

introduzione ai test di significatività

l'esperimento del *tea tasting*



R. Fisher

8 tazze di te

4 preparate versando il latte prima

4 preparate versando il latte dopo

variabile dipendente: numero di successi nel classificare le tazze come “prima” o “dopo”

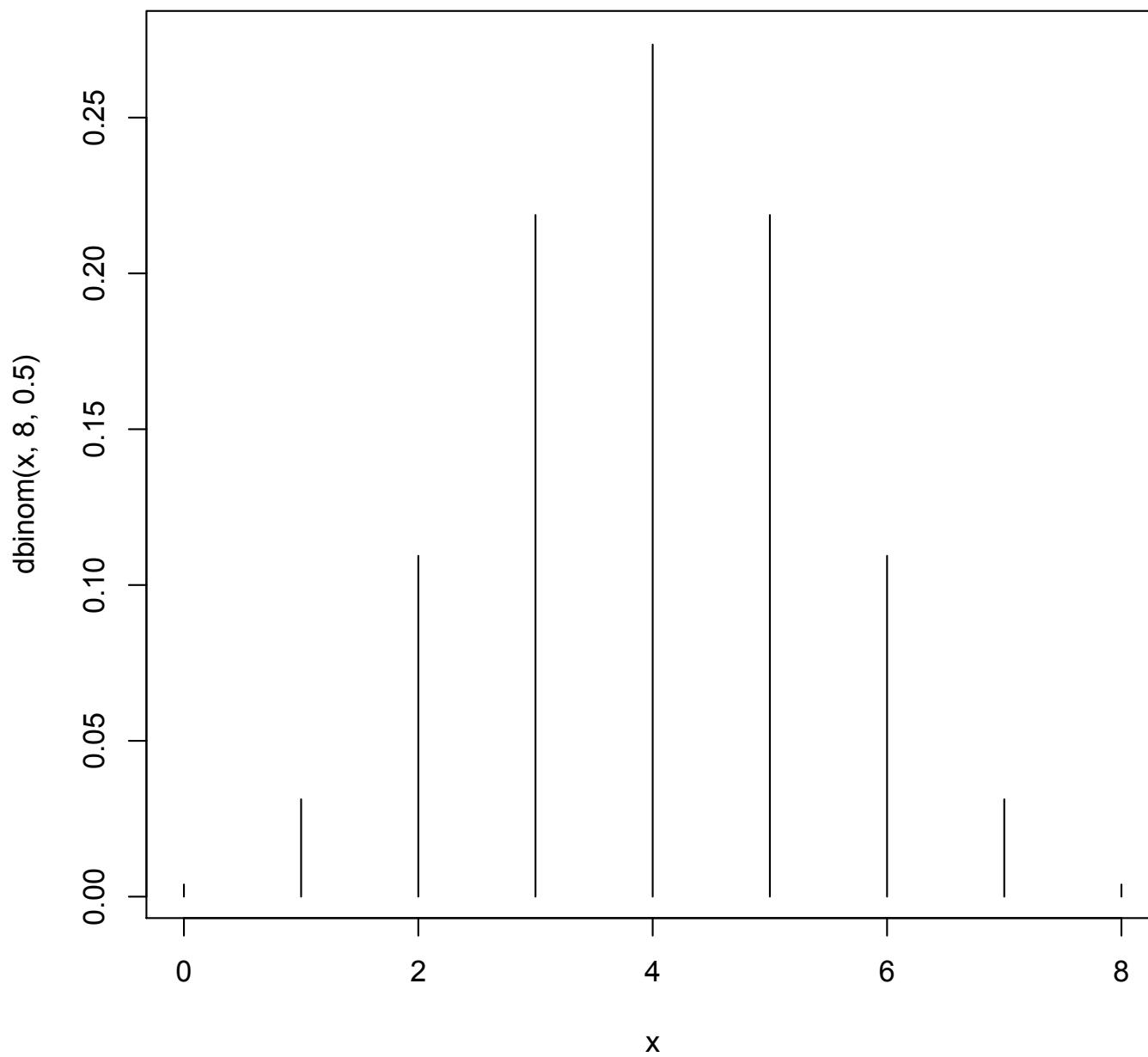
> x <- 0:8

> dbinom(x, 8, 0.5)

[1] 0.00390625 0.03125000 0.10937500 0.21875000

[5] 0.27343750 0.21875000 0.10937500 0.03125000

[9] 0.00390625



Fisher in realtà utilizzò una procedura diversa. Permise alla signora di fare dei confronti e quindi le chiese di identificare le quattro tazze “prima”.

Fisher ragionò in questa maniera: se la signora va a caso, ogni esito dell'esperimento equivale a scegliere a caso una delle maniere di combinare 8 elementi a 4 a 4.

Fisher aveva una ipotesi

dalla ipotesi aveva derivato un *modello* del processo che genera i dati, se è vera l'ipotesi

per confrontare i risultati con l'ipotesi, calcolò la probabilità di osservare quel risultato se il modello che genera i dati è quello ipotizzato

infine suggerì che $p < 0.05$ è un valore critico ragionevole per decidere che il modello non è plausibile

Le combinazioni di 8 elementi a 4 a 4 sono 70.

```
> choose(8, 4)  
[1] 70
```

Di queste solo una corrisponde all'esito in cui ci sono tutte le quattro "prima". Pertanto la probabilità di questo esito è $1/70 = 0.014$.

Fisher decise quindi che sarebbe stato disposto a rifiutare l'ipotesi che la signora vada a caso solo se ne avesse indovinate 4 su 4.

Si dice che la signora le indovinò tutte!

il NHST

perché studiarlo

il NHST è ancora oggi il metodo inferenziale più usato nella letteratura scientifica, soprattutto psicologica

per questo motivo, occorre sapere come si usa

una procedura controversa

il NHST è un metodo controverso

**voci critiche isolate sono presenti
in letteratura da almeno 50 anni**

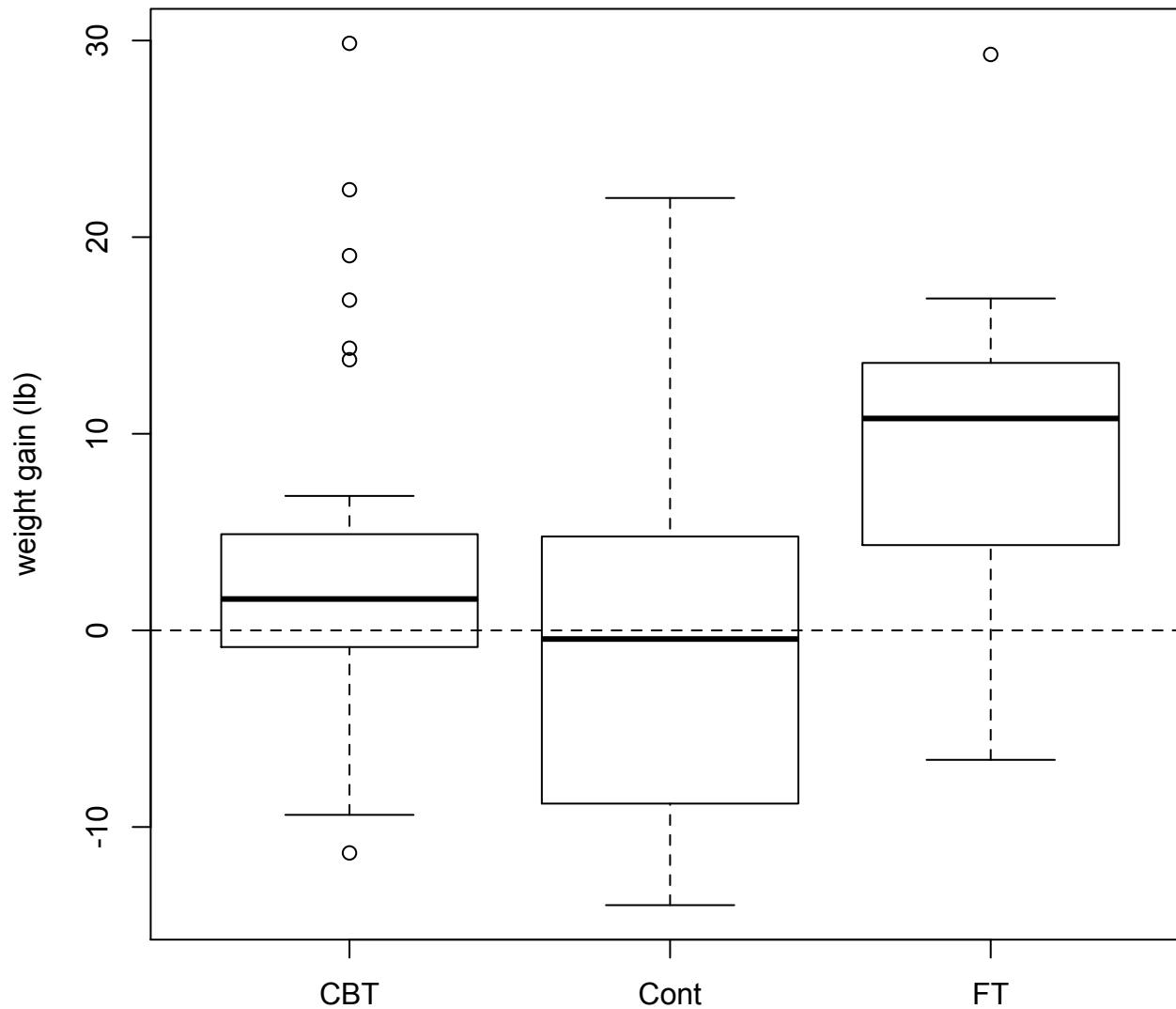
**da alcuni anni sempre di più viene
messo in discussione, e
accompagnato da un uso sempre
maggiore di CI e strumenti grafici**

