Inferenza su una media

σ^2 ignota

- Abbiamo imparato ad utilizzare la procedura classica dell'inferenza statistica sulla media μ della popolazione nel caso in cui la deviazione standard σ della popolazione sia conosciuta.
- Tale procedura fa uso del fatto che la media del campione $\bar{y}$ segue approssimativamente la distribuzione $N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$.
- Indipendentemente dalla distribuzione della popolazione, tale approssimazione diventa via via più accurata al crescere della numerosità n del campione (teorema del limite centrale).
- Se la popolazione è normale, ovvero se $Y \sim N(\mu; \sigma)$, allora $\bar{y} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$ anche se n è piccolo.

σ^2 ignota

- Quando la media della popolazione (μ) non è nota, di norma anche la sua varianza (σ^2) è ignota; di conseguenza, dobbiamo usare la varianza del campione (s^2) per stimare la varianza della popolazione.
- Con σ^2 ignota ed il ricorso all'uso di s^2 in sua sostituzione, la distribuzione delle probabilità non è più fornita dalla distribuzione normale Z ma da quella del t, detta t di Student, dallo pseudonimo di William Sealy Gosset.

σ^2 ignota

- Ci porremo ora l'obiettivo di usare la distribuzione t di Student per confrontare

  una media osservata e una media attesa, con calcolo dei limiti di confidenza di una media;

  le medie di due campioni indipendenti, con calcolo dei limiti di confidenza della media delle differenze;

  le medie di due campioni dipendenti, o per dati appaiati, con intervallo di confidenza della media delle differenze.

σ^2 ignota

- Benchè le medie dei campioni $\bar{y}$ siano distribuite normalmente con media μ e deviazione standard $\sigma/\sqrt{n}$, non possiamo fare uso di questo fatto dato che il parametro σ è ignoto.
- Possiamo però stimare la deviazione standard di $\bar{y}$ usando la deviazione standard del campione s in luogo del parametro sconosciuto σ.

La risultante deviazione standard stimata di $\bar{y}$ viene chiamata errore standard stimato di $\bar{y}$:

$$\widehat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}$$

Terminologia

- Il parametro $\sigma/\sqrt{n}$ è chiamato errore standard di $\bar{y}$;
- La statistica $s/\sqrt{n}$ è chiamata l'errore standard stimato di $\bar{y}$.

σ nota

Quando la deviazione standard σ della popolazione è nota, la verifica delle ipotesi e gli intervalli di confidenza per μ sono basati sulla media standardizzata del campione:

$$z_{=} = \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}}$$

dove $z_0 \sim N\,(0,\ 1)$.

Indice statistico

Quando σ è ignota, si calcola l'analoga statistica:

$$t_0 = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

ma questa statistica non segue la distribuzione normale.

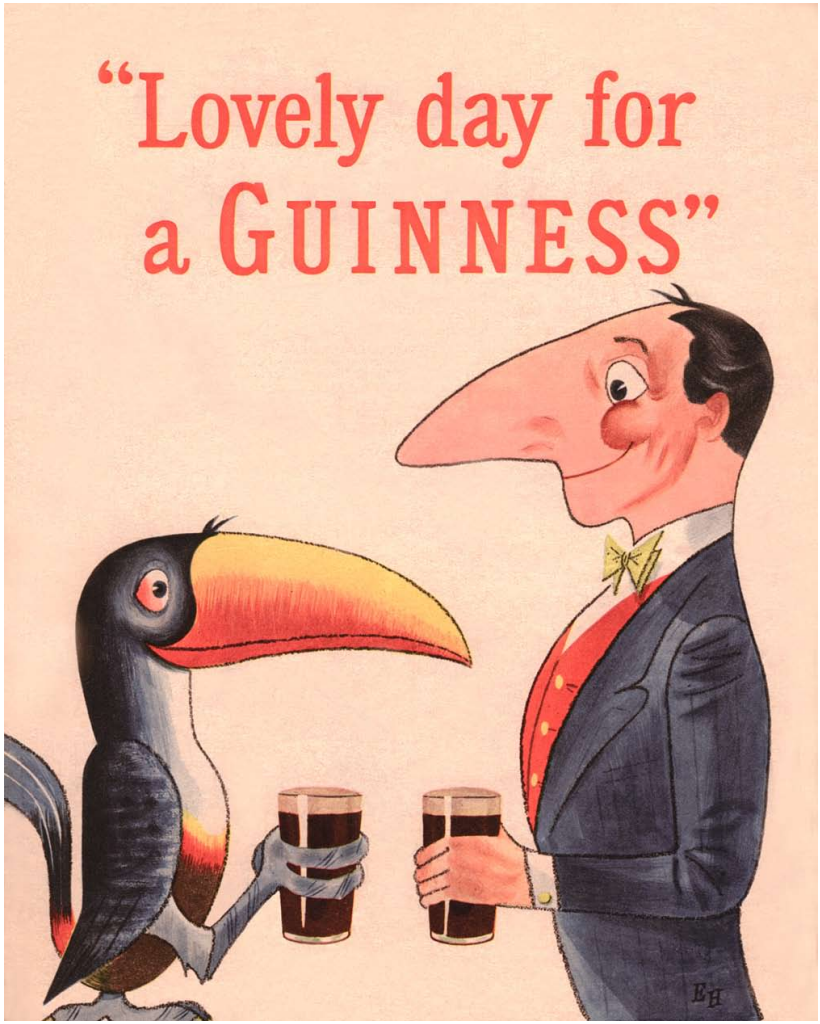
Se i valori Y sono distribuiti normalmente*, allora la statistica t_0 seguirà la distribuzione t di Student con $n-1$ gradi di libertà:

$$t_0 \sim t\,(n-1)$$

* t-Student approssima la normale per "n" grandi, conseguentemente il suo utilizzo concreto sarà limitato ai piccoli campioni, dove il Teorema del Limite Centrale non sarà applicabile.

Distribuzioni t di Student

Distribuzioni t di Student



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

Distribuzioni t di Student

- Le proprietà della distribuzione t sono particolarmente utili nello studio di fenomeni casuali relativi a piccoli campioni di osservazioni, cioè quando l'ampiezza di $n < 30$.
- La variabile casuale t è definita come rapporto tra una variabile casuale normale standardizzata e la radice quadrata di una variabile χ^2 (*che studieremo*) divisa per i suoi gradi di libertà, *ammesso che le due variabili siano tra loro indipendenti*.
- Anticipiamo il seguente fatto, *che studieremo*: *...aridaje*

$$\left(\frac{n-1}{\sigma^2}\right) s^2 \sim \chi^2_{(n-1)}$$

ossia, la varianza campionaria corretta s^2 ha una distribuzione campionaria χ^2 con $\nu = n - 1$ gradi di libertà.

