STATISTICA DI BASE III

GIUSEPPE DE NICOLAO

giuseppe.denicolao@unipv.it

Dipartimento di Informatica e Sistemistica

Università di Pavia

SOMMARIO

- Test di ipotesi
- Confrontare due trattamenti: un esempio
- Uso di distribuzioni di riferimento esterne
- L'ipotesi di campionamento indipendente
- Test non parametrici
- Randomizzazione e dati appaiati

TEST DI IPOTESI

Idea: Formulo un'ipotesi (H_0: "ipotesi nulla") e poi uso i dati a disposizione per respingerla o meno

Da dove viene H_0?

- Modello/teoria
- dati passati
- specifiche di prodotto

Esempio: Carta di controllo; in base ai dati precedenti risulta che il processo (in controllo) è gaussiano con media μ e varianza σ^2. Arrivano n nuovi dati $x_1,\ldots,x_n$. Decidere se il processo è ancora in controllo.

H_0: Il processo e' in controllo $\Leftrightarrow \overline{X} \sim N(\mu, \sigma^2/n)$

Idea (intuitiva): Se $|\overline{X}-\mu|$ è "grande", respingo H_0.

Problema: Cosa vuol dire "grande"? Anche se il processo è in controllo, può capitare di tanto in tanto che $|\overline{X}-\mu|$ sia "grande".

Due tipi di errore:

- Respingo H_0 anche se è giusta (F+, errore di tipo I, rischio del produttore, "condanna dell'innocente")

$$\alpha := P(F+)$$

- Non respingo H_0 anche se sbagliata (F-, errore di tipo II, rischio dell'acquirente, "assoluzione del colpevole")

$$\beta := P(F-)$$

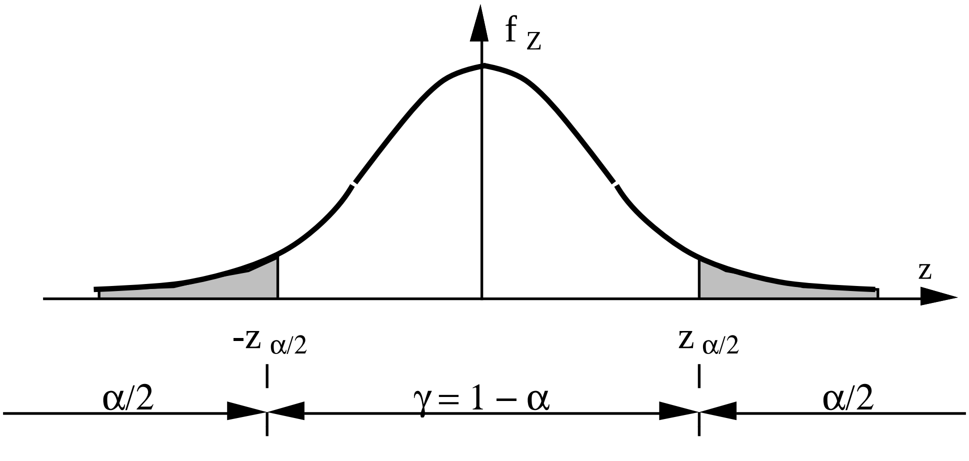
Vediamo come risolvere il problema se fisso α *(livello di significatività):*

$$Z := \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\Downarrow$$

Cerco $z_{\alpha/2}$ tale che

$$P(|Z| \leq z_{\alpha/2}) = 1 - \alpha = \gamma$$



Regola:

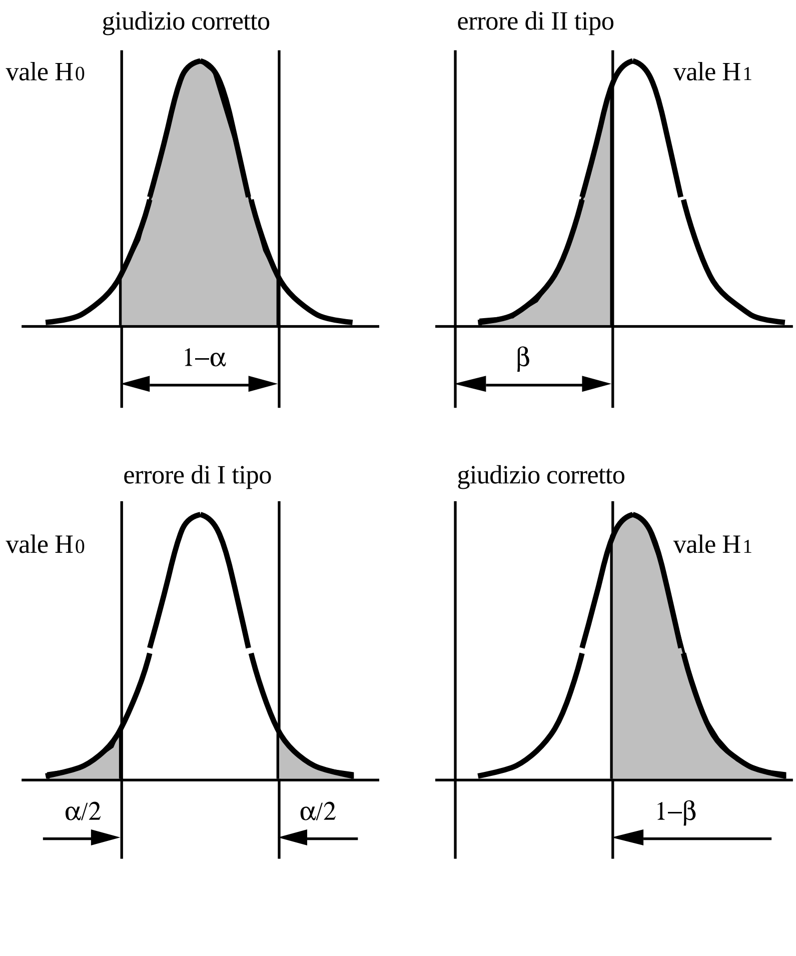
- Se $|Z| > z_{\alpha/2}$ respingo H_0
- Altrimenti, non respingo ("accetto") H_0

Osservazioni:

- Vedo subito che $P(F+) = \alpha$
- Come valutare $\beta := P(F-)$?

$$\Downarrow$$

Ho bisogno di sapere a quale distribuzione appartiene $\bar{X}$ quando il processo non è in controllo



"***P*-value**"

Definizione: Il P-value è il più piccolo α in corrispondenza del quale H_0 viene respinta.

Esempio: Test di gaussianità. H_0: X è gaussiana

- L'ipotesi viene respinta con $\alpha = 0.05 \Rightarrow X$ non è gaussiana.

- Può darsi che con $\alpha = 0.049$ l'ipotesi non venga respinta $\Rightarrow$ sono sul limite (con $\alpha = 0.05$ respingo, ma per un pelo).

