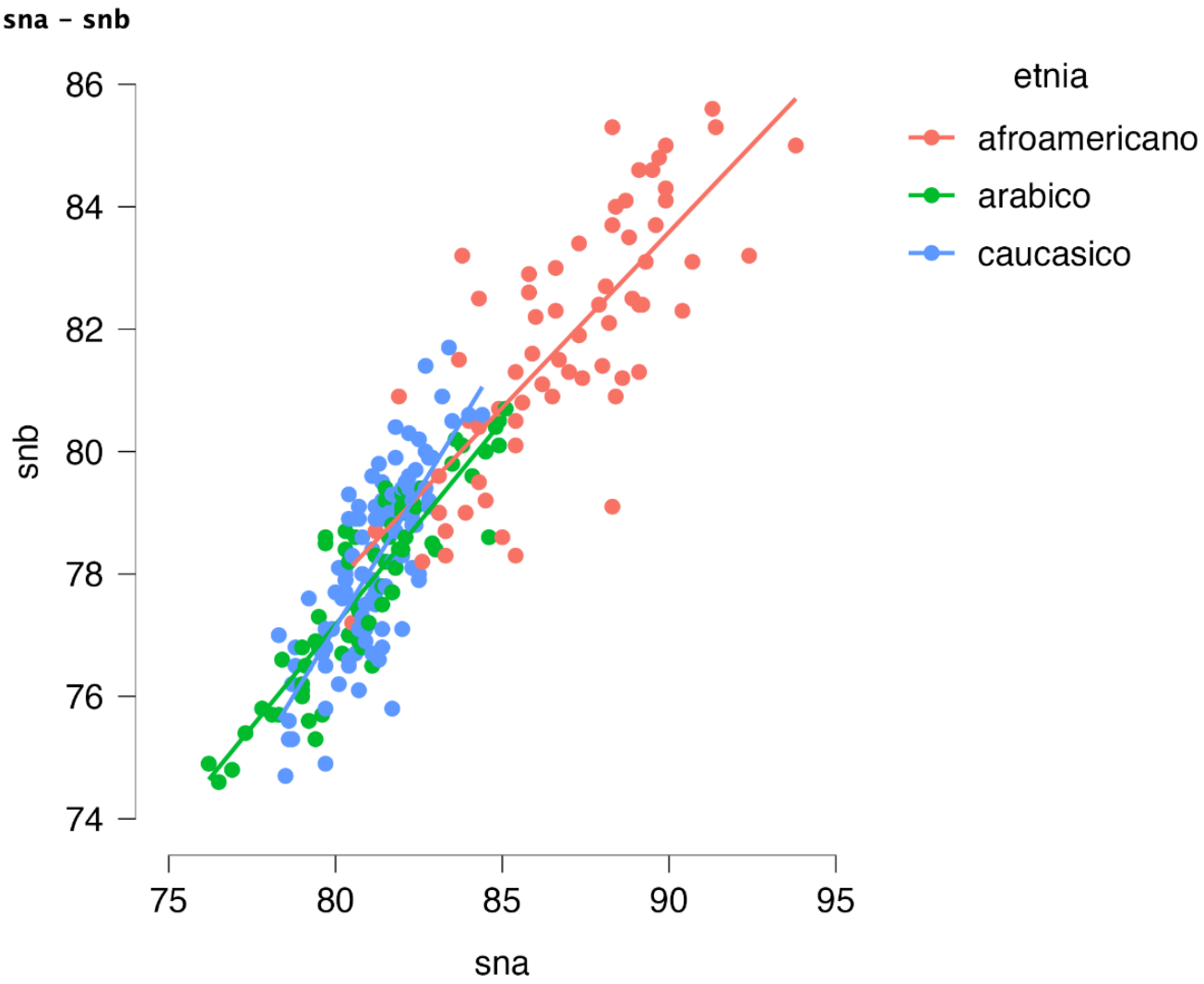
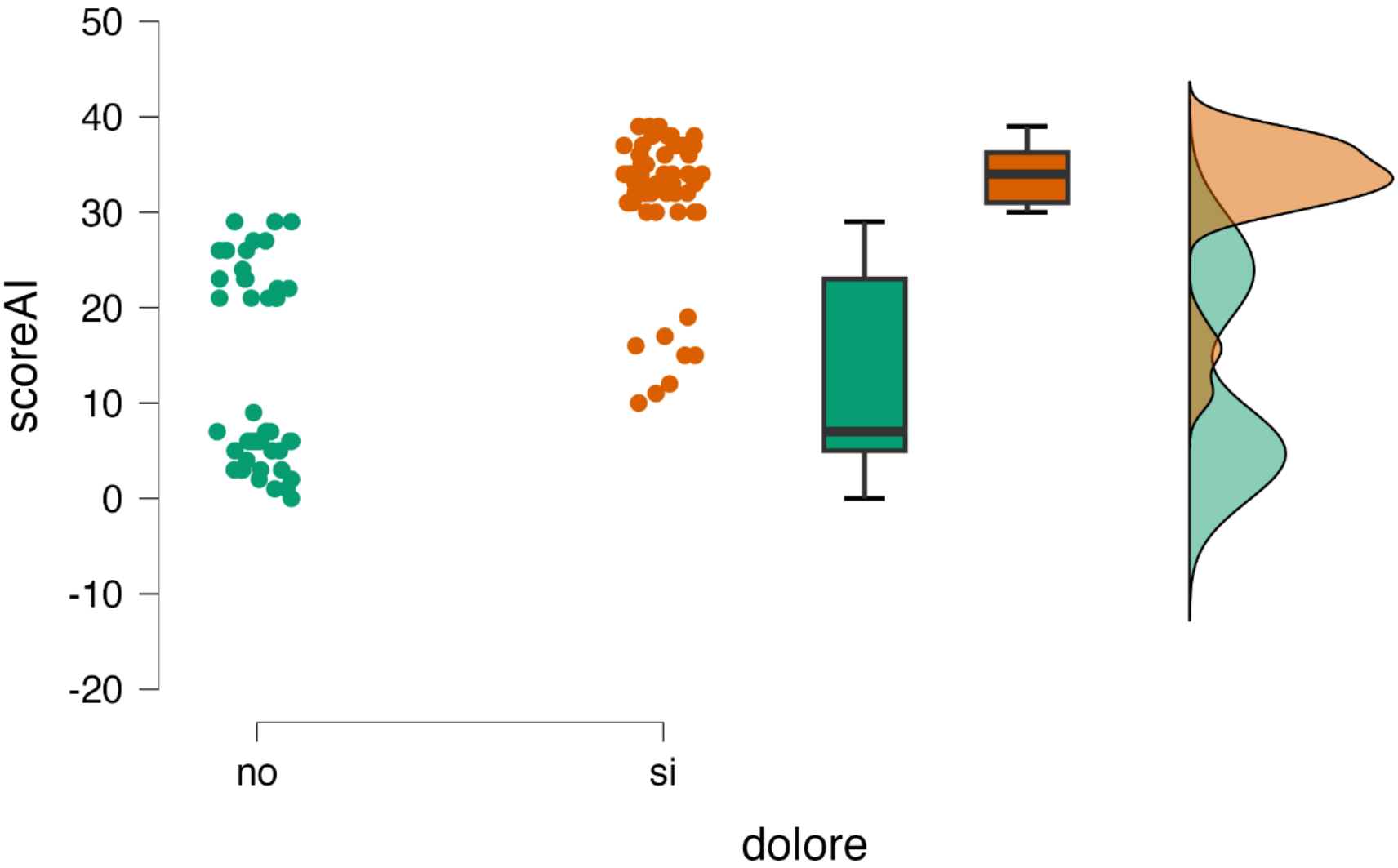