Distribuzioni t di Student

- Abbiamo detto che la variabile aleatoria t viene espressa dal seguente rapporto

$$t_{(\nu)} = \frac{z}{\sqrt{\frac{\chi^2_{(\nu)}}{\nu}}},$$

- che possiamo costruire con le quantità note

$$t_{(n-1)} = \frac{z}{\sqrt{\frac{\chi^2_{(n-1)}}{n-1}}} = \frac{\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\,s^2}{\sigma^2}/\,(n-1)}} = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{s} = \boxed{\frac{\bar{y}-\mu}{s/\sqrt{n}}}$$

- dato che media e varianza campionaria sono tra loro indipendenti[1]

[1] $(y_i - \bar{y})$ è indipendente da $\bar{y}$: infatti se $\bar{y} + a \rightarrow [(y_i + a) - (\bar{y} + a)] = [y_i + \cancel{a} - \bar{y} - \cancel{a}]$

Distribuzioni t di Student

- Il rapporto ottenuto non dipende più dalla deviazione standard σ della popolazione e può essere interpretato coma la standardizzazione della media campionaria $\bar{Y}$, che diviene t Student, e tale sostituzione comporta una maggiore incertezza circa il valore di μ.
- La distribuzione t può quindi essere usata per compiere inferenze sulla media μ della popolazione usando esclusivamente valori campionari, quali $\bar{Y}$, s^2 ed n.

Distribuzioni t di Student

- La funzione di densità di probabilità t

$$f(t) = C\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

viene specificata dalla costante C, che rende l'area totale sotto la curva uguale ad uno, naturalmente dai valori t e sopratutto dal numero di gradi di libertà ν

- ha valore atteso $E(t) = 0$, come nella normale standard, dal momento che il valore $f(t)$ decresce simmetricamente attorno a tale valore:

$$f(t = 0) = C/\sqrt{1^{(\nu+1)}} > f(\mathbf{t = -1}) = C/\sqrt{\left(\frac{\nu+1}{\nu}\right)^{(\nu+1)}}$$ [...> 1]

$$f(t = 0) = C/\sqrt{1^{(\nu+1)}} > f(\mathbf{t = 1}) = C/\sqrt{\left(\frac{\nu+1}{\nu}\right)^{(\nu+1)}}$$

Distribuzioni t di Student

- mentre la varianza (per $\nu > 2$) dipende dal numero di gdl ν ed è data da

$$Var(t) = \frac{\nu}{\nu - 2},$$

che tende ad 1 (il valore della normale standard) al crescere di ν:

$$\frac{3}{1} = 3, \quad \frac{5}{3} = 1.67, \quad \frac{15}{13} = 1.15, \quad \frac{30}{28} = 1.07, \quad \frac{100}{98} = 1.02,$$

- Le curve t sono tutte *campanulari* e *simmetriche*. Quelle che hanno un ν piccolo sono notevolmente più appiattite di quella normale , ma al crescere di ν la distribuzione di Student tende alla distribuzione normale

Distribuzioni t di Student

I gradi di libertà sono dati dal denominatore della deviazione standard s del campione:

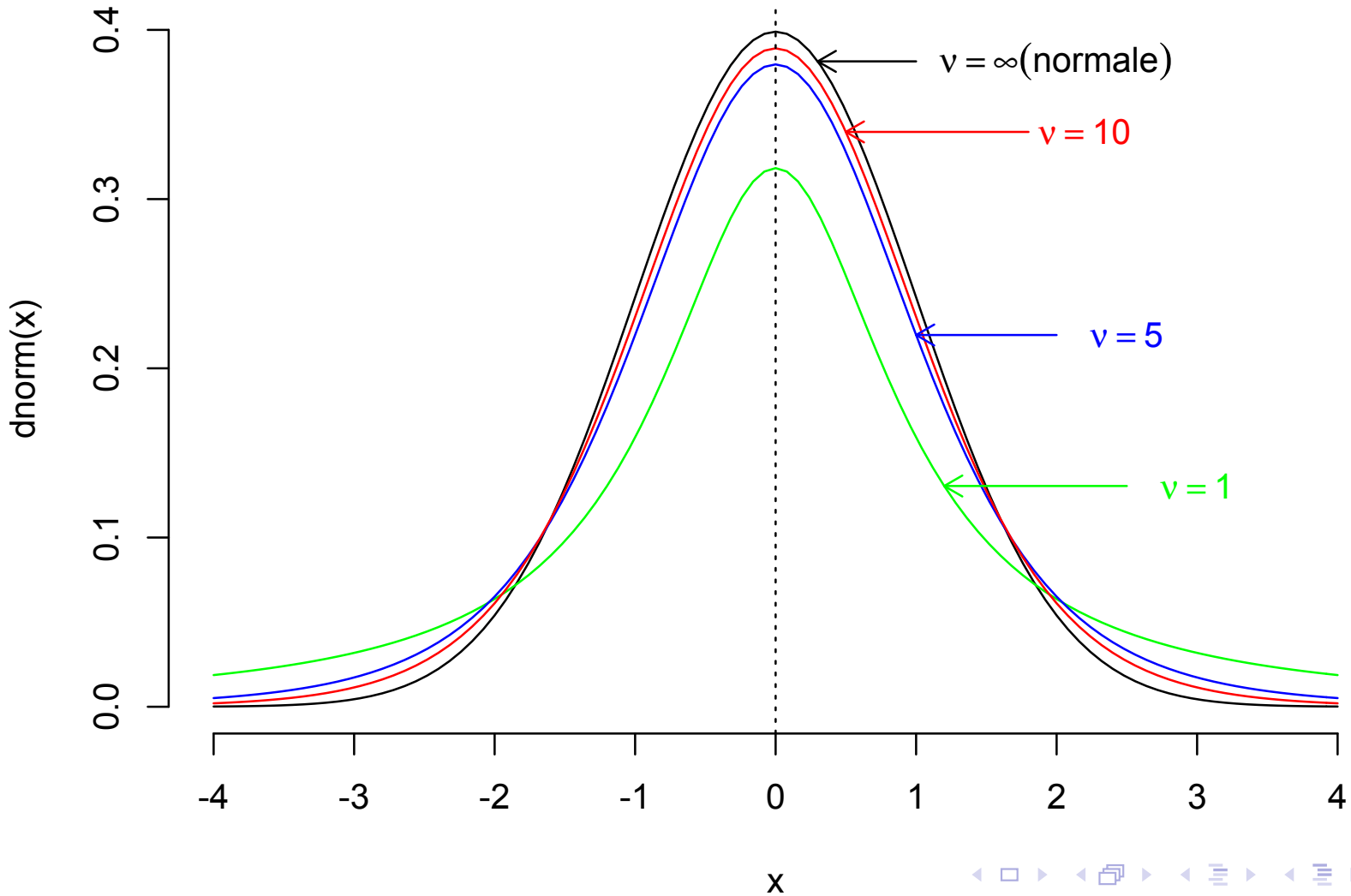
$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

All'aumentare di n, s diventa via via una stima sempre migliore di σ e si ha la convergenza dei valori della distribuzione t di Student verso la distribuzione normale standardizzata.