- Viceversa, se H_0 viene respinta anche con $\alpha = 0.001$, capisco che X è "decisamente non gaussiana"

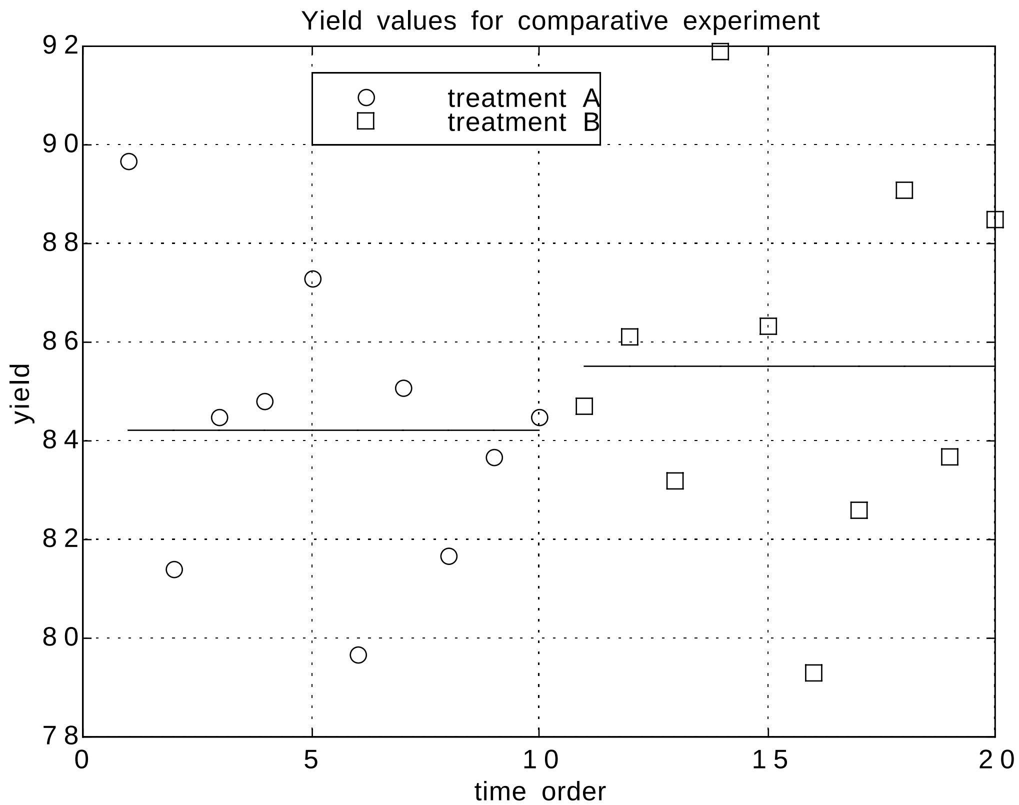
$$\Downarrow$$

P-value << 0.05 $\Leftrightarrow H_0$ "decisamente falsa"

P-value >> 0.05 $\Leftrightarrow$ poca evidenza che H_0 sia falsa

CONFRONTARE DUE TRATTAMENTI: UN ESEMPIO

Esperimento industriale: Confronto tra le rese usando due trattamenti A e B



$\bar{y}_A = 84.24$ $\bar{y}_B = 85.54$ $\bar{y}_B - \bar{y}_A = 1.30$

(G.E.P. Box, W.G. Hunter, J. Stuart Hunter, Statistics for Experimenters, Wiley 1978)

Domanda: B è veramente meglio di A? Non potrebbe accadere che ripetendo l'esperimento si abbia una maggior resa con il trattamento A?

USO DI DISTRIBUZIONI DI RIFERIMENTO ESTERNE

Esempio (il prezzo è giusto?): Mario Rossi ha cambiato città e deve acquistare una casa. Come capire se il prezzo è giusto?

Soluzione: Vedere il maggior numero possibile di case sul mercato in modo da costruirsi un "insieme di riferimento", cioè una distribuzione di prezzi.

$$\Downarrow$$

usando la distribuzione di riferimento, si può
valutare se una casa ha un prezzo sopra,
entro o sotto la media.

Osservazione: L'insieme di riferimento deve essere coerente con il tipo di casa che si deve valutare: inutile considerare villette se si vuole acquistare un appartamento in condominio.

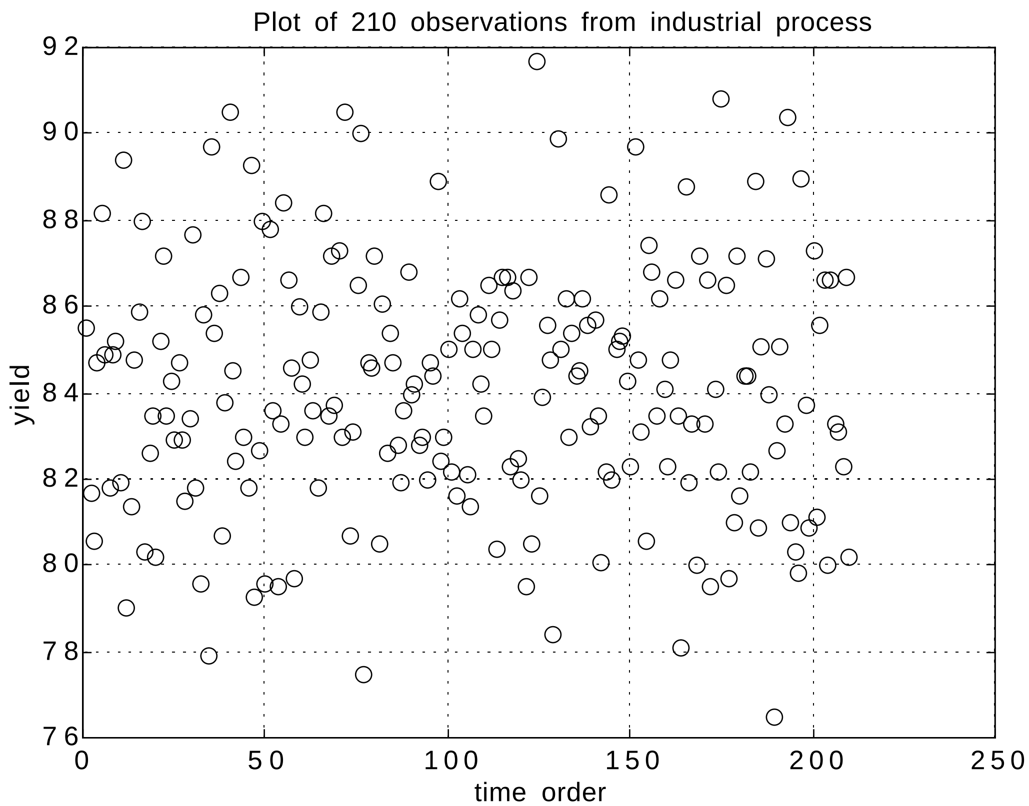
Torniamo all'esperimento industriale: prendiamo le parti dell'avvocato del diavolo e formuliamo

L'ipotesi nulla: I due trattamenti A e B hanno il medesimo effetto sulla resa e la differenza $\bar{y}_B$-$\bar{y}_A$ = 1.30 è solo frutto del caso.

Idea: Per provare che c'è differenza devo screditare l'ipotesi nulla mostrando che, alla luce dei dati, essa è assai poco probabile.

Importanza della distribuzione di riferimento: Se avessi una distribuzione di riferimento potrei capire se, sotto l'ipotesi nulla, la differenza $\bar{y}_B$-$\bar{y}_A$ = 1.30 ricade nella norma oppure no.

