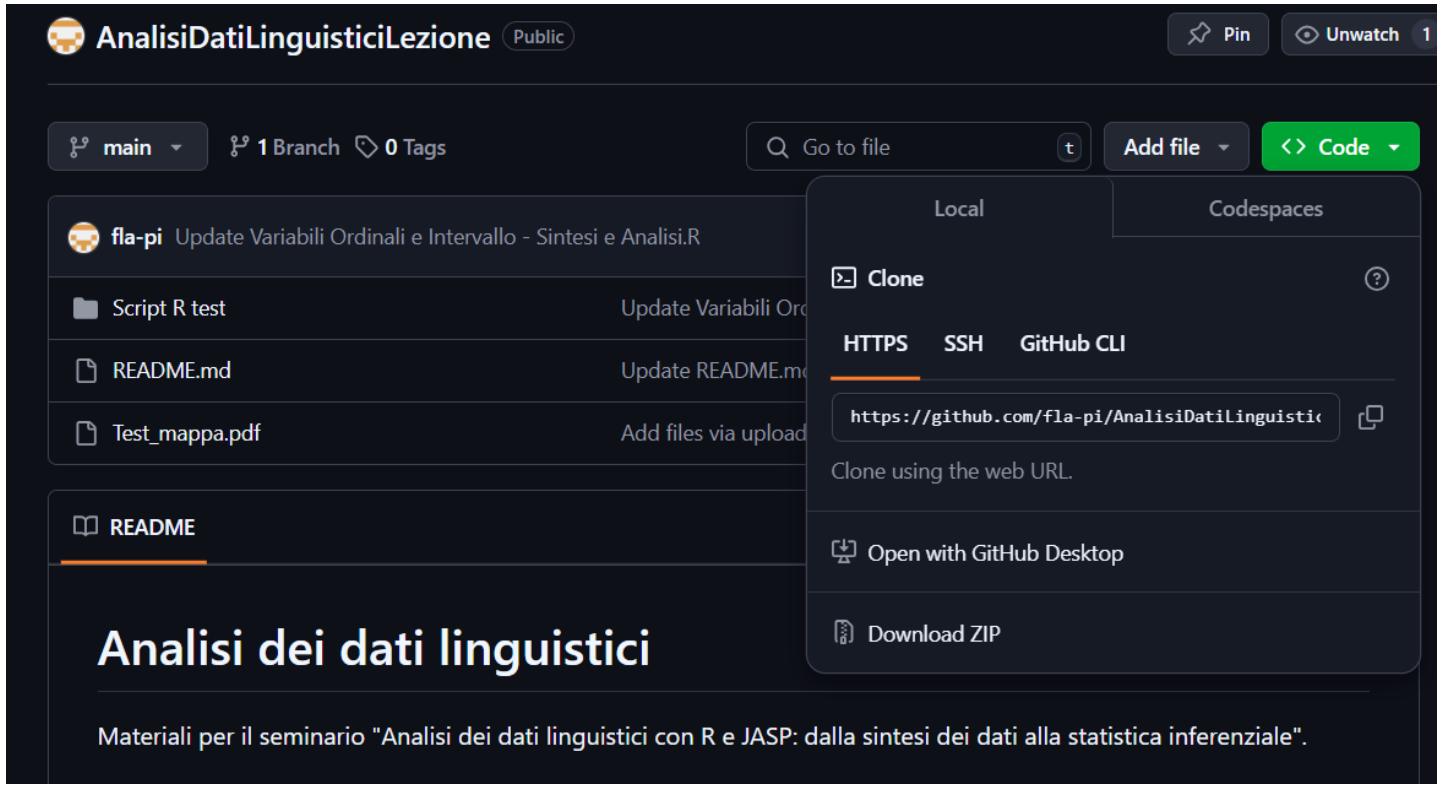
Analisi dei dati linguistici con R e JASP:
dalla sintesi dei dati alla statistica inferenziale

Flavio Pisciotta

Università degli Studi di Salerno

Software e dati

github.com/fla-pi/AnalisiDatiLinguisticiLezione/



Software e dati

In omaggio nella repository:

Per oggi:
- Tre dataset da Stefanowitsch, A. 2020, *Corpus linguistics: A guide to the methodology*
- Tre file .R con script per test e grafici
- Queste slide

Per sempre:
- *Decision tree* per scegliere i test statistici
- *Cheatsheet* per R
- *Cheatsheet* per JASP

Che cosa faremo oggi?

1. Che cosa NON faremo oggi?
2. Che cos'è la statistica? (e perché ci serve?)
3. Statistica descrittiva e statistica inferenziale
4. Come impostare un'analisi quantitativa
   - Quali tipologia di dati vogliamo analizzare? I tipi di variabili
   - Che cosa vogliamo testare?
5. Dalla sintesi dei dati alla statistica inferenziale
   - Variabili continue
   - Variabili categoriali
   - Variabili ordinali

Che cosa NON faremo oggi?

- Diventare maghi della statistica
- Diventare programmatori provetti
- Imparare nozioni statistiche avanzate

piuttosto…

- Cercare di assorbire la logica dietro l'analisi quantitativa

→ sapere che cosa è possibile fare (e quando ha senso farlo) per poterlo cercare consapevolmente su Google

→ Utilizzare strumenti *user-friendly* (perché già fare ricerca è difficile di suo…)

Che cosa faremo oggi?

- All’incirca:
  - Prima ora: Nozioni di base di statistica
  - Seconda ora: Introduzione a R
  - Terza ora: Mani sul dataset!
  - Quarta ora: Qualche accenno a JASP

Che cos'è la statistica?

- *Scienza che ha per oggetto lo studio dei fenomeni collettivi suscettibili di misurazione e di descrizione quantitativa (spec. quando il numero degli individui interessato è talmente elevato da escludere la possibilità o la convenienza di seguire le vicende di ogni singolo individuo)*[1]

  ...in condizioni di incertezza o non determinismo, cioè di non completa conoscenza di esso o di una sua parte.[2]

- Il linguaggio è un fenomeno collettivo in tutte le sue manifestazioni (insieme di parlanti, insieme di enunciati)
- Difficilmente troviamo leggi deterministiche nel linguaggio
- Necessità di analisi quantitative (non introspettive, non aneddotiche) per verificare la presenza o meno di un fenomeno

[1] Treccani
[2] Wikipedia

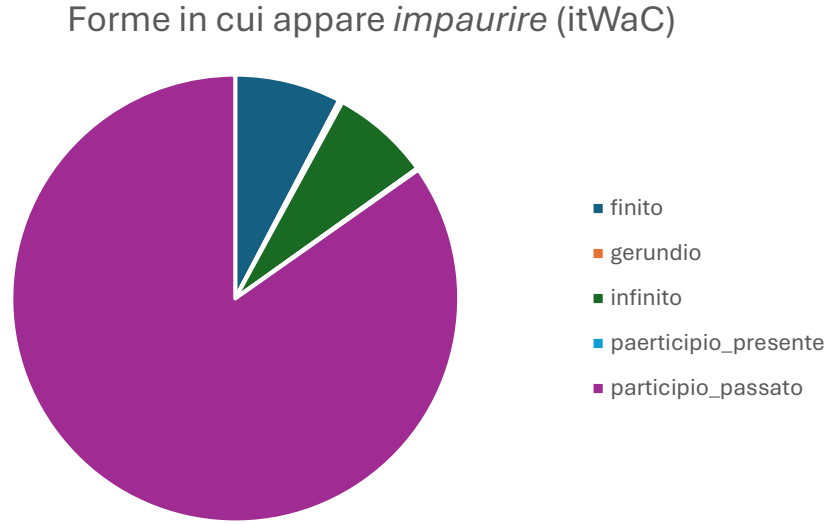
Perché ci serve la statistica?

1. Riassumere efficacemente una serie di informazioni e caratteristiche del campione che studiamo, evitando il più possibile il *cherry picking*

Il verbo *impaurire* compare perlopiù come participio passato:

1) *il cane bassotto , visibilmente impaurito, stava cercando di attraversare l' autostrada*
(itWaC)

forma verbale	n
finito	443
gerundio	15
infinito	413
participio_presente	7
participio_passato	4883



Perché ci serve la statistica?

2. Per verificare se un fenomeno osservato si verifica **significativamente** di più/di meno di quanto atteso:

Ci sono delle differenze tra i soggetti di verbo1 e verbo2?

soggetto	verbo1	verbo2
umano	55 (67%)	84 (64%)
animato	14 (17%)	30 (23%)
inanimato	13 (16%)	17 (13%)
Totale	82 (100%)	131 (100%)

Ma un test statistico ci mostra che questa variazione può essere attribuita ragionevolmente al caso

Chi-quadrato, *p-value* = 0.56

Statistica descrittiva e inferenziale

- Riprendendo i punti 1 e 2, possiamo tracciare una differenza tra statistica descrittiva e inferenziale:

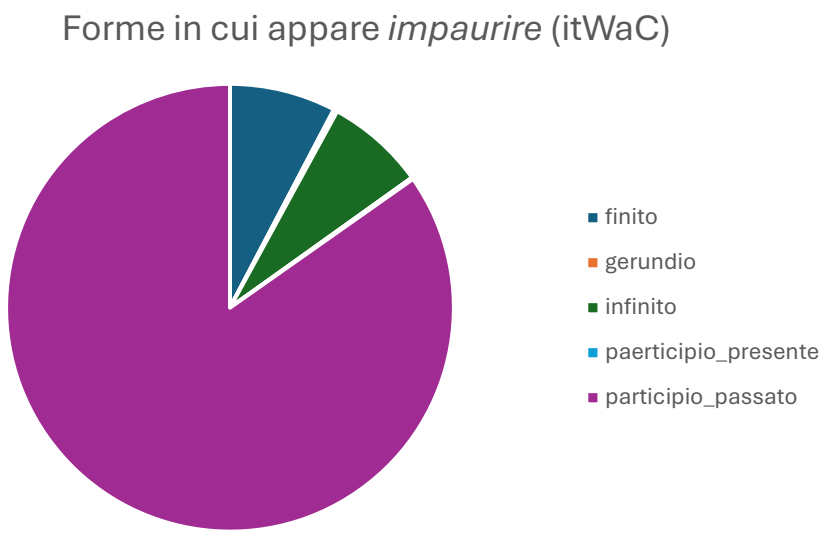
  - La statistica descrittiva si occupa della **sintesi e della rappresentazione dei dati**
  - ➢ **descrivere il nostro campione** attraverso grafici, tabelle, percentuali, medie, etc…

  - La statistica inferenziale permette di compiere inferenze/trarre conclusioni sulla popolazione a partire dal campione:
  - ➢ **esiste una differenza tra x e y nel nostro campione**, ma è abbastanza rilevante per poterla estendere a tutta la popolazione? Ci permette di trarre conclusioni generalizzabili?