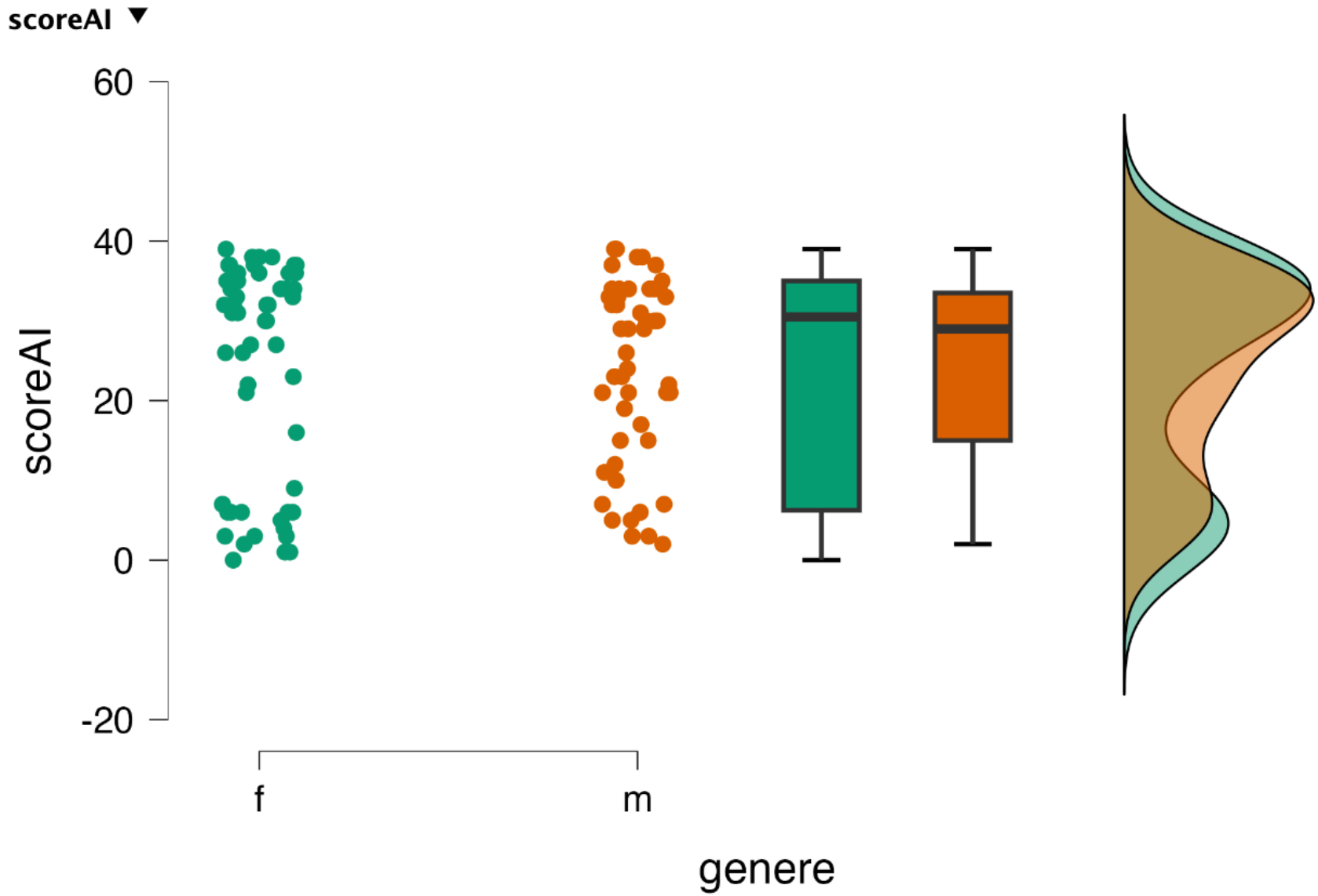
T-Test: perché e percóme