Distribuzioni t di Student

- La forma della distribuzione t di Student è simmetrica e a campana come la normale, ma con una dispersione maggiore.
- Ciò riflette l'incertezza aggiuntiva che deriva dal fatto che il parametro σ è stato stimato mediante s.

Distribuzioni t di Student



Intervallo di confidenza

Intervallo di confidenza

- Gli intervalli di confidenza e i test delle ipotesi statistiche basati sulla distribuzione t di Student sono molto simili a quelli basati sulla distribuzione normale.
  - Anziché usare i valori critici della distribuzione normale, però, dobbiamo usare i valori critici t.

Questi valori possono essere trovati usando la Tavola 2 in Appendice del libro di Luccio & Caudek, per diversi valori dei gradi di libertà.

Intervallo di confidenza

Illustrazione. Consideriamo nuovamente l'esperimento fittizio discusso in precedenza

- Uno studio di pedagogia sperimentale intende stabilire se l'utilizzo del computer in un corso introduttivo di statistica migliori l'apprendimento.
- Vengono contattati dieci docenti che insegnano un corso introduttivo di statistica presso la Facoltà di Psicologia. Ciascun docente divide in maniera casuale i suoi studenti in due gruppi: un gruppo userà il computer, mentre l'altro gruppo non lo userà.
- A fine corso, lo stesso esame viene somministrato a tutti gli studenti.

Illustrazione

I due metodi d'insegnamento producono in media una differenza pari a 9.40 con deviazione standard uguale a 8.38.

Il punteggio medio all'esame per ciascun gruppo è fornito nella tabella seguente, insieme alla differenza tra i punteggi dei gruppi che hanno utilizzato i due metodi d'insegnamento.

Illustrazione

Docente	Metodo computer	Metodo tradizionale	Differenza
1	94	71	23
2	75	70	5
3	75	58	17
4	84	80	4
5	79	70	9
6	85	67	18
7	73	66	7
8	75	80	-5
9	75	72	3
10	83	70	13
		media	$\bar{x} = 9.4$
		deviazione standard	$s = 8.38$

Illustrazione

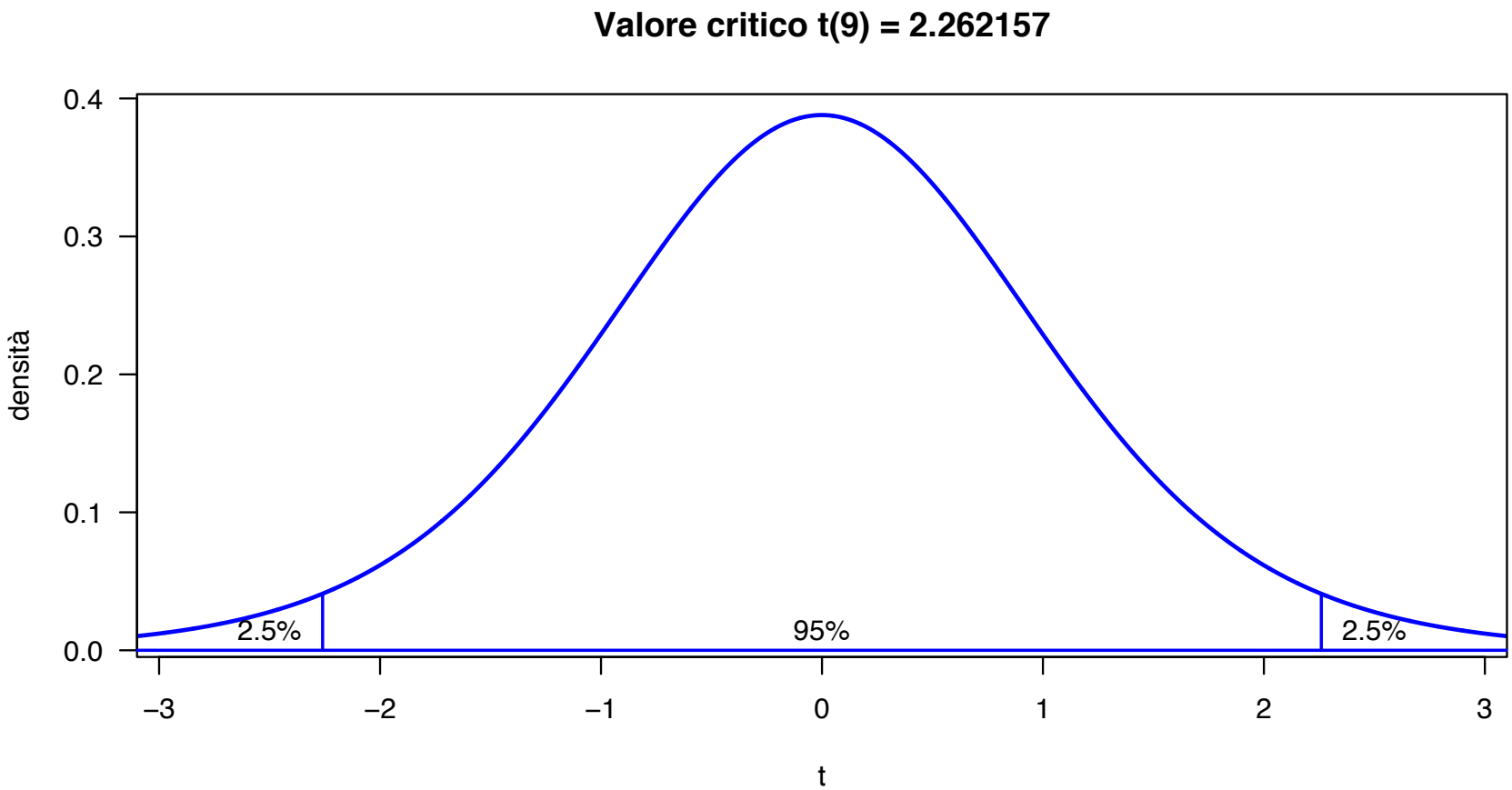
- In precedenza abbiamo analizzato questi dati in maniera inappropriata, fingendo che la deviazione standard s del campione fosse uguale alla deviazione standard σ della popolazione.
- Usiamo ora la distribuzione t di Student per costruire l'intervallo di confidenza al 95%.
  - Gradi di libertà $\longrightarrow n - 1 = 10 - 1 = 9$.
  - Il valore critico $t = 2.262$ è il valore della distribuzione t con 9 gradi di libertà alla probabilità $\alpha/2 = 0.025$.
  - Si noti che il valore $t_{(\alpha/2;\, n-1)}$ è più grande del corrispondente valore critico $z = 1.96$, e $t_{(\alpha/2;\, n-1)}$ e produrrà quindi un intervallo di confidenza più grande.