Supponiamo di avere 210 rese ottenute nel passato con il trattamento A:



Costruzione di una distribuzione di riferimento (I): Prendo tutte le sottosequenze di 20 osservazioni consecutive e calcolo la differenza tra le seconde 10 e le prime 10:

Tabella: 210 valori di resa consecutivi

time order	*obs*	*average 10 obs*
1	85.5	
2	81.7	
3	80.6	
4	84.7	
5	88.2	
6	84.9	
7	81.8	
8	84.9	
9	85.2	
10	81.9	**83.94**
11	89.4	*84.33*
12	79.0	84.06
13	81.4	84.14
14	84.8	84.15
15	85.9	83.92
16	88.0	84.23
17	80.3	84.08
18	82.6	83.85
19	83.5	83.68
20	80.2	**83.51**
21	85.2	*83.09*
22	87.2	83.91
23	83.5	84.12
24	84.3	84.07
25	82.9	
…	…	…
210	80.2	83.55

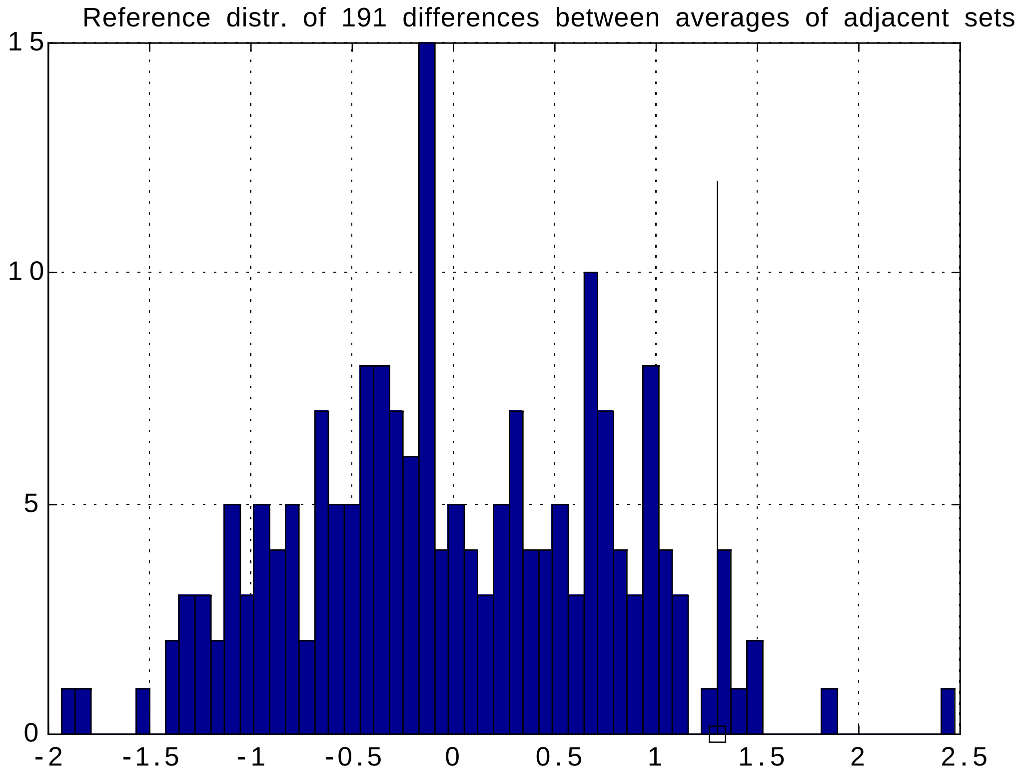
Domanda: Usando il trattamento A, quanto spesso accade che le differenze tra la medie di gruppi successivi di 10 osservazioni siano più grandi di 1.30?

Tabella: Insieme di riferimento delle 191 differenze tra le medie di due gruppi adiacenti di 10 valori successivi (le differenze superiori a +1.30 sono in grassetto)

-0.43	-0.91	0.38	0.74	0.37
-1.24	0.64	-0.22	0.98	0.03
-0.15	-0.17	0.20	**1.87**	0.75
-0.02	-0.17	-0.37	0.66	0.44
-0.08	0.96	-0.16	-0.04	0.17
-0.15	0.78	0.12	-0.60	-0.23
-0.79	-0.13	0.80	-0.93	0.97
-0.38	0.30	0.54	0.02	0.72
-0.26	-0.34	0.08	-0.50	0.98
-0.10	0.71	-1.01	-0.51	-0.21
0.82	0.68	-0.55	-0.67	-0.81
0.90	0.53	-0.05	-0.78	0.29
-0.68	1.01	-0.30	-1.15	0.49
-0.66	**1.46**	0.33	-1.07	-0.58
-1.25	0.76	0.79	-0.30	-0.30
-0.27	1.04	-0.11	0.78	-0.01
0.13	**1.35**	-0.42	0.95	-0.61
0.21	**1.37**	0.30	-0.17	0.40
0.24	0.88	1.13	0.61	-1.06
0.29	-0.12	1.25	0.74	-0.13
-0.18	0.20	0.97	0.67	-0.52
0.43	-0.12	0.68	0.79	-1.07
1.47	-0.37	0.68	0.66	-1.40
1.33	-1.38	-0.45	1.00	0.11
2.48	-0.90	-0.62	-0.11	0.46
1.01	-0.80	-0.03	-0.40	-0.01
1.33	-1.04	0.54	-0.45	0.33
0.29	-1.94	-0.43	0.10	-0.87
0.57	-0.90	-1.24	-0.30	-0.18
0.95	-0.76	-0.64	-0.97	0.51
-0.42	-0.63	-0.86	-0.82	**1.39**
-0.36	-0.94	-1.10	-1.53	0.61
-0.52	-0.32	-0.16	-1.20	0.50
-1.33	-0.21	1.09	-1.10	0.64
-1.81	-0.36	0.87	-0.43	-0.53
-0.36	-0.93	1.11	-1.32	
-1.02	-0.75	-0.12	-1.30	
0.21	0.13	0.67	-0.64	
-0.29	0.39	1.01	-0.58	

Spiegazione:

- -0.43 = 83.51 - 83.94
- -1.24 = 83.09 - 84.33
- …



Risposta: Se vale l'ipotesi nulla la differenza tra le medie di due gruppi consecutivi di 10 osservazioni è maggiore di 1.30 con frequenza 9/191 = 0.047 (non è impossibile, ma accade raramente)

$$\Downarrow$$

respingo l'ipotesi nulla:
la differenza osservata è statisticamente
significativa con P-value = 0.047

Costruzione di una distribuzione di riferimento (II): Procedo come nel caso (I) ma considero solo sottosequenze disgiunte.

$$\Downarrow$$

da 210 osservazioni ricavo 10 sottosequenze di 20 elementi di cui calcolo la differenza tra la media dei primi 10 e i secondi 10

Tabella: 10 differenze circa indipendenti

risultati osservati	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_2$-$\bar{y}_1$
da precedenti registrazioni	83.94	83.51	-0.43
	84.33	84.15	-0.18
	83.73	84.37	0.64
	84.90	84.78	-0.12
	83.84	84.04	0.20
	83.74	84.42	0.68
	84.26	84.92	0.66
	83.85	84.85	1.00
	83.75	83.92	0.17
	83.62	84.08	0.46
	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_B$-$\bar{y}_A$
dall'ultimo confronto:	84.24	85.54	1.30
var. stimata delle differenze	$\dot{s}^2 = 0.2820$		
SD stimata delle differenze	$\dot{s} = 0.5310$		

Spiegazione:

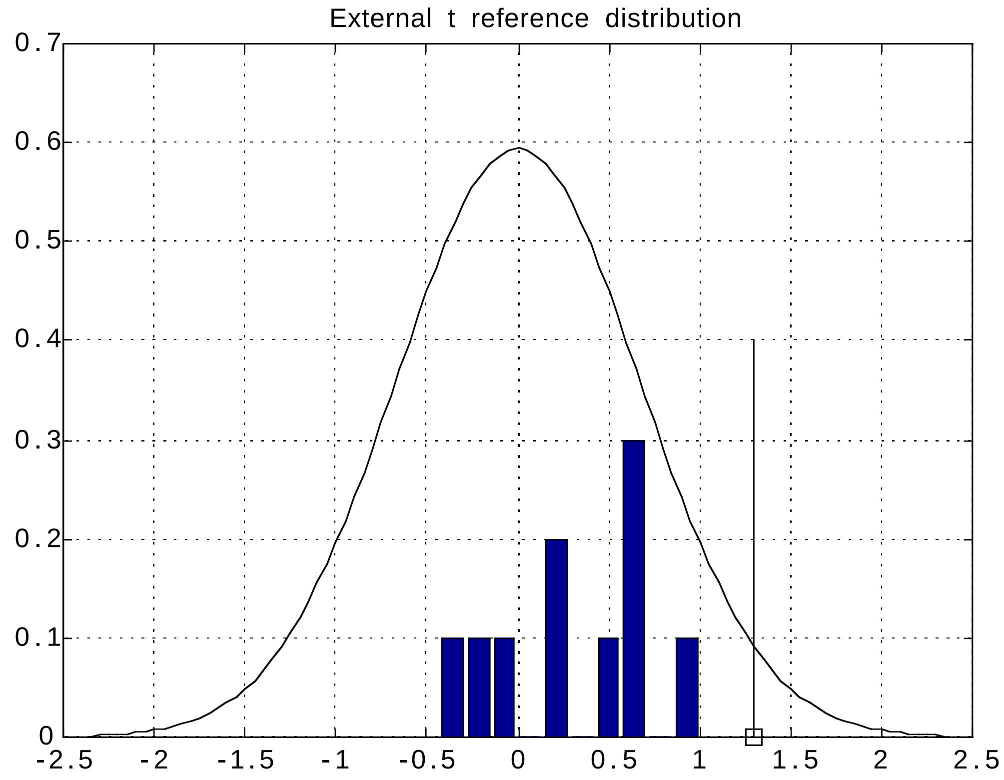
- 83.94 è la media dei dati da 1 a 10
- 83.51 è la media dei dati da 11 a 20
- 84.33 è la media dei dati da 21 a 30
- …

In questo modo ottengo 10 differenze tra di loro circa indipendenti:

- "indipendenti" perché basate su dati diversi
- "circa" perché potrebbero ereditare la correlazione che può esistere tra dati successivi (nel caso essi non siano i.i.d.)

Difficoltà: 10 valori non bastano per costruire una distribuzione di riferimento mediante l'istogramma.

Soluzione: Se le differenze sono indipendenti, si può dimostrare che la distribuzione di riferimento è una t di Student a 10 gradi di libertà moltiplicata per $\dot{s}$ (che è la SD stimata delle differenze):



In pratica:

- Calcolo

$$t_0 = \frac{\bar{y}_B - \bar{y}_A}{\dot{s}} = \frac{1.30}{0.5310} = 2.45$$

- Dalla tabella della t di Student a 10 gradi di libertà, vedo che

$$P(t > t_o) = P(t > 2.45) < P(t > 2.228) = 0.025$$
$$P\text{-}value < 0.025$$

(la differenza è statisticamente significativa)

Pro e contro dell'insieme di riferimento esterno

Vantaggio: La strada seguita non richiede particolari ipotesi statistiche ed è del tutto generale.

Svantaggio: Ho bisogno di un insieme sufficientemente ricco di osservazioni che caratterizzano il funzionamento del processo con il trattamento A

Tabella: Valori di t_α

(t_α *è tale che* $P(t > t_\alpha) = \alpha$*)*

n	α = 0.1	α = 0.05	α=0.025
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182
4	1.533	2.132	2.776
5	1.476	2.015	2.571
6	1.440	1.943	2.447
7	1.415	1.895	2.365
8	1.397	1.860	2.306
9	1.383	1.833	2.262
10	1.372	1.812	2.228
11	1.363	1.796	2.201
12	1.356	1.782	2.179
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	2.131
16	1.337	1.746	2.120
17	1.333	1.740	2.110
18	1.330	1.734	2.101
19	1.328	1.729	2.093
20	1.325	1.725	2.086
21	1.323	1.721	2.080
22	1.321	1.717	2.074
23	1.319	1.714	2.069
24	1.318	1.711	2.064
25	1.316	1.708	2.060
26	1.315	1.706	20.56
27	1.314	1.703	2.052
28	1.313	1.701	2.048
29	1.311	1.699	2.045
30	1.310	1.697	2.042
40	1.303	1.684	2.021
60	1.296	1.671	2.000
120	1.289	1.658	1.980
∞	1.282	1.645	1.960

L'IPOTESI DI CAMPIONAMENTO CASUALE

Domanda: E' possibile costruire una distribuzione di riferimento senza disporre di un insieme esterno di riferimento? Come confrontare due trattamenti nuovi su cui non sono disponibili dati storici?

Ipotesi di campionamento casuale: Sotto l'ipotesi nulla (= nessuna differenza tra i trattamenti), le osservazioni sono assimilabili ad estrazioni indipendenti da un'urna

Formulazioni equivalenti:

- campionamento casuale
- estrazioni ripetute da un'urna contenente i possibili valori
- variabili casuali indip. e identicamente distribuite
- l'ordinamento delle osservazioni non conta

Nota: In moltissimi casi pratici, l'ipotesi non è soddisfatta

Esempi:

- Diversi siti su un wafer (né indip. né identicam. distr.)
- Diversi wafer di un lotto (né indip. né identicam. distr.)

Modello con campionamento casuale: distribuzione di riferimento con valore esterno di σ

Sotto l'ipotesi nulla, $\sigma_A^2 = \sigma_B^2 = \sigma^2$, cosicché, *sfruttando l'ipotesi di campionamento casuale:*

$$Var[\bar{y}_B - \bar{y}_A] = Var[\bar{y}_B] + Var[\bar{y}_A] = \frac{\sigma^2}{n_B} + \frac{\sigma^2}{n_A}$$

Inoltre, per il teorema centrale del limite, la differenza $\bar{y}_B - \bar{y}_A$ è circa gaussiana.

$$\Downarrow$$

Proprietà: Sotto l'ipotesi nulla, la quantità

$$\frac{\bar{y}_B - \bar{y}_A}{\sigma\sqrt{1/n_A + 1/n_B}}$$

è circa distribuita come la normale standard Z.

Questione aperta: come fare con il valore di σ?

Possibilità #1: σ è noto (per esempio, uso la SD campionaria stimata nelle 210 osservazioni registrate nel passato: $s = 2.88$).

Nota: Dato che s è una stima non dovrei fare riferimento a Z, ma ad una t di Student con 209 gradi di libertà che però è praticamente identica ad una normale standard.

