5 – L'INFERENZA

TECNICHE DI ANALISI DI DATI I

Thus, a theory can very well be found to be incorrect if there is a logical error in its deduction or found to be off the mark if a fact is not in consonance with one of its conclusions. ***But the truth of a theory can never be proven.***

Einstein, Collected papers, vol. 7, doc. 28

"Hallmark of good science is that it uses models and "theory", but never believes them"
Wilks

Osservo un fenomeno

Formulo una teoria per spiegarlo

Progetto un esperimento per verificare la teoria

L'esperimento conferma la teoria?

NO

SI

Abbiamo fatto un passo avanti

Sento parlare di un fenomeno

Mi faccio un'idea del fenomeno

Cerco in rete dati che confermino la mia idea

Trovo i dati che cercavo?

NO

I poteri forti li hanno nascosti bene

SI

Hai visto che avevo ragione?

Osservo un fenomeno

Formulo una teoria per spiegarlo

Progetto un esperimento per verificare la teoria

L'esperimento conferma la teoria?

NO

SI

Abbiamo fatto un passo avanti

Sento parlare di un fenomeno

Mi faccio un'idea del fenomeno

Cerco in rete dati che confermino la mia idea

Trovo i dati che cercavo?

NO

I poteri forti li hanno nascosti bene

SI

Hai visto che avevo ragione?

Abbiamo usato grafici e modelli statistici per ***descrivere*** *una distribuzione univariata, e grafici e modelli* ***probabilistici*** *per associare alle modalità della distribuzione le probabilità, per quantificare la prevedibilità di un evento.*

D'ora in avanti, costruiremo modelli statistici per rappresentare ***inferenze****.*

Inferenze e metodo deduttivo

Studiare un'intera popolazione è impossibile o oneroso → individuiamo un **campione rappresentativo** e studiamolo per **estendere le conclusioni dell'indagine alla popolazione.**

Inferenza: una o più affermazioni esplicite su proprietà di un universo più ampio, basate su un insieme di osservazioni molto più ristretto

Le inferenze sono alla base del metodo **induttivo:** in seguito a ripetute e accurate descrizioni di un fenomeno, si rilevano **regolarità**, che contribuiscono a definire **leggi**, cioè asserzioni secondo cui **alcuni fenomeni sono regolarmente associati in popolazione**. Le regolarità non sono perfette (non sono **deterministiche**), a causa di molte possibili **covariate**: **le leggi definiscono** come **(molto o poco) probabili le regolarità** osservate.

Per esempio, registriamo il numero di item rievocato da una lista di parole immediatamente dopo la sua lettura e dopo 10 minuti in cui i soggetti sono stati impegnati in altri compiti cognitivi:

```
immediata<-c(9,8,6,7)
differita<-c(3,2,1,2)
```

	rievocazione	
sogg	X_1 immediata	X_2 differita
S_1	9	3
S_2	8	1
S_3	6	2
S_4	7	2

Verifica empirica

Osservazioni empiriche: a breve termine si ricordano circa 7.5 chunks, dopo qualche minuto circa 2

Modelli statistici

```
mean(immediata)
[1] 7.5
```

```
mean(differita)
[1] 2
```

Dopo qualche minuto di attività distraenti, il **numero di item rievocati diminuisce di circa 5.5 unità**

Modello statistico

```
summary(immediata-differita)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0     5.0     5.5     5.5     6.0     6.0
```

Regolarità

(con **variazioni**)

Teoria

Esiste un magazzino mnestico (**costrutto latente**) in cui si conservano per breve tempo le informazioni (WM), limitato per quantità di informazione e durata nel tempo. Se il materiale è elaborato, passa in un altro magazzino (MLT), illimitato per quantità e durata.

```
t.test(immediata, differita, paired = TRUE)
        Paired t-test
data:  immediata and differita
t = 19.053, df = 3, p-value = 0.0003157
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.581307 6.418693
sample estimates:
mean of the differences
                    5.5
```

Modello statistico

In **popolazione**, le persone ricordano circa 7 chunks per pochi minuti, se non si sforzano di apprenderli. A distanza di tempo, gli item rievocati diminuiscono **significativamente**

Legge

inferenza

campione | **popolazione**

Le leggi descrivono **relazioni tra fenomeni osservati**, mentre le **teorie** introducono **costrutti non direttamente** osservati né **misurati**, per spiegare e generalizzare tali relazioni, nonché per **prevedere** nuove leggi.

I costrutti possono essere **confermati**, **smentiti** o **modificati solo da altri dati empirici**.

Una teoria è **valida** solo se resiste ai tentativi di **falsificarla**: messa alla prova, si può dimostrare falsa, ma **non sarà mai possibile dimostrarla come vera**: ogni teoria è solo **vera fino a prova contraria**. Per falsificarla, viene **operazionalizzata** in una (o più) **ipotesi**.

ipotesi: affermazione che si **può definire come probabilmente vera o probabilmente falsa** a seguito di una **prova empirica**. Segue la forma **se A → allora B**: "se la teoria è vera, allora, in queste condizioni, il fenomeno si presenterà in tal modo".

Operazionalizzare: un costrutto è **definito attraverso le operazioni con cui è misurato**. Lo stress è definibile tramite/come livello di cortisolo salivare rilevato prima di una gara, la fame tramite/come deprivazione totale del cibo per 24h. Se stress e fame sono operazionalizzati diversamente (conduttanza cutanea, somministrazione di cibo appena sufficiente a mantenere l'80% del peso corporeo), produrremo risultati leggermente diversi.

L'inferenza integra la descrizione di un fenomeno con la sua probabilità: per rilevare regolarità, è **essenziale descrivere il fenomeno il più correttamente possibile**. Poiché sono soggette a diverse fonti di errore e a variabilità "naturale", è **necessario definirne la probabilità di manifestarsi sotto determinate condizioni, oggetto di ipotesi**.

D'altronde, se un **fenomeno si manifesta in maniera diversa da quella attesa** in base all'ipotesi, bisogna chiedersi **quale sia la spiegazione più semplice** per la differenza riscontrata.

Dice Guglielmo di Occam: "Entia non sunt multiplicanda preater necessitatem", o "Pluralitas non est ponendum sine necessitatem": non chiamate in causa più spiegazioni del necessario, spiegazioni più semplici devono essere preferite a spiegazioni più complesse di uno stesso fenomeno

La **spiegazione più semplice** è che **la differenza sia compatibile con l'effetto del caso**, ovvero **con le fluttuazioni campionarie**: quindi, concluderemo che **possa non esserci alcuna reale differenza / alcuna reale relazione.**

```
t.test(immediata, differita, paired = TRUE)
        Paired t-test
data:  immediata and differita
t = 19.053, df = 3, p-value = 0.0003157
alternative hypothesis: true difference in m
```

Questo processo di falsificazione delle ipotesi è stato messo a punto da Fisher, come vedremo

P- value approach

L'altra "faccia" dell'inferenza è la stima intervallare: un'indagine su un campione può avere l'obiettivo di **stimare il valore sconosciuto di un parametro nella popolazione** (una media, una varianza, una correlazione tra variabili, una differenza tra medie, ecc.), affiancando alla **misura puntuale** nel campione la corrispettiva **stima intervallare** in popolazione:

Intervallo di valori all'interno dei quali il parametro sconosciuto deve trovarsi, **con un grado di verosimiglianza elevato e predefinito**: **intervallo di fiducia** (**confidence interval**, **CI**).

L'**ampiezza dell'intervallo** è una misura preziosa della **precisione della stima** e **dell'utilità** della ricerca stessa.

L'*affermazione che la percentuale di promossi all'esame di Tecniche di analisi di dati I sta, con il 99% di probabilità, tra lo 0% e il 100%, non vi dovrebbe aiutare molto rispetto al tipo di atteggiamento che dovreste tenere riguardo al fare gli esercizi...*

Distribuzione campionaria e campionamento casuale (bernoulliano)

Qualche definizione…

Popolazione o **universo**: grande (o indefinito) insieme di eventi elementari; la denominazione **spazio campionario** sarebbe più descrittiva, ma è meno usuale.

Distribuzione della popolazione: distribuzione dei valori delle osservazioni possibili nello spazio campionario: le caratteristiche della popolazione (**parametri**: media, varianza, ecc.) sono conosciute raramente, se non mai.

Statistiche: valori campionari che consentono di ottenere le **stime** dei parametri corrispondenti.

Distribuzione campionaria (**sampling distribution**): distribuzione di valori di una statistica rilevata in campioni di numerosità N, casualmente estratti da una popolazione

Nell'applicazione della statistica alla ricerca, **sono le proprietà di queste distribuzioni campionarie che guidano le inferenze sulle proprietà delle popolazioni.**

Perché le inferenze siano valide, i campioni devono essere **rappresentativi.**

Campionamento **casuale o random o bernoulliano**, in cui **ogni unità della popolazione** ha la **medesima probabilità di entrare a far parte del campione**.

La **stima** di un parametro è un'approssimazione statistica ai risultati della ricerca sull'intera popolazione: è una **"verità relativamente ottimale"**, affetta da **un errore accettabile** e **proporzionato** al costo. Il campionamento random garantisce che la stima sia **preservata** da un **errore sistematico** di campionamento, e che l'errore in essa riscontrabile sia solo **casuale**.

L'errore **casuale** agisce in modo impredicibile: sottostima, sovrastima, varia da prova, e i suoi effetti **tendono a compensarsi** quando le rilevazioni sono ripetute un gran numero di volte.

Se una misura è soggetta a molte piccole sorgenti di errori casuali e a trascurabili errori sistematici, allora i **valori misurati saranno distribuiti su una curva a campana → operando infinite misurazioni, la media degli errori tende a 0**.

Non sarà più sufficiente che il modello statistico abbia un buon fit nel campione:

dovremo stabilire se ha un

buon fit rispetto alla popolazione

da cui il campione è stato estratto.

Come detto, la **sampling distribution** è la distribuzione delle statistiche relative ai campioni casualmente estratti: avremo la **distribuzione campionaria delle medie** dei campioni (***DCM***), la DC delle varianze dei campioni, la DC delle curtosi dei campioni...

Se i campioni sono stati realmente estratti maniera casuale, **ciascuna delle statistiche varierà, con una quota di errore casuale, all'interno della distribuzione**. Per **qualsiasi tipo** di distribuzione campionaria è possibile calcolare descrittori.

Tra le varie DC, **ci concentriamo su quella della media**: **distribuzione campionaria delle medie** o ***DCM***.

Usiamo R per simulare una DCM: con `rnorm` creiamo **mille distribuzioni** (**campioni**), ciascuna composta da **100 numeri** casuali ($N = 100$, campionamento **random**) da una **popolazione con $\mu = 50$ e $\sigma = 5$**. Le **mille medie costituiscono la *DCM***.

Potremmo costruire un campione per volta, fino a 1000:

```
campione1<-rnorm(N=100, mean=50, sd=5)
campione2<-rnorm(N=100, mean=50, sd=5)
campione3<-rnorm(N=100, mean=50, sd=5)
```

… Ma conviene usare `replicate(n= numero di repliche, expr= espressione da replicare)`, con cui creiamo la matrice `campioni`: 1000 colonne (1000 campioni) di 100 righe ciascuno (100 soggetti). Ogni colonna rappresenta una distribuzione normale di 100 casi, casualmente estratta (`rnorm`) da una popolazione normale con μ=50 e σ=5.

```
campioni <-matrix(replicate(n=1000, expr=rnorm(n=100, mean=50,sd=5)), nrow=100, ncol=1000)
```

Data: campioni

	v1	v2	v3	v4	v5
1	41.69385	46.23529	49.02676	53.47184	42.37971
2	39.23091	50.93346	47.63575	36.75766	42.78483
3	48.34363	50.24008	44.17493	48.73709	47.96540
4	37.45700	45.61407	56.39448	51.52599	48.31365
5	55.41832	43.98243	40.01560	49.11757	46.36983
6	53.53126	52.68814	57.30340	46.72163	55.56680
7	61.68309	48.94113	42.42669	51.06409	54.65873
8	53.48479	46.48427	49.16791	59.33800	47.78587

v997	v998	v999	v1000
53.60980	55.04090	51.83836	47.54058
39.53529	56.24085	53.22682	42.80902
51.68351	47.10793	50.26464	38.96364
49.91958	50.96100	48.00486	60.27773
45.60280	47.48565	49.19130	51.33537
58.74177	53.71841	48.01574	50.54100
56.81160	48.59389	59.83184	47.87877
53.22151	58.48470	55.03802	49.31685

Calcoliamo le medie delle 1000 colonne - campioni che saranno gli elementi dell'oggetto **DCM** con **`apply(X=matrice/dataframe, MARGIN=margine, FUN=funzione)`**, che ripete la funzione in `FUN=` applicandola alle righe (`MARGIN=1`) o alle colonne (`MARGIN=2`) di matrici o dataframe (`X`). Noi dobbiamo calcolare la media (`FUN=mean`) delle osservazioni in ognuna delle 1000 colonne (`MARGIN=2`) contenute nella matrice `campioni` (`X=campioni`):

```
DCM <- apply(X = campioni, MARGIN = 2, FUN = mean)
```

```
length(DCM)
[1] 1000
```

```
head(DCM); tail(DCM)
[1] 49.90577 50.10111 50.28119 49.63651 49.88786 50.25142
[1] 49.83510 50.34029 50.18313 49.94768 50.49777 49.55514
```

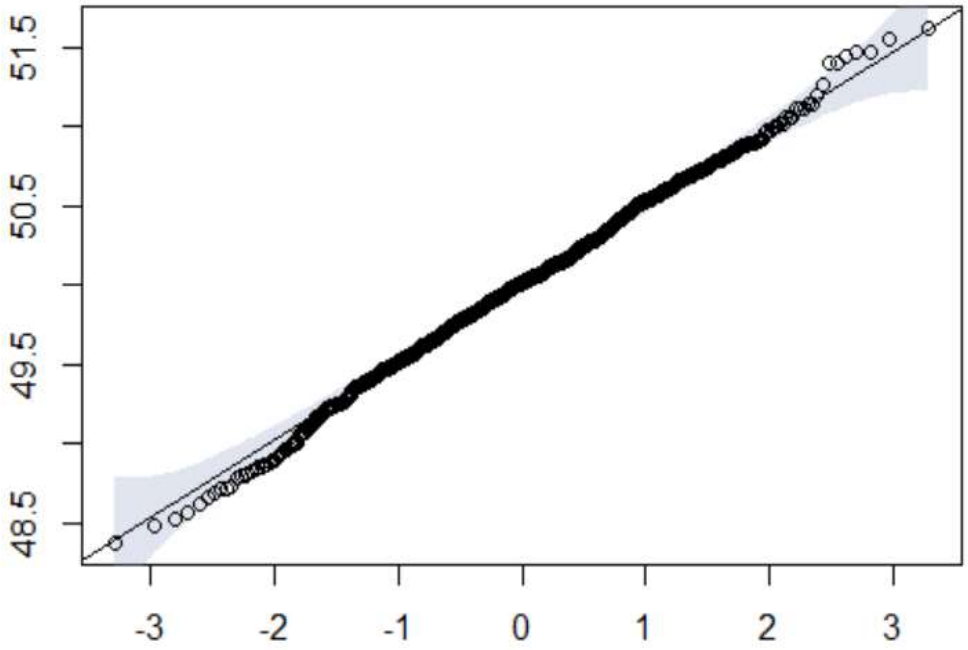
Le medie oscillano attorno a $\mu = 50$, anche se nessuna è $= 50$:

```
which(DCM==50.0)
integer(0)
```

Approssimandola, la media della DCM è pari a quella attesa : $\boldsymbol{\mu = \mu_{\bar{x}}}$

```
mean(DCM)
[1] 49.96554
```

La forma della DCM è assai simile alla forma normale della popolazione da cui abbiamo estratto i campioni:



*se provate a rifare la procedura con R, otterrete risultati simili a questi, ma leggermente **diversi**, dato che l'estrazione dei campioni è random!*

La deviazione standard di $\boldsymbol{DCM}$ si chiama **errore standard** $\boldsymbol{SE}$: $\boldsymbol{\sigma = SE_{\bar{x}}}$

Mentre s^2 e s descrivono la variabilità attorno alla media nel campione, lo $\boldsymbol{SE}$ $(\boldsymbol{s_{DCM}})$ **descrive la variabilità tra le statistiche calcolate** dei differenti campioni: è un **indicatore della variabilità del fenomeno indagato in popolazione**.

$$\boldsymbol{\mu_{\bar{x}} = \mu} \qquad \boldsymbol{SE_{\bar{x}} = \sigma}$$

Poiché non conosciamo mai s^2 o s della popolazione, ci accontentiamo della **stima** dello SE:

$$\boldsymbol{\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}}$$

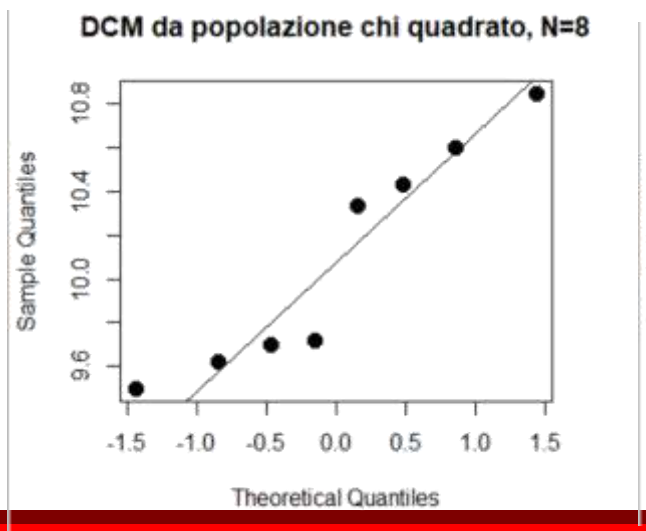
Si potrebbe supporre che $\boldsymbol{DCM}$ tenda alla distribuzione normale perché la popolazione da cui abbiamo estratto i campioni è normale, ma non è così:

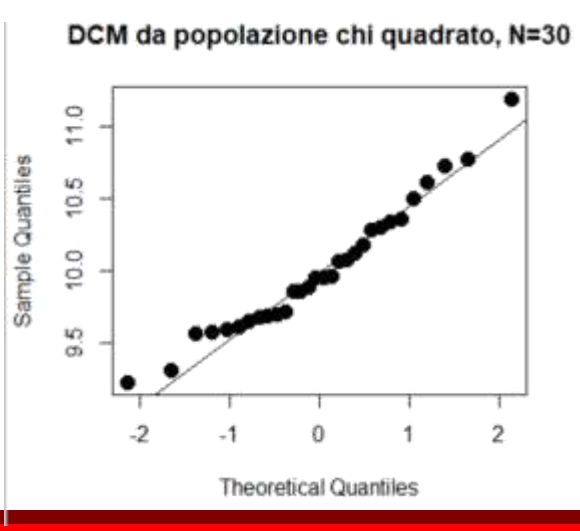
Indipendentemente da quale sia la **forma della popolazione** da cui li estraiamo, più il **numero di campioni N aumenta**, più la forma della DCM tende alla **normale**.