Illustrazione

- La formula per il calcolo dell'intervallo di confidenza è

$$
\begin{aligned}
I.C.(95\%) = & \quad \bar{y} \pm t_{(\alpha/2;\, n-1)} \frac{s}{\sqrt{n}} \\
= & \quad 9.40 \pm 2.262 \frac{8.38}{\sqrt{10}} = 9.40 \pm 5.99
\end{aligned}
$$

Illustrazione



Test t di Student

Test t di Student

- Per verificare l'ipotesi nulla secondo cui il nuovo metodo dà gli stessi risultati di quello tradizionale

$$H_0 : \mu \leq 0$$

contro l'ipotesi alternativa

$$H_a : \mu > 0$$

dobbiamo calcolare la statistica

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

Test t di Student

Illustrazione. Nel caso dell'esempio precedente, la statistica test assume il seguente valore:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = t = \frac{9.4 - 0}{8.38/\sqrt{10}} = 3.547$$

Test t di Student

- Sotto l'ipotesi nulla questa statistica segue la distribuzione t di Student con 9 gradi di libertà.
- Dato che l'ipotesi alternativa è un'ipotesi unilaterale che specifica un valore positivo per il parametro μ, dobbiamo trovare il $p-$valore per la coda destra della t di Student con 9 gradi di libertà (gdl).
- Il $p-$valore esatto della statistica test può essere trovato tramite software, ma se dobbiamo usare le tabelle non disponiamo dei valori di probabilità esatti.

Test t di Student

Con 9 gdl in un test ad una coda, nella tabella troviamo:

df	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
9	1.383	1.833	2.262	2.821	3.250

Test t di Student

- Il valore osservato $t_9 = 3.547$ è maggiore di $t_{(0.005;\, 9)} = 3.250$.
- Il $p-$valore, dunque, sarà minore di 0.005.
- Questo risultato solitamente si riporta nel modo seguente:

> *il nuovo metodo d'insegnamento produce all'esame punteggi significativamente più alti del metodo tradizionale [*$t(9) = 3.547$, $p < 0.005$, *test a una coda].*

Test t di Student

- Con un'ipotesi alternativa non-direzionale,

$$H_a : \mu \neq 0$$

il $p-$valore verrebbe raddoppiato:

... *[* $t(9) = 3.547$, $p < 0.01$, *test a due code]*

Test t di Student con R

- Oppure, più semplicemente:

```
t.test(diff, alternative="greater")

One Sample t-test

data:  diff
t = 3.5461, df = 9, p-value = 0.003127
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 4.54081     Inf
sample estimates:
mean of x
      9.4
```

Test t di Student con R

- Per un test bilaterale:

```
  t.test(diff, alternative="two.sided")

One Sample t-test

data:  diff
t = 3.5461, df = 9, p-value = 0.006254
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  3.403505 15.396495
sample estimates:
mean of x
      9.4
```

Condizioni di validità

Condizioni di validità

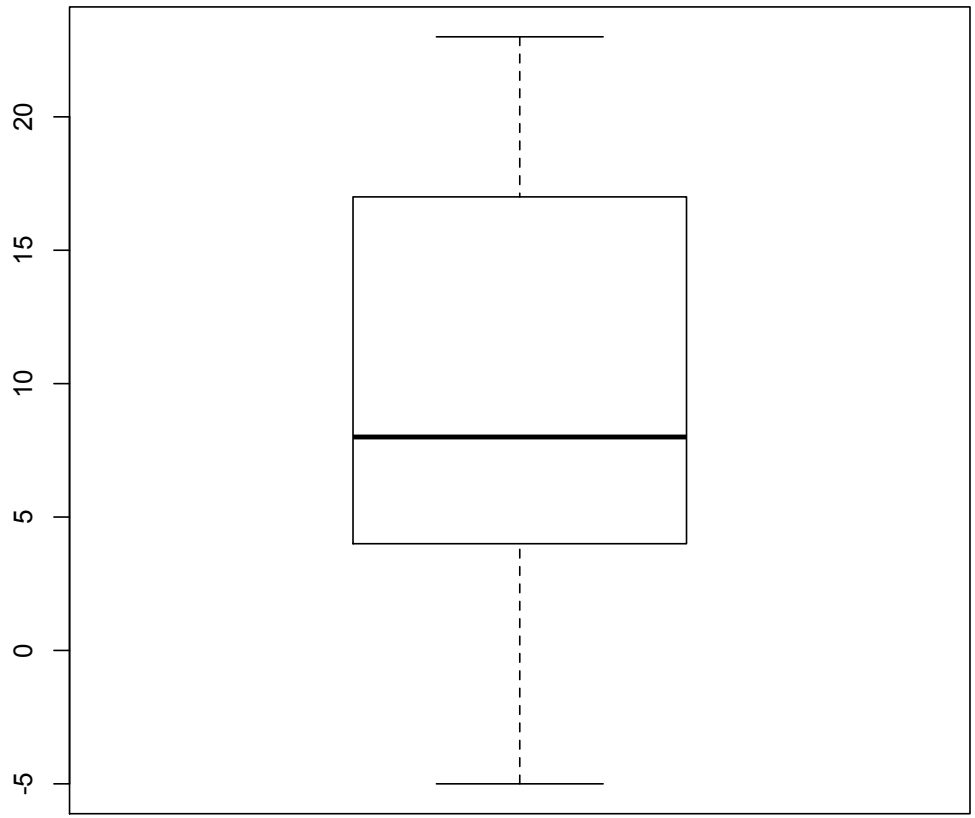
- Affinché possa essere ritenuta valida, l'applicazione del test t di Student richiede che siano rispettate alcune condizioni essenziali, valide anche per l'analisi della varianza, che ne rappresenta la generalizzazione.
- Le ipotesi sono essenzialmente tre:
  1. l'indipendenza dei dati entro e tra campioni: i dati provengono da un campione casuale semplice estratto da una popolazione molto più grande;
  2. l'omogeneità della varianza: il confronto tra due medie è valido se e solo se le popolazioni dalle quali i campioni sono estratti hanno varianze simili;
  3. la normalità: i dati sono distribuiti normalmente.