Massimo Borelli

Maggio 2024


sna - snb
86
84
82
80
78
76
74
snb
75
80
85
90
95
sna
etnia
afroamericano
arabico
caucasico


50
40
30
20
10
0
-10
-20
scoreAI
no
si
dolore

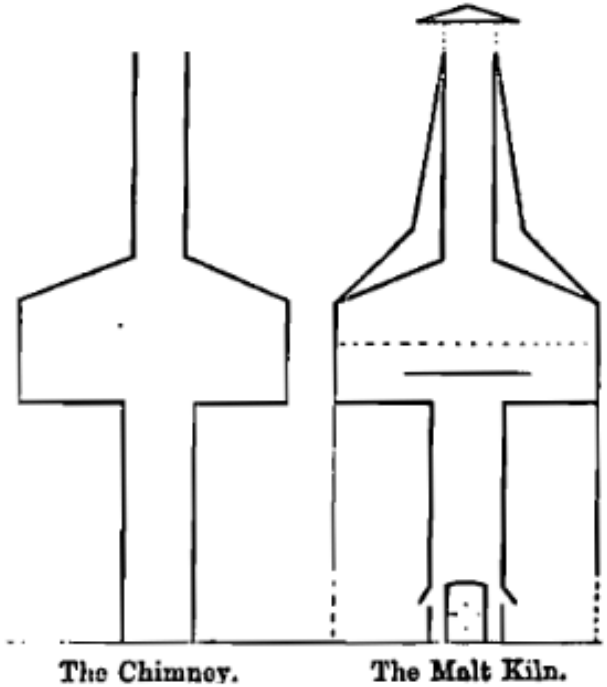

scoreAI
60
40
20
0
-20
scoreAI
f
m
genere

Analisi inferenziale univariata

esempio di importanza storica:

- il dataset `gossett.ods`

Arthur Guinness



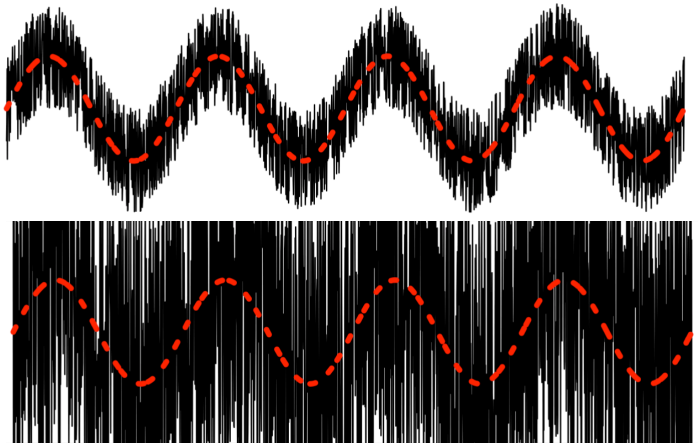
The Kiln Drying of Malt.