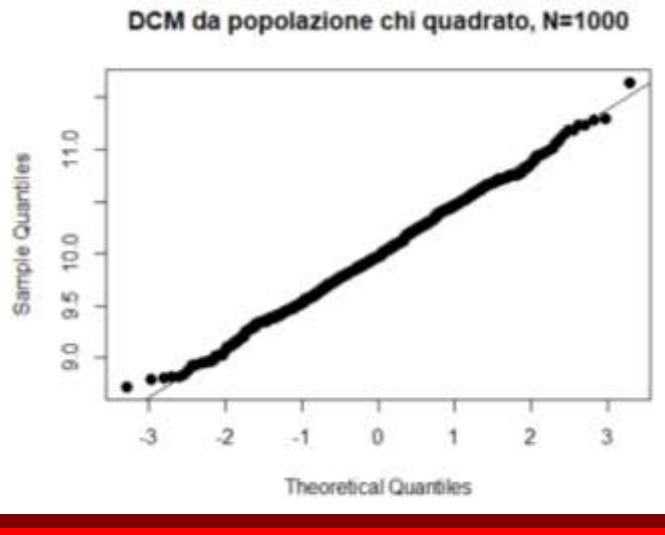
Teorema centrale del limite

Per N **variabili aleatorie indipendenti**, con $\bar{x}_{DCM} = \mu$ e $SE = \frac{\sigma}{\sqrt{N}}$, indipendentemente dalla forma delle singole distribuzioni, la successione delle variabili aleatorie standardizzate tende ad avere distribuzione normale

Gli script con cui sono state create le tre DCM da popolazione con forma χ^2 sono nella dispensa







Ogni campione è composto da 100 osservazioni

Inferenza e stime intervallari: l'intervallo di fiducia

(confidence interval, CI)

Tutti gli approcci alle **stime intervallari** (Fisher: *fiducial intervals*; Bayes: *credible intervals*…) cercano di **stimare un parametro in popolazione tenendo conto dell'incertezza della misura**, cioè **della variabilità campionaria**: cercano un **range di valori per il parametro**, invece di un singolo valore. Tra le stime intervallari, i ***confidence intervals*** **o** CI, basati sulle stime campionarie, sono raccomandati come base per la migliore statistica inferenziale

Il CI di un parametro Θ (**theta**: media, varianza, correlazione...), è un **intervallo** generato da una **procedura** (***confidence procedure***, CP) che, procedendo a **ripetuti campionamenti** (*sampling*), ha una **probabilità prefissata** di contenere il parametro Θ.

"Un CI **al** $X\%$ **di probabilità** per un parametro Θ è un intervallo, delimitato da un limite inferiore (**LL**) e un limite superiore (**UL**), generato da una procedura tale per cui, **in ripetuti campionamenti**, il CI ha una probabilità pari al $X\%$ di contenere il vero valore di Θ" (Neyman).

$X\%$ **di probabilità del** CI: **confidence level**, arbitrario (per pura tradizione, **.95** o **.99**), esprime l'incertezza associata al campionamento.

confidence procedure - CP

qualsiasi procedura che generi CI che comprendono Θ in una data percentuale X di campionamenti ripetuti

Processo casuale

confidence interval - CI

UNO dei CI generati dalla *confidence procedure*

Osservato e fisso

Per esempio, **ripetiamo** la *confidence procedure*, impostata in modo che ogni CI generato abbia il **95% di probabilità** di contenere Θ, per **100 volte**: si **generano 100 CI, 95 (circa) dei quali contengono Θ e 5 (circa) non contengono Θ** :

Ogni CI contiene O non contiene Θ, dicotomicamente

La **potenza** (*power*) delle CP è la frequenza con cui sono esclusi i valori falsi di Θ ($\boldsymbol{\Theta}$') : diverse CP escludono questi $\boldsymbol{\Theta}$' con tassi differenti → se la CP_A esclude Θ' più spesso, in media, della procedura CP_B, allora CP_A è meglio di CP_B, per **quel parametro**.

La CP **più usuale** è basata sulla **conoscenza descrittiva di un dato campionario**.

$$\boldsymbol{CI = statistica\ campionaria \pm SE \times {}^{\alpha}/_{2}}$$

*Non è l'unica CP: tra le altre, la **likelihood theory**, che vedremo in TAD 2, e il ricampionamento **bootstrapping**.*

Il CI è **centrato** sulla statistica campionaria; la sua **ampiezza** ($2w$: **precisione della stima**) è determinata **da $SE \times$ metà *confidence level* $\alpha/2$**. Si estende per una **distanza w** (***width***: **margine di errore**), delimitata dal limite inferiore LL e dal limite superiore UL.

$$\boldsymbol{CI = LL{}^{\alpha}/_{2} < parametro < UL{}^{\alpha}/_{2}}$$

${}^{\alpha}/_{2}$: **quantile z corrispondente alla probabilità cumulata prefissata**, divisa a **metà**, della distribuzione **normale standardizzata**; in realtà si usano perlopiù i **quantili t, per $df = N - 1$**.

Attenzione: lavoreremo perlopiù con CI basati su questa formula, **ma non sempre.** Per esempio, il CI di una **proporzione** secondo Wald è dato da:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

proporzione dell'evento atteso

Diversi confidence levels creano diverse ampiezze → al crescere di α, aumenta la dimensione dei quantili t (o z) in valore assoluto.

$$\bar{x} = 20.5, SE = 2.2$$

90% CI

```
(t_05<-qt(p = .05,df = 50-1,lower.tail = FALSE))
[1] 1.676551
CI_90<-c(20.5-(2.2*t_05), 20.5+(2.2*t_05))
CI_90
[1] 16.81159 24.18841
CI_90[2]-CI_90[1]
[1] 7.376824
```

95% CI

```
(t_025<-qt(p = .025,df = 50-1,lower.tail = FALSE))
[1] 2.009575
CI_95<-c(20.5-(2.2*t_025), 20.5+(2.2*t_025))
CI_95
[1] 16.07894 24.92106
CI_95[2]-CI_95[1]
[1] 8.84213
```

99% CI

```
(t_.005<-qt(p = .005,df = 50-1,lower.tail = FALSE))
[1] 2.679952
CI_99<-c(20.5-(2.2*t_005), 20.5+(2.2*t_005))
CI_99
[1] 14.60411 26.39589
CI_99[2]-CI_99[1]
[1] 11.79179
```

*il CI al 99% dà una **sicurezza maggiore, ma una precisione minore**, di trovare θ in più campionamenti*

*Tratteremo perlopiù CI **simmetrici** attorno alla statistica ($w_{LL} = w_{UL}$), ma **non è una regola generale**.*

***Vedremo CI asimmetrici** in proporzioni (test della binomiale), soprattutto se prossime a 0 o 1, e coefficienti di correlazione di Pearson.*

Vediamo un esempio di CP per CI_{media}, costruendo **100 repliche** di **un'estrazione random (`rnorm`)** di **100 casi** da una popolazione **normalmente distribuita**, con $\mu = 10$ e $SE = 1.5$.

Con `replicate(n, expr)` creiamo una matrice di 100 colonne e 100 righe.

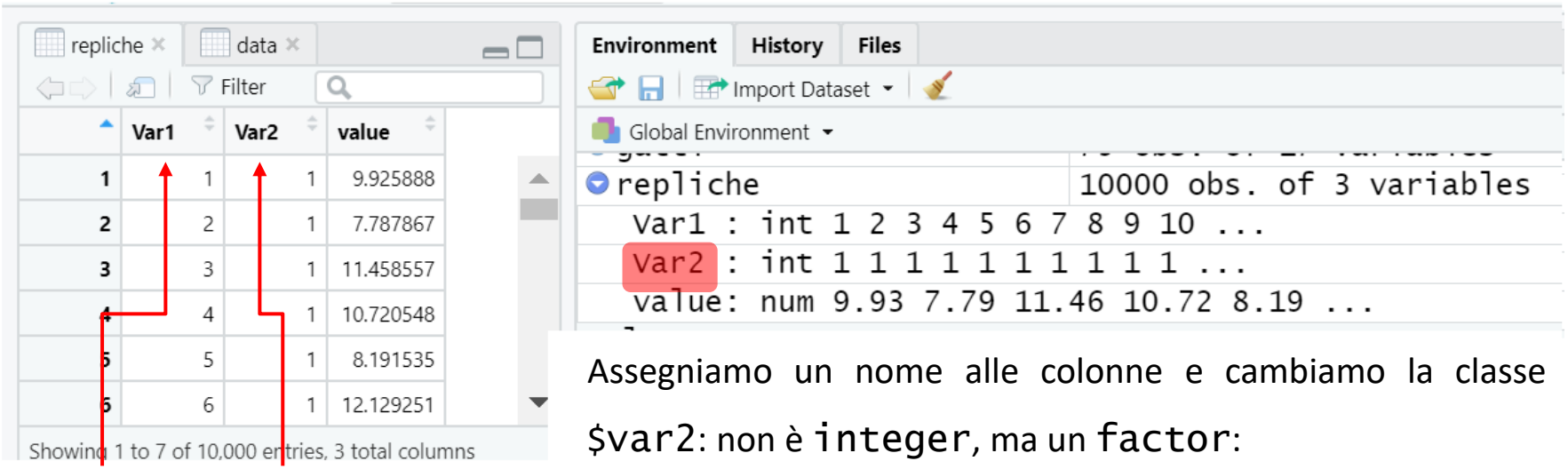
```
data<-matrix(replicate(n = 100, expr = rnorm(n = 100, mean = 10, sd = 1.5)),
nrow = 100, ncol = 100)
```

	V1	V2	V3	V4	V5
1	11.604410	9.587470	12.292145	11.042754	11.123950
2	10.280864	10.354469	7.213487	9.275642	10.136905
3	10.666558	10.336166	9.382910	8.446710	12.724747
4	9.877420	10.459004	11.110415	8.745525	12.235764
5	11.686325	13.746141	9.259746	8.697233	11.900826

*La media di ogni colonna è la **media campionaria** di una replica dell'estrazione.*

Per plottare le 100 medie con il rispettivo $95\%CI$ serve **una** colonna che contenga tutte osservazioni, contraddistinte dal numero della replica in cui sono ricavate → cambiamo il formato della matrice da ***wide*** a ***long*** (ricordate?) con `melt(data= dataframe/matrice da trasformare, measures.var= "colonne da trasporre")` di `reshape2`:

```
repliche<-melt(data, measure.vars = c("V1":"V100"))
```



Soggetto *Replica*

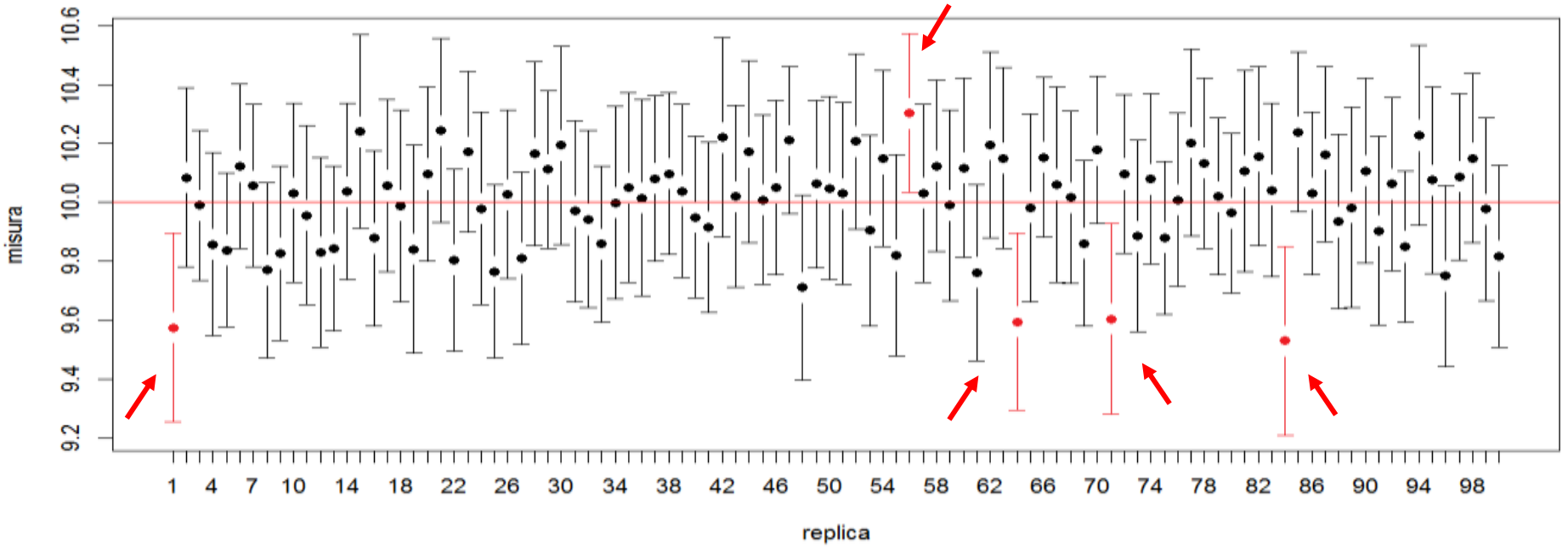
Assegniamo un nome alle colonne e cambiamo la classe di $var2: non è `integer`, ma un `factor`:

```
names(repliche)<-c("caso","replica","misura")
repliche$replica<-as.factor(repliche$replica)
```

Ora plottiamo le 100 medie con *CI*: `plotmeans(misura~fattore, bars= TRUE)` di `gplots`.

```
plotmeans(repliche$misura~repliche$replica, connect = FALSE, bars = TRUE, xlab = "replica", ylab
 = "misura", barcol = "black", use.t = TRUE, pch=19, n.label = FALSE)
 abline(h=10, col="red")
```

Di default, a= .95 (p= .95) è stimato con i quantili t, che possiamo sostituire con z (use.t= FALSE). Non connettiamo le medie (connect= FALSE), raffigurate da pallini neri (col= "black", pch=19); tracciamo in nero le barre (di default, barcol="blue"). Togliamo il numero della replica dal grafico (n.label=FALSE). Poi, tracciamo la linea corrispondente a $\Theta = 50$ *con abline(h= 50).*



Le medie campionarie "ballano" attorno a μ ("CI dance"; Cumming, 2008)

Cinque CI non comprendono μ : se stimassimo μ basandoci sui questi esperimenti, **sbaglieremmo** → fare un esperimento **equivale a incappare in uno solo tra i CI che originano da una sequenza di potenziali CI**: solo una data percentuale conterrà μ, quindi possiamo dire che c'è una probabilità pari a quella prescelta che il CI sperimentale includa μ.

Provate a fare altre simulazioni: naturalmente, dato che le estrazioni sono casuali, i risultati saranno leggermente diversi per ciascuno di voi:

- *replicate le 100 simulazioni con N=100, μ= 10, SE= 1.5: cosa osservate?*
- *a parità di tutti gli altri parametri, cambiate il confidence level con a .99: cosa osservate?*
- *per a= .95 e a* ***parità di μ e SE****,* ***diminuite N a 50 e poi a 30: cosa osservate?***

Nota bene: per ottenere la **stessa sequenza di numeri casuali** (pseudocasuali!) in diverse estrazioni random, si può fissare il "seme" di partenza, con `set.seed(numero casuale)`.

Per esempio, **senza** `set.seed` avremo:

```
round(rnorm(n = 10, mean = 0,sd = 1),3)
 [1]  1.224  0.360  0.401  0.111 -0.556  1.787  0.498 -1.967  0.701
round(rnorm(n = 10, mean = 0,sd = 1),3)
 [1] -1.068 -0.218 -1.026 -0.729 -0.625 -1.687  0.838  0.153 -1.138
```

eccetera, mentre fissando il "seme":

```
set.seed(seed = 123)
round(rnorm(n = 10, mean = 0,sd = 1),3)
 [1] -0.560 -0.230  1.559  0.071  0.129  1.715  0.461 -1.265 -0.687
set.seed(seed = 123)
round(rnorm(n = 10, mean = 0,sd = 1),3)
 [1] -0.560 -0.230  1.559  0.071  0.129  1.715  0.461 -1.265 -0.687
```

Nelle prossime slide useremo ancora il dataframe **adolescenti**, che abbiamo già descritto nelle trasformazioni non lineari: scaricatelo da Elly, se non l'avete ancora fatto, e rinominatelo come **ad**.

Ricordate: è un campione molto **numeroso** (oltre 1000 casi), ma ci sono anche molti NA.

Questa volta, ci concentreremo sul **temperamento** degli adolescenti.

```
length(ad$soggetti)
[1] 1274
summary(ad$eta);sd(ad$eta)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   16.00   17.00   17.24   18.00   22.00
[1] 1.014791
```

```
round(prop.table(table(ad$genere))*100,1)
   F    M
64.7 35.3
table(ad$istituto)
BOC BOD GIP   L  MA  ME ROM RON  SV  TO
 52  91 120  48 138 132 118  88 124 363
```

*Qui esploreremo tre scale temperamentali del Temperament and Character Inventory (**TCI**, Cloninger, 1998): **Novelty Seeking** – **NA** (necessità di alti livelli di stimolazione,, inclinazione all'impulsività), **Harm Avoidance** – **HA** (cautela, apprensività e sensibilità alle critiche ed alle punizioni), **Reward Dependence** – **RD** (tendenza a rispondere intensamente ai segnali di approvazione sociale e affettivi).*

Il manuale del test fornisce i **punteggi normativi di ragazzi di pari età**:

NS: μ= **20.2 ± 6.6**

HA: μ=**14.9 ± 7.7**

RD: μ=**17.4 ± 3.9**.

Come si pongono questi ragazzi rispetto alla popolazione di pari età? Sono un **modello attendibile della personalità degli adolescenti**? Per rispondere, stimiamo la media campionaria, tracciamone il $95\%CI$ e confrontiamola con la media attesa.