NHST

- 1. si definisce una *ipotesi nulla* (H_0)**
- 2. si calcola p di osservare un risultato almeno pari a quello osservato, se è vera H_0**
- 3. se p è minore di un criterio (alfa), si rifiuta H_0**

```
> library(MASS)
> data(anorexia)
> str(anorexia)
'data.frame': 72 obs. of 3 variables:
 $ Treat : Factor w/ 3 levels "CBT","Cont","FT": 2 2 2 2
 2 2 2 2 2 2 ...
 $ Prewt : num 80.7 89.4 91.8 74 78.1 88.3 87.3 75.1
 80.6 78.4 ...
 $ Postwt: num 80.2 80.1 86.4 86.3 76.1 78.1 75.1
 86.7 73.5 84.6 ...
>d <- within(anorexia, pcd <- (Postwt - Prewt)/Prewt *
 100)
```

```
> with(d, boxplot(PCD ~ Treat, ylab = "weight gain (lb)"))  
> abline(h = 0, lty = "dashed")
```



```
> mns <- with(d, tapply(PCD, Treat, mean))  
> mns
```

	CBT	Cont	FT
	3.72397187	-0.00655586	8.80129215

```
> sds <- with(d, tapply(PCD, Treat, sd))  
> ns <- with(d, tapply(PCD, Treat, length))  
> ses <- sds/sqrt(ns)  
> ses
```

	CBT	Cont	FT
	1.732465	1.972014	2.217835

```
> with(d, t.test(PCD[Treat == "Cont"], PCD[Treat ==  
"CBT"]))
```

Welch Two Sample t-test

data: PCD[Treat == "Cont"] and PCD[Treat == "CBT"]
t = -1.4212, df = 51.233, p-value = 0.1613

alternative hypothesis: true difference in means is not
equal to 0

95 percent confidence interval:

-8.999718 1.538663

sample estimates:

mean of x mean of y

-0.00655586 3.72397187

```
> with(d, t.test(PCD[Treat == "Cont"], PCD[Treat == "FT"]))
```

Welch Two Sample t-test

data: PCD[Treat == "Cont"] and PCD[Treat == "FT"]

t = -2.9678, df = 36.642, p-value = 0.005258

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-14.823094 -2.792602

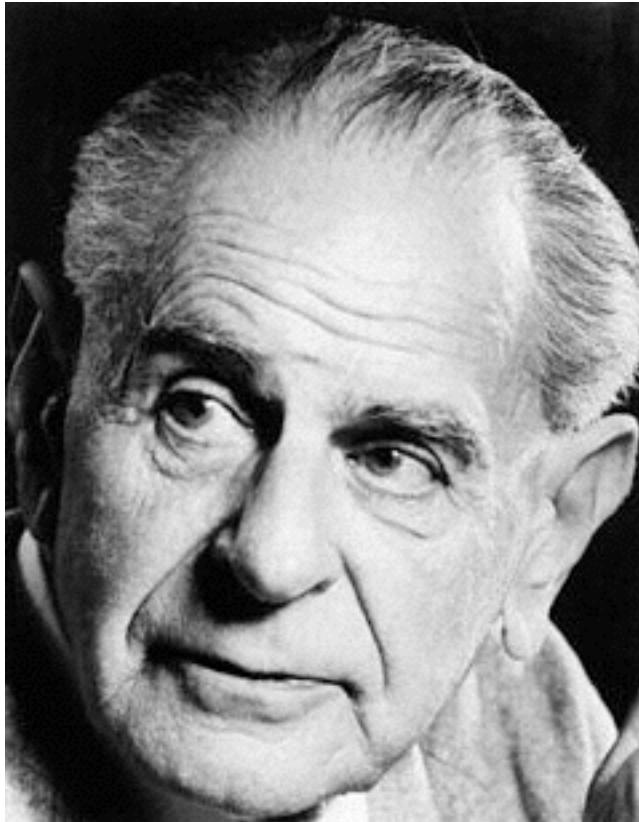
sample estimates:

mean of x mean of y

-0.00655586 8.80129215

logica

falsificazionismo



**Karl Popper
(1902-1994)**

deduzione

**derivare un'conclusione da
premesse più generiche, dentro cui
tale conclusione è implicita**

**un esempio è il sillogismo
aristotelico, che ha la forma:**

**premessa 1 ("maggiore")
premessa 2 ("minore")
conclusione**

implicazione

$p \rightarrow q$

“se p, allora q”

“ogni volta che si verifica p, si verifica q”

“se piove, il tetto perde”

tabella di verità

p \rightarrow q

p

q

TRUE

p

NON q

FALSE

NON p

q

TRUE

NON p

NON q

TRUE

implicazione stretta

$p \leftrightarrow q$

p

q

TRUE

p

NON q

FALSE

NON p

q

FALSE

NON p

NON q

TRUE

conclusioni valide

due conclusioni deduttive valide:

modus ponens:

p -> q, p, dunque q

modus tollens:

p -> q, NON q, dunque NON p

**due tipi di test del
chi-quadro**

distribuzione chi-quadrato

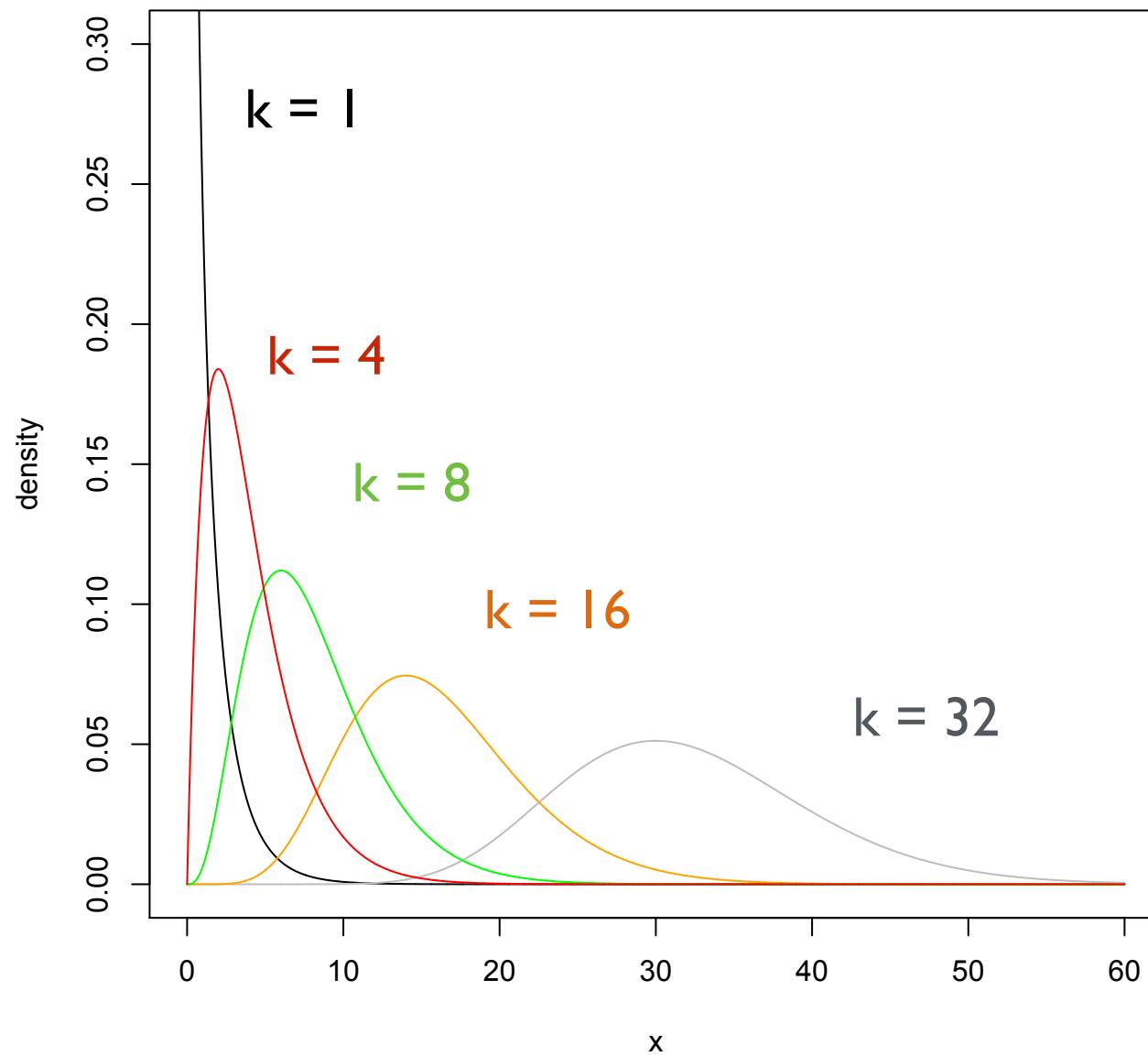
la distribuzione teorica di probabilità della somma di k scarti normalizzati dalla media di una distribuzione normale (z^2)