By H. M. Chubb.

Figura: William Sealy Gossett

Not Kiln-Dried	Kiln-Dried	Difference
1903	2009	+106
1935	1915	-20
1910	2011	+101
2496	2463	-33
2108	2180	+72
1961	1925	-36
2060	2122	+62
1444	1482	+38
1612	1542	-70
1316	1443	+127
1511	1535	+24

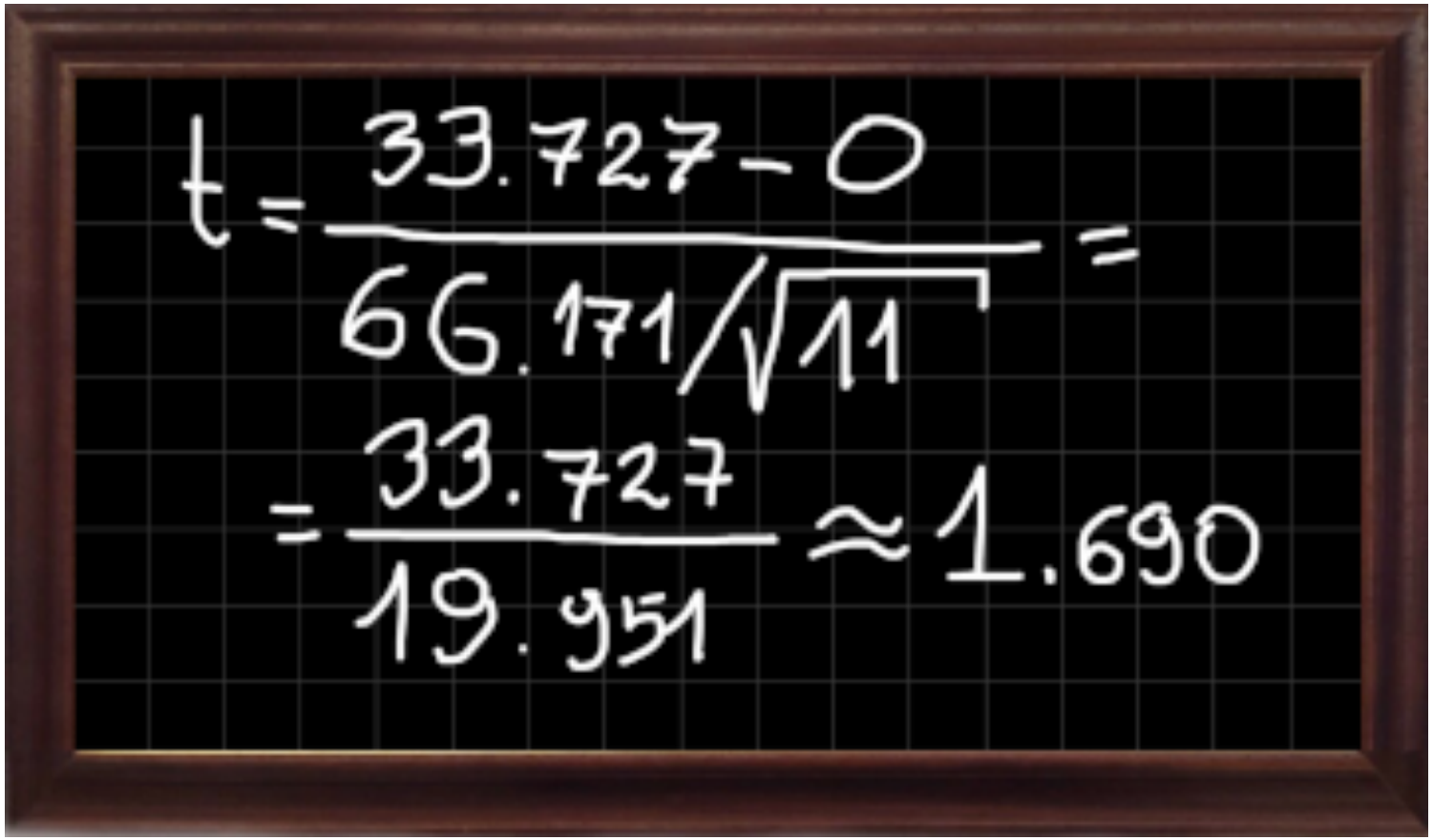
Not Kiln-Dried	Kiln-Dried	Difference
1903	2009	+106
1935	1915	-20
1910	2011	+101
2496	2463	-33
2108	2180	+72
1961	1925	-36
2060	2122	+62
1444	1482	+38
1612	1542	-70
1316	1443	+127
1511	1535	+24