Esercitiamoci nel calcolo passo passo con NS; per le altre due, ricorreremo alle funzioni.

Per comodità, togliamo gli NA: `ado_NS<-ad$NS_tot[is.na(ad$NS_tot)==FALSE]`

```
length(ado_NS)
[1] 1269
```

```
mean(ado_NS)
[1] 17.46493
```

```
sd(ado_NS)
[1] 4.679953
```

*Ci serve lo **SE**: calcoliamolo dalla sd o con MeanSE di DescTools*

```
(SE_NS<- sd(ado_NS)/sqrt(length(ado_NS)))
[1] 0.1313744
```

```
MeanSE(ado_NS)
[1] 0.1313744
```

*Ora il **confidence level**: scegliamo il $\mathbf{95\%}$ e usiamo i quantili $\boldsymbol{t}$, per $\boldsymbol{df = N - 1}$*

```
(t_95<-qt(p = 0.025,df = 1269-1,lower.tail = FALSE))
[1] 1.961837
```

Calcoliamo lower limit e upper limit:

```
LL<-mean(ado_NS)-(SE_NS*t_95)
```

```
UL<-mean(ado_NS)+(SE_NS*t_95)
```

Visualizziamo LL, media campionaria e UL, cioè il CI:

```
round(c(LL, mean(ado_NS), UL),2)
[1] 17.21 17.46 17.72
```

Nel campione, $\bar{x}_{NS} = 17.46 \pm 4.7$; il **parametro** NS atteso nella popolazione da cui abbiamo estratto questi ragazzi è compreso, con una verosimiglianza del 95%, **tra 17.21 e 17.72**.

L'ampiezza *w_2 del CI è ristretta (circa mezzo punto): la precisione della stima è piuttosto buona*

```
(w2<-(mean(ado_NS)-LL)*2)
[1] 0.5154702
```

Non ci sono funzioni di base per i CI, ma potremo usare `DescTools`; già conosciamo `Desc(distribuzione numeric/integer)`, che dà anche il CI (usa quantili t).

Più mirata è `MeanCI(distribuzione)`, che dà CI attorno alla media (default `conf.level=.95`): si può cambiare il metodo di calcolo (default `method= "classic"`, ma anche il *bootstrapping*: `method= "boot"`), usare una media *trimmed* (`trim= proporzione da troncare`), usare lo SE della popolazione, se noto, invece della sua stima (`sd= valore di σ`; saranno usati i quantili z); con NA, si aggiunge `na.rm= TRUE`.

```
MeanCI(ado_NS)
    mean   lwr.ci   upr.ci
17.46493 17.20720 17.72267
```

```
MeanCI(a$NS_tot, na.rm=TRUE)
    mean   lwr.ci   upr.ci
17.46493 17.20720 17.72267
```

Il manuale del test dà la sd della popolazione normativa: sd=**6.6**. Usatela per calcolare il 95%CI con la distribuzione di probabilità normale standardizzata

Possiamo tornare alla domanda di partenza: questo campione di adolescenti è **rappresentativo** della popolazione di coetanei, rispetto al parametro Novelty Seeking?

Se nel CI (per una verosimiglianza di almeno .95) è compreso il parametro Θ, il campione probabilmente è estratto dalla popolazione attesa e la sua $\bar{x}$ è solo una **fluttuazione casuale di μ**

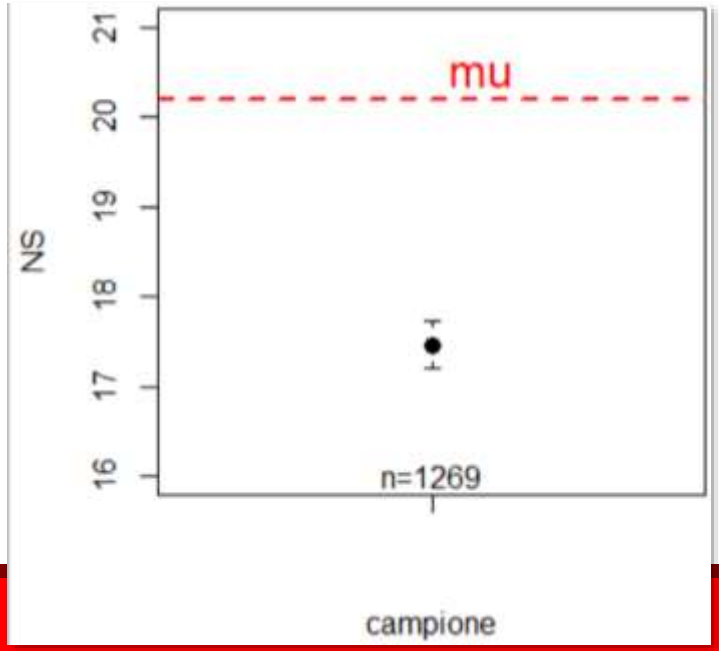
Se nel CI (per una verosimiglianza di almeno .95) non è compreso il parametro Θ, il campione probabilmente è estratto da una diversa popolazione.

Verifichiamo la **significatività della differenza tra un campione e una popolazione**

Il manuale del TCI dice che $\mu_{NS} = 20.2$*: poiché il* $95\%CI = 17.21; 17.72$ ***non comprende*** *il valore atteso, i nostri adolescenti* ***non sembrano appartenere alla popolazione normativa****, ma a una diversa popolazione, meno amante del rischio.*

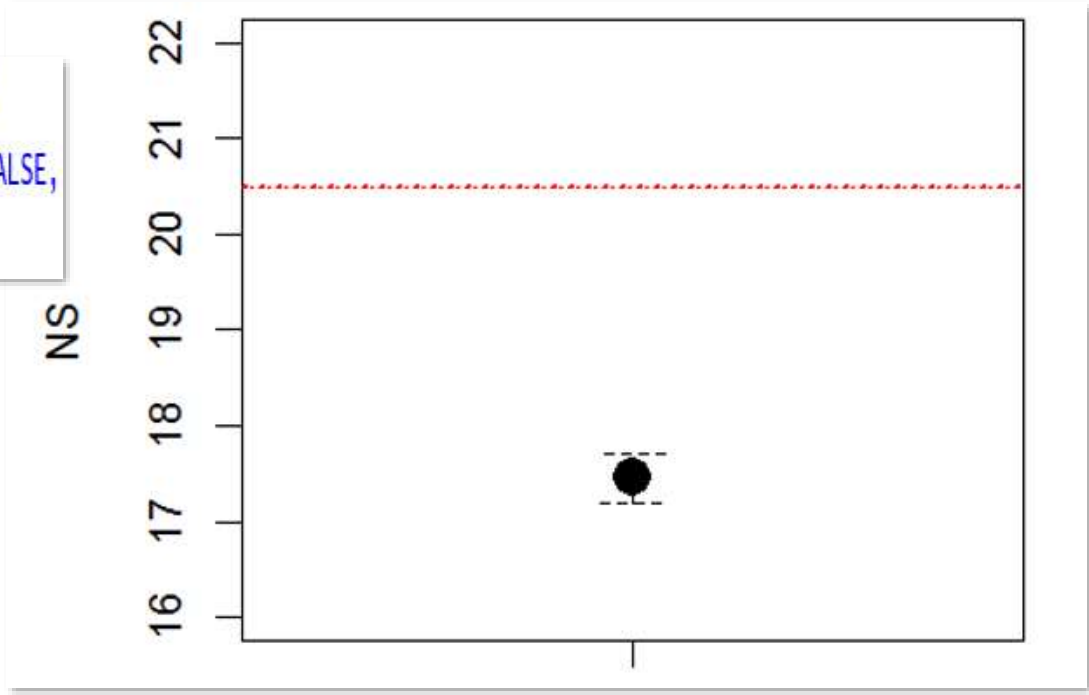
`plotmeans` di `gplots` plotta medie con CI (default .95), ma si aspetta un fattore nella formula `misura ~fattore`, che non abbiamo: creiamone uno falso.

```
plotmeans(ado_NS~as.factor(rep("",1269)), pch=19, ylim=c(16, 22),
lwd=2, barcol = "black", ylab="NS", xlab="campione",
cex=2)abline(h=20.2, lty=2, lwd=2, col="red")
text(labels = "mu",x = 1,y=20.5, pos = 4, col="red", cex=1.5)
```



Possiamo anche usare `plotMeans(response, factor1, error.bars= "conf.int", level=.95, connect= TRUE/FALSE)`, di `RcmdrMisc`, che abbiamo usato per rappresentare la sd attorno alla media. Ora sostituiamo la sd con il CI:

```
plotMeans(response = ado_NS, factor1 = as.factor(rep("",
 1269)), error.bars = "conf.int", level = .95, connect=FALSE,
 pch = 19, xlab="",ylab="NS", ylim=c(16,22))
```



`response=` indica la misura, `factor1=` i livelli del fattore per ciascuno dei quali si calcola la media (si può aggiungere anche un `factor2`); `error.bars= "conf.int"` per tracciare il CI, per una verosimiglianza specificata da `level` (default `=.95`); `connect=TRUE/FALSE` indica se connettere le medie con una linea (di default, `TRUE`).

- *Calcolate e rappresentate graficamente anche i 95%CI delle due dimensioni **HA e RS**.*

- *Conoscete i **valori normativi** anche di questi due tratti : quali conclusioni potete trarre, rispetto alla loro appartenenza alla popolazione attesa?*

- *Calcolate e rappresentate graficamente i 95% CI delle tre dimensioni in **maschi e femmine**. **Confrontate i due generi** in ogni dimensione: potete anticipare alcune delle conclusioni che trarremo tra un po'?*

Spoiler…

Possiamo conoscere il CI della media di una distribuzione anche con la funzione di base **`t.test(distribuzione, conf.level= .95)`**, che useremo molto. Il suo scopo non è fornire (solo) il CI, ma per ora potete ignorare tutto il resto dell'output.

```
> t.test(ado_NS)

        One Sample t-test

data:  ado_NS
t = 132.94, df = 1268, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 17.20720 17.72267
sample estimates:
mean of x
 17.46493
```

Riprenderemo l'uso dei CI in aggiunta – o in sostituzione – dei $p-value$ e torneremo sulla loro interpretazione grafica in TAD 2 (t-test per dati indipendenti e dati appaiati).

Inferenza e verifica delle ipotesi: P-value approach, Fixed alpha approach, Null Hypothesis Significance Test (NHST)

Sic enim se profecto res habet, ut ***numquam perfectam veritatem casus imitetur****.*

Cicerone, De Divinatione, Libro I

Statistical rituals largely eliminate statistical thinking *in the social sciences […] What I call the "**null ritual**" consists of three steps: (1) set up a statistical null hypothesis, but do not specify your own hypothesis nor any alternative hypothesis, (2) use the 5% significance level for rejecting the null and accepting your hypothesis, and (3) always perform this procedure.*

Gigerenzer, ***Mindless statistics****, 2004*

L'approccio alla verifica delle ipotesi oggi più utilizzato e criticato nelle scienze sociali è il **Null Hypothesis significance Test** (**NHST**), metodo **ibrido**, fusione infelice di due approcci facenti capo a statistici con rapporti tesi: da una parte Ronald **Fisher** (**P-value approach** – **PVA**, 1925), dall'altra **Egon Pearson** e Jerzy **Neyman** (**Fixed Alpha Approach - FAA**, 1928).

PVA e FAA sono superficialmente simili, tanto da poter essere confusi, anche se i rispettivi Autori erano combattivamente pronti a rivendicarne le differenze.

Questo ha facilitato la loro fusione nell'approccio **NHST**, **concettualmente confuso**, dato che "ibrida" costrutti in buona parte differenti, ma **tecnicamente *appealing***, perché facilita una – **impropria - decisione tutto–o–nulla sulla propria ipotesi** alla luce dei risultati ottenuti, **nonostante** PVA e FAA insistano sulla necessità di **prendere decisioni e fare scelte ragionate** in tutti i passi del processo di verifica delle ipotesi.

Descriveremo brevemente PVA, FAA e NHST, poi approfondiremo le critiche a NHST e le opportune integrazioni (o alternative).

Per chi voglia approfondire: Perezgonzalez, J.D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology, doi: 10.3389/fpsyg.2015.00223

Fisher: p-Value Approach

Dal mondo **deterministico** del ragionamento **deduttivo**, la **dimostrazione per assurdo** (**reductio ad absurdum**) è adattata al mondo empirico e **probabilistico**) di quello **induttivo**:

Se, data una **premessa** [ipotesi] come vera, ne discende una **conclusione** logica contraddittoria, allora la premessa [ipotesi] è falsa.

Se, dato un modello come vero [**l'ipotesi nulla**], i risultati empirici sono in contraddizione con le sue previsioni [sono troppo **poco probabili** secondo le previsioni del modello], allora il modello [ipotesi nulla] è respinto - ***disproved***.

Ipotesi nulla – H_0: una **serie di affermazioni su come funzionano le cose in popolazione**, in particolare su un determinato valore di un parametro della popolazione.

*"Note le caratteristiche di un **universo [popolazione]**, se, tratto da esso un **campione**, **questi viola le nostre aspettative**, possiamo inferire che è stato tratto da una **diversa** popolazione"*

(1925)

Significance testing: si calcolare il **valore di probabilità del risultato empiricamente osservato** o di uno **ancora più estremo** (con probabilità ancora minore], sotto condizione di ipotesi nulla, ovvero **ponendo come assunto che H_0 sia vera**.

Questo valore di probabilità si chiama p-value

È **molto probabile** che il **modello** [l'ipotesi nulla] **sia adeguato** alla realtà

Grande p-value

Piccolo p-value

È **poco probabile** che il **modello** [l'ipotesi nulla] **sia adeguato** alla realtà

*"Every experiment may be said to exist only in **order to give the facts a chance of disproving the null hypothesis**"*

Vediamo i **cinque passi** di cui consiste il PVA.

1. Scegliere il test adatto

La scelta dipende dagli **scopi della ricerca e dalla scala delle misure**;

Per stimare relazioni simmetriche tra due variabili sceglieremo tra: test χ^2 a due vie se nominali, coefficienti rho di Spearman o tau di Kendall se ordinali, coefficiente r di Pearson se metriche.

2. Fissare H_0

La formulazione di H_0 dipende **dagli obiettivi/dal test**.

Per valutare la differenza tra le medie di due gruppi, H_0 affermerà che la differenza tra le due medie è uguale a zero, cioè che i gruppi non sono differenti: $\boldsymbol{H_0: \bar{x}_1 - \bar{x}_2 = 0}$.

Il **parametro** oggetto di H_0 è **theta Θ**: **non** è sempre vero che $H_0: \Theta = 0$.

*Per esempio, possiamo porre come H_0 che la differenza **standardizzata** tra le medie di due gruppi, **non sia superiore a 1**:* $H_0: z_{\bar{x}_1 - \bar{x}_2} \leq |1|$.

Rispetto a Θ, H_0 può essere espressa come:

- **Bidirezionale / a due code:** $\boldsymbol{H_0: \Theta_1 = \Theta_2}$
- **Monodirezionale destra / a una coda, destra:** $\boldsymbol{H_0: \Theta_1 \leq \Theta_2}$
- **Monodirezionale sinistra / a una coda, sinistra:** $\boldsymbol{H_0: \Theta_1 \geq \Theta_2}$

3. Calcolare la probabilità del risultato ottenuto sotto condizione di H_0

Assumendo che H_0 sia vera, si assegna la **probabilità al risultato verificatosi o a uno più estremo** → il $\boldsymbol{p - value}$ è una **probabilità cumulata.**

Empiricamente osservo che $\bar{x}_A = 2$ e $\bar{x}_B = 7.5$: se il gruppo A e il gruppo B appartengono alla stessa popolazione ($\boldsymbol{H_0}$: $\boldsymbol{\Theta_A = \Theta_B}$ ovvero $\boldsymbol{\Theta_A - \Theta_B = 0}$), qual è la probabilità di osservare un $\boldsymbol{\Delta \geq |5.5|}$?

4. Definire la significatività / eccezionalità del risultato

Fisher ritiene **interessanti** i risultati **che hanno una bassa probabilità di verificarsi come semplici fluttuazioni casuali** di H_0. Giudica **conveniente** un $\boldsymbol{p - value = .05}$ come limite per stabilire se l'evento sia davvero eccezionale

*"It's usual and **convenient** for experimenters to take this point as a limit in judging whether a deviation is to be considered significant or not […], as a standard level of significance, in the sense that **they are prepared to ignore all results which fail to reach this standard** and [….] to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced in their experimental results. We shall **not often be astray** if we draw a **conventional** line at .05" (pp. 47,79).*

D'altronde: "*if one in twenty [.05] **does not seem high enough odds**, we may […] draw the line at one in fifty (**.02**), or one in a hundred (**.01**).*

5. Interpretare l'eccezionalità del risultato

Se il $p - value$ è inferiore allo "standard level of significance" precedentemente stabilito, Fisher direbbe che: "either an **exceptionally rare** chance has occurred, or the **theory is not true**" → il **dato empirico non dà sufficiente conferma ad H_0**.

Comunque, anche se i dati **non contraddicono** il modello / H_0, **non è possibile affermare di aver dimostrato che il modello/H_0 è vero**, ma solo che i dati campionari **non offrono sufficiente evidenza per rifiutare** H_0 : ***Absence of evidence is not evidence of absence***

D'altronde, un $p - value$ <livello di significatività disconferma H_0, ma **non dimostra che sia vera** un'ipotesi **alternativa** a quella disconfermata.

Quindi, a che serve realmente un $p - value$? A **poco o nulla**, dicono in molti.

Fisher stesso afferma che l'unico modo per fare inferenze sensate basate sui risultati di un test di significatività è il **controllo sul disegno di ricerca** (1955), e che un **unico** risultato significativo deve essere considerato solo un punto di partenza, da **replicare** in ulteriori ricerche (1954) e confermare con **meta-analisi** (1960).

Facciamo un esempio basato sulla **rappresentatività di un campione rispetto alla popolazione** da cui è estratto, come nel caso dei CI:

Se i risultati del campione saranno in contraddizione con quelli attesi in popolazione, cioè **troppo poco probabili** per essere variazioni casuali rispetto a quanto previsto, allora si **dovrà respingere l'ipotesi nulla che il campione sia rappresentativo**.

Compriamo 36 barattoli magnum di Nutella **scelti a caso** per verificare se il peso dichiarato in etichetta ($3Kg$ → peso della popolazione) corrisponde a quello atteso.