Statistica descrittiva e inferenziale

In che forma verbale appare maggiormente *impaurire*?

forma verbale	n
finito	443
gerundio	15
infinito	413
participio_presente	7
participio_passato	4883



Come possiamo sapere se questa differenza è davvero rilevante?

Come impostare un'analisi quantitativa?

- Una serie di *step* formalizzati per condurre una ricerca
- Osservazione: ad es. ho notato che il verbo *impaurire* appare molto spesso al participio
- Individuare le variabili: in questo caso, l'unica caratteristica che varia è la forma verbale
- Produrre un'ipotesi! *Impaurire* appare più spesso al participio rispetto alla frequenza generale delle forme verbali in italiano

➢ Dobbiamo rigettare l'ipotesi contraria alla nostra (l'ipotesi nulla, detta anche H0)

  - H0: *Impaurire* NON appare **significativamente** più spesso al participio rispetto alla frequenza…
  - H1: *Impaurire* appare **significativamente** più spesso al participio rispetto alla frequenza…

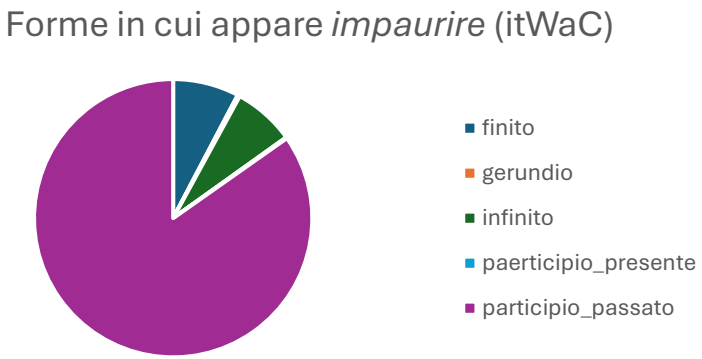
Come impostare un'analisi quantitativa?

- Raccolta dati e annotazione

word	Frequenza	verb form
impaurito	1837	participio_passato
impauriti	1341	participio_passato
impaurita	1224	participio_passato
impaurire	292	infinito
impaurite	292	participio_passato
impaurisce	175	finito
Impaurito	71	participio_passato

- possiamo rappresentare i nostri dati

forma verbale	n
finito	443
gerundio	15
infinito	413
participio_presente	7
participio_passato	4883



Come impostare un'analisi quantitativa?

- Testare la nostra ipotesi:

➢ *abbiamo un riferimento su che frequenza hanno normalmente le forme verbali in italiano?*

	occorrenze di verbi in itWaC
tempi finiti	52%
infinito	18%
gerundio	3%
participio_presente	1%
participio_passato	26%

➢ *Impaurire* si discosta significativamente da questa distribuzione? Dobbiamo applicare un test

```
chisq.test(forme_impaurire, p = c(0.52, 0.18, 0.03, 0.01, 0.26))

Chi-squared test for given probabilities data: forme_impaurire
X-squared = 11211, df = 4, p-value < 2.2e-16
```

Che cos’è questo p-value?

- Quando applichiamo un test statistico, stiamo assumendo l’ipotesi nulla come vera

(ad es. *impaurire* NON ha una distribuzione per forme verbali differente rispetto all’insieme dei verbi italiani)

- Il *p-value* rappresenta **la probabilità che la distribuzione osservata sia ottenibile mantenendo come vera l’ipotesi nulla**

Ovvero

quanto probabile è che, pur NON essendoci una differenza significativa tra le distribuzioni di impaurire e dei verbi italiani in generale, osserviamo una distribuzione come quella nel nostro campione

- Dunque, se vogliamo rigettare l’ipotesi nulla, il questa probabilità deve essere molto bassa
- La soglia convenzionale è $p < 0.05$
- Esistono più soglie di significatività: $p < 0.01$, $p < 0.001$, $p < 0.0001$

Limiti del p-value

- Non dovrei dirvelo ma…

Il p-value e la decisione della soglia di 0.05 sono talvolta oggetto di controversie, sia dal punto di vista teorico che della prassi accademica.

Questo non vuol dire che il concetto in sé è da abbandonare, ma ci dice che:

- Non solo quello che è significativo merita di essere preso in considerazione
- Non dobbiamo avere paura dei risultati non significativi: non rigettare l'ipotesi nulla è un risultato!

Come impostare un'analisi quantitativa?

- Osservazione
- Ipotesi → replicabilità!!
- Collezione dei dati
- Rappresentare/sintetizzare i dati → statistica descrittiva
- Testare se *l'ipotesi nulla* può essere rigettata → statistica inferenziale

Che cosa potremmo voler testare?

- Con i metodi statistici che presenteremo oggi, possiamo decidere di testare:

1. Se la distribuzione o i valori di **una variabile** rispettano quelli attesi

2. Se esiste una relazione tra **due variabili (correlazione, associazione)**

   - Nel caso dell'associazione, testiamo **se il valore di una variabile (detta dipendente) dipende dal valore di un'altra variabile (detta indipendente)**

ad es. se la produzione di una variante fonetica (dipendente) dipende dalla provenienza dei parlanti (indipendente)

   - Questo ci permette di trovare **delle differenze** tra due o più gruppi

Gioco: trova la variabile dipendente

- L'individuazione della variabile dipendente e indipendente dipendono dalla nostra ipotesi teorica di partenza…cos'è che influenza cosa?

Studio la relazione tra l'età dei parlanti e la loro F0

Studio la relazione tra i generi testuali e la presenza di frasi subordinate

Studio la relazione tra un'alternanza sintattica
(ad es. frasi non marcate vs frasi con dislocazione a sinistra)
e l'animatezza dei soggetti nelle due costruzioni

Quali tipi di variabili ci sono?

- Distinzione tra variabili dipendenti e indipendenti → relativa all'ipotesi
- Come sono misurate le variabili? → natura del dato

- Tre tipi di scale di misurazione delle variabili:
  - *Ratio*
  - Intervallo/Ordinali
  - Nominali

Quali tipi di variabili?

- **Variabili *ratio***

➢ si tratta di variabili quantitative in cui i valori possono essere sottoposti a delle operazioni algebriche (misure fisiche):

misure acustiche, tempi di reazione (ms), n. di sillabe in una parola, età

- **Variabili intervallo/ordinali**

➢ si tratta di variabili in cui i valori sono ordinati su una scala arbitraria (*rank*)

Intervallo: i gradi della scala sono equidistanti vs

Ordinali: non conosciamo la distanza tra i gradi della scala

➢ Esempi di variabili ordinali sono: scale di valutazione (ad es. Likert), livello CEFR

Quali tipi di variabili?

- **Variabili Nominali**

➢ si tratta di variabili in cui i valori sono categorie discrete assegnate in base ai degli attributi/proprietà

ad es. POS, classe semantica, animatezza, provenienza dei parlanti, genere

- Il tipo di una variabile può dipendere da **come la annotiamo e concettualizziamo**

➢ l'animatezza, se pensata come una scala ordinata, può essere ordinale

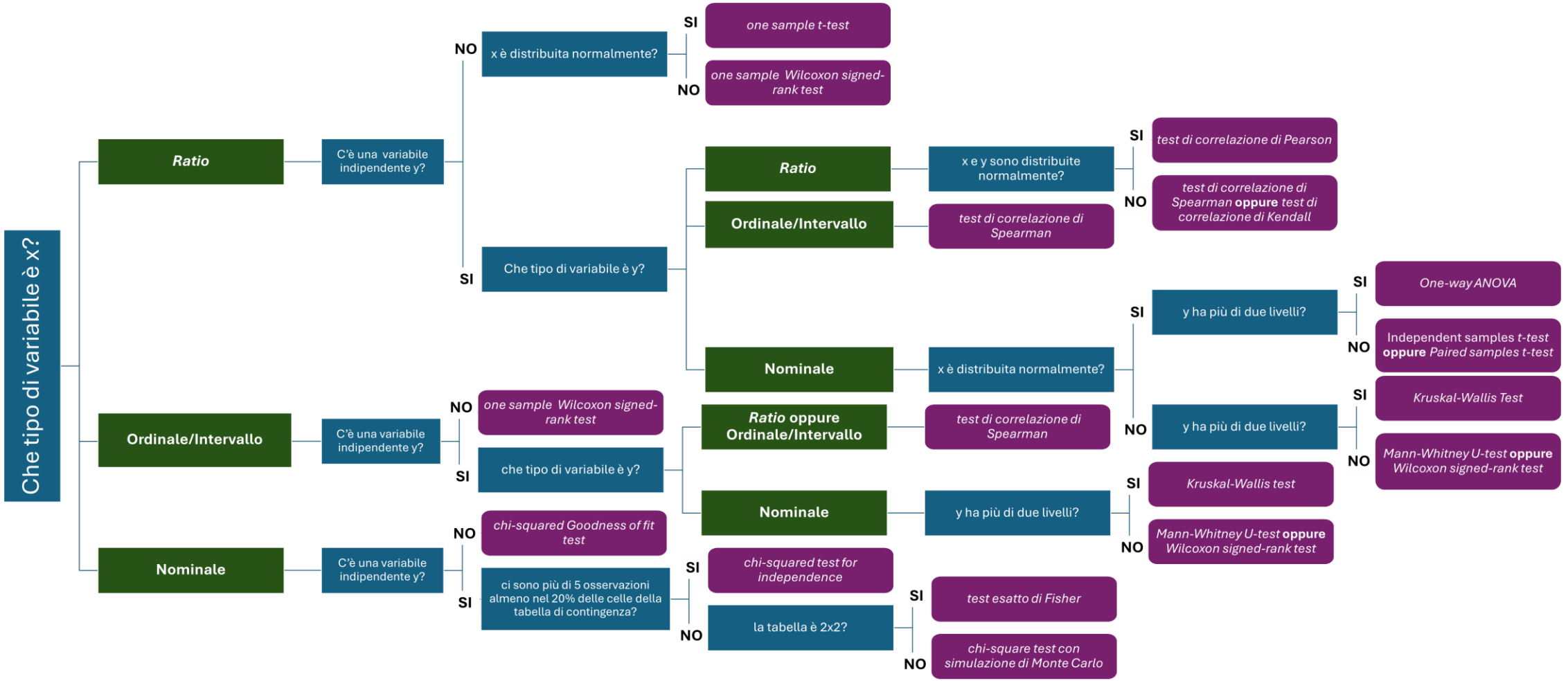
➢ l'età può essere annotata per fasce: in tal caso può essere ordinale o categoriale

Quali test per quali variabili?

- Abbiamo almeno tre tipi di informazioni sulle variabili in uno studio:
  - Quante variabili ci sono
  - Di che natura sono queste variabili
  - Qual è la variabile dipendente e quale indipendente?
- Questo ci aiuta a decidere come rappresentarle e quali test applicare

Dalla sintesi all'inferenza
una (non)veloce carrellata

Quali test per quali variabili?