In pratica, se σ è nota:

- Calcolo

$$z_o = \frac{\bar{y}_B - \bar{y}_A}{\sigma\sqrt{1/n_A + 1/n_B}} = \frac{1.30}{2.88\sqrt{1/10+1/10}} = \frac{1.30}{1.29} = 1.01$$

- Dalla tabella della normale standard vedo che

$$P(Z>z_o) = 1 - P(Z\leq z_o) = 1 - F_Z(1.01) = 1 - 0.8438 = 0.1562$$

$$P\text{-}value = 0.1562$$

(la resa B non risulta significativamente maggiore della resa A)

Osservazione: Ho ancora bisogno di un valore *esterno* per σ. In molti casi esso non è disponibile. Il prossimo passo consiste nel fare riferimento ad una stima *interna* (basata sui 20 dati dell'esperimento)

Tabella: Valori di $F_Z(z) = P(Z \leq z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Modello con campionamento casuale: distribuzione di riferimento con stima interna di σ

Possibilità #2: Stimo σ^2 combinando insieme le varianze campionarie $S_A^2 = 8.42$, $S_B^2 = 13.32$:

$$S^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$$

Proprietà: Sotto l'ipotesi nulla, la quantità

$$\frac{\bar{y}_B - \bar{y}_A}{S\sqrt{1/n_A + 1/n_B}}$$

è circa distribuita come una *t* di Student con $n_A + n_B - 2$ gradi di libertà.

In pratica, se σ non è nota:

- Calcolo

$$S^2 = \frac{(n_A\text{-}1)S_A^2+(n_B\text{-}1)S_B^2}{n_A + n_B - 2} = \frac{9\times 8.42+9\times 13.32}{10 + 10 - 2} = 10.87$$

$$t_o = \frac{\bar{y}_B\text{-}\bar{y}_A}{S\sqrt{1/n_A+1/n_B}} = \frac{1.30}{3.30\sqrt{1/10+1/10}} = \frac{1.30}{1.47} = 0.88$$

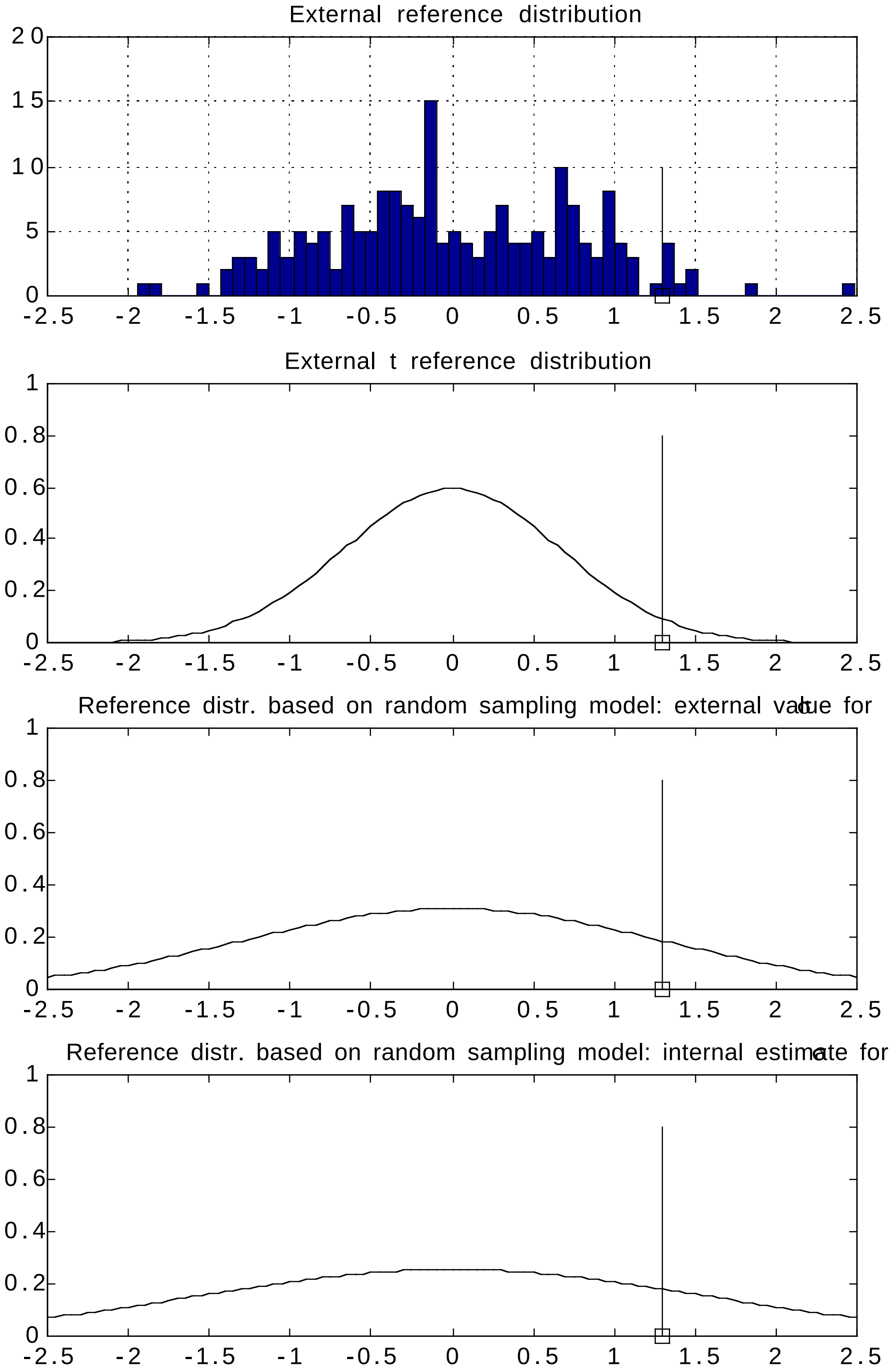
- Dalla tabella della t di Student a $n_A + n_B - 2 = 18$ gradi di libertà vedo che

$$P(t > t_o) = P(t > 0.88) > P(t > 1.330) = 0.1$$

$$P\text{-}value > 0.1$$

(la resa B non risulta significativamente maggiore della resa A)

Confronto tra tutte le distribuzioni di riferimento considerate



TEST NON PARAMETRICI

Dilemma: Per confrontare due trattamenti devo richiedere:

- Una lunga serie di osservazioni precedenti che non è detto siano disponibili.

oppure

- La validità dell'ipotesi di campionamento casuale, che spesso non vale.

Qualcuno crede (erroneamente) che il dilemma sia risolto usando i cosiddetti *test non parametrici* (nonparametric tests, distribution free tests), come, per esempio, il test di Wilcoxon.

Fatti:

- Anche i test non parametrici richiedono l'ipotesi di campionamento casuale
- L'unico loro vantaggio è che non fanno ipotesi sulla distribuzione delle singole osservazioni.
- In virtù del teorema centrale del limite, i test parametrici (come il test *t*) danno delle approssimazioni ragionevoli anche quando le osservazioni non sono gaussiane *(purché valga l'ipotesi di campionamento casuale)*.

RANDOMIZZAZIONE E DATI APPAIATI

Obiettivo: Costruire una distribuzione di riferimento senza dati esterni e senza ipotesi di campionamento indipendente!

Esempio: Un contadino vuole stabilire se il fertilizzante B fa produrre più pomodori del fertilizzante A. Ha 11 piante tutte in fila. Ne tratta 5 con il fertilizzante A e 6 con il fertilizzante B.