	difference
Valid	11
Mean	33.727
Std. Deviation	66.171
Std. Error of Mean	19.951

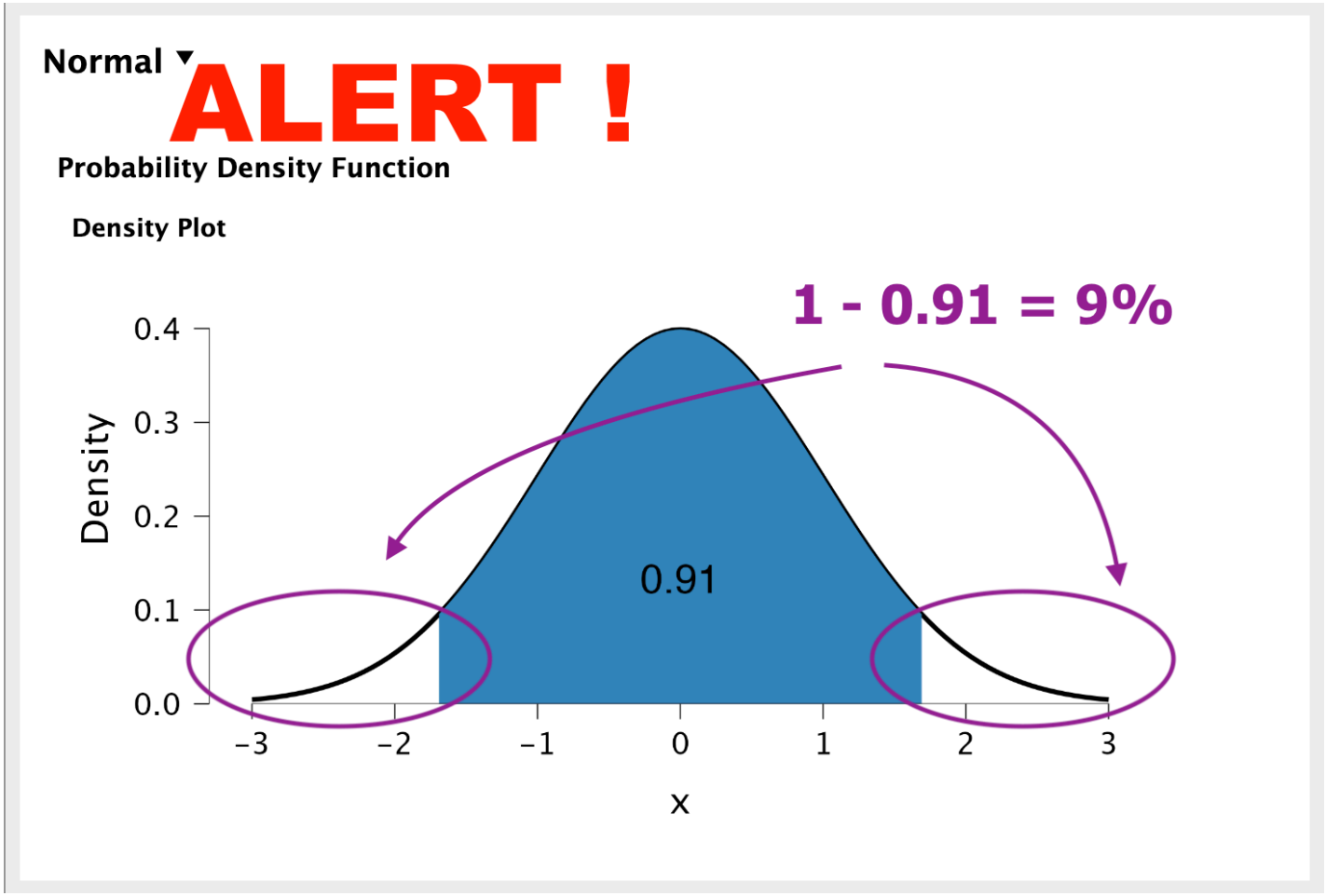
- il rapporto *segnale* - *rumore*

$$t = \frac{m - \mu}{s/\sqrt{n}}$$



il numero t = 1.69 è il **quantile**. Ma di quale variabile aleatoria?

cosa scopre William Gosset? /1



la distribuzione normale non è 'giusta', è troppo 'ottimistica'

cosa scopre William Gosset? /2

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

1. (**independenza**) in un campione di numeri casuali estratti dalla distribuzione gaussiana $N(\mu, \sigma)$, la stima della media campionaria m non fornisce alcuna informazione sulla deviazione standard campionaria s, e viceversa.