La scelta del test non dipende solo dalla natura delle variabili, ma anche dalle assunzioni dei test → test non parametrici

Variabili nominali

- **Sintesi**

➢ Generalmente, il miglior modo per riportare dei dati relativi a categorie è mostrarne le frequenze in una tabella di contingenza

	cxn_attiva	cxn_passiva
Subj_animato	50	22
Subj_inanimato	34	46

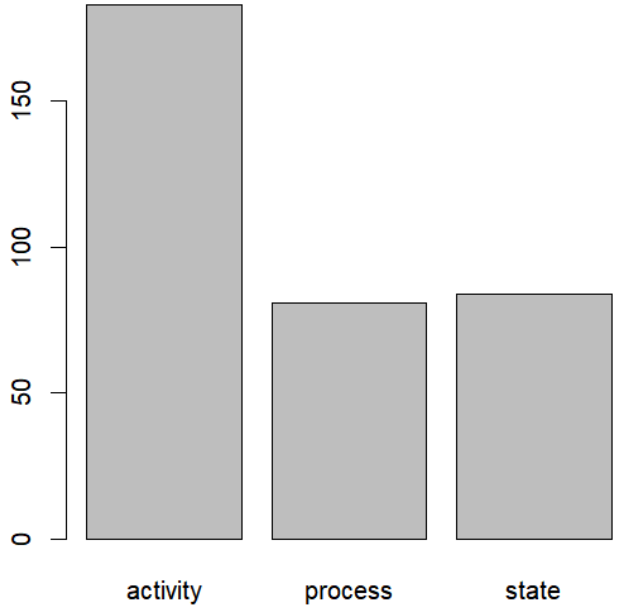
➢ ...mostrando anche (ma non solo) le percentuali

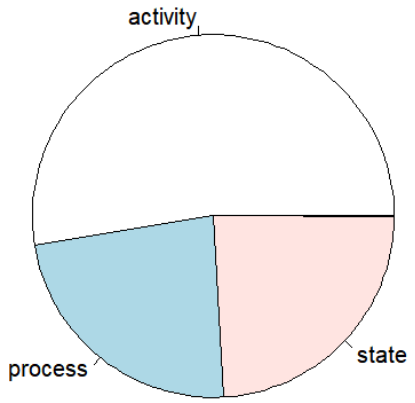
	cxn_attiva	cxn_passiva
Subj_animato	59,5%	32,4%
Subj_inanimato	40,5%	67,6%

Variabili nominali

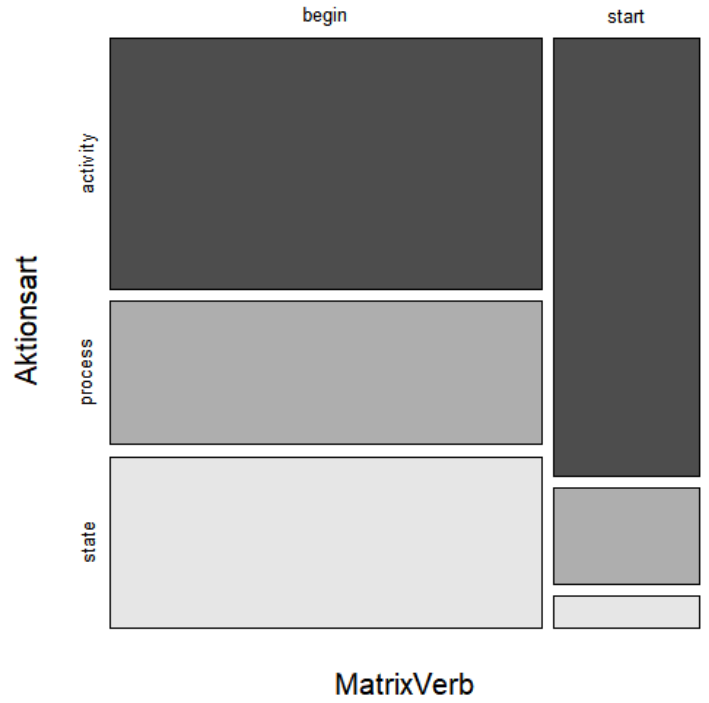
- **Rappresentazione**

  - Grafici a torta (per una variabile),
    Grafici a barre (per una o più variabili)...





  - Mosaicplot (quando si incrociano due variabili)



Variabili nominali

- **Test**

➢ Una sola variabile (vs una distribuzione attesa): *chi-square Goodness of fit test*

➢ Due variabili: dipende! Generalmente il *chi-square test, ma*...

	cxn_attiva	cxn_passiva
Subj_animato	50	22
Subj_inanimato	34	2

← Più del 20% delle celle ha meno di 5 osservazioni

➢ Test non parametrici: Test esatto di Fisher se è una tavola 2x2, mentre se è più ampia conviene utilizzare il Chi-square con simulazione del p-value (buona approssimazione del test di Fisher)

Variabili nominali

- **Test**

➢ Quando una tabella di contingenza è più grande di una 2x2 come facciamo a sapere quali livelli delle due variabili sono associati significativamente?

	activity	process	state
start	115	66	79
begin	68	15	5

➢ Residui standardizzati → in sostanza il contributo di ogni cella alla significatività del test (+/-2)

	activity	process	state
start	-5.36	1.60	4.68
begin	5.36	-1.60	-4.68

Variabili ratio

- **Sintesi dei dati**

- Generalmente, per presentare una variabile *ratio,* si indica un valore di tendenza centrale
  - misure di tendenza centrale → media, mediana, moda

- La misura generalmente utilizzata è la **media**

  (somma dei valori della variabile/numero di elementi nella variabile)

- Ma la media da sola non basta…

Non mi fido molto delle statistiche,
perché un uomo con la testa nel forno acceso e i piedi nel congelatore
statisticamente ha una temperatura media. (Charles Bukowski)

Ci tocca insegnare la statistica a Bukowski…

Variabili ratio

- **Sintesi dei dati**

- È importante riportare le **misure di dispersione**
  - **quanto si discostano le singole osservazioni dalla media**

  - varianza, ma soprattutto **deviazione standard**

St.dev Città 1 = 11.12
St. dev Città 2 = 3.15

media Città 1 = 10.5 °
media Città 2 = 9.8 °

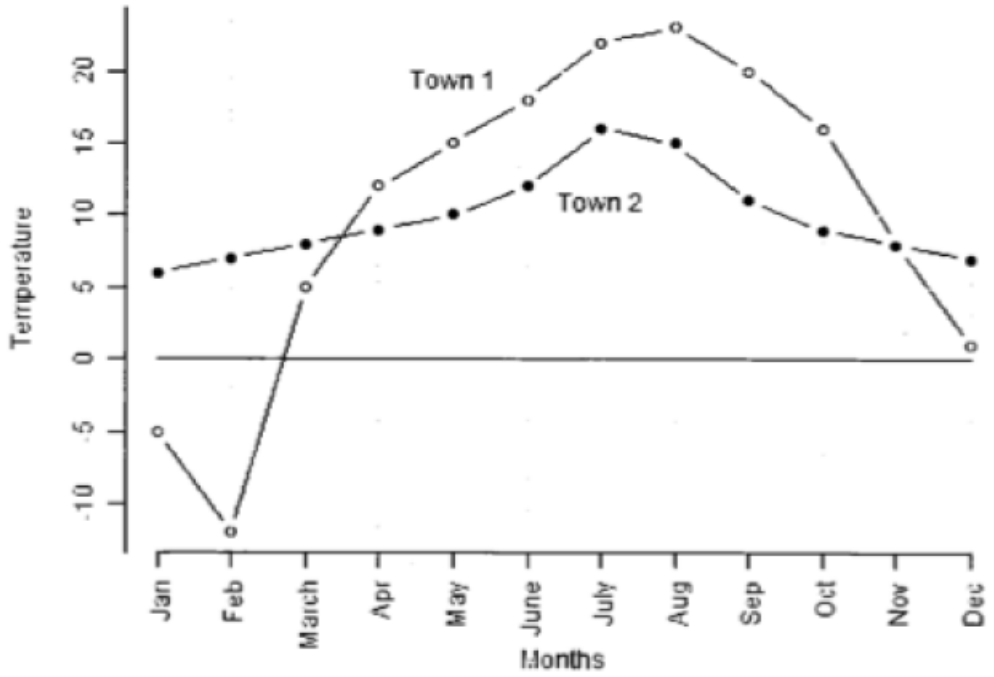
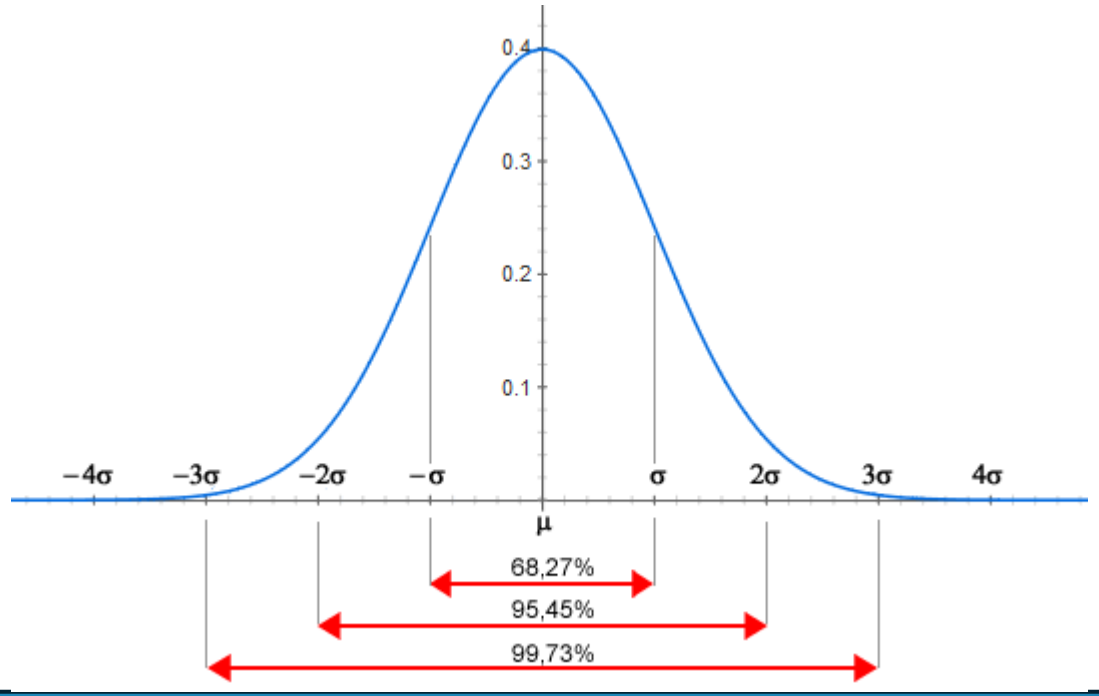