H_0 : il peso medio dei 36 barattoli è uguale a quello della popolazione: $\bar{x} = \mu$.

Pesiamo i 36 barattoli: il peso **medio** della **distribuzione campionaria è $\bar{x} = 2.92Kg$**

È evidente che $2.92 \quad 3$ (infatti, nella **realtà empirica**, H_0, in un'accezione restrittiva di perfetta identità, è **sempre falsa**); dobbiamo però tener conto della **variabilità campionaria** e degli errori di misura, per cui riformuliamo:

La **differenza** tra la **statistica** campionaria osservata ($2.92\ Kg$) e il corrispondente **parametro** nella popolazione ($3Kg$) è abbastanza **piccola** da poter considerata la media campionaria una **fluttuazione casuale del parametro**?

Quanto è probabile che si possa osservare **solo per caso** una **differenza** $|\bar{x}_{DCM} - \mu|$ **pari o superiore** a quella osservata, se i campioni appartengono effettivamente alla popolazione (ovvero sotto condizione di H_0)?

Tanto più grande è questa probabilità, tanto più probabilmente il campione è effettivamente **tratto dalla popolazione**.

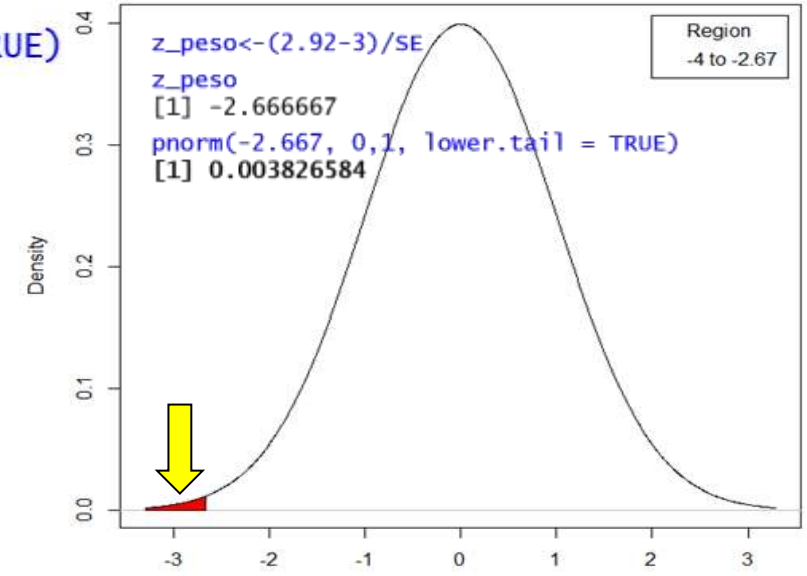
Per conoscere la probabilità di $\bar{x}_{DCM} = 2.92$ di verificarsi sotto condizione di H_0, ricorriamo alle proprietà della DCM. La sd **dei barattoli è** $sd = .18$. Quindi, iniziamo a **calcolare lo** SE**:**

$$SE = \frac{s}{\sqrt{N}}$$

```
(SE<-.18/sqrt(36))
[1] 0.03
```

Ora possiamo **conoscere la probabilità cumulata** pari al **quantile 2.92 o a un peso inferiore**, in una distribuzione normalmente distribuita con $\mu = 3$ e $SE = .03$, con la **funzione di ripartizione:**

```
pnorm(q = 2.92, mean = 3, sd = .03, lower.tail = TRUE)
[1] 0.003830381
```



La probabilità che da una popolazione con peso $\mu = 3$ siano casualmente tratti barattoli con peso medio $\bar{x} \leq 2.92$ è **molto bassa**.

La probabilità che un peso medio $\bar{x} = 2.92$ sia solo una **fluttuazione casuale** di $\mu = 3$ è **molto bassa**

Il risultato è "noteworthy", interessante, significativo? Sembrerebbe un risultato davvero eccezionale, ma prima di fare causa all'azienda dovremmo seguire il suggerimento di Fisher e ***replicare il risultato****, oppure controllare in letteratura se e quante altre verifiche del genere siano state fatte, e quali siano stati i loro risultati, combinandoli in una meta-analisi.*

Fixed Alpha Approach - FFA

Il PVA è stato criticato per la mancanza di un'esplicita ipotesi **alternativa:** sembra inutile rifiutare H_0 se non è disponibile una spiegazione alternativa (Gigerenzer, 2004), anche se Fisher considera **implicitamente** ipotesi alternative come negazioni di H_0 ($\neg H_0$).

Il Fixed Alpha Approach (FFA, 1928) è stato proposto come **integrazione** del PVA, prima di diventare una vera alternativa, introducendo una **esplicita ipotesi alternativa H_A**

"Assunto che un dato campione sia tratto da una popolazione, il fenomeno che consideriamo non sarà significativamente differente dal corrispondente Θ in popolazione se è vera H_M (**main hypothesis** o ipotesi **principale**), mentre **differirà nel caso sia vera H_A"**

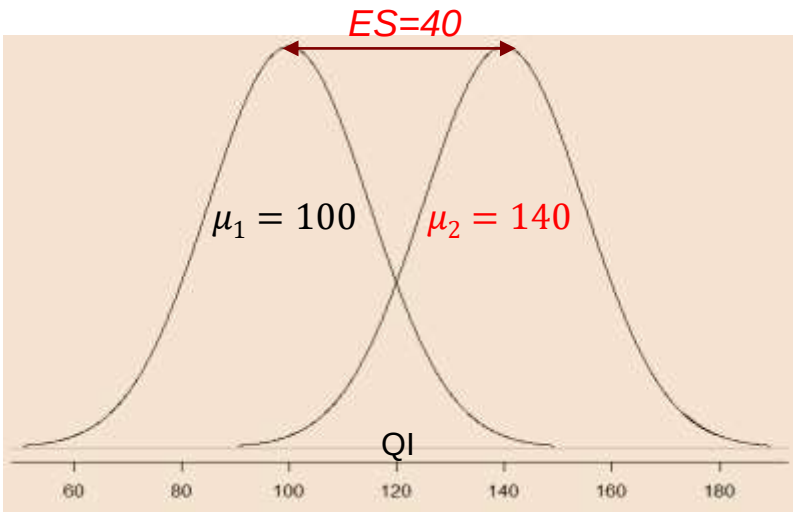
Il FAA prevede alcuni passaggi da stabilire a priori, prima di raccogliere i dati, e altri passaggi (molto minoritari) a posteriori.

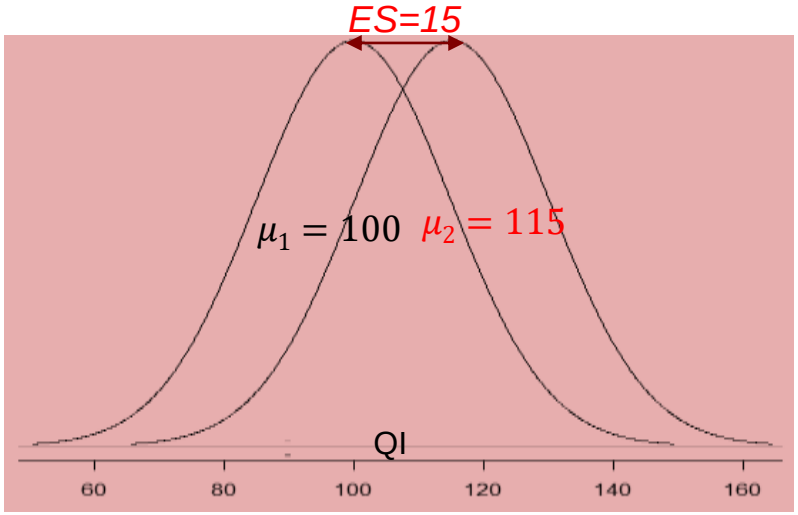
Prima di raccogliere i dati:

1. dichiarare Effect Size
2. Stabilire il test
3. Stabilire H_M
4. Stabilire H_A
5. Stabilire N
6. Calcolare il valore critico del test

1. Si **dichiara l'effect size atteso in popolazione**

La H_A rappresenta una **seconda popolazione** che si dispone a fianco della popolazione oggetto di H_M, sullo stesso continuum di valori. Le due popolazioni **differiscono sul continuum di una certa quantità**: **effect size o ES. Beta – β** è il **minimum effect size** (***MES***), sotto il quale non possiamo distinguere l'effetto da 0 (H_M).





2. Si seleziona il test ottimale

Oltre alle restrizioni e agli obiettivi visti nel PVA, il FAA predilige il test **più potente** → quello con **maggiori probabilità di rilevare l'ES.**

3. Si definisce H_M

$H_M: \mu_1 - \mu_2 = 0 \pm MES$

Si **definisce l'ipotesi principale** H_M da verificare: i valori compresi entro $0 \pm MES$ sono più probabili secondo H_M; quelli esterni sono più probabili sotto condizione di H_A.

Pearson e Neyman chiamano H_M "null hypothesis", facilitando l'equivoco con H_0: ma H_M è definita utilizzando il MES, che non fa parte del PVA, ed è solo una di due ipotesi che competono per emergere come esplicative della ricerca, mentre H_0 è l'unica protagonista nel PVA.

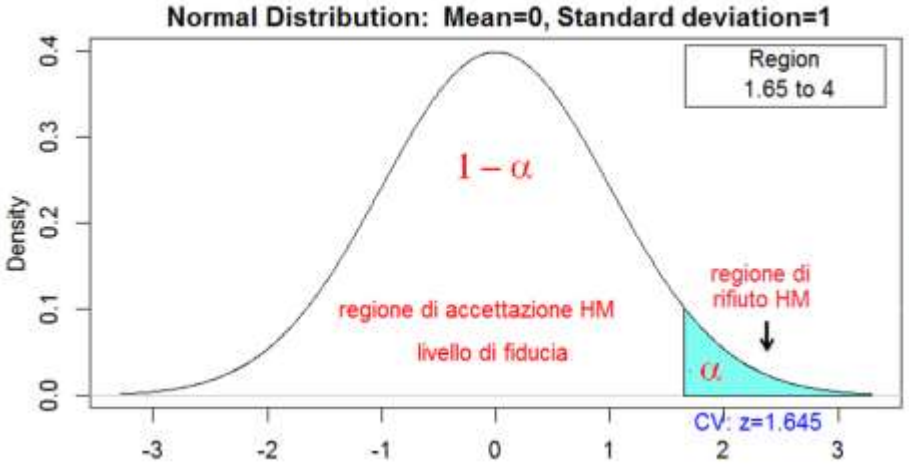
Si stabilisce anche il **controllo dell'errore di tipo I** (***Type I error***): errore compiuto **ogni volta** che H_M **è erroneamente respinta**. **Alfa - α** rappresenta la **soglia di assunzione del rischio di commettere questo errore a lungo termine**. Convenzionalmente, si usano α pari al 5% o all'1% (.05 o .01: 5 o 1 errore su 100 decisioni) o 1‰ (.001: un errore su 1000 decisioni).

Pearson e Neyman chiamano α "test significance level", facilitando l'equivoco con il livello di significatività del PVA, ma:

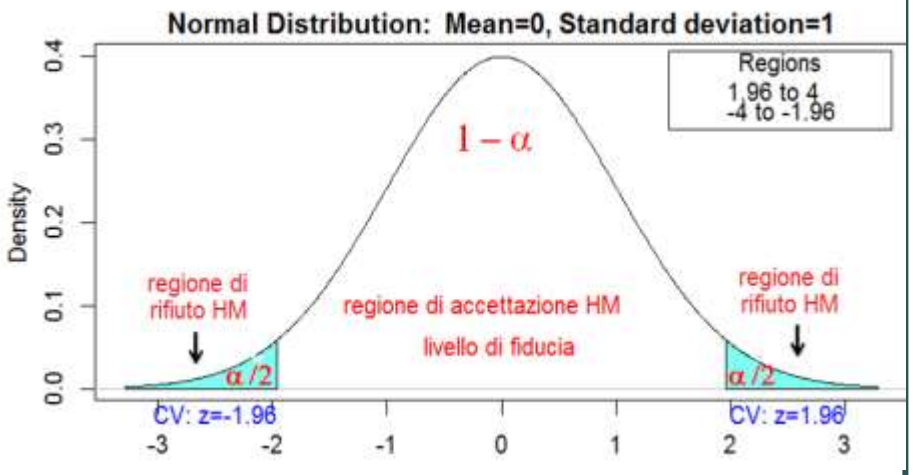
- *il $\boldsymbol{fixed}$ α si stabilisce prima della ricerca, la significatività nel PVA si interpreta dopo;*
- *FAA è un test di **accettazione**, non di significatività, perché non è interessato a stabilire quanto è forte l'evidenza **contro** H_M, ma a decidere se accettare H_A invece di H_M;*
- *il fixed alpha non ammette una gradazione di eccezionalità: si sceglie un α O un altro; dire che "il risultato è altamente significativo" perché ha un $p - value = 0.00001$ ha senso nel PVA, che ammette vari livelli di eccezionalità, ma non nel FAA.*

$H_M: \mu_1 - \mu_2 = 0 \pm MES, \alpha = .05$

α delimita la **regione critica**, o di **rifiuto**, sulla distribuzione di probabilità di H_M: una statistica che cada **fuori** dalla regione di rifiuto è **ragionevolmente probabile sotto H_M**; una statistica che cada **entro** la regione di rifiuto è **ragionevolmente probabile sotto H_A**.



Ipotesi monodirezionale



Ipotesi bidirezionale

La regione di rifiuto assomiglia a quella stabilita dal livello di significatività nel PVA come confine tra risultato eccezionale e non eccezionale. Però:

- *il PVA è centrato sul $p - value$ del risultato ottenuto, mentre la regione di rifiuto è indipendente dalla statistica del test osservata;*
- *La regione di rifiuto è fissata a priori una volta per tutte,*
- *quindi non è graduabile, mentre nel PVA si possono delineare diverse regioni critiche più estreme, come aree di maggior evidenza.*

4. Si definisce H_A

$H_A: \mu_1 - \mu_2 \neq 0 \pm MES$

Si **definisce l'ipotesi alternativa** H_A: coerentemente con H_M, può essere:

- **Bidirezionale / a due code:** $H_M: \Theta_1 = \Theta_2$ versus $\boldsymbol{H_A: \Theta_1 \neq \Theta_2}$
- **Monodirezionale destra / a una coda, destra:** $H_M: \Theta_1 \leq \Theta_2$ versus : $\boldsymbol{H_A: \Theta_1 > \Theta_2}$
- **Monodirezionale sinistra / a una coda, sinistra:** $H_M: \Theta_1 \geq \Theta_2$ versus : $\boldsymbol{H_A: \Theta_1 < \Theta_2}$

*H_A **bidirezionale** è accettata se $\Theta_1 > \Theta_2$ **e** $\Theta_1 < \Theta_2$. H_A **monodirezionale** è accettata **solo** se la **differenza va nella direzione prevista**: se $H_A = \Theta_1 > \Theta_2$, si accetta H_M sia se $\Theta_1 = \Theta_2$, sia se $\Theta_1 < \Theta_2$.*

Si stabilisce anche il **controllo dell'errore di tipo II** (***Type II error***): errore compiuto **ogni volta** che $\boldsymbol{H_A}$ **è erroneamente respinta**. **Beta -** $\boldsymbol{\beta}$ rappresenta la **soglia di assunzione del rischio di commettere questo errore a lungo termine**.

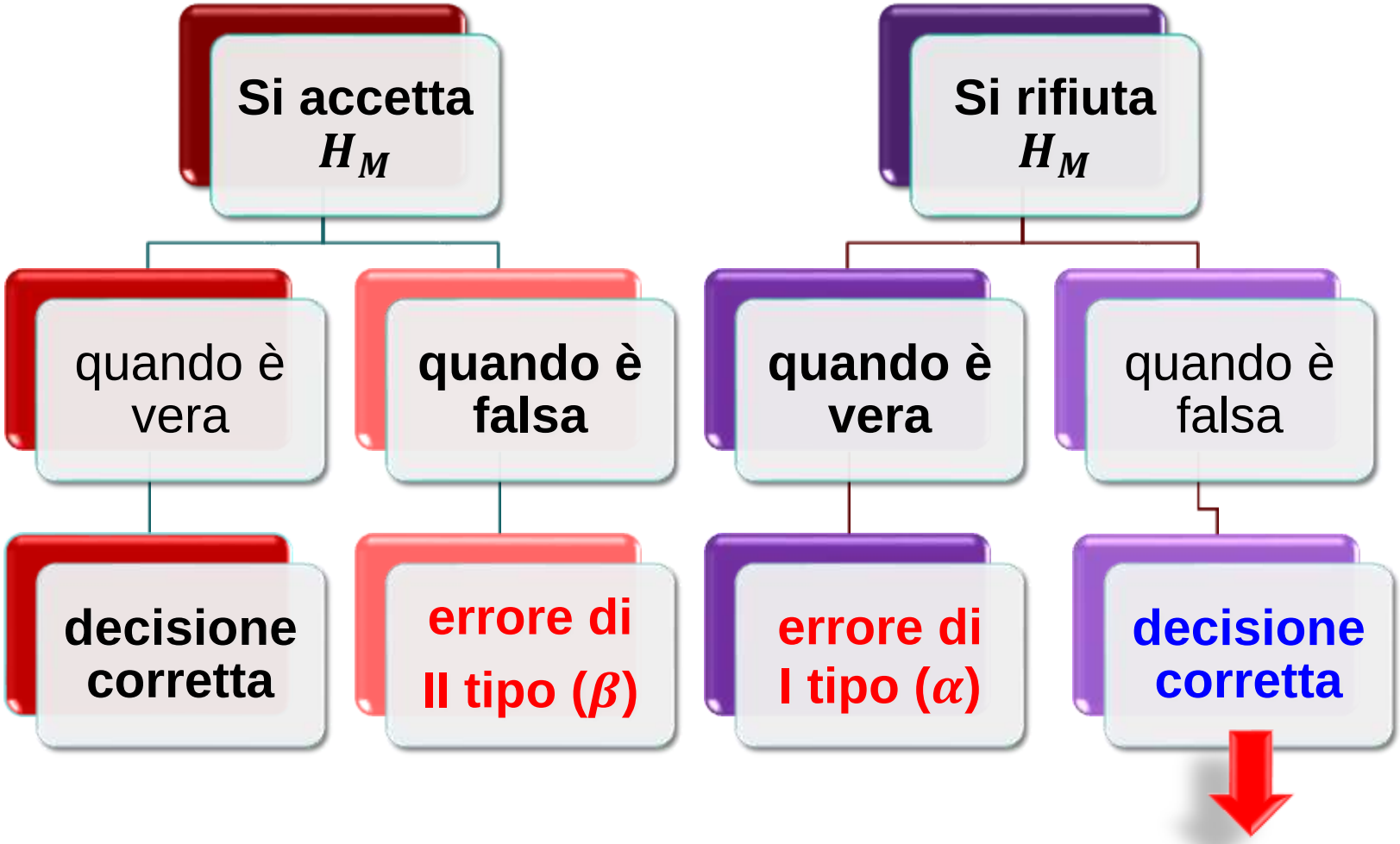
Per minimizzare β, Cohen (1969) propone che il rischio di commettere un errore di I tipo sia valutato al **massimo quattro volte maggiore** di un errore di II tipo: $\boldsymbol{\beta = 4 \times \alpha}$.