2. (**novità**) la variabile aleatoria $t = \frac{m-\mu}{s/\sqrt{n}}$ possiede una propria (.. difficile ..) funzione densità, che non è la gaussiana, ma che si può calcolare (con approssimazione).

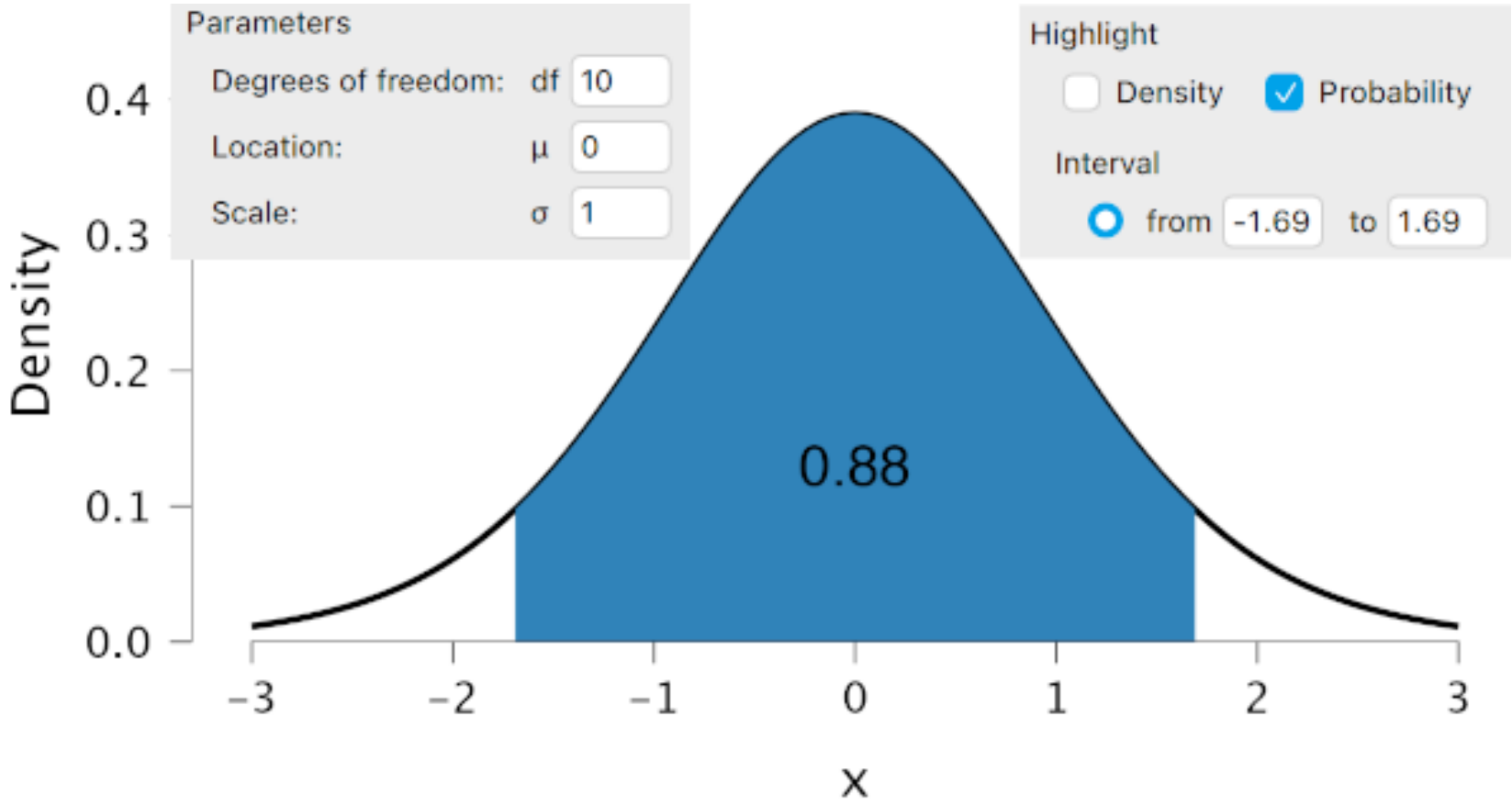
VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

JASP: **Scaled Shifted Student's t**



.. dunque, ecco il **risultato** del test T di Student

JASP: **Classical** One Sample T-Test

Tabella: One Sample T-Test

	t	df	p
difference	1.690	10	0.122

ci rimane da capire: **come** si interpreta questo risultato? Conviene o non conviene essicare i semi?