k = gradi di libertà

ogni membro della famiglia ha media = k e varianza = $2k$

si utilizza per la stima intervallare della varianza della popolazione, e per due tipi di test di significatività

distribuzione teorica di chi-quadrato



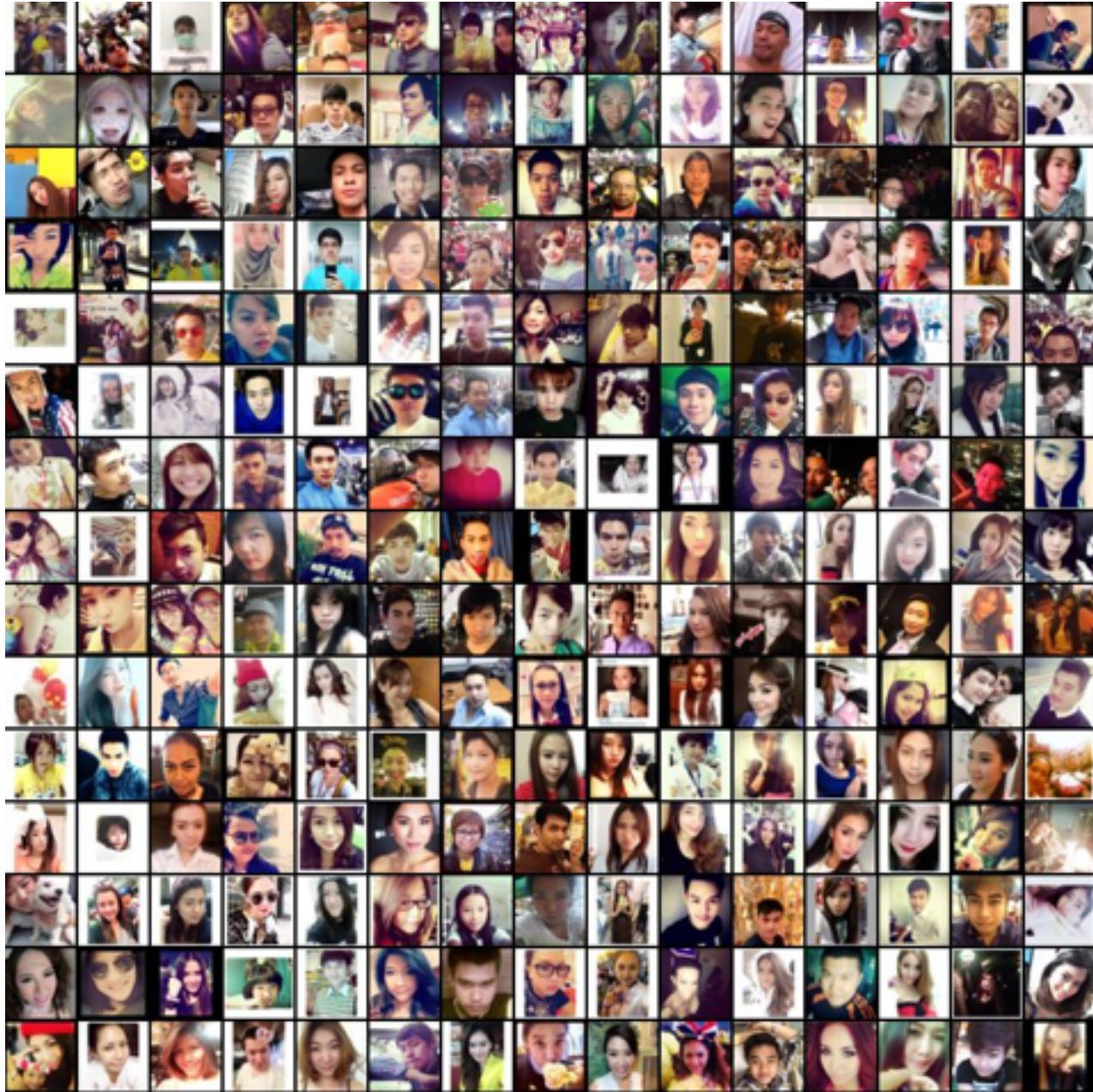
test di associazione

dati: tabella di contingenza

H0: indipendenza

distribuzione campionaria:
chi-quadrato con $(\text{righe} - 1) \times (\text{colonne} - 1)$ gradi di libertà

statistica: chi-quadrato (frequenze osservate - frequenze attese in base ad H0)