Figure 29. Temperature curves of two towns

(da Gries 2009: 111)

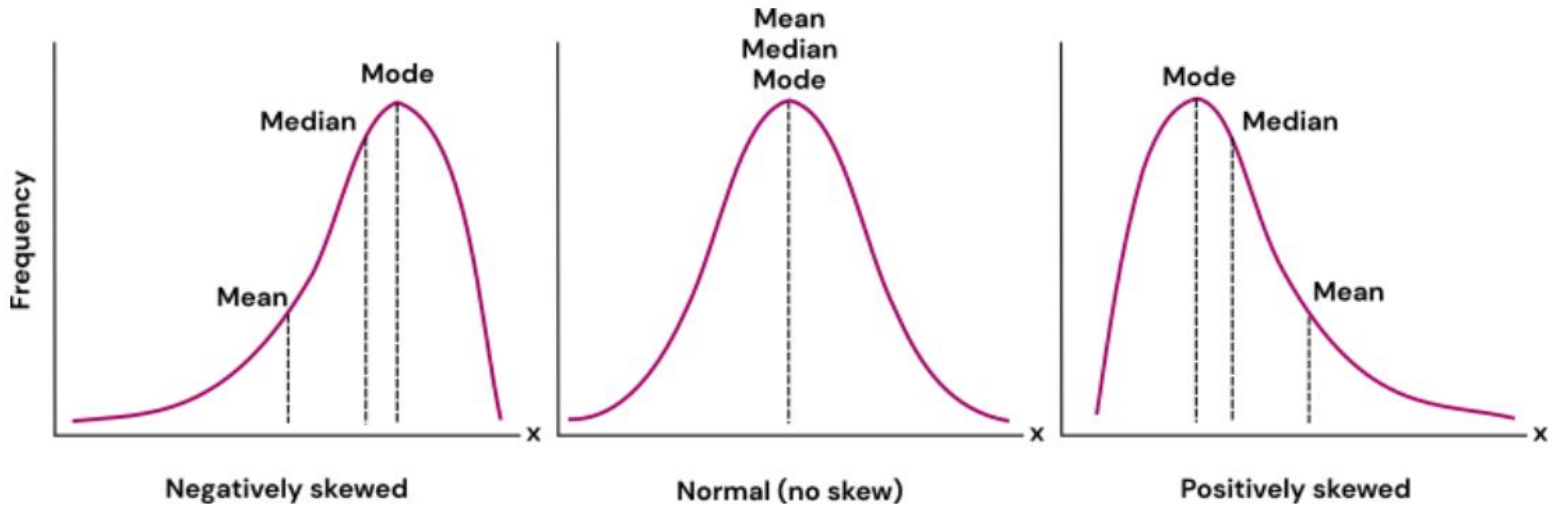
Variabili ratio

- Piccolo passo indietro → Distribuzione normale
- approssimazione per descrivere variabili i cui valori tendono a concentrarsi attorno a un singolo valore medio → la maggior parte dei dati tende a essere intorno alla media



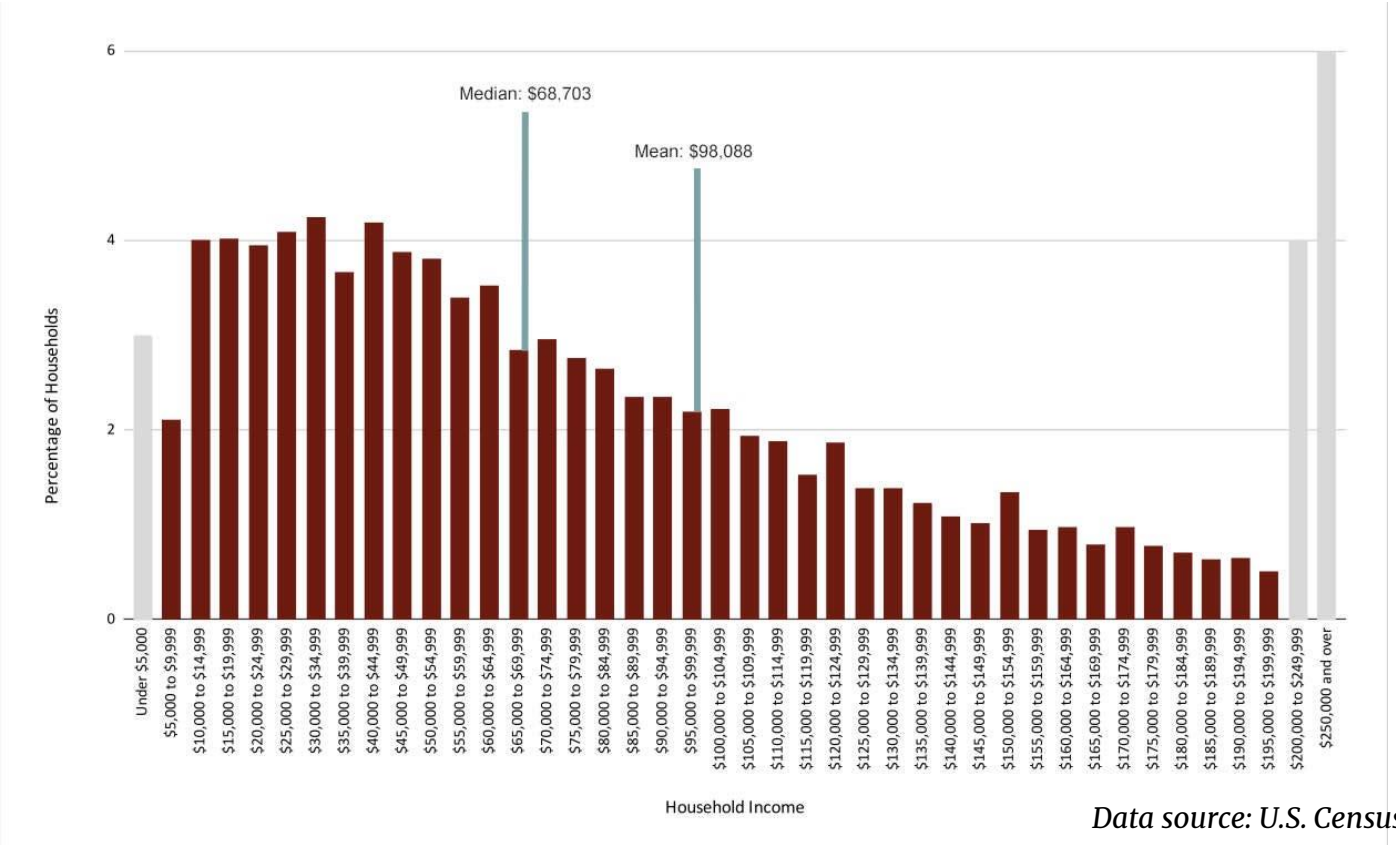
Variabili continue

- Perché ci importa?



Variabili continue

- Perché ci importa?

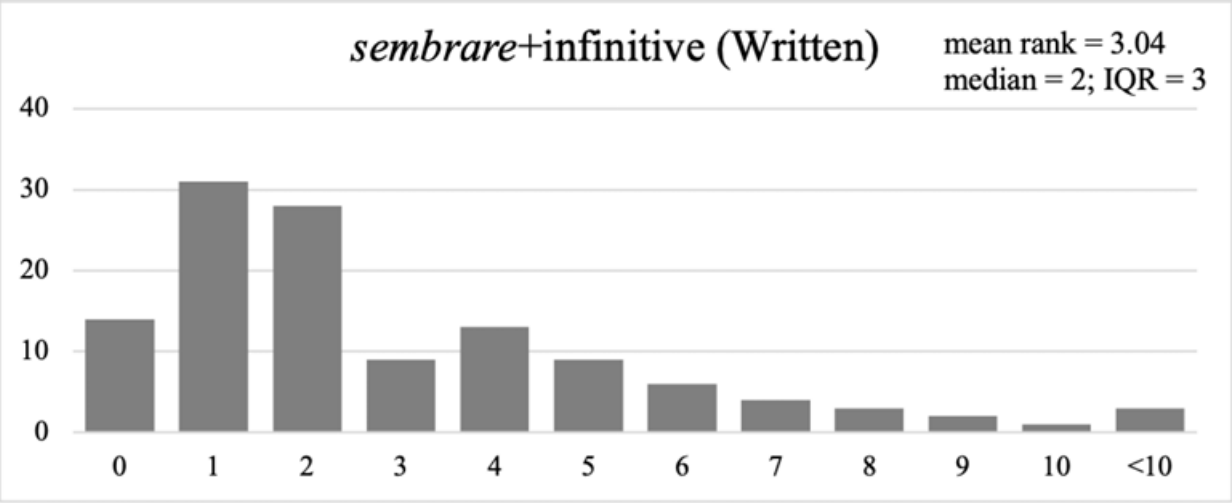


Data source: U.S. Census Bureau, Annual Social and Economic Supplement (2019)

Variabili ratio

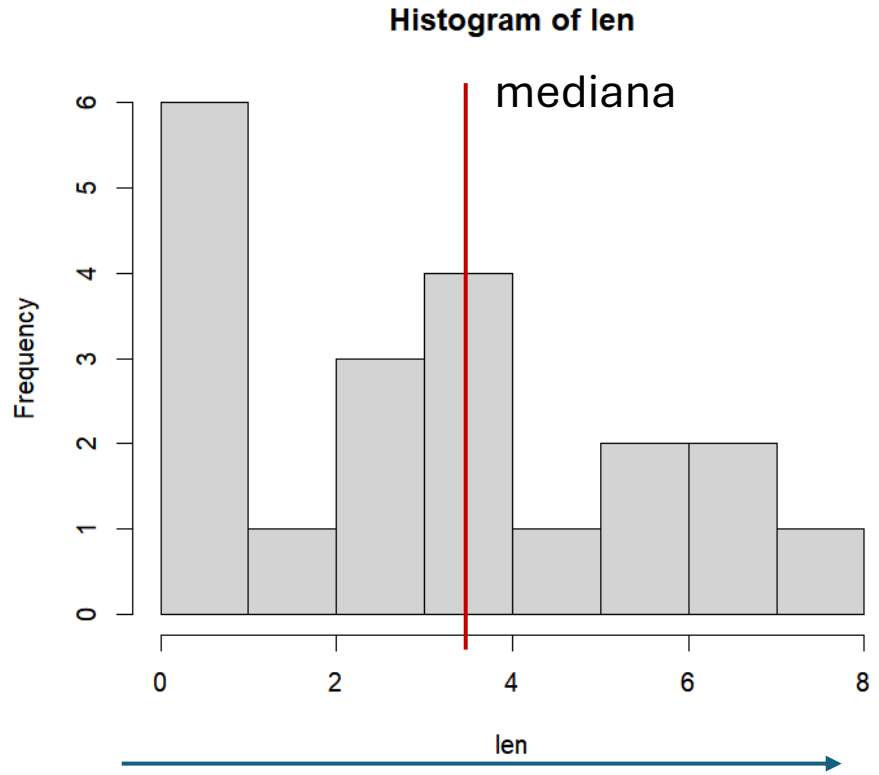
- **Sintesi dei dati**
- Quindi, se la nostra variabile non ha una distribuzione normale?