Condizioni di validità

- In campioni di piccole dimensioni, i test delle ipotesi e gli intervalli di confidenza basati sulla distribuzione t di Student sono corretti se la popolazione è distribuita normalmente.
- L'assunzione di normalità dovrebbe essere controllata esaminando i dati ma, nel caso di piccoli campioni (dove quest'assunzione è veramente importante), la violazione di questa assunzione è difficile da verificare.
- È però importante verificare l'eventuale presenza di dati anomali (*outliers*).

Condizioni di validità

Per i dati dell'esempio discusso in precedenza, possiamo dire che "tutto è in ordine".



Campioni dipendenti

Campioni dipendenti

- Gli intervalli di confidenza e i test di ipotesi vengono comunemente applicati a dati provenienti da campioni dipendenti.
- L'esperimento fittizio che abbiamo discusso in precedenza fornisce un esempio di dati appaiati: ciascun docente è impegnato con due gruppi di studenti (un gruppo per il metodo tradizionale e uno per il metodo nuovo).

Campioni dipendenti

- Il piano dell'esperimento sarebbe diverso se vi fossero 20 diversi docenti per i 20 gruppi.
  - In quel caso, a 10 docenti scelti in maniera casuale verrebbe assegnato il nuovo metodo d'insegnamento e ai rimanenti il metodo tradizionale.
  - Sarebbero così definiti due campioni indipendenti.

Campioni dipendenti

Altri comuni esempi di campioni dipendenti:

- studi che utilizzano un pre-test e post-test;

  potremmo misurare, per esempio, il numero medio $\bar{y}$ di incidenti sul lavoro in $n = 10$ fabbriche prima e dopo l'approvazione di una legge che impone un maggior numero di controlli;

- campionamento di osservazioni naturalmente associate;

  potremmo misurare la differenza del livello di depressione di mariti e mogli che hanno subito un evento traumatico.

  Tale studio è diverso da quello in cui mariti e mogli vengono campionati in maniera indipendente.

Campioni dipendenti

Il confronto tra le medie di due campioni di osservazioni appaiate è semplice:

l'analisi è applicata ad una nuova serie di dati, quelli risultanti dalle differenze d_i tra i valori delle u.s.* di ciascuna coppia.

* u.s. : Unità Sperimentali ("singoli" dati nel data set, a volte frutto di precedenti elaborazioni).

Campioni dipendenti

- Nel caso di un test bilaterale, l'ipotesi nulla H_0 è che la media δ della popolazione delle differenze sia uguale a 0, $H_0 : \delta = 0$, mentre l'ipotesi alternativa H_a è $H_a : \delta \neq 0$.
- In un test unilaterale, l'ipotesi nulla H_0 è che la media della popolazione delle differenze sia maggiore o uguale a 0, $H_0 : \delta \geq 0$, mentre l'ipotesi alternativa H_a afferma che la differenza è minore di 0, $H_a : \delta < 0$.
- Nel caso opposto, l'ipotesi nulla H_0 è che la media della popolazione delle differenze sia minore o uguale a 0, $H_0 : \delta \leq 0$, mentre l'ipotesi alternativa H_a afferma che la differenza è maggiore di 0, $H_a : \delta > 0$.

Campioni dipendenti

La significatività della media delle differenze è verificata con il rapporto

$$t_{n-1} = \frac{\bar{d} - \delta}{\frac{s_d}{\sqrt{n}}}$$

- $\bar{d}$ è la media delle differenze $d = d_1; d_2; \ldots; d_n$, dove ciascuna d_i è lo scarto fra le osservazioni corrispondenti della i-esima unità nei due campioni, ovvero $d_i = y_{1_i} - y_{2_i}$.
- δ è la differenza media attesa: spesso, ma non necessariamente, è uguale a 0,
- s_d è la deviazione standard campionaria corretta delle differenze campionarie d.
- n è il numero di differenze, corrispondente anche al numero di coppie di dati.

Campioni dipendenti

- L'intervallo di confidenza della media delle differenze $\bar{d}$ tra i due campioni dipendenti al livello α è calcolato mediante

$$\bar{d} \pm t_{(\alpha/2;\, n-1)} \frac{s_d}{\sqrt{n}}$$

dove $t_{(\alpha/2;\, n-1)}$ indica il valore della distribuzione t con $n-1$ gradi di libertà alla probabilita $\alpha/2$.

Due campioni indipendenti

Due campioni indipendenti

Il confronto tra le medie di due campioni indipendenti rappresenta il caso più comune di analisi di dati empirici. Si procede sulla base delle seguenti assunzioni.

- Due campioni casuali indipendenti sono stati estratti da due popolazioni diverse.

  I campioni appaiati sono un esempio di campioni *dipendenti*.
- Ciascuna popolazione è distribuita normalmente con media e varianza sconosciute.

Due campioni indipendenti

La seguente notazione verrà usata per descrivere le due popolazioni:

Popolazione	Variabile	Media	Deviazione standard
1	Y_1	μ_1	σ_1
2	Y_2	μ_2	σ_2

Due campioni indipendenti

- Ci poniamo il problema di confrontare le medie delle popolazioni μ_1 e μ_2.
- Possiamo costruire un intervallo di confidenza della differenza media $\mu_1 - \mu_2$, oppure verificare l'ipotesi nulla:

$$H_0 : \mu_1 = \mu_2$$

ovvero

$$H_0 : \mu_1 - \mu_2 = 0$$

Due campioni indipendenti

Le informazioni fornite dai due campioni sono:

Popolazione	Dimensione del campione	Media	Deviazione standard
1	n_1	$\bar{y}_1$	s_1
2	n_2	$\bar{y}_2$	s_2

È naturale usare la differenza $\bar{y}_1 - \bar{y}_2$ per stimare la differenza tra le due medie delle popolazioni, $\mu_1 = \mu_2$.

Due campioni indipendenti

- Dato che le deviazioni standard σ_1 e σ_2 delle due popolazioni sono sconosciute, useremo le deviazioni standard dei campioni s_1 e s_2 per stimare questi due parametri.
- Per eseguire un test statistico sulla differenza $\mu_1 = \mu_2$ dobbiamo conoscere le proprietà della distribuzione campionaria di $\bar{y}_1 - \bar{y}_2$.

