

Piano della presentazione

- Inferenza su una media
- Distribuzioni t di Student
- Intervallo di confidenza
- Test t di Student
- Condizioni di validità
- Campioni dipendenti
- Due campioni indipendenti
- Due campioni indipendenti con σ uguali

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.
- 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.

- 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.

- Benchè le medie dei campioni \bar{y} siano distribuite normalmente con media μ e deviazione standard σ/\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} :

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

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

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'analogia 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:

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

Distribuzioni t di Student

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*:

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

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

- 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

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

- dato che media e varianza campionaria sono tra loro indipendenti¹
- Il rapporto ottenuto non dipende più dalla deviazione standard σ della popolazione e può essere interpretato come 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 .
- ha valore atteso $E(t) = 0$, come nella normale standard, dal momento che il valore $f(t)$ decresce simmetricamente attorno a tale valore:
- 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 :

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

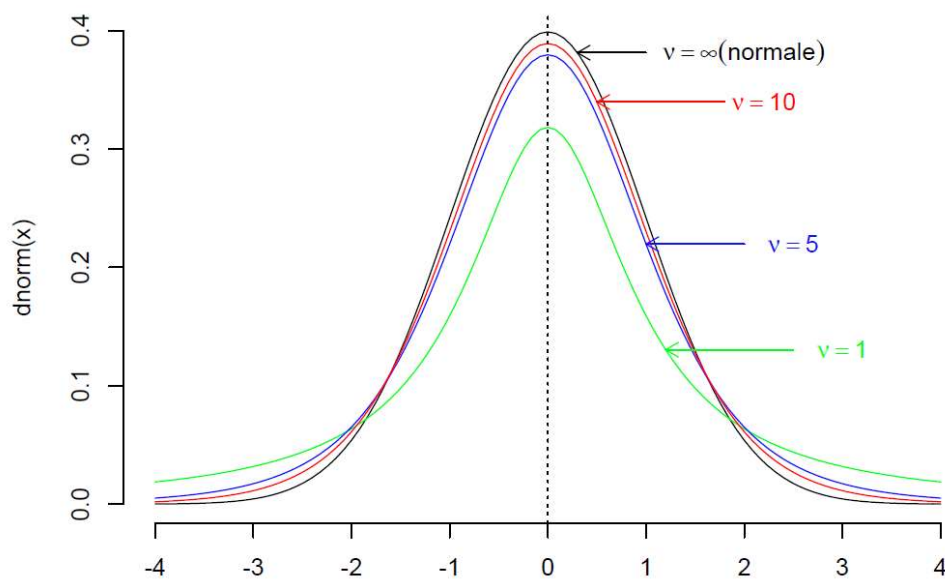
¹ $(y_i - \bar{y})$ è indipendente da \bar{y} : infatti se $\bar{y} + a \rightarrow [(y_i + a) - (\bar{y} + a)] = [y_i - \bar{y}]$

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.

- 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 .



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 .

Illustrazione

Illustrazione. Consideriamo nuovamente l'esperienza fittizia discussa 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.

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	$s = 8.38$

- 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 $\rightarrow n - 1 = 10 - 1 = 9$.
 - 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 $t_{(\alpha/2; n-1)}$ produrrà quindi un intervallo di confidenza più grande.
- La formula per il calcolo dell'intervallo di confidenza è

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

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

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

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

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

- 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.

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

t-Student				
gdl	0,050	0,025	0,010	0,005
9	1,833	2,262	2,821	3,250

- 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].

- Con un'ipotesi alternativa non-direzionale,

$$H_a : \mu \neq 0$$

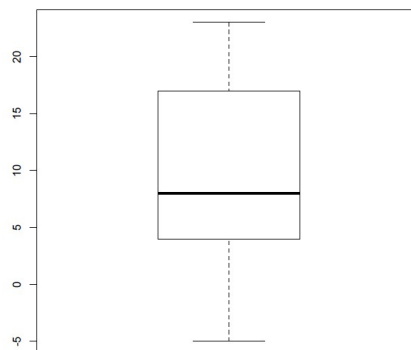
il p -valore verrebbe raddoppiato:

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

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.
- 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*).

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



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).
- 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.

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.

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**.

- 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, $H_a : \delta > 0$.

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 $d = d_1; d_2; \dots; d_n$, dove ciascuna d_i è lo scarto fra le osservazioni corrispondenti della i -esima unità nei due campioni, ovvero $d_i = y_{1i} - y_{2i}$.
- 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.
- 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 probabilità $\alpha/2$.

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.

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

- 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

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

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, $\mu_1 - \mu_2$.

- 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 $\bar{y}_1 - \bar{y}_2$.

- 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

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

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

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

- 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, $N(0; 1)$.

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.

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

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

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

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

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

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*".
- 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

- 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.
- L'intervallo di confidenza al 95% è :

$$\begin{aligned}
 & (\bar{y}_1 - \bar{y}_2) \pm t_{(0.025, 35)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
 & (1.2162 - 0.8308) \pm 2.0301 \sqrt{\frac{0.4901^2}{36} + \frac{0.4433^2}{46}} \\
 & = 0.3853 \pm 0.2124 = [0.1729; 0.5977]
 \end{aligned}$$

- 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:

$$\begin{aligned}
 & (1.2162 - 0.8308) \pm 1.9937 \sqrt{\frac{0.4901^2}{36} + \frac{0.4433^2}{46}} \\
 & = 0.3853 \pm 0.2086 = [0.1768; 0.5940]
 \end{aligned}$$

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

- 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

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

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

- 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 **augmenta quindi la potenza** del test statistico.
- 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)      qt(0.975,df=71.40065)
## [1] 2.030108      ## [1] 1.993749
```

Due campioni indipendenti con σ uguali

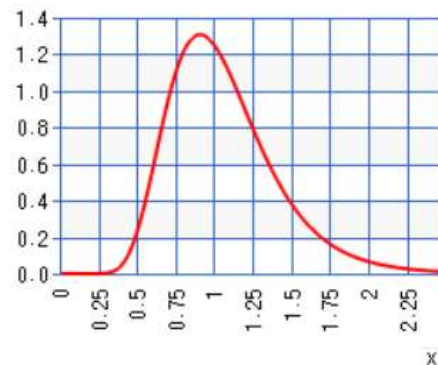
- Solo nel caso in cui le due popolazioni, da cui sono estratti i due campioni indipendenti di ampiezza n_1 e n_2 , **si assumono avere la stessa varianza**, allora è possibile definire un test t di Student con $\nu = n_1 + n_2 - 2$ gradi di libertà.

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

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

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

test F			
n	s^2	s_1^2/s_2^2	$F(45;35)_{\alpha=0.05}$
46	0,197	1,217	1,718
36	0,240		



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

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

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

- 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 :

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

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

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

- L'indice statistico t_{80} diventa

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

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.
- 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.