Quindi, se $\boldsymbol{\alpha = .05, \beta = 4 \times .05 = 0.20}$.

$H_A: \mu_1 - \mu_2 \neq 0 \pm MES, \alpha = .05, \beta = .20$

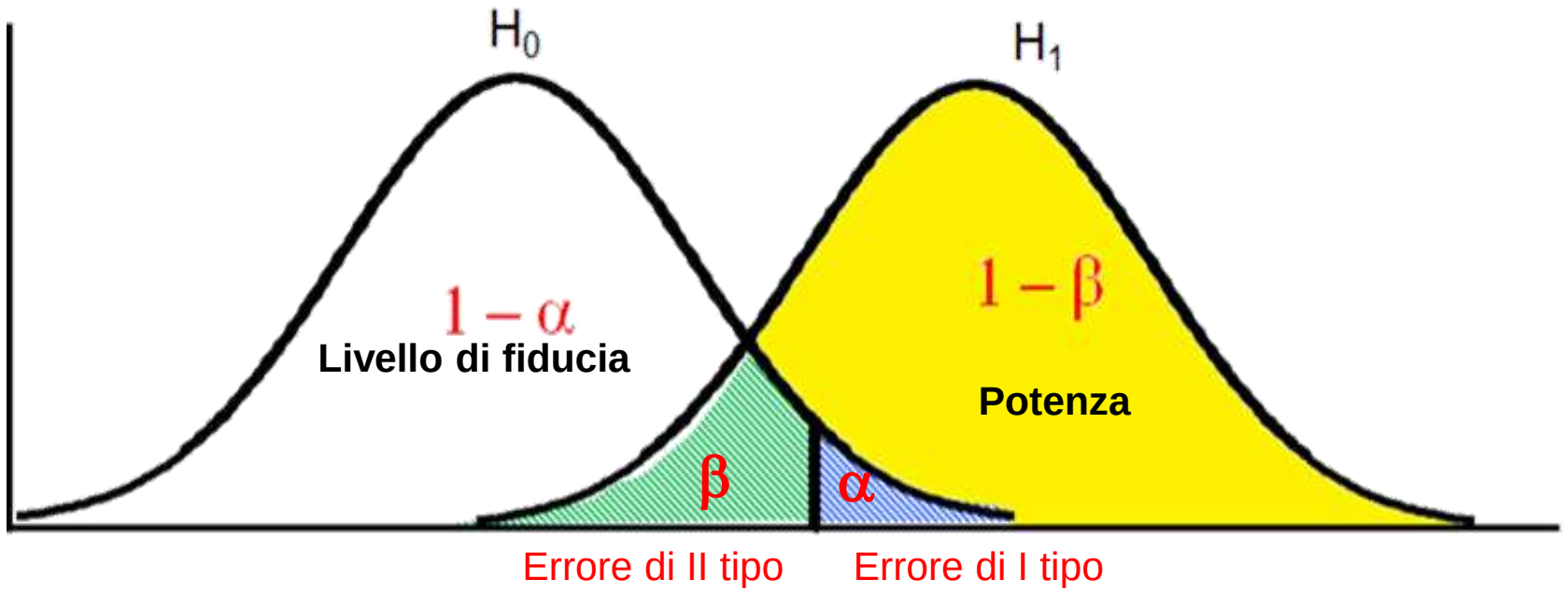
*Fisher è esplicito: "The phrase 'Errors of the second kind,' although apparently only a harmless piece of technical jargon, is useful as indicating the **type of mental confusion in which it was coined**" (1955).*

L'albero delle possibili decisioni su H_M porta a quattro alternative:



Potenza del test: **capacità di individuare in un campione un effetto, qualora questo si verifichi in popolazione**. È data da $\mathbf{1-\beta}$, quindi β è scelta a priori, proprio come α.

Di solito: $\boldsymbol{potenza}$: $\mathbf{1-.20=.80}$.


H_0
H_1
$1-\alpha$
Livello di fiducia
$1-\beta$
Potenza
β
α
Errore di II tipo
Errore di I tipo

5. Si stabilisce N

Definire la numerosità campionaria è importante, perché la potenza è strettamente **legata a N**. Come vedremo nella **power analysis**, la potenza dipende da: $\boldsymbol{N}$ (campioni più ampi aumentano la potenza), **tipo di test** (i parametrici sono più potenti), **effect size** (la potenza è funzione **crescente** di $\Theta_1 - \Theta_2$), **variabilità** dei dati (la potenza è funzione **decrescente** di σ^2), ovviamente $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$, **direzionalità** dell'ipotesi (a parità di tutti gli altri fattori, l'ipotesi monodirezionale è più potente della bidirezionale).

FAA e PVA differiscono soprattutto perché H_A dà informazioni esplicite al test su come considerare ES e β: se li ignoriamo, stiamo usando, più o meno consapevolmente, l'approccio di Fisher.

6. Si calcola il valore critico del test

Si calcola il **valore critico del test** ($\boldsymbol{CV}$), che costituirà il cut off per decidere tra H_M e H_A, in base al test identificato come ottimale, a N (in molti test partecipa al calcolo dei gradi di libertà) e alfa.

Dopo aver raccolto i dati:

7. Calcolare il valore ottenuto del test, da confrontare con il valore critico
8. Decidere a favore di H_M o H_A

7. Si calcola il valore ottenuto del test

Si calcola dai dati il **valore del test**: **valore ottenuto** o **research value**, $\boldsymbol{RV}$: è tanto più prossimo a zero quanto più il dato campionario è prossimo al valore previsto da H_M.

Si può usare il $p-value$ associato al RV, perché la verifica del dato sotto condizione di H_M è ancora valutata come probabilità del risultato empirico, data per vera H_M: $P(D|H_M)$. Quindi, il $p-value$ è tanto maggiore quanto più il dato è prossimo al valore previsto da H_M.

8. Si sceglie tra H_M e H_A

La decisione è piuttosto meccanica:

a. Il RV cade **nella regione di rifiuto** → si **respinge $\boldsymbol{H_M}$** e si accetta H_A
b. Il RV cade **fuori dalla regione di rifiuto / entro la regione di fiducia, e** il test ha **una buona potenza** → si accetta H_M;
c. Se il RV cade **fuori dalla regione di rifiuto / entro la regione di fiducia, e** il test ha una **scarsa potenza**, non si **conclude nulla**... non si dovrebbero eseguire ricerche con potenza insufficiente.

FAA è più potente di PVA per testare sul lungo periodo, quindi è particolarmente adatto a ricerche che campionano ripetutamente dalla stessa popolazione. L'approccio è deduttivo, meccanico una volta prese le decisioni a priori, e di conseguenza meno flessibile di quello di Fisher, a cui è superficialmente simile.

Usiamo il FAA per replicare l'esperimento precedente: ricompriamo random altri barattoli di Nutella per verificare se il peso dichiarato in etichetta ($\mathbf{3Kg}$) corrisponde a quello atteso.

1. $\boldsymbol{ES}$: pochi grammi di differenza tra media attesa e media campionaria non sarebbero interessanti: ci aspettiamo **almeno 70 grammi di differenza secondo** $\boldsymbol{H_A}$ $\rightarrow$ **minimum effect size**. Se ci fossero altre ricerche, ci baseremmo sui loro risultati per determinare ES.

2. **Test**: usiamo la **distribuzione di probabilità normale** *[è uno z test per campione unico]*

3. $\boldsymbol{H_M}$ **e** $\boldsymbol{\alpha}$: $\boldsymbol{\bar{x} - \mu = 0 \pm MES}$**.** Controlliamo il rischio di respingere erroneamente questa H_M a vantaggio di H_A (errore di I tipo) con un margine di errore pari al 5%: $\boldsymbol{\alpha = 0.05}$

4. $\boldsymbol{H_A}$ **e** $\boldsymbol{\beta}$: $\boldsymbol{\bar{x} < \mu \pm MES}$: ci preoccupa un peso campionario **inferiore** a quello atteso, quindi fissiamo una H_A monodirezionale **sinistra**. Controlliamo il rischio di respingere erroneamente H_A (errore di II tipo) ponendo questa probabilità ($\boldsymbol{\beta}$) come quattro volte maggiore rispetto ad α: $\boldsymbol{\beta = 0.05 \times 4 = .20}$. La **potenza** del test sarà $\mathbf{1 - .20 = .80}$.

5. Calcoliamo N alla luce di: test, α, β ed ES → la **power analysis** è fuori programma (chi vuole la trova, con alcune funzioni di R, anche nell'Appendice 2), sfioriamola solo.

*Qui usiamo `power.norm.test` di **pwr**, che effettua analisi di potenza quando si usa lo z test per campione unico. Gli argomenti sono **d**=$(\Theta_A - \Theta_M)/s$, ovvero MES standardizzato),* `sig.level= α, power= 1-β, alternative = "less / greater / two.sided"`, `n= numerosità`.

Inserendo solo **quattro di questi argomenti, sarà restituito il quinto**

Recuperiamo la $s = .18$ del precedente esperimento e chiediamo.

```
pwr.norm.test(d = (2.93-3)/.18,sig.level = .05,alternative = "less", power = .80)
     Mean power calculation for normal distribution with known variance
              d = -0.3888889
              n = 40.88058
      sig.level = 0.05
          power = 0.8
    alternative = less
```

Questa volta dovremo comprare **41 barattoli**.

Verificate ***come cambia la numerosità attesa****, a parità di tutti gli altri parametri e per lo stesso test, quando il* ***minimum effect size*** *è pari a* ***20*** *grammi e a* ***100*** *grammi. Cosa ne deduciamo?*

*E se mantenessimo fissa la numerosità della precedente ricerca (****N=36****), a parità di tutti gli altri fattori,* ***quale potenza otterremmo****?*

*Provate con i tre MES (****20, 70 e 100 grammi****): sarebbe soddisfacente in tutti i casi? Perché?*

6. Calcoliamo il valore critico CV del test, per tipo di H_A e $\alpha\rightarrow$ è un quantile z di una distribuzione normale, corrispondente ad $\alpha = .05$ nella coda **sinistra** ($H_A: \bar{x} < \mu$).

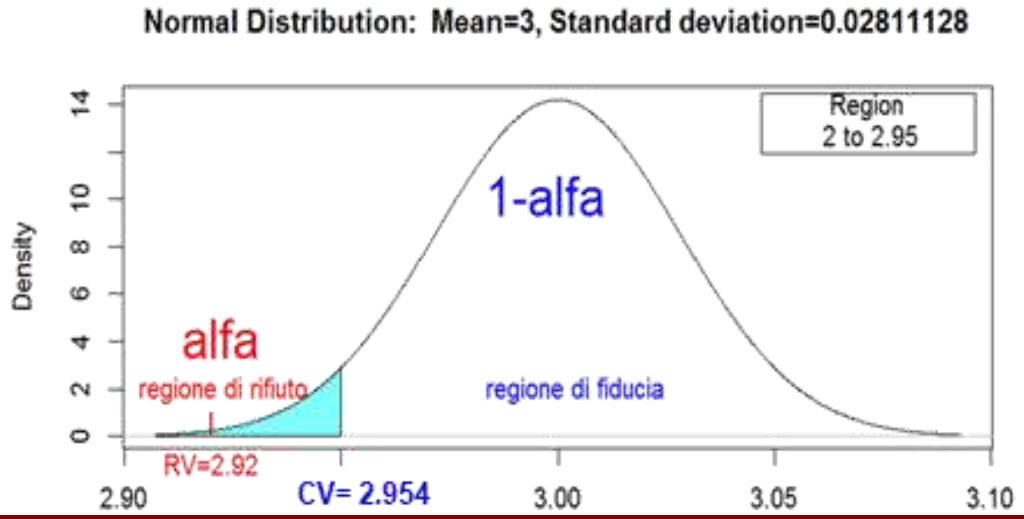
```
qnorm(p = 0.05, mean = 3, sd = .18/sqrt(41), lower.tail = TRUE)
[1] 2.953761
```

Un peso medio $\leq 2.954\ Kg$ *cadrà nella regione di rifiuto di* H_M .

7. Calcoliamo il valore ottenuto RV: compriamo i 41 barattoli nel modo più randomizzato possibile e pesiamoli nel modo più accurato possibile, per limitare errore casuale e sistematico: il loro peso medio è (ancora) $\bar{x} = 2.92$.

8. Scegliamo tra H_M e H_A: RV cade nella regione di rifiuto: il peso medio dei barattoli **non** è una fluttuazione casuale della media della popolazione

Il campione non è rappresentativo della popolazione.



Approccio NHST

Il Null Hypothesis Significance Testing (NHST) è attualmente la procedura più comune di verifica delle ipotesi, nonostante pesanti critiche: è un **miscuglio di PVA e FAA**, nonostante questi siano incompatibili in più punti.

Usa test per attribuire una probabilità al dato sotto condizione di H_0, considerando in competizione **ipotesi nulla** ($\boldsymbol{H_0}$) e **ipotesi alternativa** ($\boldsymbol{H_1}$). H_1 è concepita come "non H_0", è subordinata ad H_0, e solo saltuariamente si vede associati a priori l'effect size atteso e beta.

L'indicatore d'interesse è il $p - value$, confrontato con un cut off identificato dalla soglia α decisa a priori; questo punto è particolarmente confuso, dato che in molti testi si interpreta il $p - value$ **graduandolo** ("molto significativo") come nel PVA. Si dà priorità all'errore di I tipo, ma è consuetudine considerare anche quello di II tipo: quindi la valutazione della potenza è abbastanza frequente e spesso si calcola la numerosità campionaria a priori.

Se il risultato del test (RV) cade al di fuori della regione critica, si conferma H_0; se cade entro la ragione critica, H_0 viene considerata disconfermata e si accetta H_1.

Riassumendo, il percorso logico di un'inferenza statistica secondo NHST è:

Un **risultato non significativo non disconferma** H_0

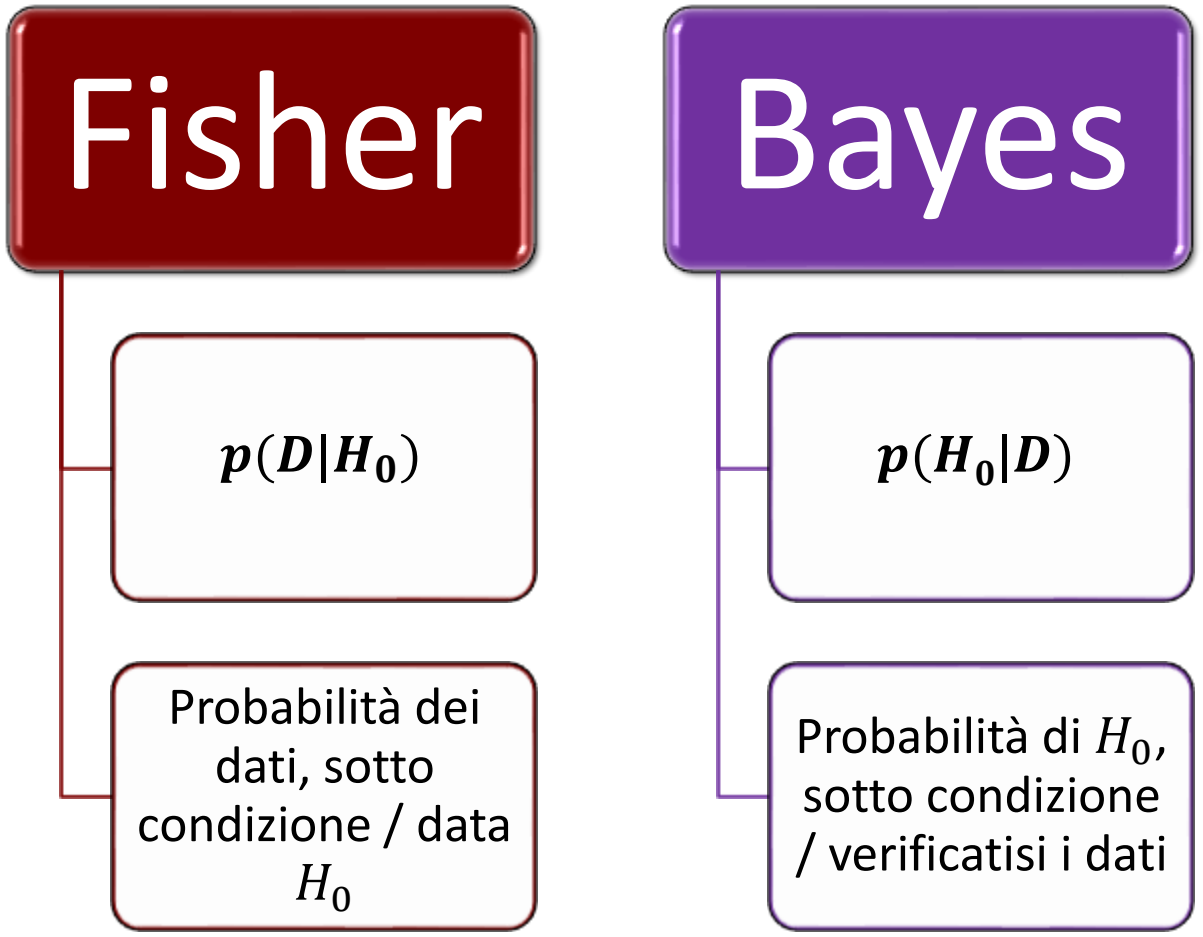
Adattiamo un modello statistico ai dati e decidiamo su H_0

Formuliamo H_1, che si esprime con i segni $\neq$, $>$ o $<$. H_0 sarà rifiutata solo se l'evidenza empirica contraria sarà forte

Sottoponiamo a falsificazione H_0, riferita a un **parametro,** che si esprime solo con il segno $=$

È possibile commettere, anche se raramente, un **errore**

ATTENZIONE: Non è vero che stabilire il livello di significatività vuol dire valutare la probabilità di H_0 (in NHST è =1, evento certo), equivoco comune. È **l'approccio di Bayes** che permette effettivamente di **stimare la probabilità dell'ipotesi nulla**,



*Beh, più precisamente la formula di **Bayes** sarebbe*:

$$P(H|D) = \frac{(D|H) \times p(H)}{p(D|H) \times p(H) + p(D|\neg H) \times p(\neg H)}$$

Critiche all'approccio NHST

"...Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on ***merely refuting the null hypothesis*** *as the standard method for corroborating substantive theories [...]* ***is a terrible mistake****, is* ***basically unsound, poor scientific strategy****, and* ***one of the worst things*** *that ever happened in the history of psychology"*

Meehl*, 1978, pag.817*

Problemi logici

Cohen critica **le basi logiche del ragionamento** di Fisher, basato sul *modus tollens* (negare le premesse negando le conseguenze), che nella logica formale è valido (da due premesse discende una sola conclusione), ma **nel mondo probabilistico dell'inferenza** no:

Fin qui tutto bene:

Se H_0 è vera, allora il dato non può verificarsi.
Tuttavia, questo dato si è verificato.