  ad es. la lunghezza in parole degli NP soggetto in una costruzione
- È più rappresentiativo usare la **mediana**! (e il range interquartile come misura di dispersione)

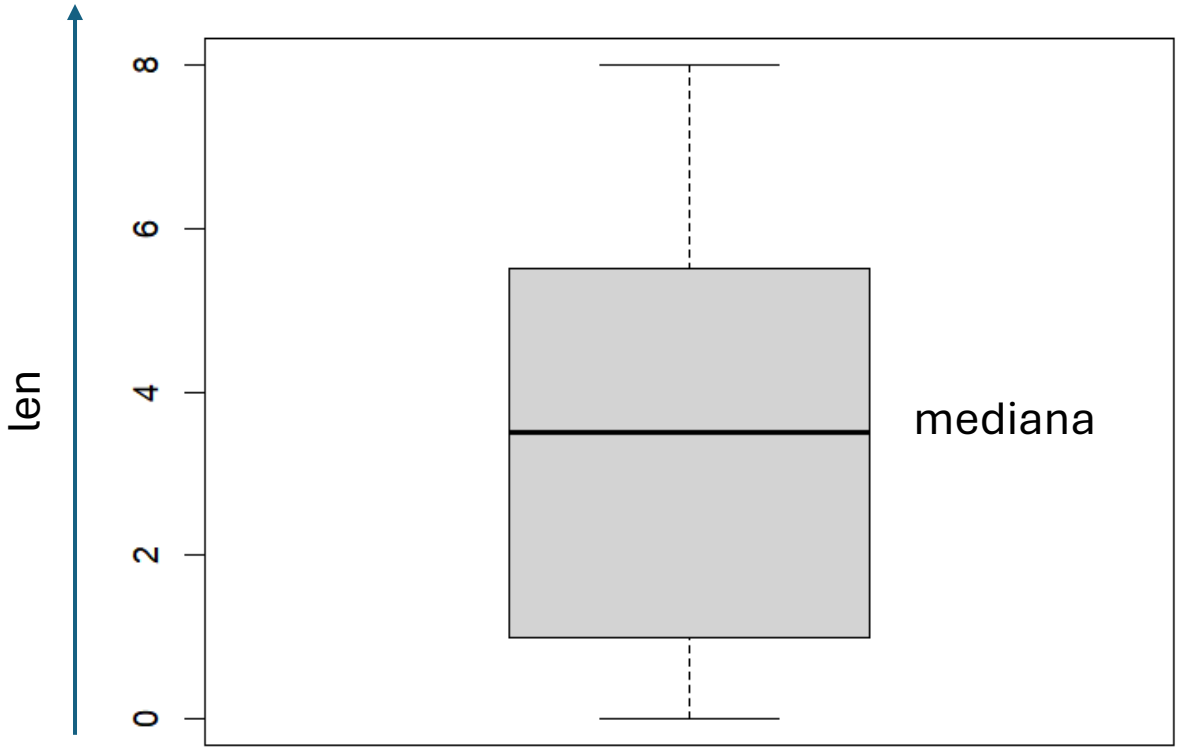


Variabili ratio

- **Rappresentazione (una sola variabile)**
  - Istogrammi:



  - Boxplot



Variabili ratio

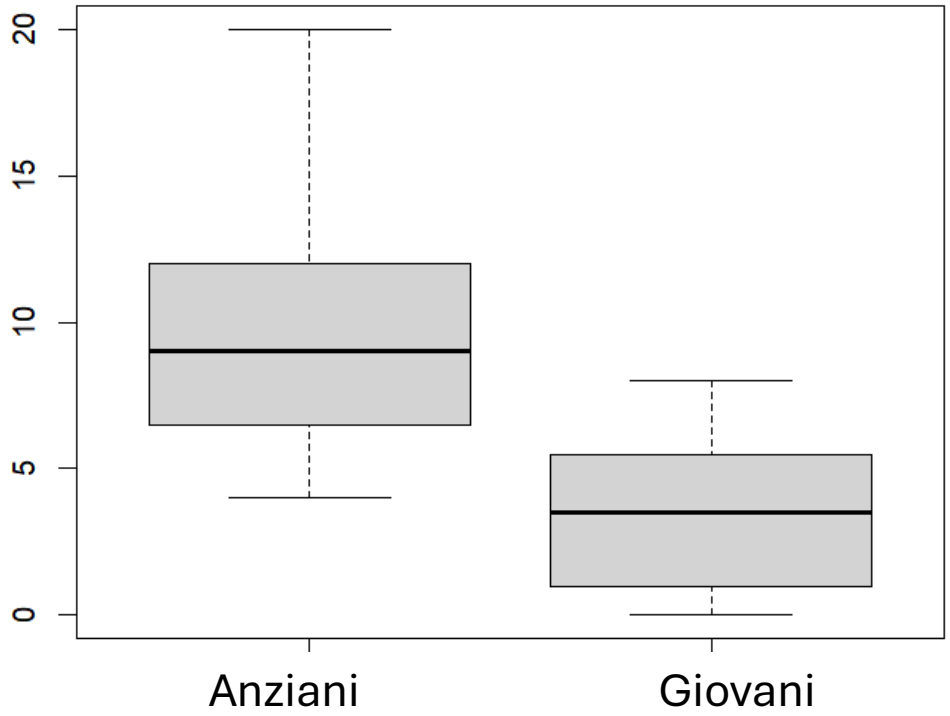
- **Rappresentazione (due variabili)**

➢ Variabile dipendente ratio, indipendente categoriale

(esistono differenze tra i valori in due gruppi?)

ad esempio, confrontiamo il numero di risposte esatte ad un test in due gruppi di età

Conviene utilizzare **un boxplot!**



Variabili ratio

- **Quali test utilizzare?**

➢ Cosa controllare: Numero di variabili e Distribuzione (normale o non normale)

➢ Dati distribuiti normalmente: confronto delle medie (*t-test*)
  - ➢ Una variabile → *one sample t-test*
  - ➢ Una variabile dipendente ratio e una indipendente categoriale → *Student's t-test*

➢ Test non parametrici: dati distribuiti non-normalmente: confronto dei *rank* (*Wilcoxon signed-rank test*)
  - ➢ Una variabile → *one sample Wilcoxon signed-rank test*
  - ➢ Una variabile dipendente ratio e una indipendente categoriale → *Wilcoxon signed-rank test*

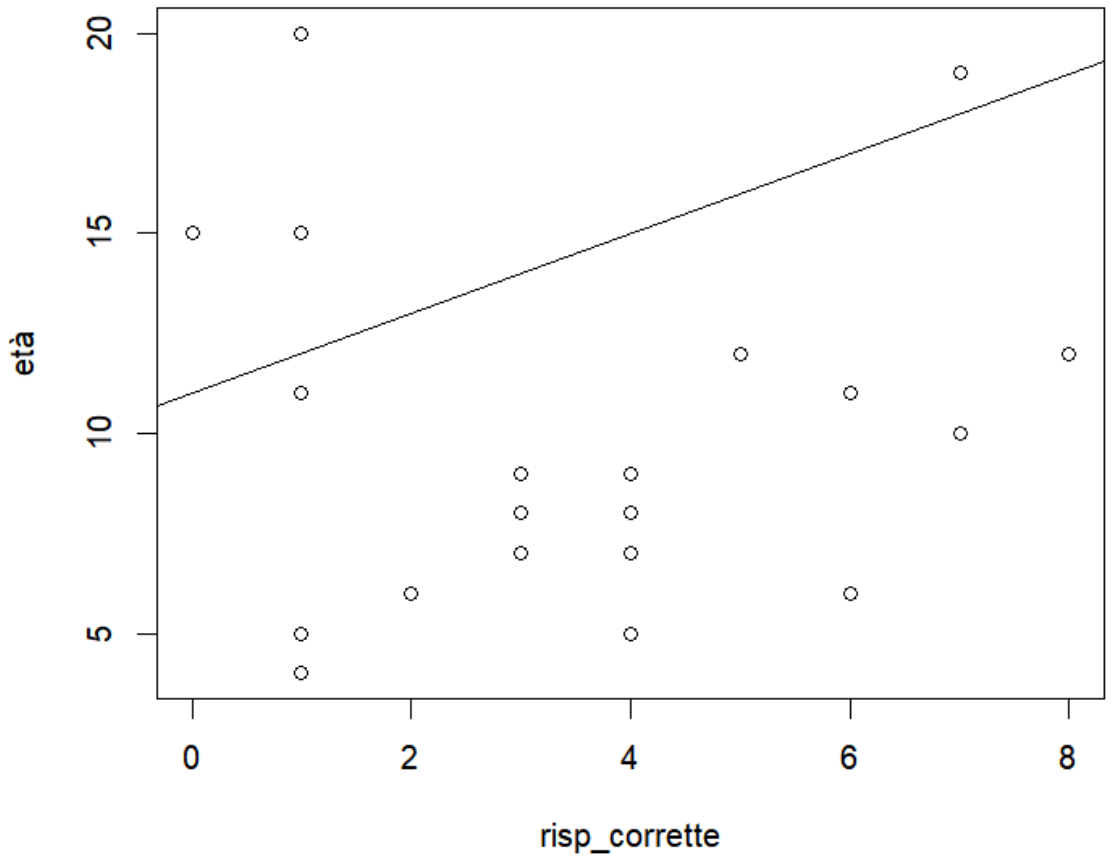
Variabili ratio

- **E se volessimo controllare se due variabili *ratio* correlano?**

➢ Al variare di un valore, varia anche il valore dell'altra

ad esempio, incrociamo l'età dei parlanti
col numero di risposte corrette ad un test

Rappresentazione: Utilizzare uno **scatterplot!**



Variabili ratio

- **E se volessimo controllare se due variabili *ratio* correlano?**

➢ Al variare di un valore, varia anche il valore dell'altra

ad esempio, incrociamo l'età dei parlanti
col numero di risposte corrette ad un test

Rappresentazione: Utilizzare uno **scatterplot!**

Come testare una correlazione?

Distribuzione normale
Coefficiente di correlazione di Pearson

Distribuzione non-normale
Coefficiente di correlazione di Spearman
Coefficiente di correlazione di Kendall

Variabili ratio

- **E se volessimo controllare se due variabili *ratio* correlano?**

➢ Al variare di un valore, varia anche il valore dell'altra

ad esempio, incrociamo l'età dei parlanti
col numero di risposte corrette ad un test

Rappresentazione: Utilizzare uno **scatterplot!**

Come testare una correlazione?