JASP: **Classical** One Sample T-Test

Tabella: One Sample T-Test

	t	df	p
difference	1.690	10	0.122

vox populi, vox Dei

- se $p < 0.05$ la differenza è significativa ..
- se $p \geq 0.05$ la differenza non è significativa ..

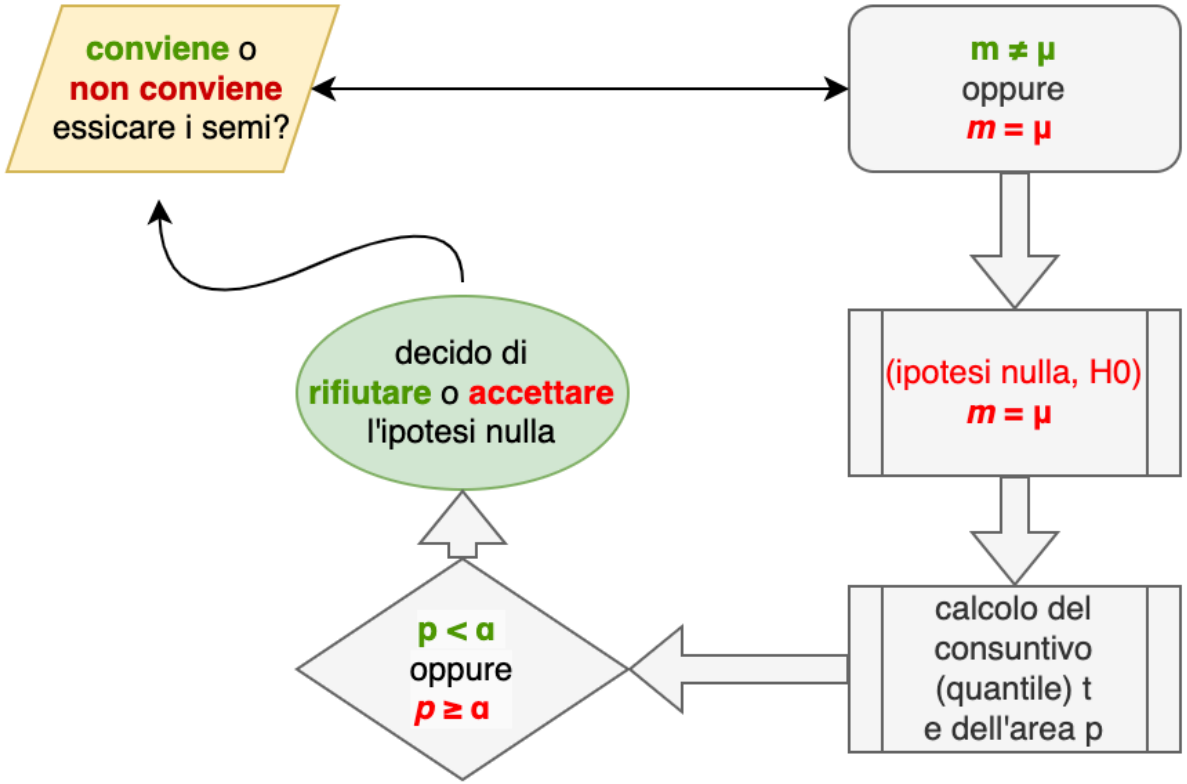
la decisione di William Gosset

To test whether it is of advantage to kiln-dry barley seed before sowing, seven varieties of barley were sown (both kiln-dried and not kiln-dried) in 1899 and four in 1900; the results are given in the table.

It will be noticed that the kiln-dried seed gave on an average the larger yield of corn and straw, but that the quality was almost always inferior. At first sight this might be supposed to be due to superior germinating power in the kiln-dried seed, but my farming friends tell me that the effect of this would be that the kiln-dried seed would produce the better quality barley. Dr Voelcker draws the conclusion "In such seasons as 1899 and 1900 there is no particular advantage in kiln-drying before sowing." Our examination completely justifies this

William Gossett non ha rifiutato l'ipotesi (H_0) che la differenza media osservata $m = +33.7$ sia diversa dalla differenza teorica $\mu = 0$

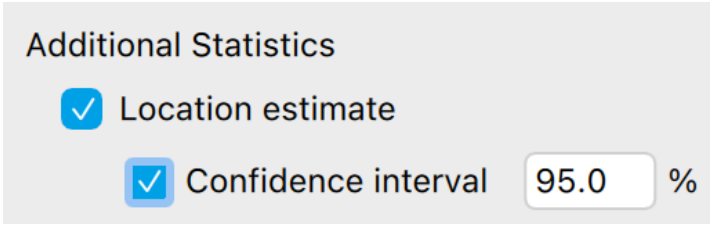
la decisione di William Gosset



William Gossett *non ha rifiutato* (? ha accettato ?) l'ipotesi nulla H_0 che la differenza media osservata $m = +33.7$ sia diversa dalla differenza teorica $\mu = 0$

c'è un altro modo di dirlo .. ..