Due campioni indipendenti

- Il valore atteso di $\bar{y}_1 - \bar{y}_2$ è $\mu_1 - \mu_2$. La differenza delle medie dei campioni fornisce una stima centrata sul parametro della differenza delle medie delle popolazioni:
  $E(\bar{y}_1 - \bar{y}_2) = E(\bar{y}_1) - E(\bar{y}_2) = \mu_1 - \mu_2$.
- La varianza di $\bar{y}_1 - \bar{y}_2$ è

$$\sigma^2_{\bar{y}_1 - \bar{y}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

e dunque l'errore standard di $\bar{y}_1 - \bar{y}_2$ è

$$\sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Due campioni indipendenti

- Se le due popolazioni sono distribuite normalmente, allora anche la statistica $\bar{y}_1 - \bar{y}_2$ si distribuirà normalmente.
- Se le deviazioni standard σ_1 e σ_2 delle due popolazioni fossero conosciute, potremmo basare l'inferenza statistica sulla distribuzione normale in quanto il valore standardizzato

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

segue la distribuzione normale, $N\,(0;\ 1)$.

Due campioni indipendenti

Viene invece usata la statistica

$$t_\nu = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Se i due campioni casuali semplici provengono da due popolazioni indipendenti distribuite normalmente, allora la statistica precedente segue approssimativamente la distribuzione t con ν gdl.

Due campioni indipendenti

I gradi di libertà ν possono essere calcolati usando l'approssimazione di Welch-Satterthwaite:

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

Una stima conservativa di ν è data dal minore di $n_1 - 1$ e $n_2 - 1$.

Intervallo di confidenza

Per costruire l'intervallo di confidenza al livello $1 - \alpha$, si calcola

$$\bar{y}_1 - \bar{y}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

laddove $t*$ è l'appropriato valore critico della distribuzione t con gradi di libertà uguali al minore di $n_1 - 1$ e $n_2 - 1$ (o al ν dell'equazione di Welch-Satterthwaite).

Per verificare l'ipotesi nulla $H_0 : \mu_1 = \mu_2$, si calcola la statistica

$$t_\nu = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Si noti che il numeratore della statistica test deriva da $(\bar{y}_1 - \bar{y}_2) - 0$, ovvero, 0 è il valore ipotizzato per $\mu_1 - \mu_2$.
- Il test sarà a una coda o a due code a seconda che l'ipotesi alternativa sia unilaterale

$$H_a : \mu_1 > \mu_2 \quad \text{oppure} \quad H_a : \mu_1 < \mu_2$$

o bilaterale

$$H_a : \mu_1 \neq \mu_2$$

Due campioni indipendenti

Illustrazione. Si consideri il seguente esempio basato sui dati contenuti nel *data frame* `Ginzberg`.

- La variabile dipendente (`adjdep`) è il punteggio sulla *Beck self-report depression scale*.
- I soggetti sono stati divisi in due gruppi a secondo del punteggio ottenuto su una scala (`adjsimp`) che misura il *bisogno di vedere il mondo "in bianco e nero"*
- I soggetti con punteggi minori o uguali alla media del campione su questa variabile sono stati assegnati al gruppo *"semplicità = bassa"*; gli altri sono stati assegnati al gruppo *"semplicità = alta"*.

Illustrazione

- Per ciascun gruppo sono state calcolate la media e la deviazione standard dei punteggi sulla *Beck self-report depression scale.*
- I risultati sono i seguenti:

Depressione	n	$\bar{y}$	s
semplicità bassa	46	0.8308	0.4433
semplicità alta	36	1.2162	0.4901

Illustrazione

- Il minore di $n_1 - 1$ e $n_2 - 1$ è $36 - 1$=35.
- Il valore critico di $t_{(0.025,\ 35)}$ è 2.0301.

```
qt(0.975,df=35)
## [1] 2.030108
```

- L'intervallo di confidenza al 95% è :

$$(\bar{y}_1 - \bar{y}_2) \pm t_{(0.025,\ 35)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(1.2162 - 0.8308) \pm 2.0301 \sqrt{\frac{0.4901^2}{36} + \frac{0.4433^2}{46}}$$

$$= 0.3853 \pm 0.2124 = [0.1729;\ 0.5977]$$

Illustrazione

- usando l'equazione di Welch-Satterthwaite:

$$\nu = \frac{(0.4901/36 + 0.4433/46)^2}{\frac{(0.4901/36)^2}{35} + \frac{(0.4433/46)^2}{45}} = 71.40065$$

- Il valore critico di $t_{(0.025,\ 71.40065)}$ diventa ≈ 1.9937.

```
qt(0.975,df=71.40065)
## [1] 1.993749
```

- e l'intervallo di confidenza al 95% di Welch-Satterthwaite risulta essere:

$$(1.2162 - 0.8308) \pm 1.9937\sqrt{\frac{0.4901^2}{36} + \frac{0.4433^2}{46}}$$

$$= 0.3853 \pm 0.2086 = [0.1768;\ 0.5940]$$

con un margine di errore $0.2086 < 0.2124$ legato all'aumento dei gdl da $\nu = 35$ a $\nu \approx 71$

Illustrazione

- Per verificare l'ipotesi nulla

$$H_0 : \mu_1 \geq \mu_2$$

contro l'ipotesi alternativa unilaterale

$$H_a : \mu_1 < \mu_2$$

(individui con punteggi più bassi sulla scala "semplicità" hanno punteggi di depressione più bassi), si calcola

$$t_\nu = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{1.2162 - 0.8308}{\sqrt{\frac{0.4901^2}{36} + \frac{0.4433^2}{46}}} = 3.6832$$

Illustrazione

- Il valore critico t con $\alpha = 0.05$ per un test unilaterale con $\nu 35$ gdl è

```
qt(0.95,df=35)
## [1] 1.689572
```

- usando il valore critico t di Welch-Satterthwaite con $\nu = 71.40065$, otteniamo un valore critico meno restrittivo

```
qt(0.95,df=71.40065)
## [1] 1.666476
```

- che riduce l'area di protezione dell'ipotesi nulla e aumenta quindi la potenza del test statistico.