Quindi, H_0 è falsa

$$\frac{A \rightarrow \neg B \quad B}{\neg A}$$

Ma quando si passa alla probabilità, il ragionamento porta a conclusioni errate:

Se H_0 è vera, allora questo dato è **altamente improbabile**

Tuttavia, questo dato si è verificato.

Quindi, H_0 è altamente improbabile

Se uomo suona la chitarra, è **molto improbabile** che faccia parte degli AC/DC

Tuttavia, Angus Young fa parte degli AC/DC

Quindi probabilmente Angus Young non suona la chitarra

Questa fallacia è stata definita "the **illusion of attaining improbability**" (Falk e Greenbaum, 1995) o "the odds-against-chance fantasy" (Carver, 1978).

Non offre le informazioni desiderate

NHST e **inferenza** hanno in realtà obiettivi diversi:

- **Inferenza**: $\boldsymbol{P(H_0|D)}$, conoscere la **probabilità di $\boldsymbol{H_0}$** alla luce dei dati (approccio di **Bayes**)
- **NHST**: $\boldsymbol{P(D|H_0)}$, conoscere la **probabilità di ottenere dati** ugualmente o più discrepanti da quelli effettivamente verificatisi, assunto che H_0 sia vera.

Paradosso di Lindley (1957): per gli stessi dati, in alcune condizioni $p(H_0|D)$ tende a 1, mentre $p(D|H_0)$ si approssima 0: il $p - value$, dunque, non riflette la probabilità che H_0 sia scorretta.

Non consente di verificare teorie

Anche quando H_0 non è supportata dai dati, è sempre necessario escludere una serie di H_1 concorrenti, prima di affermare la validità di H_1, che può essere confermata **solo** da una **solida base teorica**, un **disegno** di ricerca **sperimentale** **appropriato** e ripetute **repliche** del risultato. Inoltre (Tuckey, 1969, 1991), servono informazioni sulla **direzione** della differenza tra parametri e sulla sua **grandezza**: NHST dice qualcosa solo sulla sua direzione, ma **non** sulla **grandezza** e l'importanza pratica **degli effetti** (**significatività empirica**).

"The fallacy of replication"

Gigerenzer, 1993): spesso si sbaglia interpretando il **valore complementare di p**, cioè $\mathbf{1-p}$, come probabilità della **replicabilità** dei risultati: tanto più piccolo è il $p-value$ di un risultato, tanto più probabile sarebbe ritrovarlo replicando l'esperimento. **Se** il $p-value$ rivelasse la verità su Θ, e si replicasse l'esperimento con condizioni identiche tranne un diverso campione random, il $p-value$ della replica dovrebbe confermare la stessa verità, **ma** dati basati su **simulazioni** di repliche dimostrano che il $p-value$ cambia drammaticamente da una simulazione all'altra: non si può predire se una replica successiva avrà un $p-value$ molto, poco o per nulla inferiore ad α. Cumming (2008; 2014) ha efficacemente definito il fenomeno **$p-value$ dance**:

https://www.youtube.com/watch?v=5OL1RqHrZQ8.

Questo è un problema per il PVA, che ammette diverse gradazioni di eccezionalità e quindi si confronta con risultati altamente incoerenti, mentre per il FAA l'incoerenza è meno grave, finchè i $p-value$ portano alla stessa decisione su H_0, dato che la loro soglia di riferimento è fissa.

La conclusione di Commings è di **usare i CI, non i $p-value$**, per interpretare i risultati.

Il vero **indice di replicabilità è la potenza**, non la significatività.

Rigidità della decisione

"*Kendall mentioned that Fisher produced the tables of significance levels to save space and to avoid copyright problems with Karl Pearson, whom he disliked".* Good, 1971

Adottando un livello a fisso, si **converte** un **continuum di incertezza** (probabilità da 0 a 1), in una **decisione dicotomica** su H_0: un'intera ipotesi può essere disconosciuta anche se il *p*-value del risultato è di poco superiore alla **convenzionale** soglia alfa .05, sebbene: "***surely, God loves the .06 nearly as much as the .05 level of significance***" (Rosnow e Rosenthal,1989).

Fisher ha solo **suggerito** un criterio pari al 5%, e per motivazioni non matematiche, ma decisamente arbitrarie e probabilmente editoriali; era aspramente critico su α e il suo rigido utilizzo: "**no scientific worker has a fixed level of significance at which, from year in year and in all circumstances, he rejects hyptheses: he rather gives his mind to each particular case in the light of the evidence and his ideas**" (1956).

Opportuni correttivi all'approccio NHST

Un buon numero di autori altrettanto autorevoli ha difeso la validità e l'utilità dell'approccio NHST : molte **strategie**, proposte come completamente sostitutive, sono state suggerite anche come **complementari** alla verifica della significatività. Vediamo, per usarle da ora in poi, quelle il cui uso è stato proposto dalla **Task Force on Statistical Inference** dell'American Psychological Society (**APA**; Wilkinson and TFSI, 1999).

Calcolare e interpretare i CI

Calcolare i CI **attorno alle stime campionarie** è un'eccellente **integrazione**, e in realtà un vero **sostituto**, della verifica della significatività:

- ✓ Come abbiamo visto, **quando il** CI **(almeno al** 95%**) contiene il valore previsto da** H_0**, qualunque esso sia, accettiamo** H_0 con una probabilità prefissata.
- ✓ I CI danno anche **informazioni sulla precisione della stima** dei parametri.
- ✓ i *CI* relativi alla differenza o alla relazione tra parametri, oltre a includere o meno il valore previsto da H_0 ($\bar{x}_1 - \bar{x}_2 = 0$; $^{pr_1}/_{pr_2} = 1$, $r_{x1x2} = 0$), indicano anche **direzione e grandezza della differenza o della relazione tra parametri**.
- ✓ Inoltre, la stima puntuale nel campione e la stima intervallare nella popolazione usano la stessa unità di misura, rendendo **facile l'interpretazione** dei risultato

Se hanno tanti pregi, **perché sono riportati piuttosto raramente** negli articoli di molte discipline psicologiche? Secondo alcuni (e.g., Cohen, 1994) la **rarità** dei *CI* negli articoli sarebbe dovuta proprio alla loro **imbarazzante ampiezza**...

Indici di intensità dell'effetto – effect size

Cohen (1988): l'**effect size** (***ES***) è il **grado in cui il fenomeno è trovato nella popolazione.**

Snyder e Lawson (993): l'ES è il **grado in cui la variabile dipendente Y è controllata, predetta o spiegata dalla** / dalle variabili **indipendenti X.**

Gli ES consentono **di confrontare diversi studi**, perché sono riconducibili a una scala comune; sono essenziali per la **power analysis** e per le **meta-analisi**.

Una definizione pignola **distingue :**

- ✓ **indici di *ES***: **quantificano la differenza tra parametri**, che secondo H_0 è $= 0$ (le medie di un test di psicopatia di un gruppo clinico e di un gruppo di controllo): d ed f di Cohen, g di Hedges, g di Glass, indicatori robusti… e molti altri
- ✓ **misure di associazione**: **quantificano la proporzione di varianza** di Y **associata** o **spiegata** dalla **varianza di una o più X** (quanta variabilità del peso dipende dall'altezza, dall'attività, dalle calorie): r, R^2, R_M^2, OR, η^2, ϖ^2…

Vedremo i coefficienti di ES analisi per analisi.

Il più **importante** per anzianità, semplicità, comprensibilità e **coerenza** con la logica dell'ES è il **coefficiente d di Cohen(1954)**, in cui può essere convertita la gran parte degli altri coefficienti di ES. Nella sua forma – base:

*Differenza (in valore assoluto) tra le medie del gruppo **sperimentale S** e del gruppo di **controllo C***

$$d = \frac{|\bar{x}_S - \bar{x}_C|}{s}$$

*... ponderata per la **sd comune: media** delle **sd** (aritmetica se $N_S = N_C$, ponderata se $Ns \neq N_c$)*

d standardizza la differenza tra le medie due campioni esprimendola in unità di sd; il segno indica solo la direzione della differenza: la grandezza dell'effetto si legge in valore assoluto.

Per ciascun coefficiente, **criteri convenzionalmente stabiliti** definiscono l'ES **trascurabile**, **debole**, **moderato** o **forte**; le soglie per d, indicate da Cohen, sono:

|.0 - .20|	|.20- .50|	|.50- .80|	|> .80|
Effetto trascurabile	Effetto debole	Effetto moderato	Effetto forte

Attenzione però a **non interpretare rigidamente** e acriticamente le soglie: meglio leggere gli indici **comparativamente**, rispetto a risultati precedenti o al disegno di ricerca.

d esprime perfettamente **quello che tutti i coefficienti di ES valutano:**

$$d = \frac{|\bar{x}_S - \bar{x}_C|}{s}$$

$$ES = \frac{segnale}{rumore}$$

Entità della differenza / della relazione

... variabilità interna al gruppo / al soggetto

La formulazione si rifà al rapporto segnale – rumore (***signal to noise ratio***), nato nel campo delle comunicazioni radio e ma trasferito in numerosi diversi campi, e affine alla **teoria delle detezione del segnale**, che quantifica la **capacità di discriminare il segnale vero e proprio**, dotato di **significato**, **in mezzo al rumore di fondo**, privo di significato e **confondente**.

Per tutti gli indici di ES dovrebbero essere **calcolati i CI** (R lo fa volentieri per noi). Per esempio, la funzione `cohen.d` del package `effsize` che useremo per calcolare l'entità della differenza tra due medie ci fornirà i relativi CI.

Analisi di potenza o power analysis

E. Pearson e Neyman (1928) hanno dimostrato che, **nota la grandezza della differenza tra H_M e H_A** (l'**effect size**), e **fissando** il valore della probabilità di commettere un errore tipo I (α) e di tipo II (β) ovvero $1-\beta$: potenza), è possibile **determinare a priori la numerosità N necessaria per rilevare l'effetto nel campione**, se esso davvero esiste in popolazione

*Postuliamo che la differenza tra popolazione clinica e normativa nei punteggi di un test sui pensieri intrusivi sia forte ($d \geq .80$), accettiamo una soglia di rischio per l'errore di I tipo $\alpha = .05$ e per quello di II tipo $\beta = .20$, cioè una $1-\beta = .80$: la power analysis stima quale sia la **N minima** dei due campioni per ottenere una differenza significativa, se esiste realmente nelle popolazioni clinica e normativa.*

Generalizzando, fissati tre dei parametri tra α, β, ES e N, è possibile stimare il quarto: fissati ES, N e α, si stima $1-\beta$; fissati ES, N e β, si stima α; fissati N, α e β, si stima ES.

La power analysis è indispensabile quando uno studio porta a confermare H_0, per confermare se davvero l'effetto è inesistente o irrilevante in popolazione. I risultati di Cohen (1962) rispetto alla scarsa potenza di molti studi nelle discipline psicologiche sono stati confermati…

La power analysis in R ***non*** *rientra nel programma: potrebbe però esservi necessaria per la tesi, per cui trovate alcune indicazioni su* ***pwr*** *(per statistiche più complesse,* ***pwr2****) nell'Appendice II della dispensa.*

Replicabilità del risultato e meta-analisi

La conoscenza scientifica si sviluppa **attraverso la replica degli studi**: i risultati di **uno studio non replicato**, indipendentemente dalla significatività statistica, sono solo **speculativi** (Hubbard e Armstrong, 1994) e privi di significato intrinseco (Lindsay e Ehrenberg, 1993).

La replica può essere **esterna** (un nuovo esperimento) o **interna** (usando metodi come la validazione incrociata – **cross validation** – o procedure di **ricampionamento**).

Servono **almeno quattro repliche** coerenti per accertare la potenza dello studio, ovvero un minimo di quattro studi su cinque che ottengono l'esito atteso per arrivare all'80% di potenza: una sola replica, anche significativa, non è una base sufficiente per supportare o contraddire il risultato di uno studio precedente (PerezGonzalez, 2015)

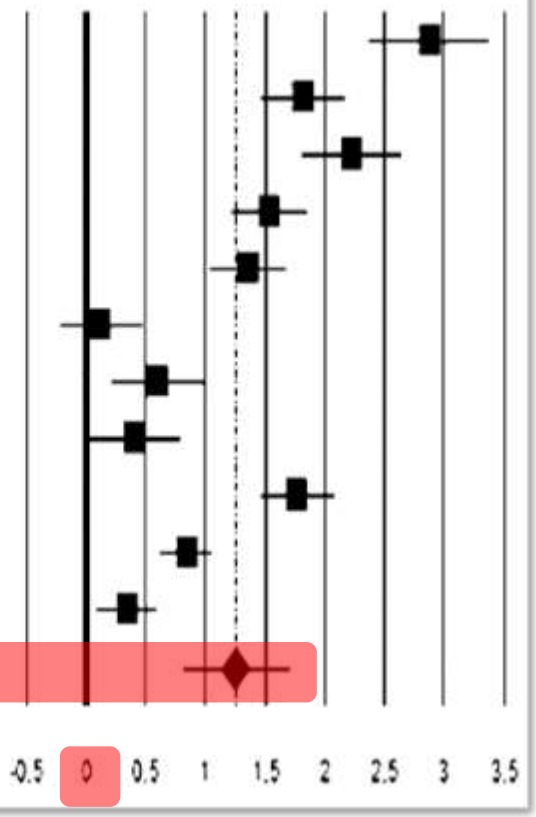
Una valutazione scientificamente rigorosa dei risultati di più repliche è fornita dalle **meta-analisi** (Glass, 1976: "***an analysis of analyses***"), che **vanno oltre i singoli studi**, combinando i dati di diverse ricerche: lo **scopo è aumentare la potenza rispetto a singoli studi e ottenere la stima migliore dell'effetto atteso (*true effect*)**.

Tracciamo un **veloce** ritratto dei passaggi delle meta-analisi, rimandando alla dispensa e ad altri testi per una trattazione più estesa (https://training.cochrane.org/handbook/current).

1. Si formula una **precisa domanda di ricerca**: "*L'approccio psicoterapeutico X ai sintomi del DOC è realmente **più efficace** dell'approccio Y applicato alla stessa sintomatologia?*"
2. Si **identificano gli studi rilevanti** per la domanda di ricerca, stilando **criteri di inclusione ed esclusione rigorosi** – ma non troppo restrittivi, arrivando al **set definitivo di studi**;
3. Si **estraggono i dati**: ES, variabili di **moderazione dell'effetto** (età, professione, esposizione a fattori di rischio..), la misura o le misure che **operazionalizzano l'esito** (punteggi a test, mortalità o guarigione, ecc.), il **tipo di disegno**,
4. Si valuta l'**eterogeneità nei dati** e la presenza di possibili *bias* metodologici, dando un **giudizio di qualità metodologica**.
5. **Analisi statistica e grafici**: si calcola una **stima sintetica dell'effetto** come **media ponderata degli ES rilevati nei singoli studi**: i **pesi** riflettono l'importanza relativa di ogni studio. La rappresentazione grafica è il **forest plot**, che mostra le stime degli effetti e relativi CI per i singoli studi e la media pesata della meta-analisi (Lewis e Clark, 2001).

Ogni studio è rappresentato da un quadrato, la cui area esprime il peso assegnato allo studio, e una barra che indica il CI dell'effect size. L'ES sintetico è rappresentato da un rombo

	ES	95% CI	W	V	SE	Sig.
Lesch 2008a/W1	2,88	2,38 . 3,38	8,62%	0,06	0,25	0,000
Lesch 2008a/W2	1,82	1,47 . 2,18	9,09%	0,03	0,18	0,000
Lesch 2008b/W1	2,24	1,82 . 2,65	8,89%	0,05	0,21	0,000
Lesch 2008b/W2	1,54	1,23 . 1,86	9,19%	0,03	0,16	0,000
Sam et al. 2008/W1	1,37	1,06 . 1,69	9,20%	0,03	0,16	0,000
Stave et al. 2007a/W1	0,14	-0,21 . 0,49	9,10%	0,03	0,18	0,448
Stave et al. 2007b/W1	0,61	0,22 . 1,00	8,97%	0,04	0,20	0,002
Stave et al. 2007c/W1	0,42	0,03 . 0,81	8,99%	0,04	0,20	0,036
Stephenson et al. 2011/W1	1,78	1,47 . 2,09	9,21%	0,02	0,16	0,000
Stephenson et al. 2011/W2	0,85	0,62 . 1,07	9,41%	0,01	0,11	0,000
Williams et al. 2010/W1	0,35	0,09 . 0,61	9,33%	0,02	0,13	0,008
Overall (random-effects model)	1,26	0,81 . 1,71	100,00%	0,05	0,23	0,000



Anche un'accurata selezione non evita che la **meta-analisi sia afflitta da *bias***: *publication bias* (i risultati che respingono H_0 sono pubblicati più facilmente), *duplicate* o *multiple publication bias* (uno stesso studio origina più pubblicazioni relative ai medesimi dati).

Come condurre una ricerca "onesta" usando i suggerimenti dei paragrafi precedenti?

1) **Formulate le domande di ricerca in termini di stima**, invece di usare espressioni che prevedono risposte dicotomiche: "Quanto grande sarà l'effetto..." o "in che misura...", invece di "verificare l'ipotesi che non ci sia alcuna differenza tra..." o "verificare se questo trattamento sia migliore di..."