The SELFIEEXPLORATORY is part of SELFIECITY

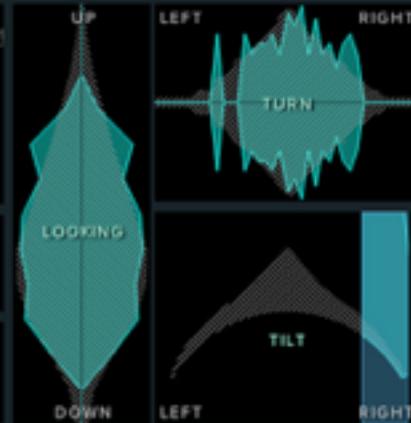
DEMOGRAPHICS



YOUNG AGE OLD



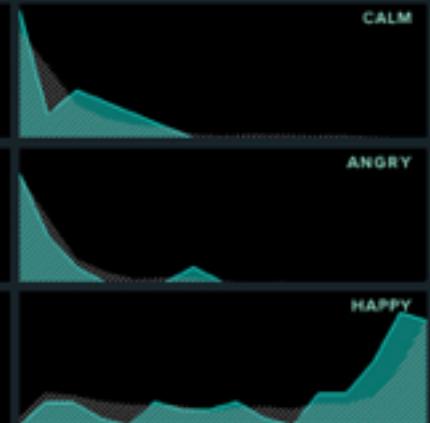
POSE



FEATURES



MOOD

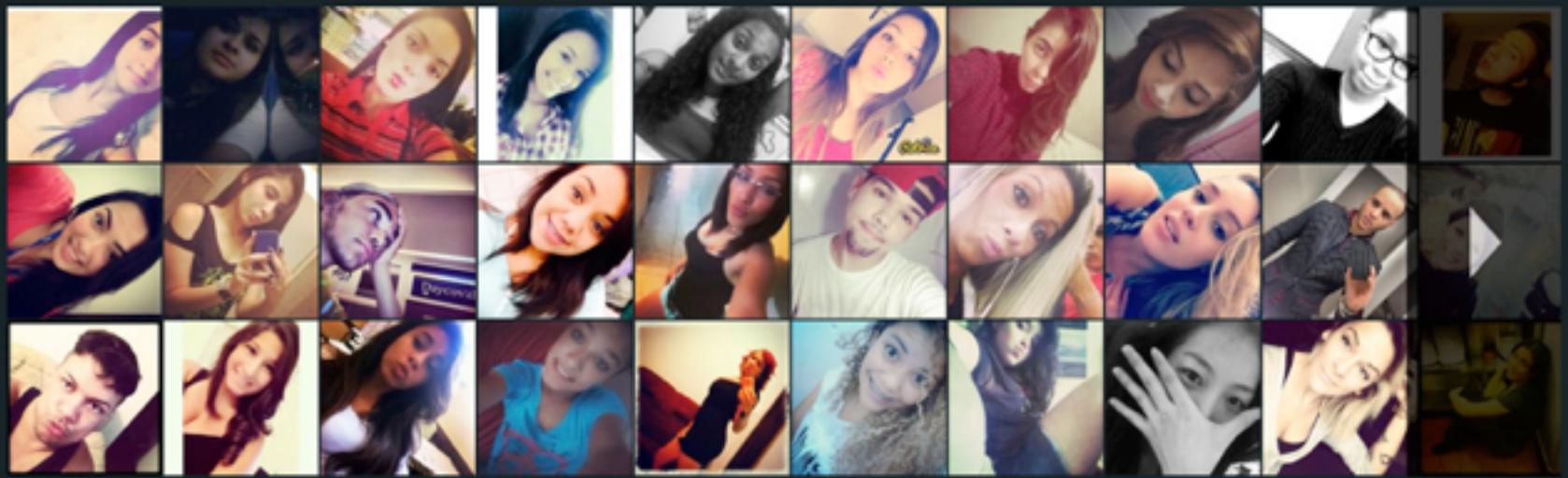


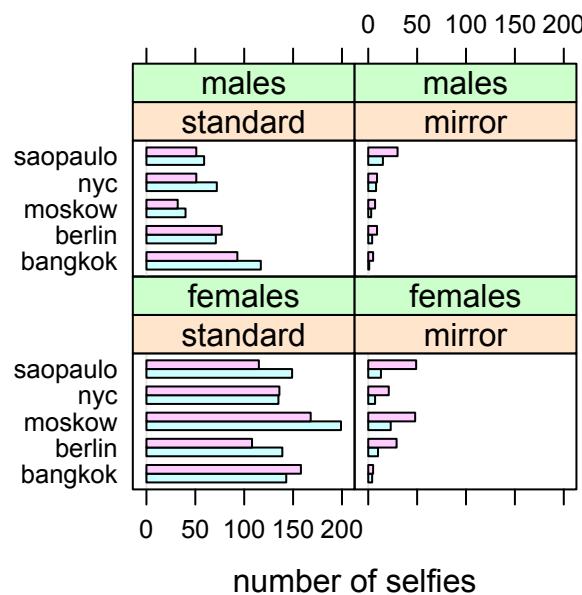
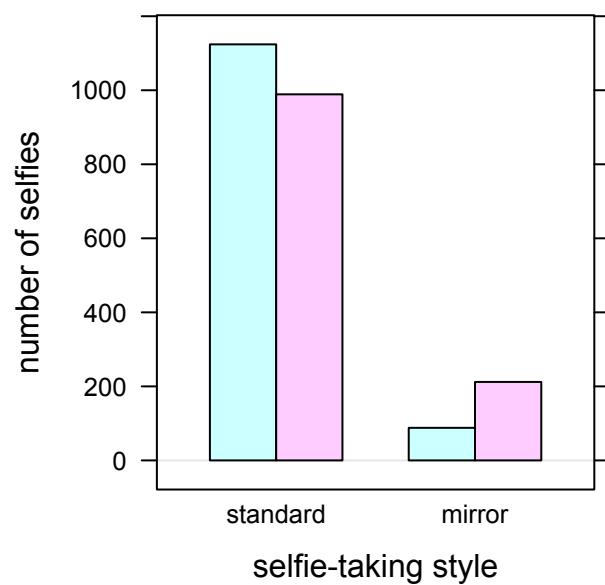
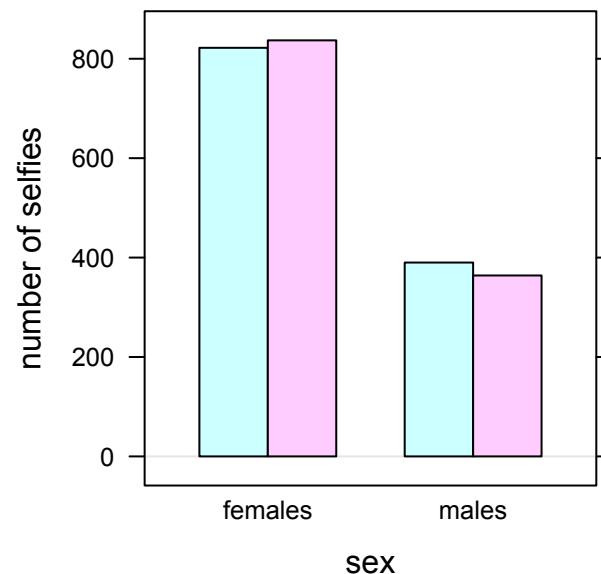
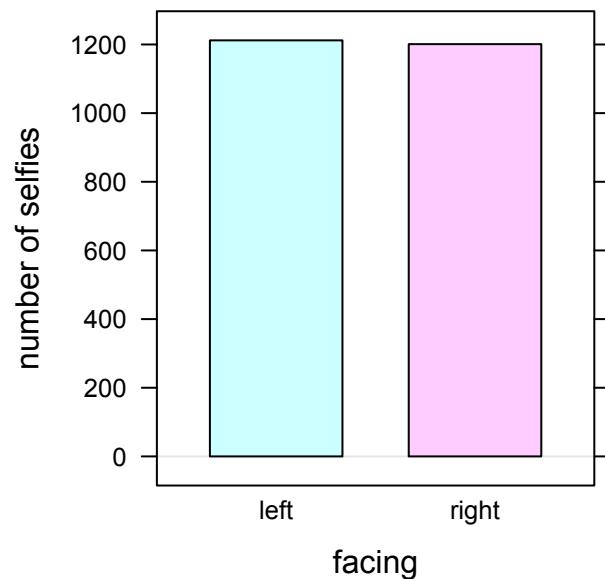
65 of 3200 faces.

Normal

Crop

Crop & rotate





```
d <- read.table("~/Dropbox/selfiecity/selfiecity.txt", 1)

> str(d)
'data.frame': 3200 obs. of 5 variables:
 $ n      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ mirror: Factor w/ 3 levels "n","u","y": 1 3 1 1 1 1 1 1 1 1
3 ...
 $ pose   : Factor w/ 6 levels "f","l","r","sl",...: 3 2 5 3 3 2
2 4 1 3 ...
 $ city   : Factor w/ 5 levels "bangkok","berlin",...: 2 2 2 2 2
2 2 2 2 2 ...
 $ sex    : Factor w/ 3 levels "f","m","u": 2 2 2 2 2 2 2 2 2 2
...
 ...

> head(d)
  n mirror pose   city sex
1 1      n     r berlin   m
2 2      y     l berlin   m
3 3      n    sr berlin   m
4 4      n     r berlin   m
5 5      n     r berlin   m
6 6      n     l berlin   m
```

```

sr <- subset(d, (s$mirror == "y" | s$mirror == "n") & (s$pose
== "l" | s$pose == "r") & s$sex != "u")
levels(sr$pose) <- c("front", "left", "right", "slight left",
"slight right", "unknown")
levels(sr$sex) <- c("females", "males", "unknown")
levels(sr$mirror) <- c("standard", "unknown", "mirror")
t <- table(droplevels(sr)$mirror, droplevels(sr)$pose)
print(t)
print(chisq.test(t))

```

	left	right
standard	1124	989
mirror	88	212

Pearson's Chi-squared test with Yates' continuity correction

```

data: t
X-squared = 58.8789, df = 1, p-value = 1.677e-14

```

test di *goodness-of-fit*

***dati:* distribuzione di frequenze osservate**

H0: goodness of fit

distribuzione campionaria:
chi-quadrato con (righe - 1) gradi di libertà

***statistica:* chi-quadrato (frequenze osservate - attese in base ad H0)**

```
> sr <- subset(d, (s$mirror == "n") & (s$pose ==  
"l" | s$pose == "r") & s$sex != "u")  
> t <- table(droplevels(sr)$pose)  
> t
```

	l	r
1124	989	

```
> chisq.test(t, p = c(0.5, 0.5))
```

Chi-squared test for given probabilities

```
data: t  
X-squared = 8.6252, df = 1, p-value = 0.003315
```

Pearson's Chi-squared Test for Count Data

Description

chisq.test performs chi-squared contingency table tests and goodness-of-fit tests.

Usage

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

Arguments

x

a numeric vector or matrix. x and y can also both be factors.

y

a numeric vector; ignored if x is a matrix. If x is a factor, y should be a factor of the same length.

correct

a logical indicating whether to apply continuity correction when computing the test statistic for 2 by 2 tables: one half is subtracted from all $|O - E|$ differences; however, the correction will not be bigger than the differences themselves. No correction is done if simulate.p.value = TRUE.

p

a vector of probabilities of the same length of x. An error is given if any entry of p is negative.

rescale.p

a logical scalar: if TRUE then p is rescaled (if necessary) to sum to 1. If rescale.p is FALSE, and p does not

```
> chisq.test(t, p = c(0.5, 0.5))
```

Chi-squared test for given probabilities

```
data: t  
X-squared = 8.6252, df = 1, p-value = 0.003315
```

```
> chisq.test(t)
```

Chi-squared test for given probabilities

```
data: t  
X-squared = 8.6252, df = 1, p-value = 0.003315
```

t test: confronti
between e within

terminologia

between (fra i gruppi), o per gruppi indipendenti, o fattoriali: test su campioni composti da soggetti diversi

within (all'interno dei gruppi), o per gruppi dipendenti, o accoppiati, o per misure ripetute: test su campioni composti dagli stessi soggetti, misurati più di una volta