Distribuzione normale
Coefficiente di correlazione di Pearson

Distribuzione non-normale
Coefficiente di correlazione di Spearman
Coefficiente di correlazione di Kendall

Table 18. Correlation coefficients and their interpretation

Correlation coefficient	Labeling the correlation	Kind of correlation
$0.7 < r \leq 1$	very high	positive correlation:
$0.5 < r \leq 0.7$	high	the more/higher …, the more/higher …
$0.2 < r \leq 0.5$	intermediate	the less/lower …, the less/lower …
$0 < r \leq 0.2$	low	
$r \approx 0$	no statistical correlation	
$0 > r \geq -0.2$	low	negative correlation:
$-0.2 > r \geq -0.5$	intermediate	the more/higher …, the less/lower …
$-0.5 > r \geq -0.7$	high	the less/lower …, the more/higher …
$-0.7 > r \geq -1$	very high	

Variabili ordinali

- Come trattare le variabili ordinali?
  - Sono variabili che non rappresentano dei veri e propri valori «numerici», ma piuttosto delle categorie, possono essere ordinate su una scala:

Ad es., in un questionario possiamo sapere che "molto soddisfatto'' è migliore di "per nulla soddisfatto''

  - Non sappiamo però quanta distanza c'è tra "molto soddisfatto'' e "soddisfatto'' (invece sappiamo quantificare la distanza tra 0,5 e 1)

  - Volendo, possono essere approcciate quindi sia come variabili nominali che variabili «numeriche» (simili alle variabili intervallo), convertendone i livelli in valori ordinali:

per nulla soddisfatto	1
mediamente soddisfatto	2
molto soddisfatto	3

Ipotizzare che sia significativa non tanto la differenza tra categorie, ma il grado maggiore o minore di quella categoria su una scala

Variabili ordinali

- **Sintesi e rappresentazione**

- anche se alcuni sostengono che la moda (la categoria più frequente) sia la misura di tendenza centrale migliore da utilizzare, può aver senso utilizzare la **mediana**
- Possiamo decidere anche di mostrare una tavola con le frequenze

	Giudizio di accettabilità									
Construction	1	2	3	4	5	6	7	8	9	10
costruzione1	17.98	12.36	0.00	2.81	15.73	8.43	15.73	5.62	1.12	20.22
costruzione2	74.21	17.19	1.36	0.45	2.26	0.45	0.00	3.17	0.45	0.45

- Per la rappresentazione, tuttavia, è più cauto utilizzare un grafico a barre (e non un istogramma), o un mosaicplot, come per le variabili nominali

Variabili ordinali

- **Test**

- Generalmente, vengono utilizzati i test (non parametrici) validi per le variabili ratio non distribuite normalmente (quindi *Mann-Whitney U test* e *Wilcoxon signed-rank test*)
- Se vogliamo studiare la correlazione tra una variabile ordinale ed una ratio/ordinale, utilizziamo il coefficiente di correlazione di Spearman

Gioco: associa i metodi ai tipi di variabile

Sintesi:

media

mediana

frequenza

Grafici:

grafico a barre

grafico a torta

istogramma

boxplot

mosaicplot

scatterplot

Variabili:

ratio

ordinali

nominali

Che cosa faremo oggi II – la vendetta

1. Che cos'è R? (e perché tutti lo amano?)
2. Tipi di dati in R
3. Funzioni (giusto qualche accenno)
4. Che cosa serve a noi di tutto questo?
5. Esercitiamoci!

Che cos'è R?

- Software e ambiente

...ma anche: **linguaggio di programmazione**

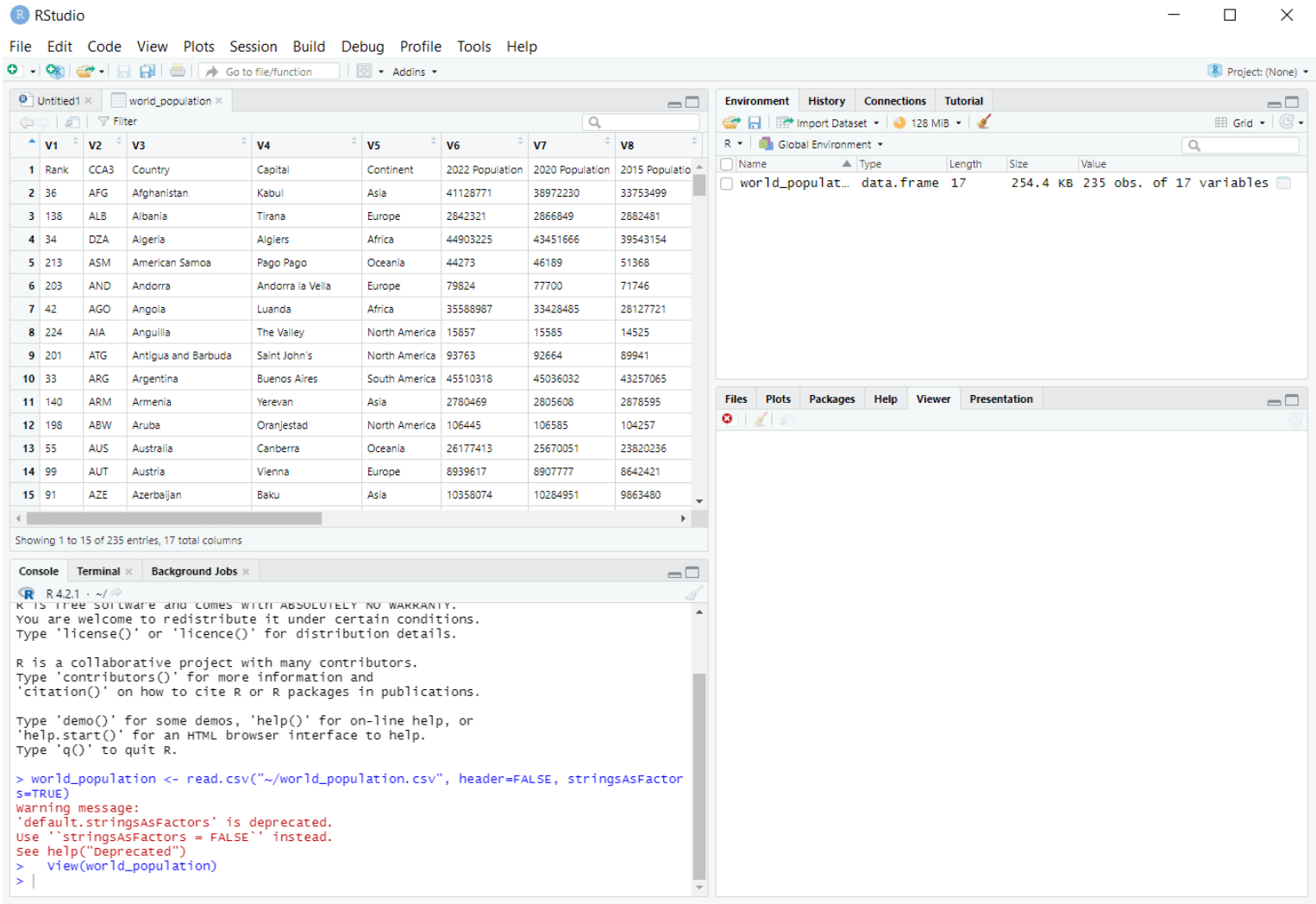
- Perché tutti lo amano? Esistono software alternativi per la statistica, ma
  - Sono software proprietari
  - Sono spesso a pagamento

- Vantaggi di R:
  - Open source, multipiattaforma
  - Comunità di utenti e sviluppatori (supporto nelle community online, *library* specializzate)
  - Flessibilità e trasparenza

Che cos’è R?

- Oggi non impareremo R (bisogna usarlo per imparare), ma introdurremo dei concetti di base del linguaggio e dell’ambiente per capire che cosa stiamo facendo
- Partiamo però da come ci appare RStudio (la UI che utilizzeremo)…

Interfaccia in RStudio



Tipi e strutture di dati in R

- Come nel mondo fisico (e cognitivo) esistono gli oggetti, esistono oggetti anche nel "mondo'' di R
  - No alberi, tazze e sassi, ma numeri, stringhe, vettori etc…
  - Ogni oggetto appartiene a una classe che ne definisce proprietà e metodi (ad es., con i numeri posso fare operazioni algebriche)

- Ci sono molti tipi di dati, e ora presenterò brevemente i tipi strettamente indispensabili in base a come sono strutturati

Strutture di dati in R

- Tipi “semplici’’:
  - 1, “casa’’, TRUE, 9.5, “postalveolare’’
  - Hanno in comune il fatto di rappresentare un solo valore
  - Possono essere di diversi tipi:
    - Numeri interi (integer): 1, 2, 99, 100
    - Numeri decimali (numeric): 10.2, 9.7, 0.5
    - Stringhe (character): “casa’’, “prefisso’’, “10’’
    - Valori logici (logical): TRUE, FALSE

Strutture di dati in R

- Possiamo concatenare questi valori in una specifica struttura, ovvero i vettori:

x <- c(1,4,6,0,8)

vettore <- c('casa', 'palla', 'zaino')

- Piccolo passo indietro…cosa sono *x* e *vettore*?
- Sono variabili (ma non in senso statistico…) → nomi che assegniamo a dei dati o strutture per memorizzarli nell'ambiente di lavoro

a <- 'miaomiao'

- Quando digiteremo x, R saprà che ci stiamo riferendo al vettore *c(1,4,6,0,8)*; se digitiamo *a*..?

Strutture di dati in R

- Un'altra struttura molto importante sono i *dataframe*, sostanzialmente i nostri dataset così come possiamo vederli in Excel

ID	Construction	Study.1..5.2.	Study.2..5.3.	Study.3..5.4.	Part.of.Speech	Animacy.Category
1	s-possessive	1	1	0	pronoun	ORG
2	s-possessive	1	0	0	pronoun	HUM
3	s-possessive	1	0	0	pronoun	HUM
4	s-possessive	1	0	0	pronoun	ORG
5	s-possessive	1	0	0	pronoun	HUM
6	s-possessive	0	0	0	proper name	HUM
7	s-possessive	1	0	0	pronoun	HUM
8	s-possessive	1	0	0	pronoun	HUM
9	s-possessive	1	0	0	pronoun	ORG
10	s-possessive	1	0	0	pronoun	HUM
11	s-possessive	1	1	0	pronoun	HUM
12	s-possessive	1	0	1	common noun	ORG
13	s-possessive	1	0	0	pronoun	ORG
14	s-possessive	0	0	0	proper name	HUM
15	s-possessive	0	0	0	proper name	HUM
16	s-possessive	1	0	0	pronoun	HUM
17	s-possessive	1	0	1	common noun	TIM

Strutture di dati in R

- I dataframe possono essere visti come strutture formate da vettori!

c('pronoun', 'pronoun', 'pronoun',...)