2) **Identificate l'effect size** che vi serve

3) **Esplicitate tutti i dettagli della procedura e dell'analisi** dei dati, **prima** di eseguire lo studio; usate la **power *analysis* per definire N**

4 e 5) Dopo aver eseguito lo studio, **calcolate le stime puntuali e i CI dei coefficienti di effect size**; rappresentateli **graficamente**

6) **Interpretate l'intensità degli ES**, che rappresenta il principale *outcome* della vostra ricerca, e **l'ampiezza dei loro CI**, che indica la precisione della stima. Discutete le implicazioni teoriche e pratiche dei risultati

7) Pensate sempre in **un'ottica meta-analitica**

8) Nel **presentare i risultati**, descrivete in maniera esaustiva e trasparente la ricerca; mettete a disposizione anche i dati grezzi.

Cumming, 2013

Ora applichiamo la verifica delle ipotesi al caso più semplice:

confrontare un campione con una popolazione,

per diversi tipi di variabili

Test per un solo campione, variabili continue

La funzione `t.test` si può usare per: **confrontare una media campionaria con la media in popolazione** (***t*-test per campione unico**), confrontare le medie di due livelli indipendenti di una variabile factor (***t* − *test*** per **campioni indipendenti);** confrontare due medie a misure ripetute, prese sullo stesso soggetto in due condizioni diverse (***t* − *test*** **per dati appaiati**).

La logica del $t - test$ **per campione unico** è che la differenza tra $\bar{x}$ e μ, rapportata alla variabilità stimata (ES), si distribuisce come un quantile di una distribuzione t, per $df = N - 1$

$$t_{df=N-1} = \frac{\bar{x} - modello}{variabilità\ del\ modello} \qquad t_{df=N-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

Il risultato del $t - test$ consente di verificare se:

$$H_0: \bar{x} = \mu \qquad H_1: \bar{x} \neq \mu \qquad H_1: \bar{x} > \mu \qquad H_1: \bar{x} < \mu$$

In **`t.test` per un campione**, **`x=`** Y, oggetto del test, **`mu=`** media in popolazione; H_1 può essere bidirezionale (**`alternative=`** `"two.sided"` di default) o monodirezionale (`"greater"` o `"less"`), il CI di default è al 95% (**`conf.level=`** .95).

*Per verificare se il **campione di adolescenti è stato estratto dalla popolazione attesa** per un dato tratto di personalità, ovvero se la sua media è un'oscillazione casuale del tratto negli adolescenti (H_0), usiamo il tratto NS. Ricordiamo che $\mu = 20.2$ e $\bar{x} = 17.46$.*

Quanto è probabile che **17.46** sia **una fluttuazione casuale** da $\mu = 20.2$ ($\boldsymbol{H_0}$)?

```
t.test(ado_NS, mu = 20.2)
        One Sample t-test
data:  ado_NS
t = -20.819, df = 1268, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 20.2
95 percent confidence interval:
 17.20720 17.72267
sample estimates:
mean of x
 17.46493
```

Uh, il nostro primo output di un test!

Scomponiamolo.

***Il quantile t** espresso dal test:*

$$t_{df=N-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

***probabilità bidirezionale di un quantile** $t \geq -20.82$ **per** $df = 1269$ sotto condizione di H_0*

```
t.test(ado_NS, mu = 20.2)

        One Sample t-test

data:  ado_NS
t = -20.819, df = 1268, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 20.2
95 percent confidence interval:
 17.20720 17.72267
sample estimates:
mean of x
 17.46493
```

H_1 bidirezionale

Media nel campione

$95\%CI$ ***della media attesa nella popolazione da cui è tratto il campione di studenti:*** *con il 95% di probabilità, la media del tratto NS nella popolazione da cui è stato tratto il campione sta tra 17.21 e 17.72*

Attenzione alla notazione scientifica usata per il p-value:

```
p-value < 2.2e-16
```

Il numero è **espresso come potenza di 10**:

- ✓ **positiva**: **2.190127e+07** = 2.190127 * 10^{7}
- ✓ **negativa**: **2.190127e-07** = **2.190127 * 10^{-7}**

Spostate la virgola per un numero di cifre equivalenti alla potenza: a **destra della prima cifra** nel caso di una potenza **positiva**, a **sinistra** nel caso di una notazione **negativa**.

- ✓ **positiva**: **2.190127e+07** = 21901270.0
- ✓ **negativa**: **2.190127e-07** = **0.0000002190127**

Per le potenze negative, immaginate di mettere davanti alla prima cifra del risultato un numero di zeri pari alla potenza indicata e poi inserite la virgola dopo il primo zero, se vi risulta più facile pensarla così.

In pratica, questo output ci evita di calcolare:

$$t_{df=N-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

```
t<-(mean(a$NS_tot, na.rm=TRUE)-20.2)/(sd(a$NS_tot, na.rm=TRUE)/sqrt(1269))
round(t,3)
[1] -20.819
```

```
pt(q = t, df = 1269-1, lower.tail = TRUE)
[1] 2.440146e-83
```

```
t.test(ado_NS, mu = 20.2)

        One Sample t-test

data:  ado_NS
t = -20.819, df = 1268, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 20.2
95 percent confidence interval:
 17.20720 17.72267
sample estimates:
mean of x
 17.46493
```

```
mean(a$NS_tot, na.rm = TRUE)
[1] 17.46493
```

```
MeanCI(a$NS_tot, na.rm=TRUE)
    mean   lwr.ci   upr.ci
17.46493 17.20720 17.72267
```

Quindi: o “an **exceptionally rare** chance has occured”, o, più verosimilmente: “the **theory is not true**": gli studenti **appartengono in realtà a un’altra popolazione**. È la **stessa conclusione cui eravamo giunti osservando il CI:**

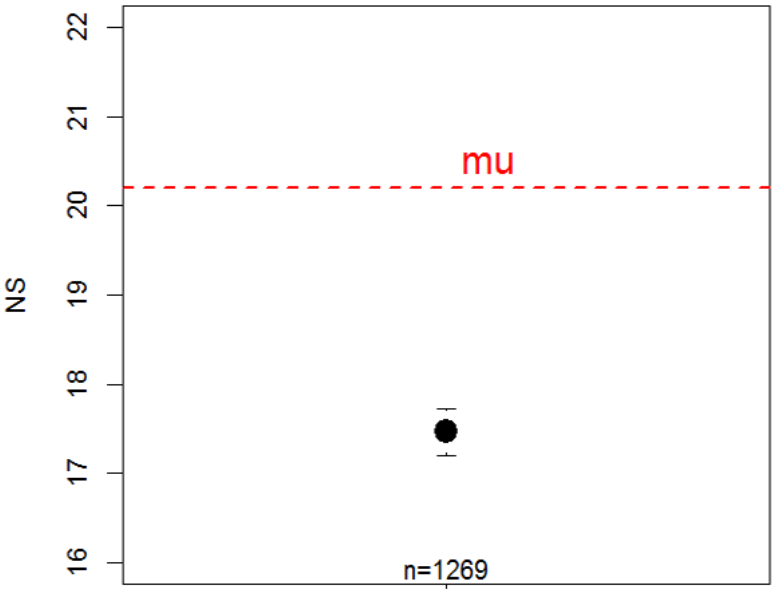
Quando nel $95\%CI$ (o maggiore) è compreso il valore previsto da H_0 (Θ), accettiamo H_0

Quando nel $95\%CI$ (o maggiore) **non** è **compreso il valore previsto da H_0 (Θ), rifiutiamo H_0**

```
95 percent confidence interval:
 17.20720 17.72267
```

$\mu = 20.2$ **non è compresa nel CI**

Questa **regola** rende praticamente **superflua l’interpretazione del $p - value$ in riferimento ad α,** a meno che per il CI non si scelga una verosimiglianza lontana da .95 **(.80, .90, ecc.)**



Non abbiamo finito: dobbiamo **calcolare il coefficiente di effect size della differenza**: d **di Cohen per una media**: differenza tra $\bar{x}$ e μ rapportata alla sd della distribuzione campionaria (notate l'affinità con il $t - test$...).

$$d = \frac{|\bar{x}_S - \bar{x}_C|}{s}$$

```
abs((mean(ado_NS, na.rm=T)-20.2)/sd(ado_NS, na.rm=T))
[1] 0.5844218
```

In `effsize` trovate `cohen.d(d=distribuzione, f=NA, mu=)`, che dà anche il CI di *d*.

in `f=` va inserito un factor, che in questo caso non c'è: scriviamo `NA`.

```
cohen.d(d=ado_NS, f = NA, mu = 20.2, na.rm = T)
      Cohen's d (single sample)
d estimate: -0.5844218 (medium)
Reference mu: 20.2
95 percent confidence interval:
     lower      upper
-0.6968928 -0.4719509
```

La differenza è al massimo moderata: la significatività è dovuta a N, ovvero alla potenza del test.

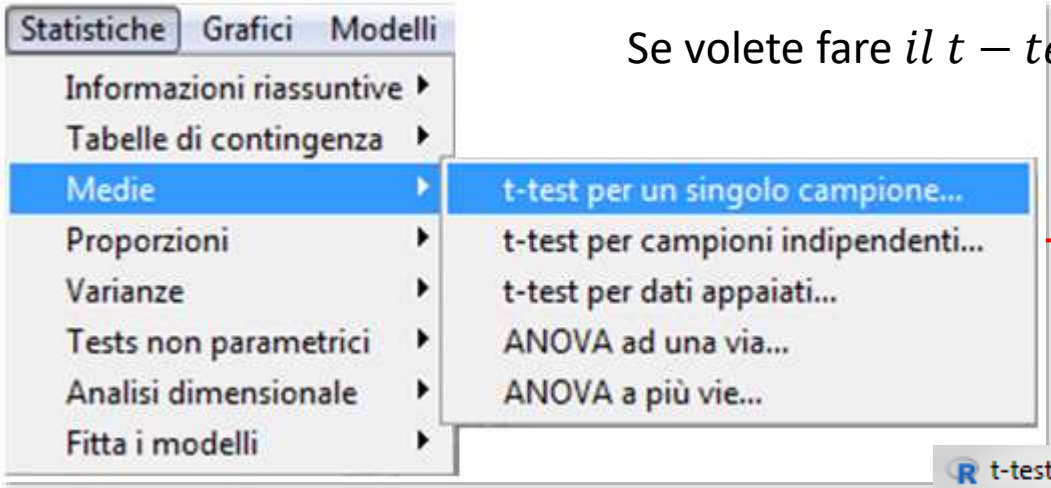
Ecco un ESEMPIO di come commentare l'analisi

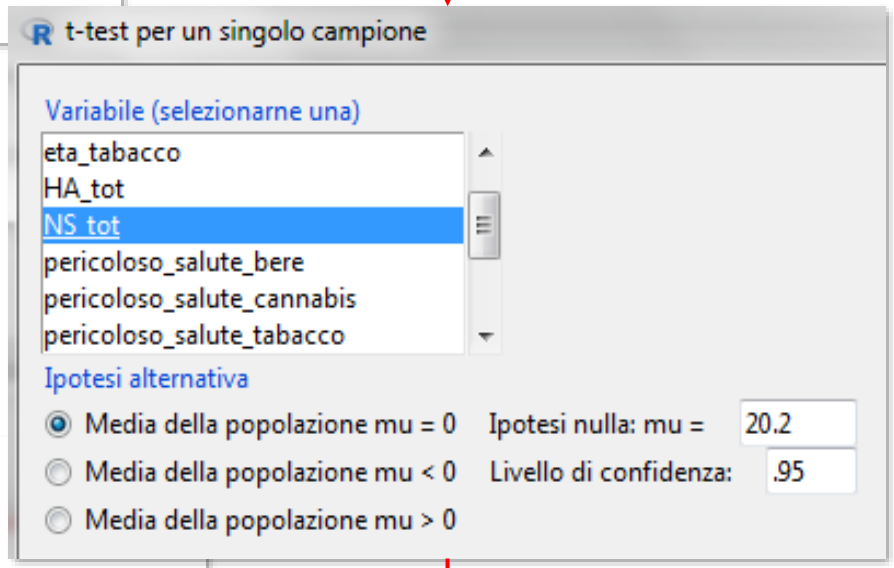
La probabilità che nella NS di campione e popolazione normativa si verifichi una differenza pari a 2.74 è $<.001$ ($t_{1268} = -20.8, p < .001$), quindi molto bassa e inferiore alla convenzionale soglia α: è **probabile** che il **campione appartenga a una diversa** popolazione.

Ovvero: con il 95% di verosimiglianza, la media della popolazione cui appartengono gli studenti varia tra 17.2 e 17.7 punti: il **valore previsto da H_0** ($\mu = 20.2$) **non è compreso nel CI**, quindi è **probabile** che la **popolazione** da cui è stato estratto il campione **non sia la stessa** dei soggetti reclutati dall'autore del test per comporre la popolazione normativa. Il CI è ristretto: la **precisione della stima è buona.**

Tuttavia: il coefficiente di ES $d = |.584|$ suggerisce che la differenza tra campione e popolazione sia modesta: la sua significatività può probabilmente essere almeno parzialmente dipendente dalla numerosità del campione. Gli adolescenti del campione non hanno un profilo temperamentale **interpretativamente** differente da quello della popolazione.

Se volete fare $il\ t-test$ per un campione con **Rcommander**:





Si sceglie la variabile, si indica μ, si stabilisce la direzione di H_1 e la verosimiglianza del CI:

```
        One Sample t-test

data:  NS_tot
t = -20.819, df = 1268, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 20.2
95 percent confidence interval:
 17.20720 17.72267
sample estimates:
mean of x
 17.46493
```

Avete costruito i CI anche per i tratti HA e RD: applicate il t- test per campione unico anche a queste dimensioni, richiamate i grafici dei CI, calcolate i coefficienti d di Cohen e interpretate tutte queste informazioni.

Se Y **avesse una distribuzione normale**, potremmo usare i quantili z di una distribuzione di probabilità normale standardizzata, invece dei quantili t, ma il corrispondente $z\ test$ non è implementato nelle funzioni di base.

È comunque molto semplice: si calcola il quantile z dato dalla differenza tra media campionaria e μ rapportata alla stima dello SE, per poi attribuirgli il corretto $p - value$:

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

Test per un solo campione, variabili discrete

Possiamo **usare distribuzioni di probabilità non continue per verificare ipotesi** che riguardano un campione e una popolazione, se la variabile oggetto di analisi è discreta.

In `attaccamento` avevamo notato che il **genere dei caregiver** era **sbilanciato**:

```
prop.table(table(attaccamento$genere))
     F      M
0.875  0.125
```

H_0: la proporzione dei successi (**donna**) nel campione è una **fluttuazione casuale della proporzione prevista in popolazione**, in cui la proporzione dei successi è uguale alla loro probabilità di verificarsi per caso, cioè **.50**.

Quanto è probabile avere un campione con una proporzione di donne pari a .875, se il campione è rappresentativo di una popolazione in cui $p_{donne} = p_{uomini} = .50$?

H_0: $p_{donne} = .5$ H_1: $p_{donne} \neq .5$ H_1: $p_{donne} > .5$ H_1: $p_{donne} < .5$

Per **variabili categoriali dicotomiche**, si usa il **test della binomiale**: `binom.test(x, n, p)`, in cui `x=` numero successi, `n=` numerosità complessiva, `p=` probabilità **teorica** del successo. `alternative=` H_1 bidirezionale (default) o monodirezionale (`"greater"`, `"less"`).

La probabilità di estrarre una proporzione di donne $\geq 87.5\%$ *da una popolazione in cui* $p_{donna} = .5$ *è*

$p = 0.00000138$

```
binom.test(x = 35, n = 40, p = .50)
        Exact binomial test
data:  35 and 40
number of successes = 35, number of trials = 40, p-value = 1.383e-06
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.7319671 0.9581404
sample estimates:
probability of success
0.875
```

H_1 bidirezionale

Proporzione dei successi nel campione

95%*CI* ***della proporzione attesa di donne*** *nella popolazione da cui è tratto il campione: con il 95% di verosimiglianza, nella popolazione di caregiver possiamo attenderci tra il 73.2% e il 95.8% di donne. CI un po' ampio e* ***non simmetrico****;* ***il valore atteso da*** $H_0 = .50$ ***non è compreso nel CI***

Attenzione

Abbiamo già usato la **funzione di ripartizione** della distribuzione di probabilità binomiale per calcolare la probabilità cumulata **da un certo quantile in su** (`lower.tail= FALSE`). La logica è esattamente la stessa del test della binomiale, ma con una **differenza essenziale**:

La funzione di ripartizione con `lower.tail= FALSE` calcola la **probabilità cumulata di ottenere un quantile maggiore di x**, cioè $P(X > x)$

`binom.test(alternative= "g")` calcola la probabilità di cumulata di ottenere un **quantile uguale o maggiore di x**, cioè $P(X \geq x)$, secondo la logica della verifica di H_0

```
pbinom(q = 35, size = 40, prob = .5, lower.tail = F)
[1] 9.285122e-08
```

```
binom.test(x = 35, n =40, p = .5,alternative = "g")
        Exact binomial test
data:  35 and 40
number of successes = 35, number of trials = 40, p-value = 6.913e-07
```

```
pbinom(q = 34, size = 40, prob = .5, lower.tail = F)
[1] 6.91306e-07
```

Se calcolassimo "a mano" il CI della proporzione con la formula di Wald, scopriremmo limiti leggermente diversi:

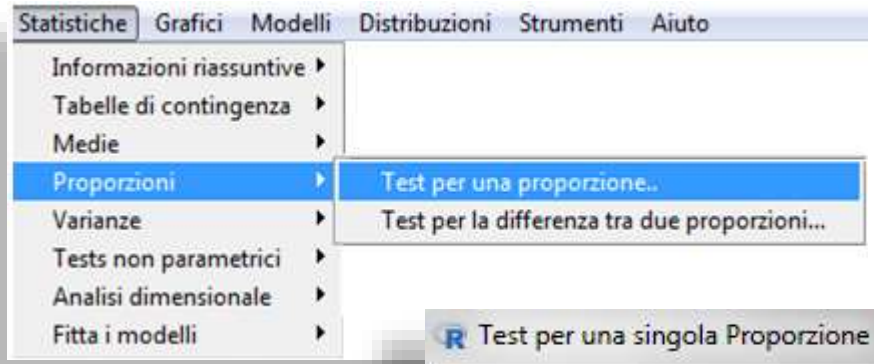
$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