fonti di variazione

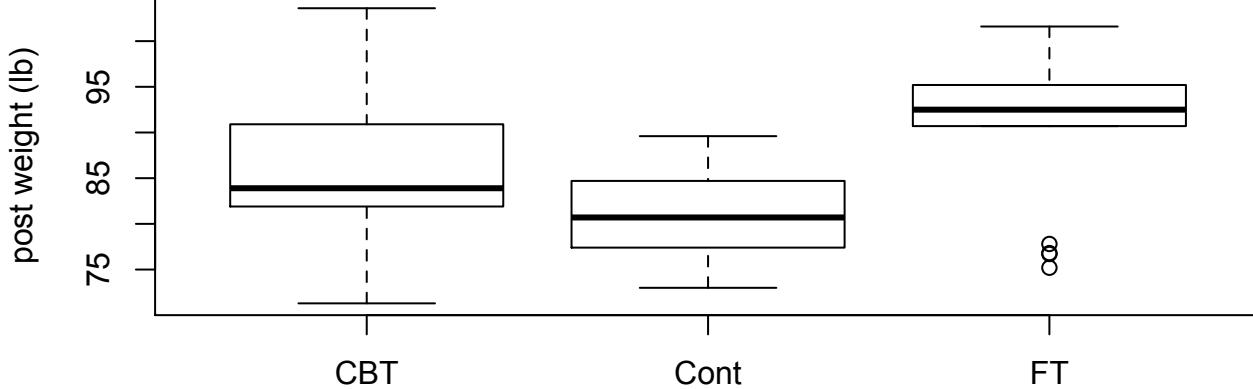
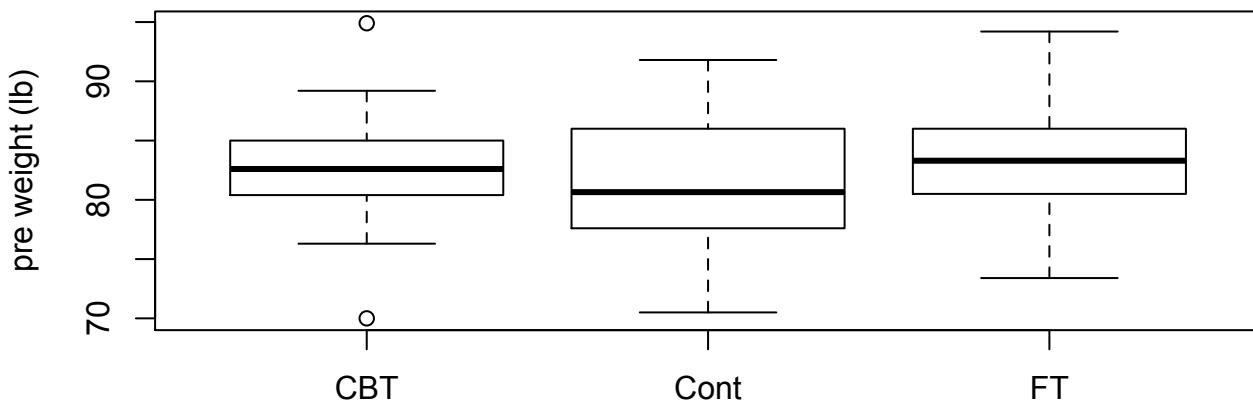
between:

dipende sia dalle differenze di trattamento fra i due gruppi, sia dalle differenze interindividuali all'interno dei gruppi

within:

le differenze individuali vengono eliminate perché ogni soggetto viene confrontato con se stesso

```
> par(mfrow = c(2, 1))
> with(d, boxplot(Prewt ~ Treat, ylab = "pre weight
  (lb)"))
> with(d, boxplot(Postwt ~ Treat, ylab = "post weight
  (lb)"))
```



```
> with(d, t.test(Prewt[Treat == "FT"], Prewt[Treat ==  
  "Cont"]))
```

Welch Two Sample t-test

```
data: Prewt[Treat == "FT"] and Prewt[Treat == "Cont"]  
t = 1.0112, df = 37.397, p-value = 0.3184  
alternative hypothesis: true difference in means is not equal to  
 0  
95 percent confidence interval:  
 -1.676823  5.020262  
sample estimates:  
mean of x mean of y  
83.22941  81.55769
```

```
> with(d, t.test(Prewt[Treat == "CBT"], Prewt[Treat ==  
"Cont"]))
```

Welch Two Sample t-test

data: Prewt[Treat == "CBT"] and Prewt[Treat == "Cont"]

t = 0.7882, df = 49.352, p-value = 0.4343

alternative hypothesis: true difference in means is not equal to
0

95 percent confidence interval:

-1.753429 4.017355

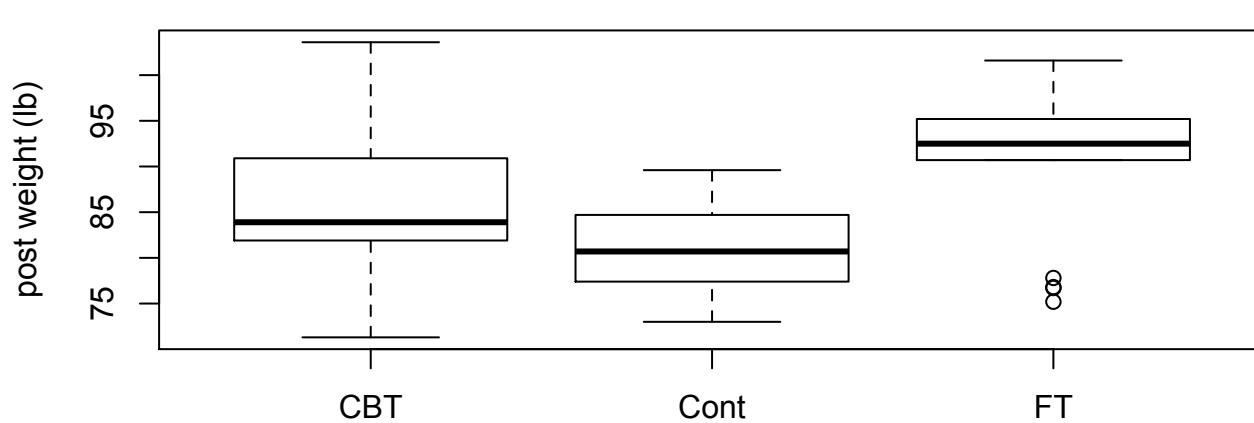
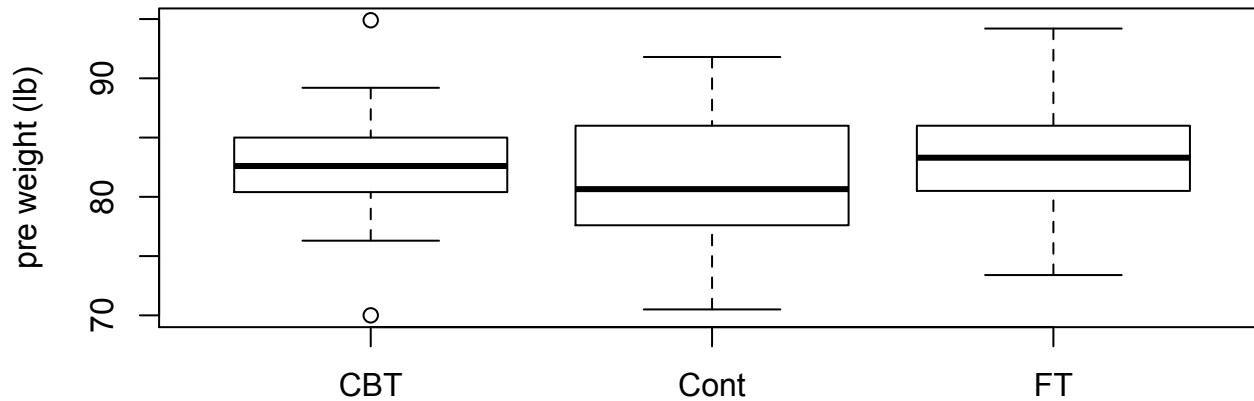
sample estimates:

mean of x mean of y

82.68966 81.55769

n.s.

n.s.



```
> with(d, t.test(Postwt[Treat == "FT"], Postwt[Treat == "Cont"],))
```

Welch Two Sample t-test

data: Postwt[Treat == "FT"] and Postwt[Treat == "Cont"]

t = 4.1601, df = 22.62, p-value = 0.0003888

alternative hypothesis: true difference in means is not equal to
0

95 percent confidence interval:

4.714618 14.058233

sample estimates:

mean of x mean of y

90.49412 81.10769

```
> with(d, t.test(Postwt[Treat == "CBT"], Postwt[Treat ==  
  "Cont"],))
```

Welch Two Sample t-test

data: Postwt[Treat == "CBT"] and Postwt[Treat == "Cont"]
t = 2.5372, df = 45.221, p-value = 0.01469

alternative hypothesis: true difference in means is not equal to
0

95 percent confidence interval:

0.9466501 8.2310687

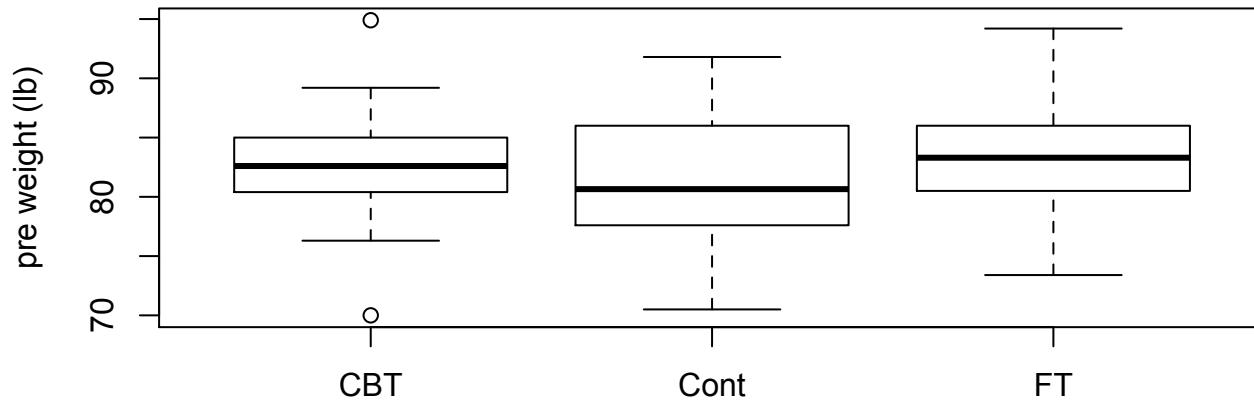
sample estimates:

mean of x mean of y

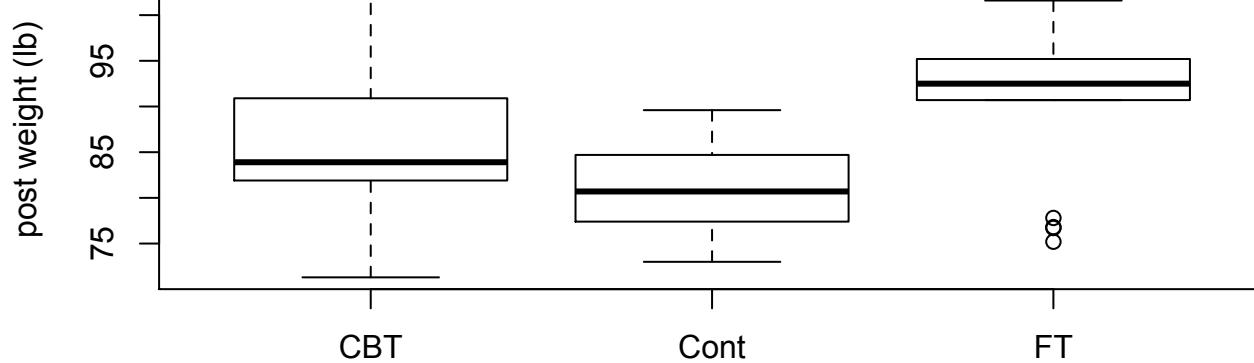
85.69655 81.10769

n.s.

n.s.

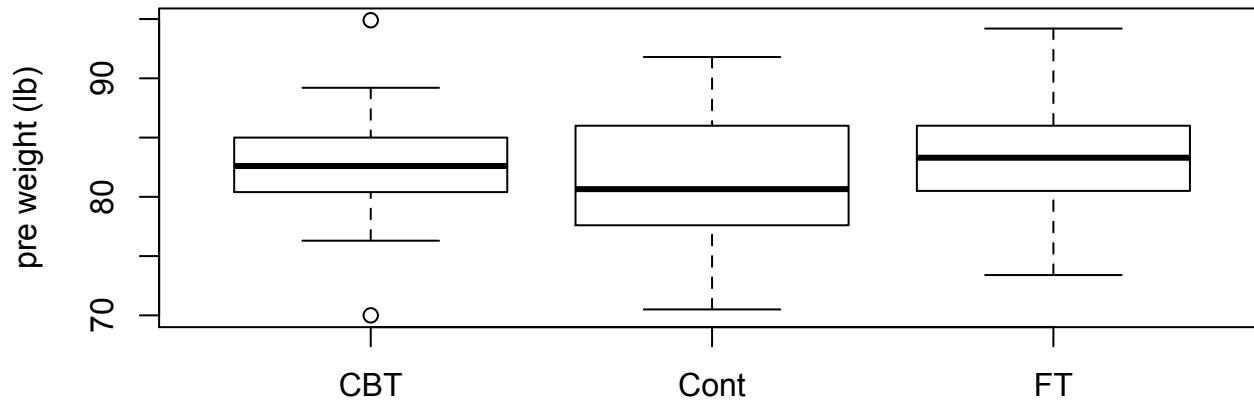


$p < 0.02$ $p < 0.0004$

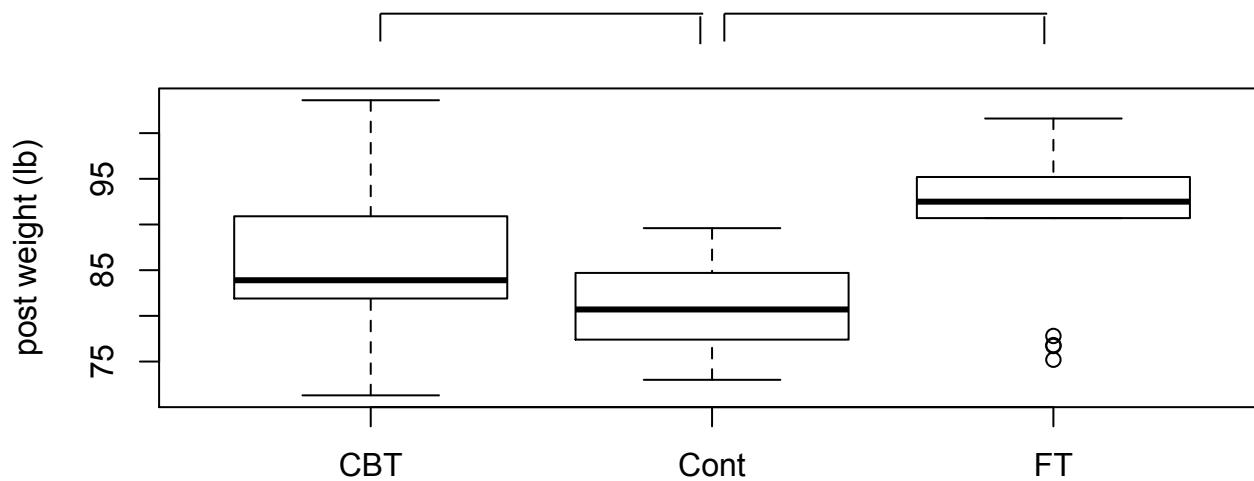


n.s.

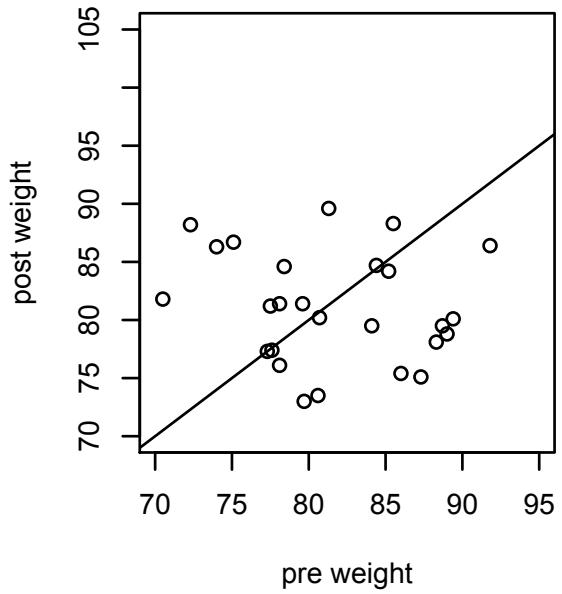
n.s.