Illustrazione

- Dato che la statistica $t = 3.6832$ osservata è maggiore dei valori critici $t_{35} = 1.689572$ e $t_{\nu=71.40065} = 1.666476$, possiamo rifiutare l'ipotesi nulla secondo cui i due gruppi di individui (con basso e alto bisogno di vedere il mondo "in bianco e nero") hanno lo stesso livello di depressione in media sulla scala Beck.
- Se fosse stata specificata un'ipotesi alternativa bilaterale, i valori critici sarebbero stati

```
 qt(0.975,df=35)
## [1] 2.030108
 qt(0.975,df=71.40065)
## [1] 1.993749
```

Illustrazione

- Questa analisi si può eseguire in R nel modo seguente

```
t.test(depressione~semplicità, alternative = c("two.sided"),
           mu = 0, paired = FALSE, var.equal = FALSE)
```

- dove il comando `alternative = c("two.sided")` imposta un test bilaterale, in cui il $p - value = 0.0002223$ verrà moltiplicato per 2.

```
Welch Two Sample t-test

data:  depressione by semplicità
t = 3.6832, df = 71.401, p-value = 0.0004446
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1767460 0.5939139
sample estimates:
 mean in group alta mean in group bassa
          1.2161608           0.8308309
```

Illustrazione

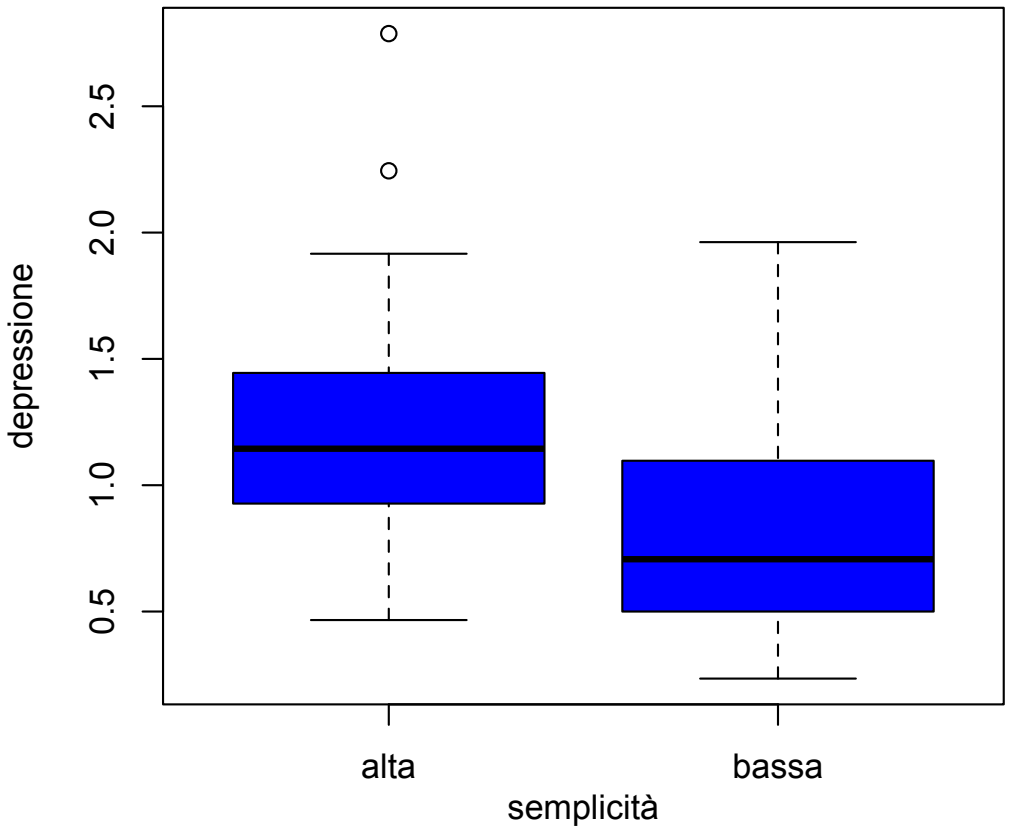
- Invece `alternative = c("greater")` imposta un test unilaterale (destro), in cui il $p - value = 0.0004446$ precedente viene diviso per 2.

```
t.test(depressione~semplicità, alternative = c("greater"),
          mu = 0, paired = FALSE, var.equal = FALSE)

Welch Two Sample t-test

data:  depressione by semplicità
t = 3.6832, df = 71.401, p-value = 0.0002223
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.210985      Inf
sample estimates:
 mean in group alta mean in group bassa
         1.2161608           0.8308309
```

Illustrazione

L'esame grafico dei dati rivela però un potenziale problema:

- ci sono due valori anomali nel gruppo "semplicità alta":



Illustrazione

Se i due dati anomali vengono eliminati

```
 outlier<-boxplot(depressione~semplicità)$out
 outlier

 posizione<-which(depressione==outlier)
 posizione
## [1] 71 80

 depressione<-depressione[-c(posizione)]
 semplicità<-semplicità[-c(posizione)]
```

Illustrazione

il test t di Student bilaterale ($p - value \times 2$) diventa:

```
  t.test(depressione~semplicità, alternative = c("two.sided"),
                mu = 0, paired = FALSE, var.equal = FALSE)

Welch Two Sample t-test

data:  depressione by semplicità
t = 3.361, df = 76.42, p-value = 0.001215
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1258512 0.4918777
sample estimates:
 mean in group alta mean in group bassa
          1.1396953           0.8308309
```

Illustrazione

il test t di Student unilaterale (destro) diventa:

```
 t.test(depressione~semplicità, alternative = c("greater"),
              mu = 0, paired = FALSE, var.equal = FALSE)

Welch Two Sample t-test

data:  depressione by semplicità
t = 3.361, df = 76.42, p-value = 0.0006075
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1558517       Inf
sample estimates:
 mean in group alta mean in group bassa
          1.1396953           0.8308309
```

Due campioni indipendenti con σ uguali

Due campioni indipendenti con σ uguali

- Solo nel caso in cui le due popolazioni, da cui sono estratti i due campioni indipendenti di ampiezza n_1 e n_2, si assumono avere la stessa varianza, allora è possibile definire un test t di Student con $\nu = n_1 + n_2 - 2$ gradi di libertà.

$$t_{(n_1+n_2-2)} = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\widehat{\sigma}_c^2/n_1 + \widehat{\sigma}_c^2/n_2}} = \boxed{\frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\widehat{\sigma}_c\sqrt{1/n_1 + 1/n_2}}}$$

- dove $\widehat{\sigma}_c$ è la stima combinata della varianza unica delle popolazioni, a partire dagli elementi dei due campioni sotto esame:

$$\boxed{\widehat{\sigma}_c = s_c = \sqrt{\frac{(n-1)s_1^2 + (n-1)s_2^2}{n_1 + n_2 - 2}}}$$