Osservazione: Se tratto le prime 5 piante con A, le differenze osservate tra i due gruppi potrebbero essere dovute alla variabilità spaziale (e non al tipo di fertilizzante): per esempio, l'inizio della fila potrebbe essere più soleggiato e/o meglio irrigato, etc. (altro esempio: ruolo della posizione dei wafer all'interno di un lotto).

Randomizzazione: Il contadino mescola 5 carte rosse e 6 carte nere e ottiene la sequenza:

1	2	3	4	5	6	7	8	9	10	11
R	*R*	*N*	*N*	*R*	*N*	*N*	*N*	*R*	*R*	*N*

$$\textit{Carta rossa} \rightarrow \textit{fertilizzante A}$$
$$\textit{Carta nera} \rightarrow \textit{fertilizzante B}$$

Tabella: Risultati dell'esperimento randomizzato (rese di pomodori)

posizione	1	2	3	4	5	6	7	8	9	10	11
fertilizzante	A	A	B	B	A	B	B	B	A	A	B
Kg pomodori	29.9	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

	fertilizzante A	*fertilizzante B*
	29.9	26.6
	11.4	23.7
	25.3	28.5
	16.5	14.2
	21.1	17.9
		24.3
	$n_A = 5$	$n_B = 6$
	$\bar{y}_A = 20.84$	$\bar{y}_B = 22.53$
differenza delle medie (fert. B - fert. A):		$\bar{y}_B - \bar{y}_A = 1.69$

Osservazione: Sotto l'ipotesi nulla (A e B hanno lo stesso effetto), le lettere "A" e "B" sono semplici etichette che non influenzano il risultato.

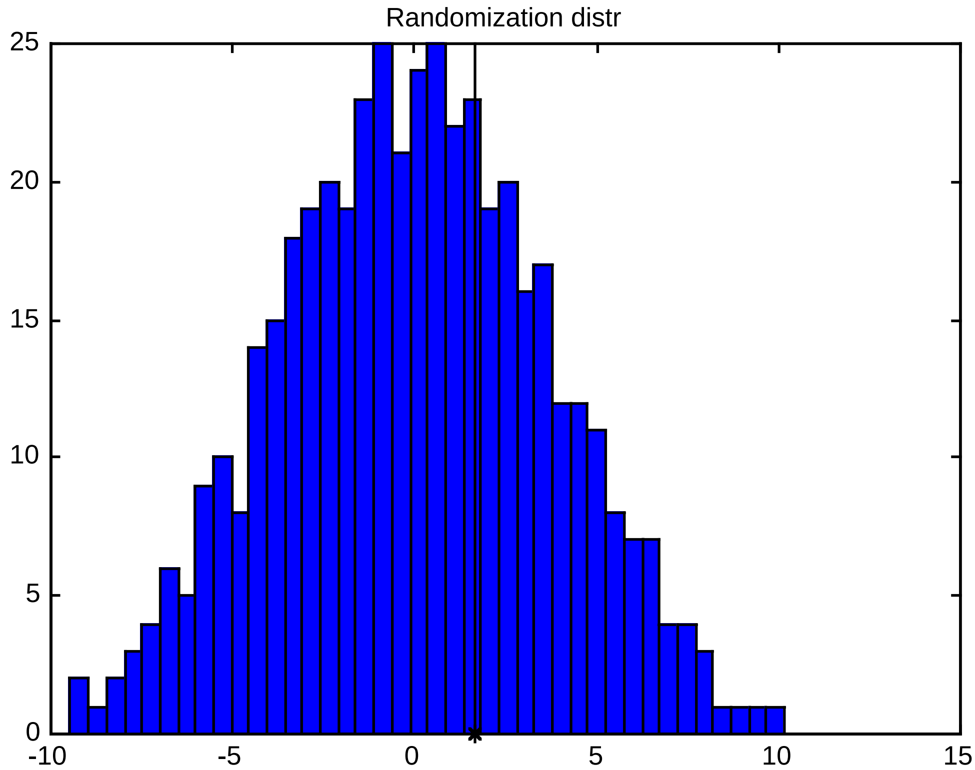
$$\Downarrow$$

Se ridispongo le etichette A e B in modo diverso ottengo una sequenza di rese che ha la stessa probabilità di quella di partenza.

Ci sono in tutto 11!/5!6! = 462 modi di allocare 5 etichette A e 6 etichette B sulle 11 piante.

$$\Downarrow$$

La distribuzione di riferimento è quella che si ottiene calcolando la differenza $\bar{y}_B$-$\bar{y}_A$ nei 462 casi e costruendo il relativo istogramma



Si trova che nel 33% dei casi
la differenza $\bar{y}_B$-$\bar{y}_A$ è maggiore di 1.69

$$\Downarrow$$

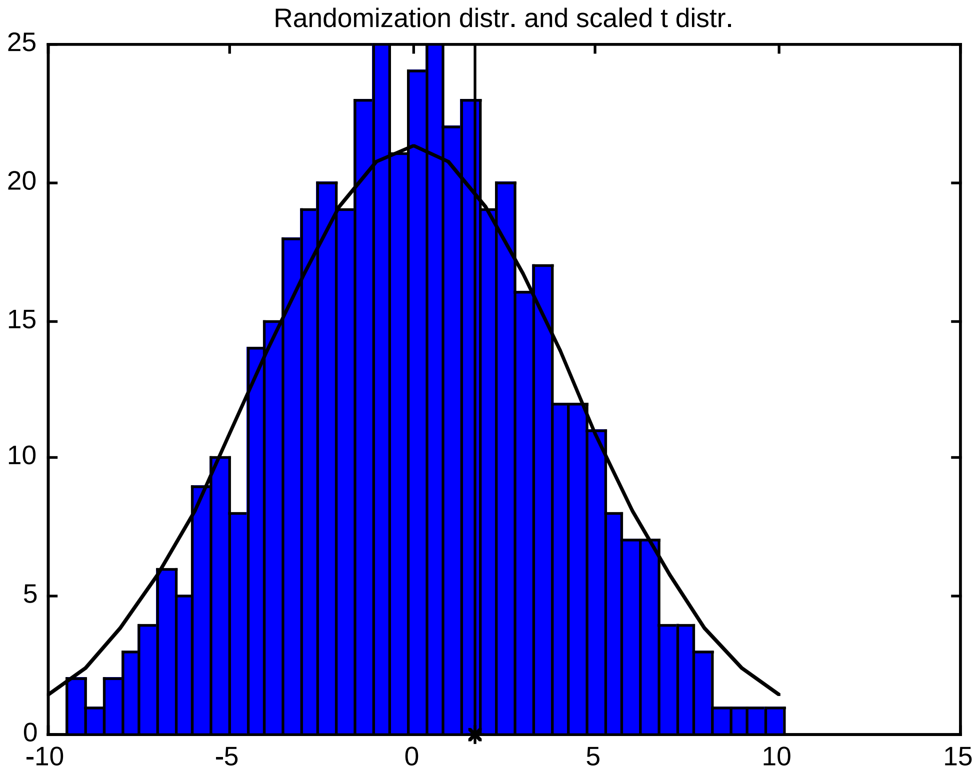
P-value = 0.33

Proprietà: La distribuzione randomizzata è ben approssimata dalla distribuzione di una t di Student a $n_A + n_B - 2$ gradi di libertà moltiplicata per

$$S\sqrt{1/n_A+1/n_B)}$$

dove

$$S^2 = \frac{(n_A - 1)S_A{}^2 + (n_B - 1)S_B{}^2}{n_A + n_B - 2}$$



In pratica:

- Calcolo

$$S^2 = \frac{(n_A-1)S_A^2+(n_B-1)S_B^2}{n_A + n_B - 2} = \frac{4\times 52.50+5\times 29.51}{5 + 6 - 2} = 39.73$$

$$t_o = \frac{\bar{y}_B-\bar{y}_A}{S\sqrt{1/n_A+1/n_B}} = \frac{1.69}{6.30\sqrt{1/5+1/6}} = \frac{1.69}{3.82} = 0.44$$

- Dalla tabella della t di Student a $n_A + n_B - 2 = 9$ gradi di libertà vedo che

$$P(t > t_o) = P(t > 0.44) > P(t > 1.383) = 0.1$$

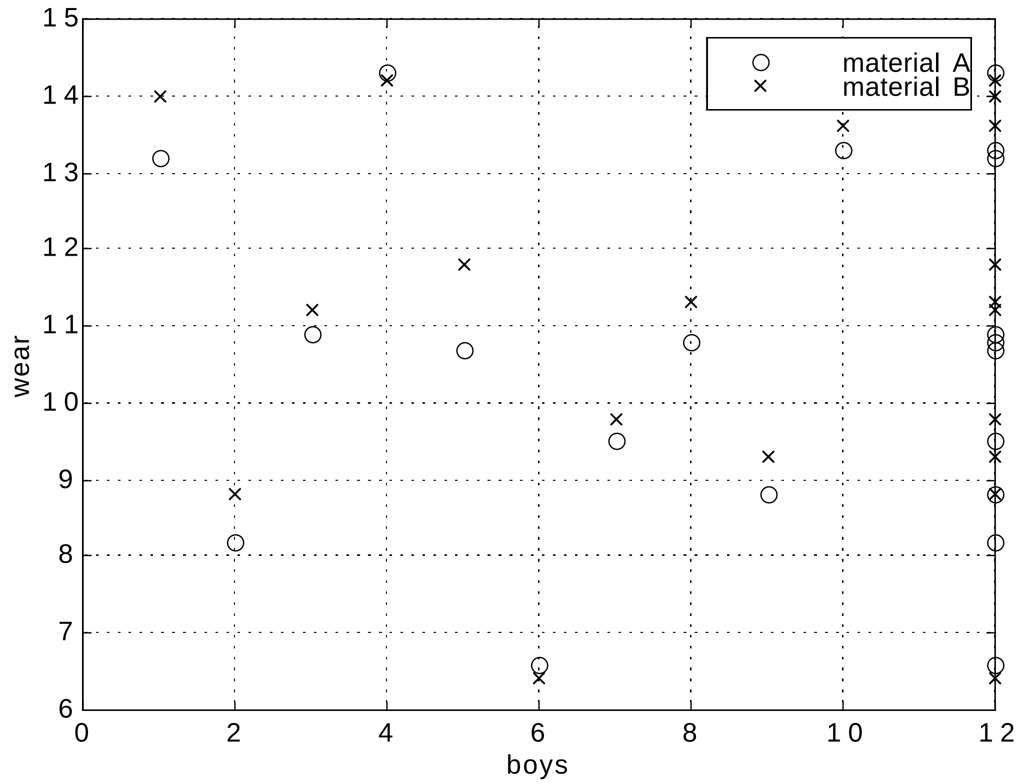
P-value > 0.1
(usando la tabella completa: *P-value* = 0.34)

(la resa B non risulta significativamente maggiore della resa A)

Morale: *Pur di randomizzare posso usare il test t (come approssimazione della distr. randomizzata) senza dover ipotizzare il campionamento casuale.*

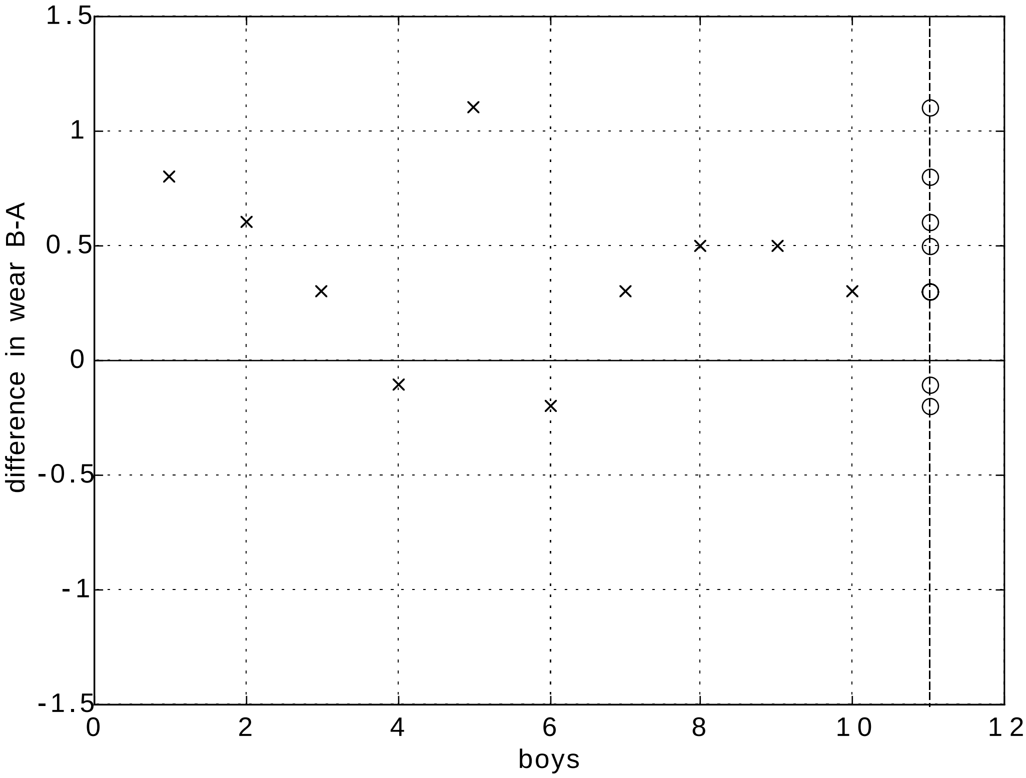
Obiettivo: Neutralizzare eventuali sorgenti di variabilità estranee al confronto.

Esempio: Due materiali A e B per fabbricare suole di scarpe. Esperimento in cui 10 ragazzi usano ciascuno un paio scarpe con suole diverse (una scarpa con A e l'altra con B). Alla fine misuro il consumo delle suole.



Se esamino il diagramma di dispersione (colonna 12) non noto grandi differenze tra i due gruppi. La variabilità che esiste da ragazzo a ragazzo oscura l'eventuale differenza tra A e B.

Idea: Se considero le differenze di consumo tra la scarpa B e quella A di ciascun ragazzo, elimino la variabilità che esiste tra i ragazzi. Ciascun paio di scarpe ha sollecitazioni che variano a seconda delle abitudini del ragazzo che le calza; all'interno dello stesso paio le due scarpe subiscono sollecitazioni simili tra loro.



Possibile fonte di disturbo: e' possibile che, mediamente, le scarpe sinistre abbiano sollecitazioni diverse da quelle destre.

Soluzione: randomizzazione!