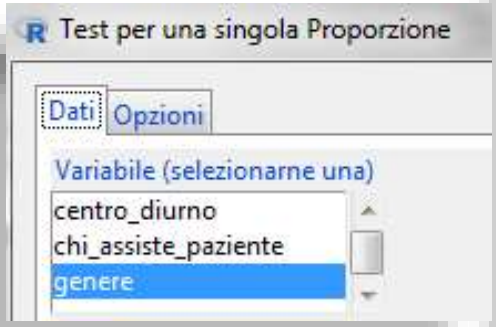
`binom.test` usa il **metodo esatto di Clopper-Pearson**, raccomandabile in caso di eventi molto rari - $\hat{p} < 1$- o molto comuni - $\hat{p} >$.99. `BinomCI(x=, N=, method=)` di `DescTools` è flessibile: `method = "wilson"` (di default), `"wald"` e `"clopper-pearson"`

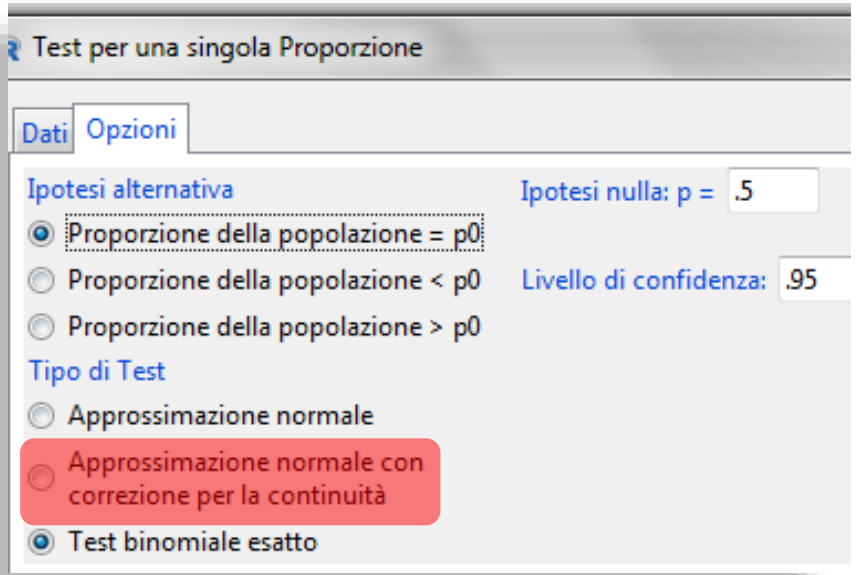
```
> BinomCI(x = 35, n = 40, method = "wilson")
      est    lwr.ci   upr.ci
[1,] 0.875 0.7388788 0.945405

> BinomCI(x = 35, n = 40, method = "wald")
      est   lwr.ci   upr.ci
[1,] 0.875 0.772511 0.977489

> BinomCI(x = 35, n = 40, method = "clopper-pearson")
      est    lwr.ci    upr.ci
[1,] 0.875 0.7319671 0.9581404
```







Se volete usare `Rcommander`:

*vedremo nel test chi quadrato **a due vie** cosa sia la correzione per la continuità*

Considerate tutto il campione degli adolescenti: stabilite se ci sono più studenti che si ubriacano e che fumano sigarette di quelli attesi in base al caso.

Poi dividete il campione in minorenni (fino a 17 anni) e maggiorenni (da 18 anni) e rifate la stessa valutazione nei due subset: cosa potete commentare?

Per **variabili categoriali** che **non seguono distribuzioni binomiali**, si verifica se le **frequenze delle categorie** nel campione si **presentino con una distribuzione** affine a quella determinata dal solo **caso** (H_0) o che invece **almeno** una delle categorie **mostri più o meno osservazioni** rispetto a quelle prevista dal caso (H_1 bidirezionale).

H_0: la **forma della distribuzione** dei dati è **rettangolare**: tutte le categorie si manifestano con la medesima frequenza → la diversità delle frequenze è solo una fluttuazione casuale

H_1: la **forma della distribuzione** dei dati **non** è **rettangolare**: categorie di eventi si presentano con frequenze non casualmente diverse

Si usa il **test del chi quadrato χ^2 a una via** (**o test di bontà dell'adattamento**; Pearson, 1900; Fisher, 1922), applicato a una **tabella di contingenza con una sola riga e K colonne (categorie)**, in cui le osservazioni sono **indipendenti**: cadono all'interno di una cella O di un'altra cella. Usa la **distribuzione di probabilità χ^2** per attribuire un $p - value$ al risultato.

$$\chi^2_{k-1} = \sum \frac{(O-A)^2}{A}$$

N/k

differenza tra la frequenza empirica **osservata** (O) in ogni cella e la **frequenza attesa** in ogni cella in base al solo caso (A): **residui di cella** (**cell residuals**).

Se H_0 **è vera**, lo **scarto in ogni cella è** $O - A = 0$, o comunque molto piccolo. Sommando i residui, quantificheremmo la **distanza dei dati dallo 0 previsto da** H_0 **- modello** → test di bontà di adattamento/ **goodness of fit**.

Però: $\Sigma_{residui} = 0$ → eleviamo i residui al quadrato, e **rapportiamo ogni** res^2 **ad** A, prima di sommarli, con una **sorta** di "standardizzazione": i res^2 sono interpretabili come il numero di valori teorici compresi nello scarto. La $\sum \frac{(O-A)^2}{A}$ si distribuisce come un **quantile** di una **distribuzione** χ^2, **con** $df = k - 1$

Una** cella deve far rispettare il vincolo **N

Generalizzando la formula:

$$\chi^2_{k-1} = \sum \frac{(osservate - modello)^2}{modello}$$

$$res_{ij} = osservate_{ij} - modello_{ij}$$

$$errore_{ij} = Y_{ij} - modello_{ij}$$

*La distribuzione dello stato civile nei caregiver (`attaccamento`, rinominato a) è **casuale** (H_0) o significativamente diversa da una casuale (H_1)?*

```
table(a$stato_civile))
   coniugato   convivente divorziato/a       single
          21             4             7             8
```

Le A sono date dal rapporto tra N e il numero di categorie k:

```
(A<-40/4)
[1] 10
```

Se la distribuzione fosse casuale (H_0), avremmo 10 caregiver per cella.

Calcoliamo gli scarti da A in ogni cella, li eleviamo al quadrato, li rapportiamo ad A e li sommiamo per ottenere la statistica χ^2 :

$$\chi^2_{k-1} = \sum \frac{(O - A)^2}{A}$$

```
(chi2<-(21-10)^2/10+(4-10)^2/10+(7-10)^2/10+(8-10)^2/10)
[1] 17
```

Stimiamo **la probabilità di ottenere un quantile $\chi^2 = 17$ o uno più grande** sotto condizione di H_0, cioè assegniamo un $p - value$ a $\chi^2 = 17$ *[beh, a uno immediatamente superiore, ma χ^2 è continua, la differenza è impercettibile]*, per $df = 3$

```
pchisq(q = 17, df = 3, lower.tail = FALSE)
[1] 0.0007067424
```

Rifiutiamo H_0: la distribuzione dello stato civile tra i caregiver non è casuale

D'ora in poi, useremo `chisq.test(table(frequenze osservate))`:

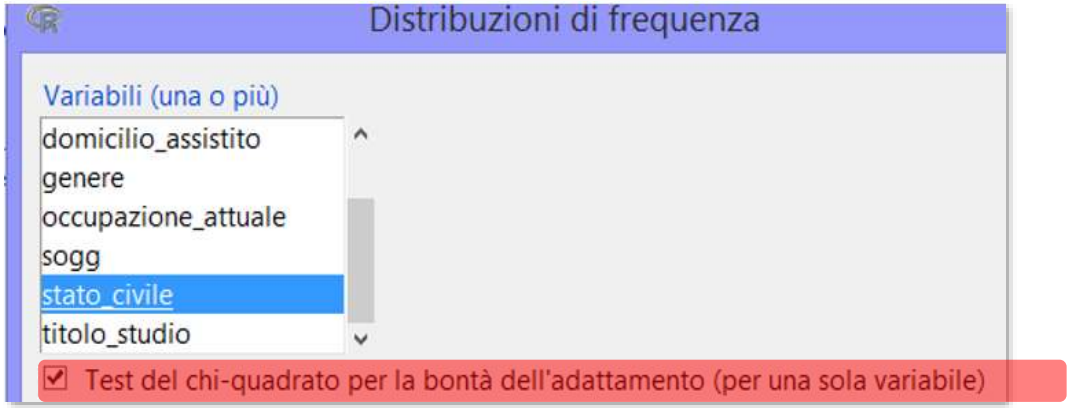
```
chisq.test(x= table(a$stato_civile))
        Chi-squared test for given probabilities
data:  table(a$stato_civile)
X-squared = 17, df = 3, p-value = 0.0007067
```

Oppure `Desc`: applicata a una tabella di contingenza, produce descrizione, grafico e **il test χ^2**:

```
Desc(table(a$stato_civile))
-------------------------------------------------
table(a$stato_civile) (table)
Summary:
n: 40, rows: 4
Pearson's Chi-squared test (1-dim uniform):
  X-squared = 17, df = 3, p-value = 0.0007067

          level  freq   perc  cumfreq  cumperc
1     coniugato    21  52.5%       21    52.5%
2    convivente     4  10.0%       25    62.5%
3  divorziato/a     7  17.5%       32    80.0%
4        single     8  20.0%       40   100.0%
```

Con `RCommander`, scegliete `Statistiche → Informazioni riassuntive → Distribuzioni di frequenza`:



Test per un solo campione, ipotesi sulla forma

Se l'interpretazione del $\boldsymbol{Q-Q\ plot}$ non basta per stabilire se la distribuzione campionaria è affine a una normale teorica, si può usare **anche** un test inferenziale: tra i molti in letteratura, vedremo **per ora** il test $\boldsymbol{W}$ di **Shapiro-Wilks**.

Il test W valuta una **relazione** di **regressione** (**predizione**) tra i **valori osservati** e i corrispondenti **quantili** di una distribuzione normale: **partendo dal quantile osservato**, possiamo predire il **corrispondente quantile di una distribuzione nomale**? Se **sì**, la correlazione al quadrato (**R^2, ovvero W**) tra valori osservati e quantili della normale **tende a 1** e la distribuzione campione è **normalmente** distribuita. Se **no**, la W **tende a 0** e la distribuzione campionaria **non** è **normalmente** distribuita.

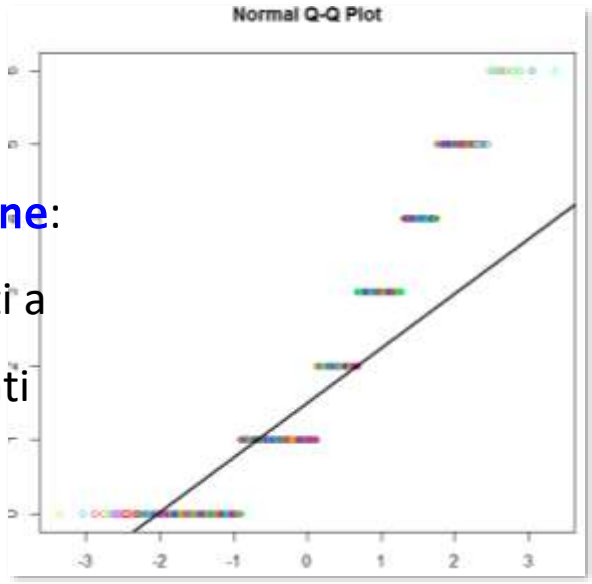
H_0: $p > .05$ → la **differenza tra W e 1 non è significativa**, la distribuzione **è** distribuita in modo affine alle **normale**

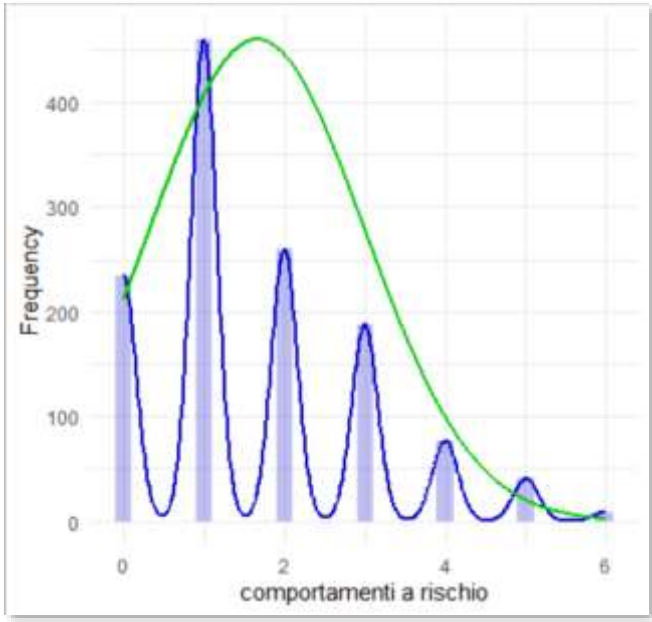
H_1: $p < .05$ → la **differenza tra W e 1 è significativa**, la distribuzione **non** è distribuita in modo affine alle **normale**

In R usiamo

`shapiro.test(distribuzione`:

applichiamola ai comportamenti a rischio ammessi dagli adolescenti






```
shapiro.test(ad$comportamenti_rischio)
        Shapiro-Wilk normality test
data:  ad$comportamenti_rischio
W = 0.89019, p-value < 2.2e-16
```

W è significativamente $\neq 1$: la forma di `$comportamenti_a_rischio` **non si sovrappone alla normale**. Perché W sia valutato come $\cong 0$, deve **davvero** tendere a 1, in genere $>.95$...

Attenzione alla potenza: con **grandi N** , anche piccole deviazioni dalla normale possono creare un risultato significativo. È **sempre** indicato affiancare un grafico al test.

Considerate tutto il campione degli adolescenti: stabilite se ci sono più studenti che si ubriacano e che fumano sigarette di quelli attesi in base al caso.

Poi dividete il campione in minorenni (fino a 17 anni) e maggiorenni (da 18 anni) e rifate la stessa valutazione nei due subset: cosa potete commentare?

Aggiunta n. 1: distribuzioni di probabilità centrali e non centrali

Distinguiamo **distribuzioni di probabilità centrali e non centrali**: quando si parla di **distribuzione di probabilità senza ulteriori specificazioni**, intendiamo in effetti distribuzione di probabilità **centrale**.

La distribuzione di **probabilità centrale** rappresenta come si distribuisce la statistica di un test **assumendo che H_0 sia vera** in popolazione

quanto è probabile che la differenza nel punteggio di ansia tra un gruppo di donne ($\bar{x}_D = 40.5$) e uno di uomini ($\bar{x}_U = 38.5$) sia $\Delta = 2$, se D e U provenissero dalla stessa popolazione ($\Delta = 0$)?

La forma della distribuzione di probabilità dipende **dai gradi di libertà**

La distribuzione di probabilità **non centrale** rappresenta come si distribuisce la statistica di un test **quando H_1 è vera** in popolazione

quanto è probabile che la differenza nel punteggio di ansia tra un gruppo di donne ($\bar{x}_D$=40.5) e un gruppo di uomini ($\bar{x}_U$=38.5) risulti $\Delta = 2$, se provenissero da popolazioni con $\bar{x}_D$= 45 e con $\bar{x}_U$=35 ($\Delta = 10$)?

La forma della distribuzione di probabilità dipende dai **gradi di libertà e del parametro di non centralità NCP**

Il **parametro di non centralità NCP** è **derivato dalla numerosità campionaria N e dall'effect size atteso**.

Per il confronto di due gruppi si usa la distribuzione di probabilità t, per $df = N - 2$, il cui parametro di centralità è :

$$t: ncp_t = \sqrt{\frac{N}{2}} \times effect\ size$$

Il NCP è l'argomento opzionale `ncp= valore` delle funzioni `pt`, `pf`, `pchisq`: di default è = 0 → la differenza attesa tra le medie in popolazione è = zero, cioè quanto previsto da H_0.

*Quanto è verosimile trovare una **differenza** tra due gruppi ($n_1 = 19$, $n_2 = 18$: $N = 37$) **superiore a** $d = 13$, se **provengono dalla stessa popolazione** → Δ=0*

```
pt(q = 13,df = 35,lower.tail = FALSE)
[1] 2.930281e-15
```

```
pt(q = 13,df = 35,lower.tail = FALSE, ncp=0)
[1] 2.930281e-15
```

*Quanto è verosimile trovare una **differenza** tra due gruppi ($n_1 = 19$, $n_2 = 18$: $N = 37$) **superiore a** $d = 13$, se provengono da **popolazioni diverse**, tali per cui NCP è:*

```
pt(q = 13,df = 35,lower.tail = FALSE, ncp=5)
[1] 2.717007e-06
```

```
pt(q = 13,df = 35,lower.tail = FALSE, ncp=12)
[1] 0.3146047
```

```
pt(q = 13,df = 35,lower.tail = FALSE, ncp=20)
[1] 0.9998951
```

Aggiunta n. 2: oggetti di classe `htest`

Se **salviamo come oggetti i prodotti di** `t.test`**,** `chisq.test` (e molte altre), otteniamo oggetti di classe `htest` (hypothesis test). Altri test producono oggetti di classe `lm` o `glm`. Tutti sono **liste**, composte dagli elementi del test: `test$values`. Alcuni elementi sono forniti nell'output, altri si scelgono nella lista dei `test$values`. Qui vediamo gli elementi di `t.test`, quelli di `chisq.test` li vedremo nel test del χ^2 a **due** vie, prossimo argomento.

```
> NS<-t.test(a$NS, mu = 20.2)
> class(NS)
[1] "htest"
> NS$
```

statistic
parameter
p.value
conf.int
estimate
null.value
stderr

conf.int
estimate
null.value
stderr
alternative
method
data.name

Gli elementi inseriti nell'output del `t.test` *sono tutti disponibili nella lista:* `$statistic` *(il quantile t),* `$parameter` *(i df),* `$p.value`*,* `$conf.int` *(il CI),* `$estimate` *(media nel campione),* `$alternative` *(direzione di* H_1*),* `$null.value` *(valore previsto da* H_0*),* `$method` *(tipo di t test);* `$data.name` *(variabile)*

```
> NS[1]
$statistic
        t
-20.81887
```

```
> NS[2]
$parameter
  df
1268
```