ID	Construction	Study.1..5.2.	Study.2..5.3.	Study.3..5.4.	Part.of.Speech	Animacy.Category
1	s-possessive	1	1	0	pronoun	ORG
2	s-possessive	1	0	0	pronoun	HUM
3	s-possessive	1	0	0	pronoun	HUM
4	s-possessive	1	0	0	pronoun	ORG
5	s-possessive	1	0	0	pronoun	HUM
6	s-possessive	0	0	0	proper name	HUM
7	s-possessive	1	0	0	pronoun	HUM
8	s-possessive	1	0	0	pronoun	HUM
9	s-possessive	1	0	0	pronoun	ORG
10	s-possessive	1	0	0	pronoun	HUM
11	s-possessive	1	1	0	pronoun	HUM
12	s-possessive	1	0	1	common noun	ORG
13	s-possessive	1	0	0	pronoun	ORG
14	s-possessive	0	0	0	proper name	HUM
15	s-possessive	0	0	0	proper name	HUM
16	s-possessive	1	0	0	pronoun	HUM
17	s-possessive	1	0	1	common noun	TIM

Strutture di dati in R

- Quando lavoriamo in R con le nostre variabili quindi, stiamo lavorando generalmente con dei vettori (aka le colonne nel nostro dataset)
- Esistono due modi per riferirsi ad un vettore (una variabile) in R:
  - *nomedataframe$nomecolonna*

    In questo modo, possiamo dire a R: "vai a vedere la colonna x nel dataframe y''

  - Il comando *attach(nomedataframe)* permette di evocare le variabili digitando direttamente il loro nome in R, senza fare riferimento al dataframe

- Diversi tipi di variabili conterrano diversi tipi di dati:

c('pronoun', 'noun phrase', 'pronoun',...) ← variabile nominale

c(1, 4, 22, 6,...) ← variabile ratio (o, volendo, ordinale)

Funzioni

- Ma in pratica, come facciamo i nostri calcoli statistici in R?
- Non li facciamo noi in prima persona, ma attraverso delle funzioni

```
> f <- function (x, y) x+y
  f(2,3)
  5
```

- Funzione → insieme di comandi con variabili libere

utile per non fare tutti questi calcoli →

```
> chisq.test
function (x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)),
    rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
{
    DNAME <- deparse(substitute(x))
    if (is.data.frame(x))
        x <- as.matrix(x)
    if (is.matrix(x)) {
        if (min(dim(x)) == 1L)
            x <- as.vector(x)
    }
    if (!is.matrix(x) && !is.null(y)) {
        if (length(x) != length(y))
            stop("'x' and 'y' must have the same length")
        DNAME2 <- deparse(substitute(y))
        xname <- if (length(DNAME) > 1L || nchar(DNAME, "w") >
            30)
            ""
        else DNAME
        yname <- if (length(DNAME2) > 1L || nchar(DNAME2, "w") >
            30)
            ""
        else DNAME2
        OK <- complete.cases(x, y)
        x <- factor(x[OK])
        y <- factor(y[OK])
        if ((nlevels(x) < 2L) || (nlevels(y) < 2L))
            stop("'x' and 'y' must have at least 2 levels")
        x <- table(x, y)
        names(dimnames(x)) <- c(xname, yname)
        DNAME <- paste(paste(DNAME, collapse = "\n"), "and",
```

```
chisq.test(forme_impaurire, p = c(0.52, 0.18, 0.03, 0.01, 0.26))

Chi-squared test for given probabilities data: forme_impaurire
X-squared = 11211, df = 4, p-value < 2.2e-16
```

Che cosa serve a noi di tutto questo?

- Che cosa serve a noi di tutto questo?
- Immaginando di avere un bel dataset annotato…

	A	B	C	D	E	F	G	H	I	J	K
1	Example ID	Construction	Study 1 (5.2	Study 2 (5.3	Study 3 (5.4	Part-of-Speech	Animacy Category	Animacy Rank	Modifier Length	Head Length	Modifier
2	1	s-possessive	1	1	0	pronoun	ORG	2	1	1	its [adminis
3	2	s-possessive	1	0	0	pronoun	HUM	1	1	2	his
4	3	s-possessive	1	0	0	pronoun	HUM	1	1	1	his
5	4	s-possessive	1	0	0	pronoun	ORG	2	1	5	its [team]
6	5	s-possessive	1	0	0	pronoun	HUM	1	1	2	his
7	6	s-possessive	0	0	0	proper name	HUM	1	2	3	Jack Krame
8	7	s-possessive	1	0	0	pronoun	HUM	1	1	6	their [boys]
9	8	s-possessive	1	0	0	pronoun	HUM	1	1	2	their [defen
10	9	s-possessive	1	0	0	pronoun	ORG	2	1	5	its [Multnon
11	10	s-possessive	1	0	0	pronoun	HUM	1	1	2	his
12	11	s-possessive	1	1	0	pronoun	HUM	1	1	2	her
13	12	s-possessive	1	0	1	common noun	ORG	2	2	14	the governn
14	13	s-possessive	1	0	0	pronoun	ORG	2	1	1	their
15	14	s-possessive	0	0	0	proper name	HUM	1	2	1	Senator Go
16	15	s-possessive	0	0	0	proper name	HUM	1	3	2	Mantle's an
17	16	s-possessive	1	0	0	pronoun	HUM	1	1	1	his
18	17	s-possessive	1	0	1	common noun	TIM	9	2	6	the year's
19	18	s-possessive	1	0	0	pronoun	HUM	1	1	1	their
20	19	s-possessive	1	0	0	pronoun	ORG	2	1	1	their
21	20	s-possessive	1	0	0	pronoun	HUM	1	1	1	his
22	21	s-possessive	1	1	0	pronoun	HUM	1	1	1	their

- I valori in ogni cella sono **dati** (stringhe di caratteri, numeri, etc..)
- La nostra tabella è un **dataframe**
- Le colonne, le nostre variabili annotate, sono **vettori (che contengono un solo tipo di dato ognuno! Es. Variabili nominali: stringhe di caratteri)**
- I metodi di sintesi, rappresentazione e inferenza sono applicati tramite **funzioni**, che prendono come argomenti i nostri vettori (ovvero le variabili)

```
mean(Head.Length)
t.test(Modifier.Length ~ Construction)
```

Esercitazione

Prima, apriamo Rstudio e proviamo a fare qualcosa di semplice..

Esercitazione

È il momento di mettere le mani su qualche dato!

- Aprite i tre file nella cartella Script R test
- caricate il dataset *possessives* → competizione tra *of-possessive* e *s-possessive* in inglese

- Che cosa analizzare?

  Domanda di ricerca: Analizziamo l'influenza di due fattori sulla scelta tra le due costruzioni possessive: 1) **Animatezza** e 2) **Lunghezza in parole** dell' NP testa

- Che cosa dovete fare?
  - Leggete attentamente il file di testo con le info sul dataset
  - Per ognuno dei due step, partite descrivendo la relazione tra le due variabili indipendenti e la dipendente (tramite grafici, tabelle, misure, ecc.).
  - Seguite il *decision tree* a vostra disposizione e applicate i test giusti
  - IMPORTANTE: annotate ogni step, i risultati e i grafici su un foglio di testo a parte, così poi ne parliamo!

Esercitazione

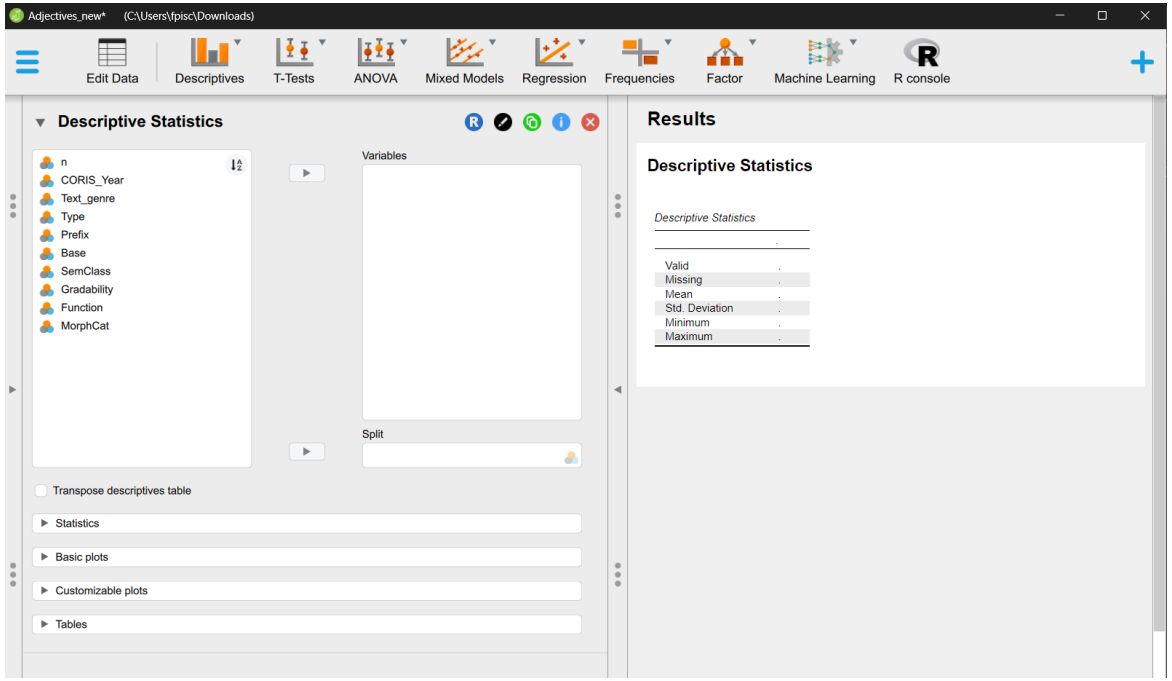
- Domanda bonus: e se provassimo a considerare l'Animatezza come una scala?
  - Quali analisi possiamo condurre?
  - Cambia qualcosa nella nostra comprensione dei dati?

Che cosa faremo oggi III – quasi finita

1. JASP
2. Caricare dataset in JASP
3. Rappresentazione dei dati in JASP: pro e contro
4. Test statistici in JASP
5. Un altro piccolo esercizio?