Esperimento con dati appaiati randomizzati: Lancio una moneta 10 volte e ottengo la sequenza:

1	2	3	4	5	6	7	8	9	10
T	*T*	*C*	*T*	*C*	*T*	*T*	*T*	*C*	*T*

$$\textit{Testa} \rightarrow (\textit{sinistra} = A, \textit{destra} = B)$$
$$\textit{Croce} \rightarrow (\textit{sinistra} = B, \textit{destra} = A)$$

Tabella: Consumo delle suole con due materiali A e B

ragazzo	*materiale A*		*materiale B*		*differenza B-A*
1	13.2	(S)	14.0	(D)	0.8
2	8.2	(S)	8.8	(D)	0.6
3	10.9	(D)	11.2	(S)	0.3
4	14.3	(S)	14.2	(D)	-0.1
5	10.7	(D)	11.8	(S)	1.1
6	6.6	(S)	6.4	(D)	-0.2
7	9.5	(S)	9.8	(D)	0.3
8	10.8	(S)	11.3	(D)	0.5
9	8.8	(D)	9.3	(S)	0.5
10	13.3	(S)	13.6	(D)	0.3

differenza media : $d = 0.41$
SD campionaria delle differenze : $s_d = 0.386$

(S) = suola sinistra; (R) = suola destra

Osservazioni:

- Sotto l'ipotesi nulla (A e B hanno la stessa resistenza all'usura), il lancio delle monete influenza solo il segno delle differenze B-A.

- Ci sono in tutto $2^n = 2^{10} = 1024$ sequenze di teste e croci (n = numero di coppie)

$$\Downarrow$$

La distribuzione di riferimento è quella che si ottiene cambiando i segni in tutti i 1024 modi possibili, calcolando la differenza delle coppie in ciascun caso e costruendo il relativo istogramma

Proprietà: La distribuzione randomizzata è ben approssimata dalla distribuzione di una t di Student a n - 1 gradi di libertà moltiplicata per $s_d/\sqrt{n}$.

In pratica:

- Calcolo

$$t_0 = \frac{d \sqrt{n}}{s_d} = \frac{0.41 \sqrt{10}}{0.386} = \frac{0.41}{0.12} = 3.4$$

- Dalla tabella della t di Student a $n - 1 = 9$ gradi di libertà, vedo che

$$P(t > t_o) = P(t > 3.4) < P(t > 2.262) = 0.025$$

$$P\text{-}value < 0.025$$

(usando la tabella completa: *P-value* = 0.004)

(la differenza è statisticamente significativa)

Morale: *Ancora una volta, pur di randomizzare posso usare il test t (come approssimazione della distr. randomizzata) senza dover ipotizzare il campionamento casuale.*

L'esperimento dei pomodori rivisitato

Osservazione: Se c'è variabilità spaziale tra le piante, è verosimile che piante consecutive siano più simili.

Idea: Supponendo che siano 12, suddivido le piante in coppie:

(1 2)	(3 4)	(5 6)	(7 8)	(9 10)	(11 12)

Randomizzazione: Lancio una moneta per decidere in ciascuna coppia se applicare il fertilizzante alla pianta dispari o a quella pari:

(1 2)	(3 4)	(5 6)	(7 8)	(9 10)	(11 12)
(B A)	(B A)	(A B)	(B A)	(A B)	(B A)

Conviene appaiare? Dipende:

- Sola randomizzazione di 12 dati: test *t* con 10 gradi di libertà.

- Dati appaiati: test *t* con 5 gradi di libertà.

$$\Downarrow$$

Conviene appaiare se la riduzione di varianza conseguente all'accoppiamento compensa il calo dei gradi di libertà (distribuzione di riferimento più "panciuta")

Blocchi

La coppia è un particolare *blocco* di dimensioni due.

Esempio: 4 diversi tipi di ferro da cavallo. Ad ogni cavallo applico tutti e 4 i tipi scegliendo però le zampe cui applicarli in modo casuale.

Osservazioni vicine nel tempo o nello spazio sono
di solito più simili rispetto a quelle distanti

$$\Downarrow$$

le osservazioni fatte nello stesso giorno
o nello stesso lotto possono costituire un blocco

Morale: *Suddividi in blocchi ciò che puoi e randomizza ciò che non puoi suddividere.*

CONCLUSIONI

- Test di ipotesi: sebbene la scelta della soglia implichi un compromesso tra F+ ed F-, viene di solito stabilita imponendo un valore piccolo di $\alpha = P(F+)$.

- Confronto di due trattamenti: usare un insieme di dati di riferimento esterni evita ipotesi poco realistiche ma richiede la disponibilità di numerosi dati passati.

- Confronto di due trattamenti: con l'ipotesi di campionamento casuale posso fare a meno di dati di riferimento esterni. Sono sicuro che l'ipotesi sia realistica?

- Confronto di due trattamenti: progetti a blocchi randomizzati garantiscono la validità del test *t* anche in assenza dell'ipotesi di campionamento casuale.

DOMANDE

V ***F***

1. In un test di ipotesi, le probabilità α e β di errore del I e II tipo soddisfano $\alpha + \beta = 1$.

   ❑ ❑

2. In un test di ipotesi con $\alpha = 0.05$, la probabilità di respingere l'ipotesi quando essa è vera è pari a 0.05.

   ❑ ❑

3. Nel confronto di due trattamenti, per usare una distribuzione di riferimento esterna è necessario che le osservazioni siano indipendenti.

   ❑ ❑

4. Se sono soddisfatte le ipotesi di applicabilità di un test per confrontare due trattamenti, posso essere tanto più sicuro della loro differenza quanto più piccolo è il *P*-value.

   ❑ ❑

5. Nel confronto di due trattamenti, se vale l'ipotesi di campionamento casuale, posso costruire una distribuzione di riferimento senza usare dati esterni.

   ❑ ❑

6. Nel confronto di due trattamenti, se vale l'ipotesi di campionamento casuale posso sempre fare riferimento alla tabella della normale standard.

   ❑ ❑

7. Nel confronto di due trattamenti, i test non parametrici non richiedono l'ipotesi di campionamento casuale.

   ❑ ❑

8. Nel confronto di due trattamenti, se ho molti dati, il teorema centrale del limite mi permette di fare a meno dell'ipotesi di campionamento casuale.

   ❑ ❑

9. Nel confronto di due trattamenti, l'unica alternativa all'uso di dati esterni è fare l'ipotesi di campionamento casuale.

   ❑ ❑

10. La randomizzazione e l'appaiamento non possono essere applicati insieme.

   ❑ ❑

ESERCITAZIONE

Si supponga di voler confrontare due trattamenti A e B, per vedere se B produce spessori significativamente maggiori. A tale scopo, si prende un lotto e si applica il trattamento A a 12 wafer e quello B ad altri 12 wafer. Da ogni wafer si ottiene una misura media di spessore.

1) Discutere come si potrebbe effettuare il confronto se fossero disponibili le misure dei wafer di diversi lotti precedentemente sottoposti al trattamento A. Se si usa un test *t*, indicare il numero di gradi di libertà.

2) Discutere se è possibile costruire una distribuzione di riferimento interna basata sull'ipotesi di campionamento casuale. Se si usa un test *t*, indicare il numero di gradi di libertà.

3) Discutere una possibile strategia di randomizzazione (senza appaiare). Se si usa un test *t*, indicare il numero di gradi di libertà.

4) Discutere una possibile strategia di appaiamento. Se si usa un test *t*, indicare il numero di gradi di libertà.