.. con il *confidence interval* = intervallo di **fiducia**



	t	df	p	Mean Diff.	95% Conf. Int. Lower	95% Conf. Int. Upper
difference	1.690	10	0.122	33.727	−10.727	78.182

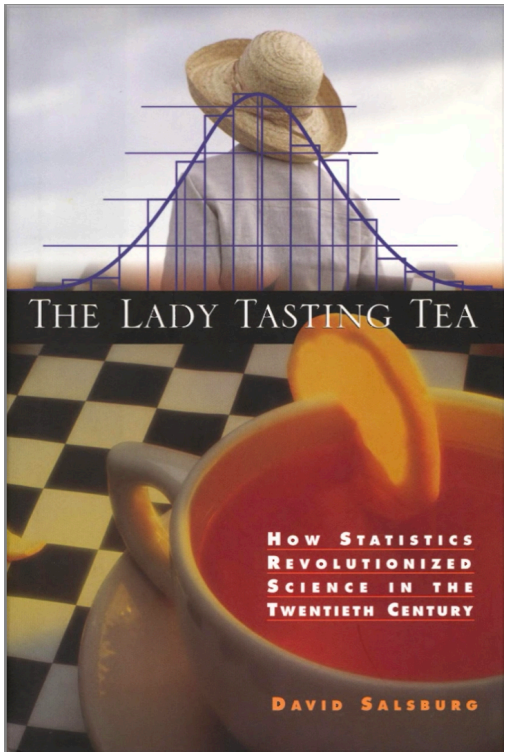
.. e siccome $0 \in [-10.7; 78.2]$ allora con un **grado di fiducia** del 95% (non una probabilità ..) decidiamo che la media m non si discosta dalla media vera $\mu = 0$; ossia, che la resa media 33.7 è dovuta al caso.

ci sono molti punti da chiarire ..

1. il *livello di significatività* al 5% è una convenzione
2. gli sperimentatori scelgono il livello di significatività α
3. il livello di significatività α ed il *sample size n* hanno un impatto sulla *potenza* del test
4. significatività statistica oppure *significatività clinica*
5. e se $p \geq 0.05$ .. *absence of evidence, or evidence of absence*?

1. l'idea di Ronald Fisher

il livello di significatività $\alpha = 0.05$ è una convenzione.



2. gli sperimentatori scelgono il livello di significatività α

Guideline 2. Define and justify a critical significance level α appropriate to the goals of your study.
For any statistical test, if the achieved significance level P is less than the critical significance level α, defined before any data are collected, then the experimental effect is likely to be real (see Ref. 9, p. 782). By tradition, most researchers define α to be 0.05: that is, 5% of the time they are willing to declare an effect exists when it does not. These examples illustrate that $\alpha = 0.05$ is sometimes inappropriate.
If you plan a study in the hopes of finding an effect that could lead to a promising scientific discovery, then $\alpha = 0.10$ is appropriate. Why? When you define α to be 0.10, you increase the probability that you find the effect if it exists.
In contrast, if you want to be especially confident of a possible scientific discovery, then $\alpha = 0.01$ is appropriate: only 1% of the time are you willing to declare an effect exists when it does not.

Douglas Curran-Everett, Dale J. Benos
Guidelines for reporting statistics in journals published by the American Physiological Society
https://journals.physiology.org/doi/full/10.1152/ajpcell.00250.2004

3. *potenza* di un test

il livello di significatività α ed il *sample size n* hanno un impatto sulla *potenza* del test

4. significatività statistica $\neq$ significatività clinica

ORIGINAL ARTICLE

Cetuximab for the Treatment of Colorectal Cancer

RESULTS

In comparison with best supportive care alone, cetuximab treatment was associated with a significant improvement in overall survival (hazard ratio for death, 0.77; 95% confidence interval [CI], 0.64 to 0.92; P=0.005)

The median survival was 6.1 months in the cetuximab group and 4.6 months in the supportive-care group.

5. .. e se $p \geq 0.05$?

? *absence of evidence, or evidence of absence* ?