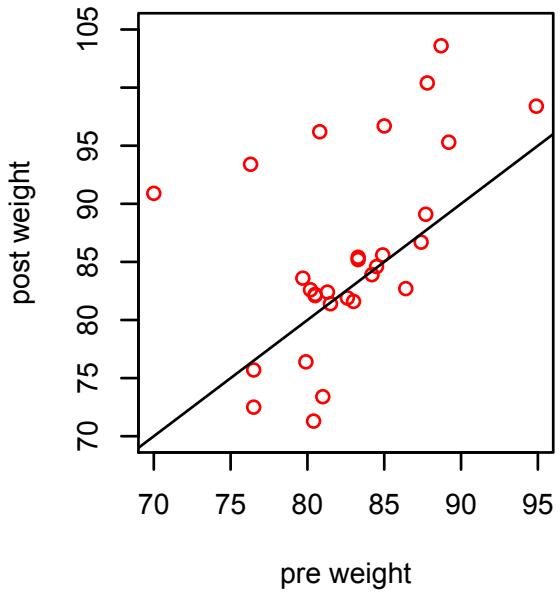
*



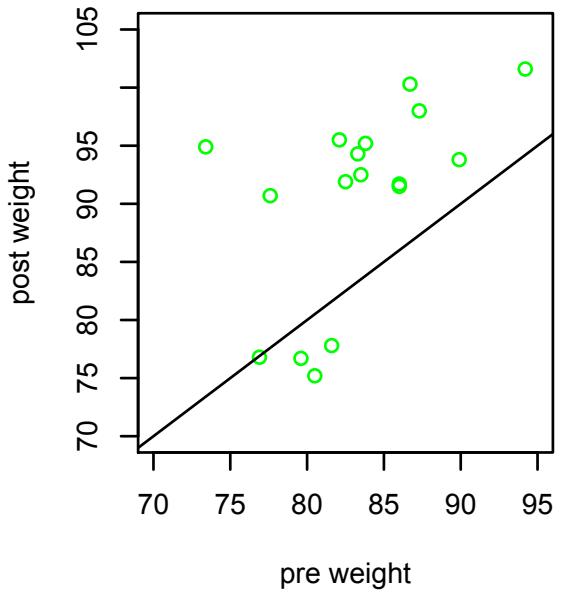
Cont



CBT

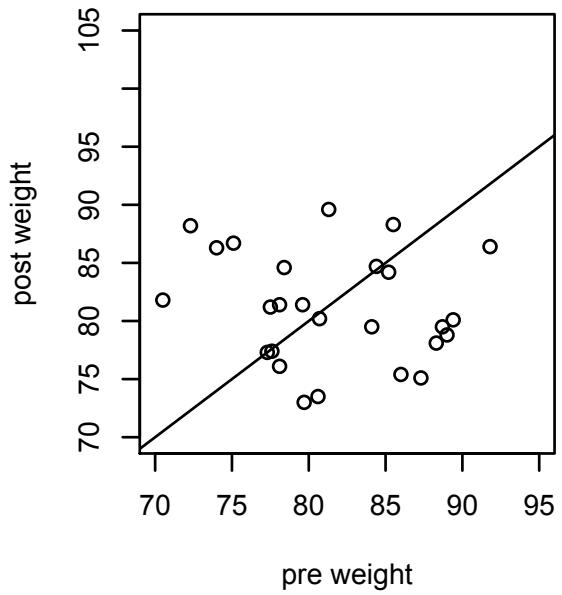


FT

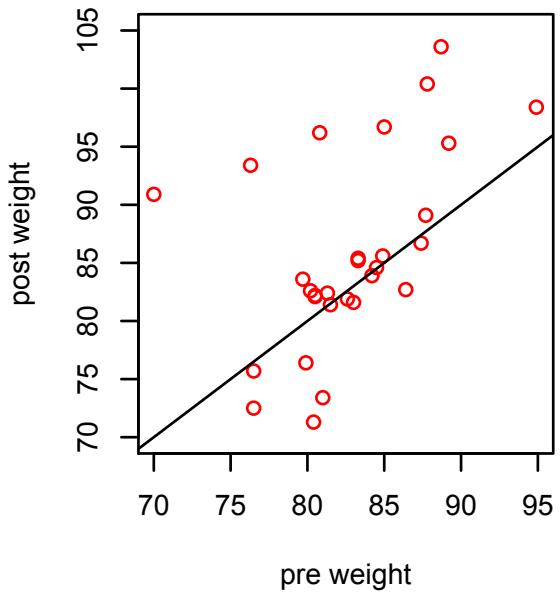


```
> par(mfrow = c(1, 3))
> with(d, plot(Postwt[Treat == "Cont"] ~ Prewt[Treat
  == "Cont"], col = "black", main ="Cont", ylab =
"post weight", xlab = "pre weight", ylim = c(70,
105), xlim = c(70, 95)))
> abline(0, 1)
> with(d, plot(Postwt[Treat == "CBT"] ~ Prewt[Treat
  == "CBT"], col = "red", main ="CBT", ylab = "post
weight", xlab = "pre weight", ylim = c(70, 105),
xlim = c(70, 95)))
> abline(0, 1)
> with(d, plot(Postwt[Treat == "FT"] ~ Prewt[Treat
  == "FT"], col = "green", main ="FT", ylab = "post
weight", xlab = "pre weight", ylim = c(70, 105),
xlim = c(70, 95)))
> abline(0, 1)
```

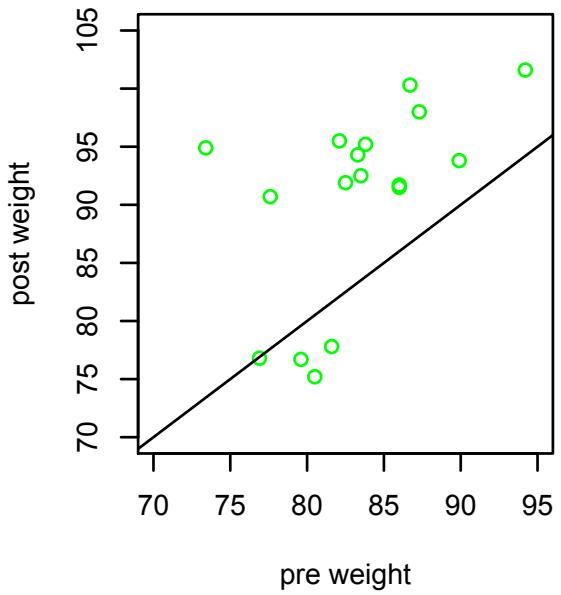
Cont



CBT



FT



```
> with(d, t.test(Postwt[Treat == "Cont"], Prewt[Treat == "Cont"], paired = T))
```

Paired t-test

data: Postwt[Treat == "Cont"] and Prewt[Treat == "Cont"]
t = -0.2872, df = 25, p-value = 0.7763
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-3.676708 2.776708

sample estimates:

mean of the differences

-0.45

```
> with(d, t.test(Postwt[Treat == "CBT"], Prewt[Treat == "CBT"], paired = T))
```

Paired t-test

data: Postwt[Treat == "CBT"] and Prewt[Treat == "CBT"]
t = 2.2156, df = 28, p-value = 0.03502
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.2268902 5.7869029

sample estimates:

mean of the differences

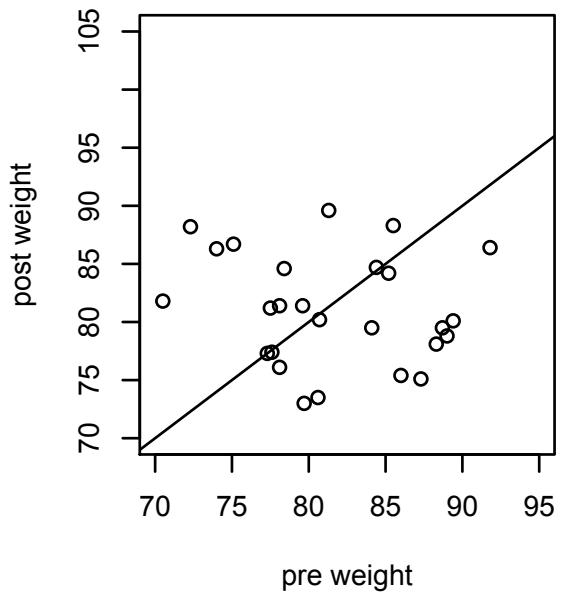
3.006897

```
> with(d, t.test(Postwt[Treat == "FT"], Prewt[Treat == "FT"], paired = T))
```

Paired t-test

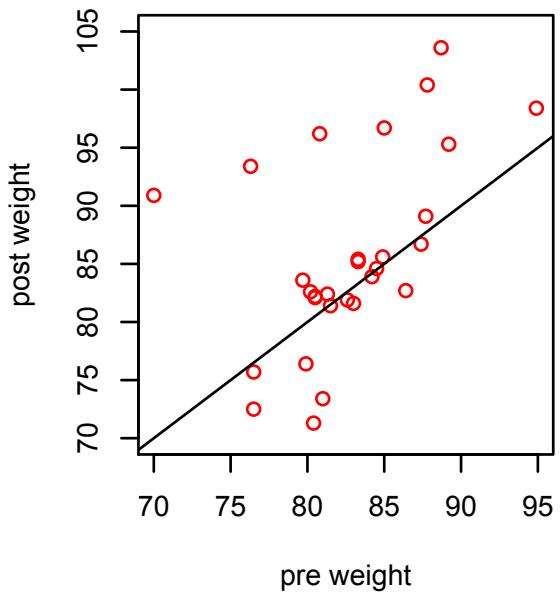
```
data: Postwt[Treat == "FT"] and Prewt[Treat == "FT"]
t = 4.1849, df = 16, p-value = 0.0007003
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 3.58470 10.94471
sample estimates:
mean of the differences
 7.264706
```

Cont



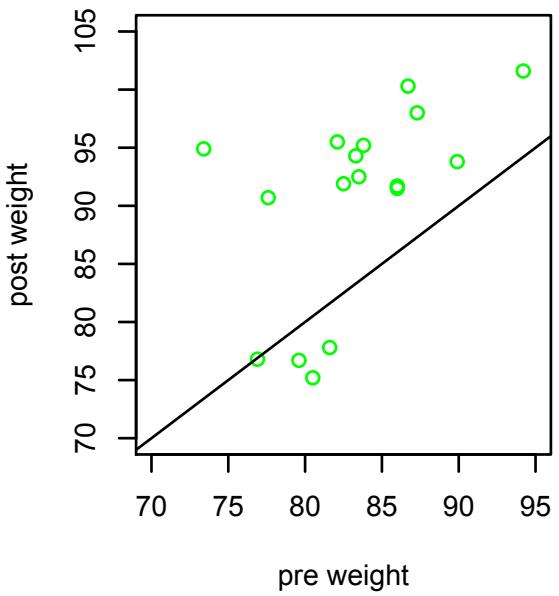
n.s.

