio usato in precedenza per calcolare rapidamente i parametri della regressione e la loro significatività:
Si vede che nel secondo caso il termine noto (intercetta) non è significativamente diverso da 0.

	H cm	P kg	x^2 cm^2	x*y cm kg
	34.32	7.15	1177.69	245.29
	54.41	13.94	2959.91	758.31
	38.16	5.20	1455.83	198.48
	48.70	12.17	2371.90	592.52
	25.31	3.85	640.48	97.40
	34.66	5.49	1201.04	190.40
	45.58	6.82	2077.24	310.90
	54.32	13.91	2950.27	755.36
	41.97	10.71	1761.72	449.38
	51.03	14.00	2604.14	714.57
	41.03	8.56	1683.35	351.26
	51.06	10.51	2606.65	536.46
	45.97	12.69	2112.94	583.24
	50.73	13.36	2573.05	677.52
	45.03	12.93	2028.08	582.20
	30.30	8.36	918.36	253.24
	39.88	8.34	1590.44	332.54
	43.10	6.59	1857.50	283.84
	51.30	11.72	2632.12	601.40
	30.91	3.68	955.70	113.71
	42.22	11.34	1782.22	478.63
	49.49	9.28	2449.41	459.27
	30.16	7.30	909.77	220.13
n=	23			
	sum(x)	*sum(y)*	*sum(x^2)*	*sum(xy)*
	979.6229	217.8723	43299.8067	9786.053
	(sum(x))^2			
	959661.1			

media X	42.59
dev.st. X	8.462244
var. X	71.60957

calcolo esplicito	
b=	-4.21735249
a=	0.32142102



Funzioni Excel			
b=	-4.21735249	+/-	2.150747
a=	0.32142102	+/-	0.049569
s_{xy}	1.967466		

t-TEST

		t_limite
t(a)=	6.48	2.08
significatività	2.00E-06	
t(b)=	1.96	2.08
significatività	6.33E-02	

Correlazione	0.816647
R^2	0.666912

OUTPUT RIEPILOGO

Statistica della regressione	
R multiplo	0.817
R al quadrato	0.667
R al quadrato corretto	0.651
Errore standard	1.967
Osservazioni	23

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	1	162.758	162.758	42.046	2.00216E-06
Residuo	21	81.289	3.871		
Totale	22	244.047			

	Coefficienti	*Errore standard*	*Stat t*	*significatività*	*Inf. 95%*	*Sup. 95%*	*Inf. 90.0%*	*Sup. 90.0%*
Intercetta	-4.217	2.151	-1.961	0.063	-8.690	0.255	-7.918	-0.516
Variabile X 1	0.321	0.050	6.484	2.002E-06	0.218	0.425	0.236	0.407

Per valutare se due rette sono significativamente diverse utilizziamo un t-test. In particolare vogliamo verificare se le differenze riscontrate nei coefficienti angolari e nelle intercette sono diverse con un livello di confidenza 1-α=95%
La variabile t-Student è definita:

$$t_{osservato} = \frac{|\text{differenza tra i parametri}|}{\text{err. standard sulla differenza tra i parametri}}$$

Quindi, per il coefficiente angolare (pendenza) si ha:

$$t_{oss} = \frac{|a_1 - a_2|}{s_{a_1 - a_2}}$$

Se il numero di osservazioni è lo stesso per ambedue le rette (è questo il caso) si ha:

$$t_{oss} = \frac{|a_1 - a_2|}{\sqrt{s_{a_1}^2 + s_{a_2}^2}}$$

L'analogo vale per l'intercetta.
Il valore della variabile t osservata deve essere confrontato con la statistica di una variabile t con ν gradi di libertà, dove: $\nu = n_1 + n_2 - 4$ (abbiamo in totale n1+n2 osservazioni ma abbiamo stimato 4 parametri a_1, b_1, a_2, b_2 dai dati sperimentali)

Otteniamo quindi la tabella seguente:

	retta 1	retta 2	t_{oss}	significatività	t_{lim}
a	0.449	0.321	1.934	0.060	2.018
s_a	0.044	0.050			
b	-6.750	-4.217	0.873	0.388	
s_b	1.946	2.151			
n	23	23			

Da cui possiamo concludere che, con un livello di confidenza del 95% non possiamo affermare che il coefficiente angolare e l'intercetta delle due rette siano diverse.

Nota: se il numero di osservazioni fosse diverso l'errore standard sulla differenza tra i parametri è un po' più complicato e non verrà trattato.