Test statistico di omogeneità delle varianze

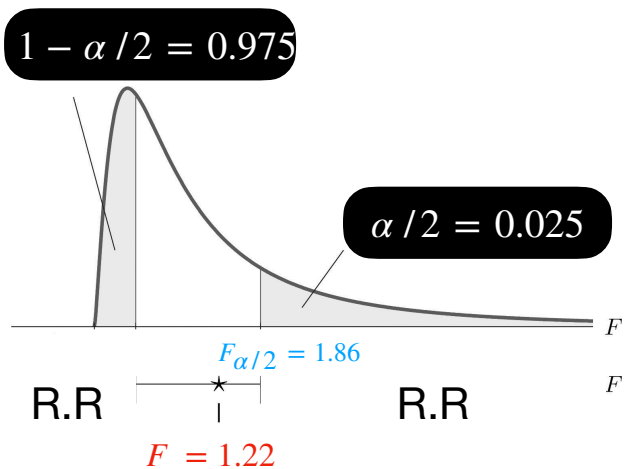
$$H_0 : \sigma_1 = \sigma_2$$

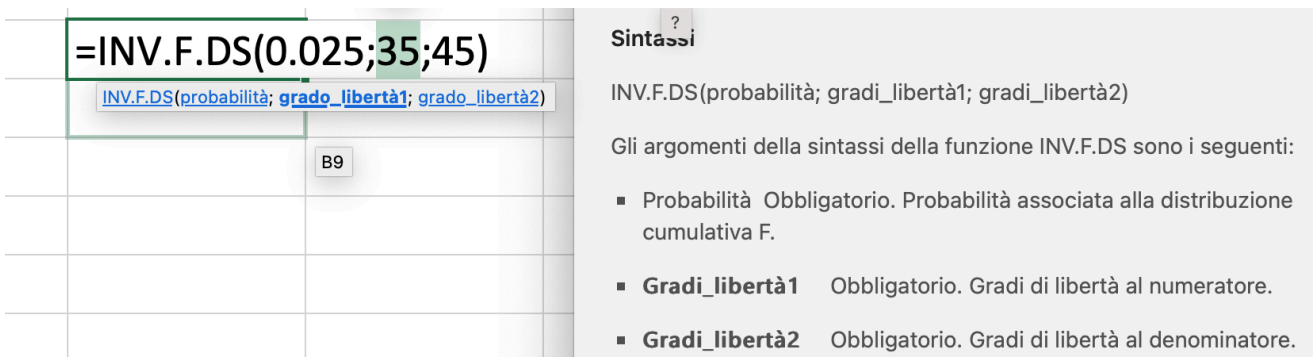
$$H_1 : \sigma_1 \neq \sigma_2$$

$$\alpha = 0.05$$

$$F = \frac{s_1^2}{s_2^2} = \frac{0.4901^2}{0.4433^2} = 1.22$$

$$F_{(\alpha/2;\nu_1=35;\nu_2=45)} = 1.86$$





$$F < F_{(\alpha/2;\nu_1=35;\nu_2=45)}$$

pertanto, non rifiutiamo H_0 e concludiamo che le due popolazioni da cui sono estratti i due campioni indipendenti, di ampiezza n_1 e n_2, si assumono avere la stessa varianza.

Due campioni indipendenti con σ uguali

- Per i dati del nostro esempio avremo che s_c assume il seguente valore

$$s_c = \sqrt{\frac{(36-1)0.4901_1^2 + (46-1)0.4433^2}{36+46-2}} = 0.4644$$

- i gradi di libertà del test diventano $36 + 46 - 2 = 80$.

Due campioni indipendenti con σ uguali

- Il valore critico di $t_{(0.025,\ 80)}$ è 1.9901.

```
qt(0.975,df=80)
## [1] 1.990063
```

- L'intervallo di confidenza al 95% è :

$$(\bar{y}_1 - \bar{y}_2) \pm t_{(0.025,\ 80)} s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(1.2162 - 0.8308) \pm 1.9901 \times 0.4644 \sqrt{\frac{1}{36} + \frac{1}{46}}$$

$$= 0.3853 \pm 0.2056 = [0.1797;\ 0.5909]$$

Due campioni indipendenti con σ uguali

- L'indice statistico t_{80} diventa

$$t_{80} = \frac{1.2162 - 0.8308}{0.4644\sqrt{\frac{1}{36} + \frac{1}{46}}} = 3.729$$

il cui $p-$valore (unilaterale e bilaterale) ottenuto con R è

```
 1-pt(3.729,df=80)
## [1] 0.0001787555
 (1-pt(3.729,df=80) )*2
## [1] 0.000357511
```

Due campioni indipendenti con σ uguali

- Un t test unilaterale con 80 gradi di libertà si ottiene impostando l'assunzione di omogeneità della varianza nelle due popolazioni da cui sono tratti i dati `var.equal=TRUE`:

```
t.test(depressione~semplicità,alternative=c("two.sided"),
                                    var.equal=TRUE)
```

```
Two Sample t-test

data:  depressione by semplicità
t = 3.729, df = 80, p-value = 0.0003576
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1796882 0.5909717
sample estimates:
 mean in group alta mean in group bassa
          1.2161608           0.8308309
```

Due campioni indipendenti con σ uguali

- Un t test bilaterale con 80 gradi di libertà si ottiene impostando l'assunzione di omogeneità della varianza nelle due popolazioni da cui sono tratti i dati: `var.equal=TRUE`:

```
t.test(depressione~semplicità,alternative=c("greater"),
                                    var.equal=TRUE)
```

```
Two Sample t-test

data:  depressione by semplicità
t = 3.729, df = 80, p-value = 0.0001788
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.2133689       Inf
sample estimates:
 mean in group alta mean in group bassa
          1.2161608           0.8308309
```

Robustezza del test t di Student

- Nel caso di campioni con $n > 30$, il test t di Student e gli intervalli di confidenza sono approssimativamente corretti anche se le distribuzioni di partenza sono asimmetriche (e quindi non normali).
- Si dice che il test t di Student è robusto rispetto alla violazione dell'ipotesi di normalità.
- In tali circostanze, però, è più opportuno usare le mediane dei gruppi, anziché le medie, quali misure di tendenza centrale.

Robustezza del test t di Student

- Il test t di Student e gli intervalli di confidenza, invece, non sono robusti rispetto alla presenza di dati anomali (outliers).
- E' dunque importante stabilire se dati anomali sono presenti e, nel caso, eliminarli seguendo le procedure appropriate.

Conclusioni

- Il test e l'intervallo di confidenza sulla media di un campione si basano sulla distribuzione t di Student, dato che la deviazione standard σ della popolazione è ignota (lo stesso si può dire del confronto tra due medie).

- Dato che $t \to N$ al crescere di n, quando i campioni sono grandi, le inferenze basate su t diventano indistinguibili da quelle basate su z.

- Se l'intervallo di confidenza della media delle differenze tra campioni contiene il valore 0, non si può rifiutare l'ipotesi $H_0 : \mu_1 = \mu_2$ al livello $1 - \alpha$.

Conclusioni

- Nel caso di due campioni dipendenti, ciascuna osservazione nel primo campione forma una coppia con un'osservazione nel secondo campione. In questo caso, il test t di Student viene applicato alla media delle differenze dei punteggi appaiati.

- L'uso della distribuzione t di Student richiede che i dati provengano da distribuzioni normali.

- L'uso del test t di Student non è appropriato in presenza di dati anomali (outliers).