CBT



$p < 0.04$

FT



$p < 0.0008$

**il p-value
osservato: il suo
significato e
alcuni errori da
evitare**

alfa = p(errore I)

		stato di cose reale	
		HO	H1
decisioni	rifiuto HO	errore I	corretta
	non rifiuto HO	corretta	errore II

alcuni aspetti negativi

**fraintendimenti sul significato di
alfa:**

viene confuso con $p(H_0)$

**o con il grado di “credibilità
scientifica” del risultato**

altri aspetti negativi

arbitrarietà di alfa:

e se i risultati sono $p < 0.051$?

incentivo ad adottare strategie di analisi scorrette, anche inconsapevolmente

ancora altri aspetti negativi

illusione di obiettività:

**rifiutare HO viene confuso con la
certezza della presenza di un
effetto**

**impone una logica tutto o niente,
tralasciando il ruolo della replica
degli esperimenti**

**la danza dei p-
value**

la danza dei p-value

```
> x <- rnorm(n = 30, mean = 10, sd = 5)
> y <- rnorm(n = 30, mean = 10, sd = 5)
> t.test(x, y)
```

sample estimates:
mean of x mean of y
8.374315 11.918865

t = -3.1387, df = 57.686, p-value = 0.002674

la danza dei p-value

```
> x <- rnorm(n = 30, mean = 10, sd = 5)
> y <- rnorm(n = 30, mean = 10, sd = 5)
> t.test(x, y)
```

sample estimates:
mean of x mean of y
10.67682 10.51481

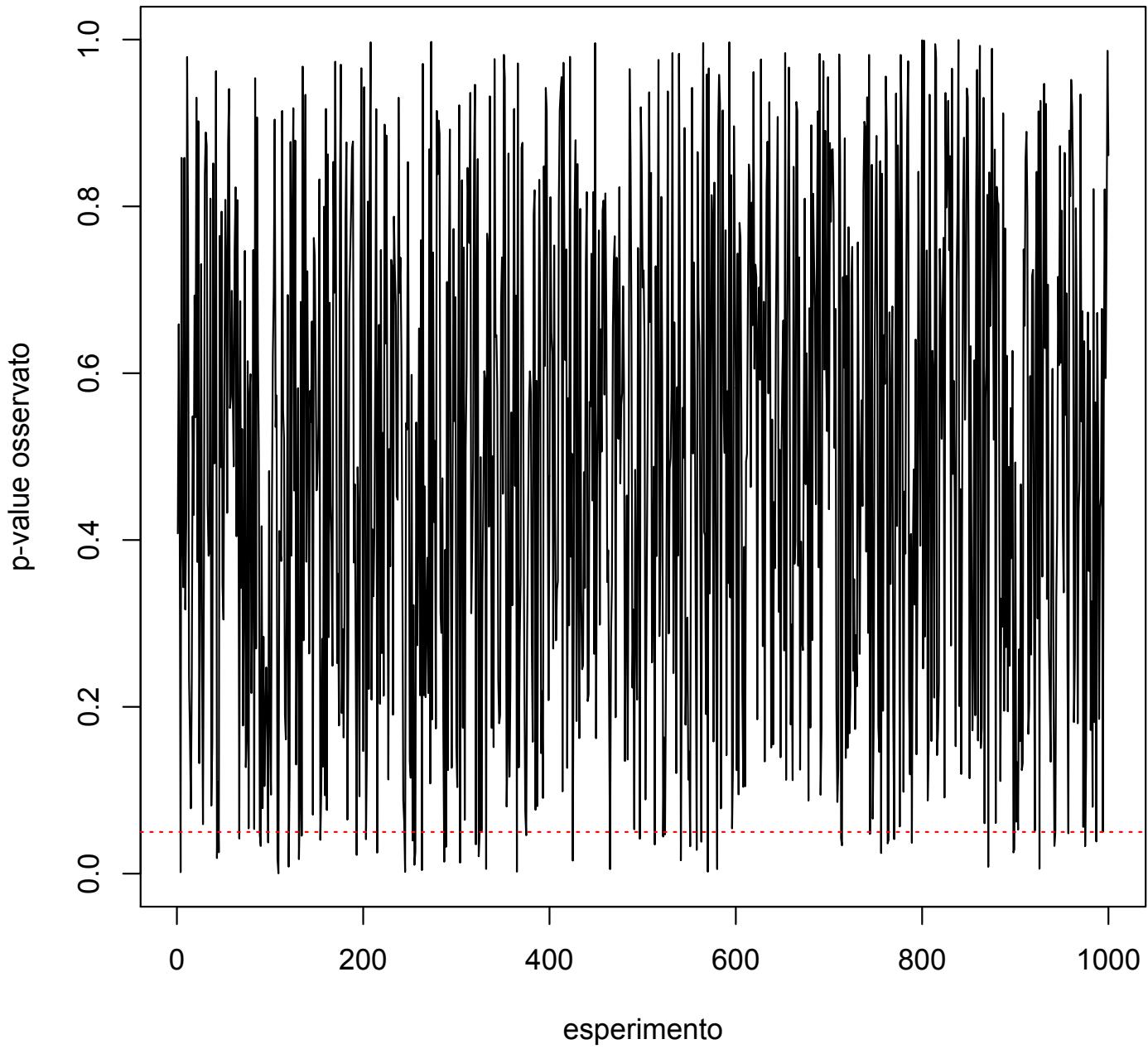
t = 0.11384, df = 56.817, p-value = 0.9098

la danza dei p-value

```
> x <- rnorm(n = 30, mean = 10, sd = 5)
> y <- rnorm(n = 30, mean = 10, sd = 5)
> t.test(x, y)
```

sample estimates:
mean of x mean of y
10.544181 9.425629

t = 0.79922, df = 53.169, p-value = 0.4277



58

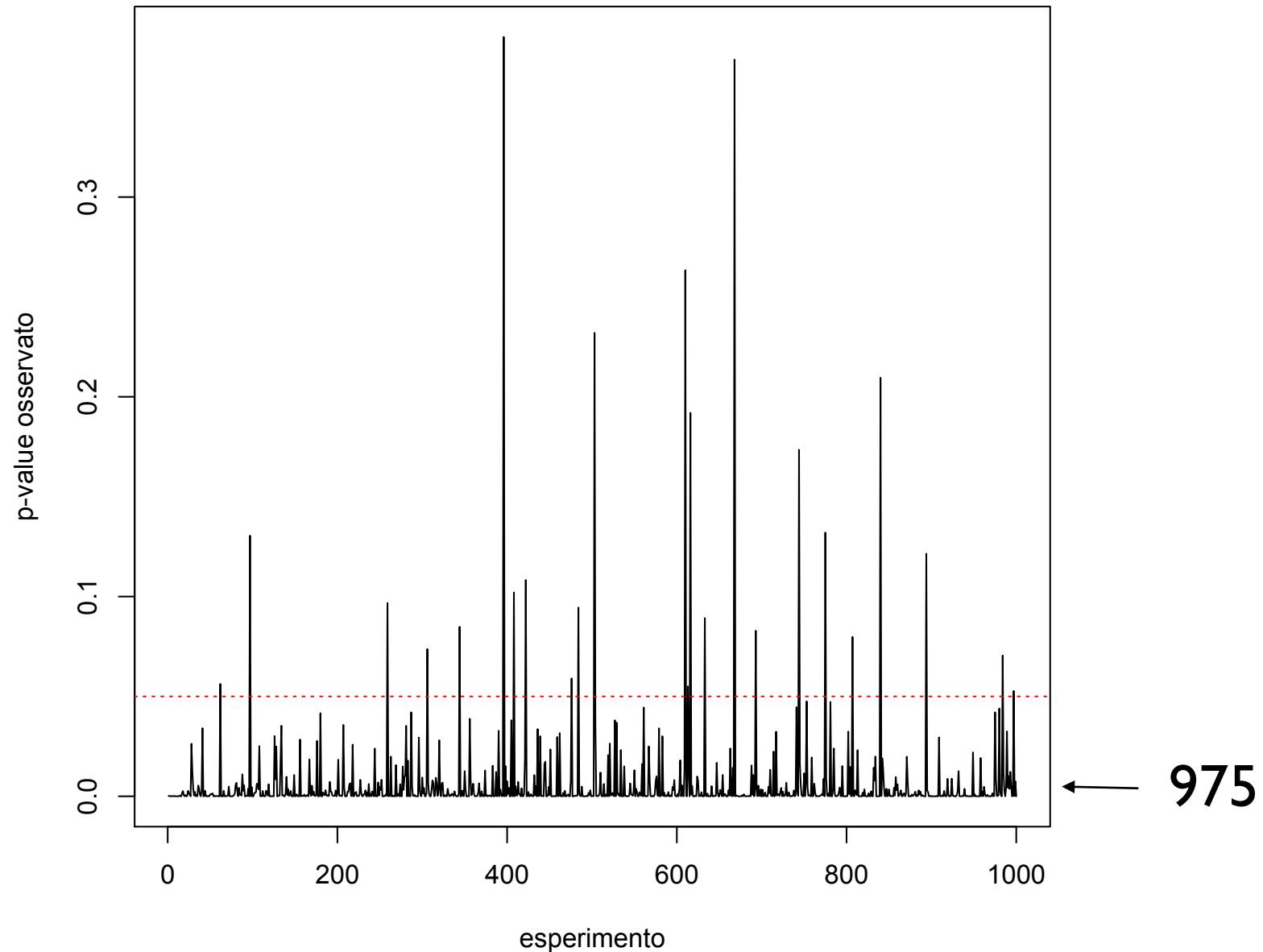
la danza dei p-value 2

```
> x <- rnorm(n = 30, mean = 10, sd = 10)
> y <- rnorm(n = 30, mean = 20, sd = 10)
> t.test(x, y)
```

sample estimates:

mean of x mean of y
9.444616 17.648267

t = -2.6025, df = 54.878, p-value = 0.01188



la danza dei p-value 3

```
> x <- rnorm(n = 30, mean = 10, sd = 20)
> y <- rnorm(n = 30, mean = 20, sd = 20)
> t.test(x, y)
```

sample estimates:

mean of x mean of y
4.69695 15.87700

t = -2.3513, df = 35.782, p-value = 0.02434