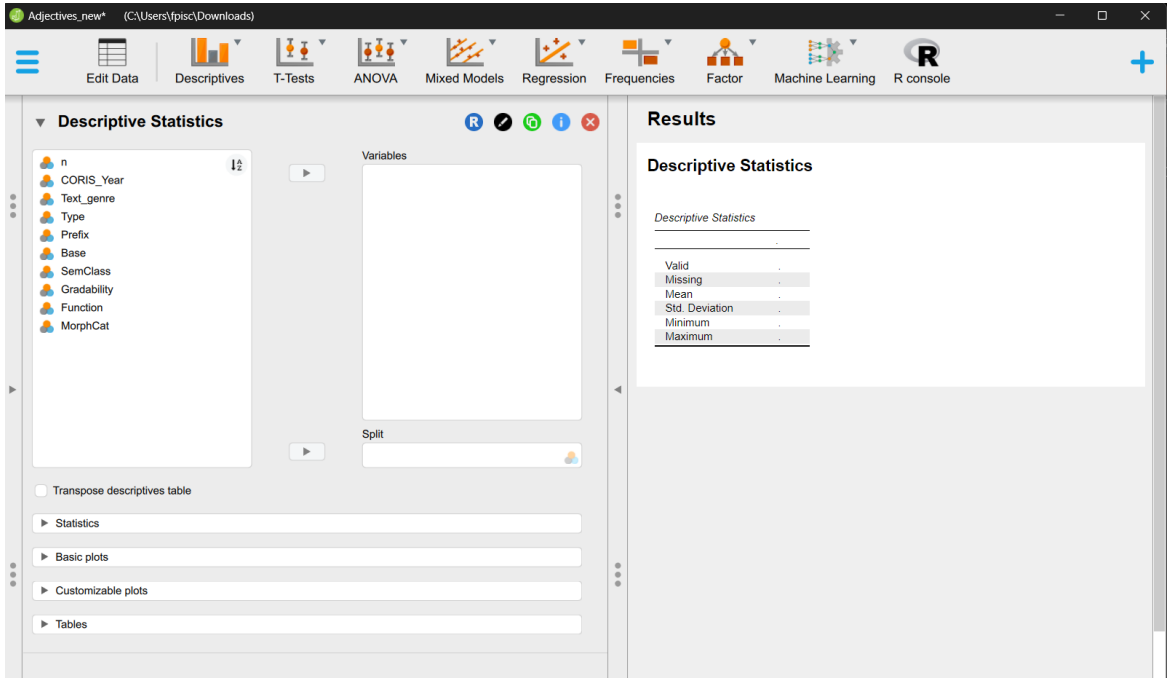
JASP

- Programma *open source* e gratuito, supportato dall'Università di Amsterdam
- Alternativa *user-friendly* → nessuna necessità di scrivere codice, (quasi) tramite l'interfaccia grafica

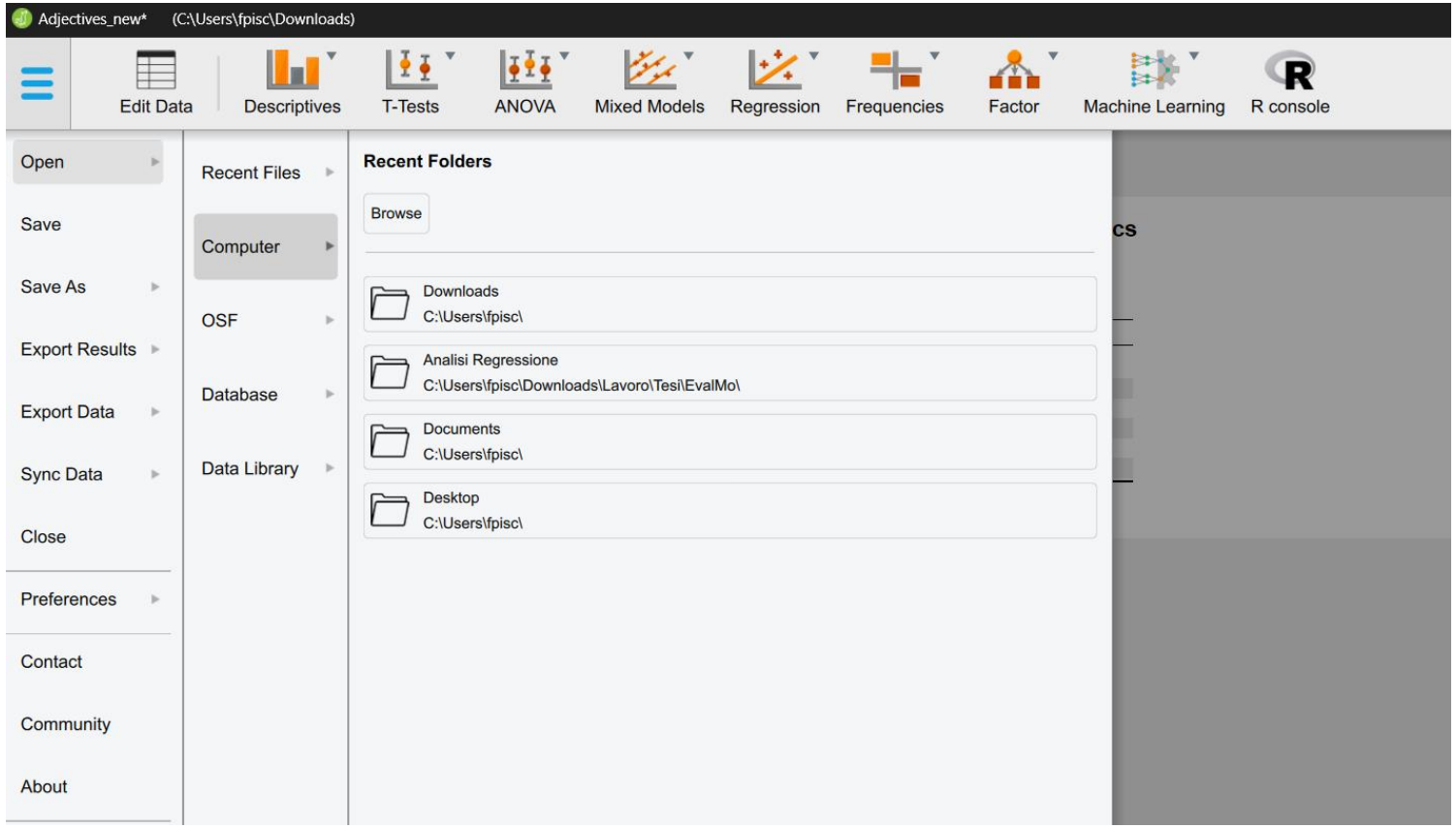


JASP

- Non è importante che tutti imparino a programmare, ma è importante che tutti cerchino di utilizzare metodi rigorosi e riproducibili di rappresentazione e analisi dei dati!
- Contro: minore flessibilità di R (almeno in superficie), ma ottimo per la maggior parte dei task statistici di base



Caricare un dataset in JASP



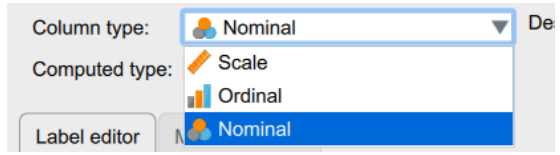
Caricare un dataset in JASP

Adjectives_new* (C:\Users\fpisc\Downloads)

Analyses | Synchronisation | Resize Data | Insert | Remove | Undo | Redo

	n	CORIS_Year	Text_genre	Type	Prefix	Base	SemClass	Gradability	Function	MorphCat
1	semi_291	MONITOR2001_04	MISC	semiabbandonata	0 semi	abbandonato	spatial	upper_closed_scale	attributive	participle
2	mezzo_55	CORIS1980_2000	NARRAT	mezzo abbandonare	1 mezzo	abbandonato	spatial	upper_closed_scale	attributive	participle
3	mezzo_129	MONITOR2005_07	NARRAT	mezzo abbassare	1 mezzo	abbassato	spatial	lower_closed_scale	attributive	participle
4	mezzo_293	MONITOR2001_04	NARRAT	mezzo abbattere	1 mezzo	abbattuto	relational	totally_closed_scale	attributive	participle
5	semi_61	MONITOR2017_20	NARRAT	semi - abitabile	0 semi	abitabile	relational	binary	attributive	complex
6	semi_188	CORIS1980_2000	NARRAT	semi - accecare	0 semi	accecato	body	lower_closed_scale	predicative	participle
7	semi_265	CORIS1980_2000	STAMPA	semiaccecati	0 semi	accecato	body	lower_closed_scale	attributive	participle
8	mezzo_124	MONITOR2011_13	NARRAT	mezzo accecare	1 mezzo	accecato	body	lower_closed_scale	predicative	participle
9	mezzo_57	MONITOR2014_16	NARRAT	mezzo acciaccata	1 mezzo	acciaccato	body	lower_closed_scale	attributive	participle
10	semi_124	CORIS1980_2000	NARRAT	semiaccucciata	0 semi	accucciato	spatial	totally_closed_scale	predicative	participle
11	semi_141	CORIS1980_2000	NARRAT	semiaddormentato	0 semi	addormentato	body	upper_closed_scale	attributive	participle
12	semi_167	MONITOR2005_07	NARRAT	semiaddormentati	0 semi	addormentato	body	upper_closed_scale	attributive	participle
13	semi_179	CORIS1980_2000	PRACC	semiaddormentato	0 semi	addormentato	body	upper_closed_scale	predicative	participle
14	semi_185	CORIS1980_2000	NARRAT	semiaddormentate	0 semi	addormentato	body	upper_closed_scale	attributive	participle
15	mezzo_19	CORIS1980_2000	NARRAT	mezzo addormentato	1 mezzo	addormentato	body	upper_closed_scale	predicative	participle
16	mezzo_46	CORIS1980_2000	NARRAT	mezzo addormentato	1 mezzo	addormentato	body	upper_closed_scale	predicative	participle
17	mezzo_51	MONITOR2014_16	NARRAT	mezzo addormentato	1 mezzo	addormentato	body	upper_closed_scale	predicative	participle
18	mezzo_59	CORIS1980_2000	NARRAT	mezzo addormentare	1 mezzo	addormentato	body	upper_closed_scale	predicative	participle
19	mezzo_71	MONITOR2017_20	STAMPA	mezzo addormentato	1 mezzo	addormentato	body	upper_closed_scale	predicative	participle
20	mezzo_79	CORIS1980_2000	NARRAT	mezzo addormentato	1 mezzo	addormentato	body	upper_closed_scale	predicative	participle
21	mezzo_117	CORIS1980_2000	NARRAT	mezzo addormentato	1 mezzo	addormentato	body	upper_closed_scale	attributive	participle
22	mezzo_133	CORIS1980_2000	NARRAT	mezzo addormentato	1 mezzo	addormentato	body	upper_closed_scale	attributive	participle
23	mezzo_145	MONITOR2017_20	NARRAT	mezzo addormentato	1 mezzo	addormentato	body	upper_closed_scale	predicative	participle

Assegna automaticamente un tipo ad ogni variabile:

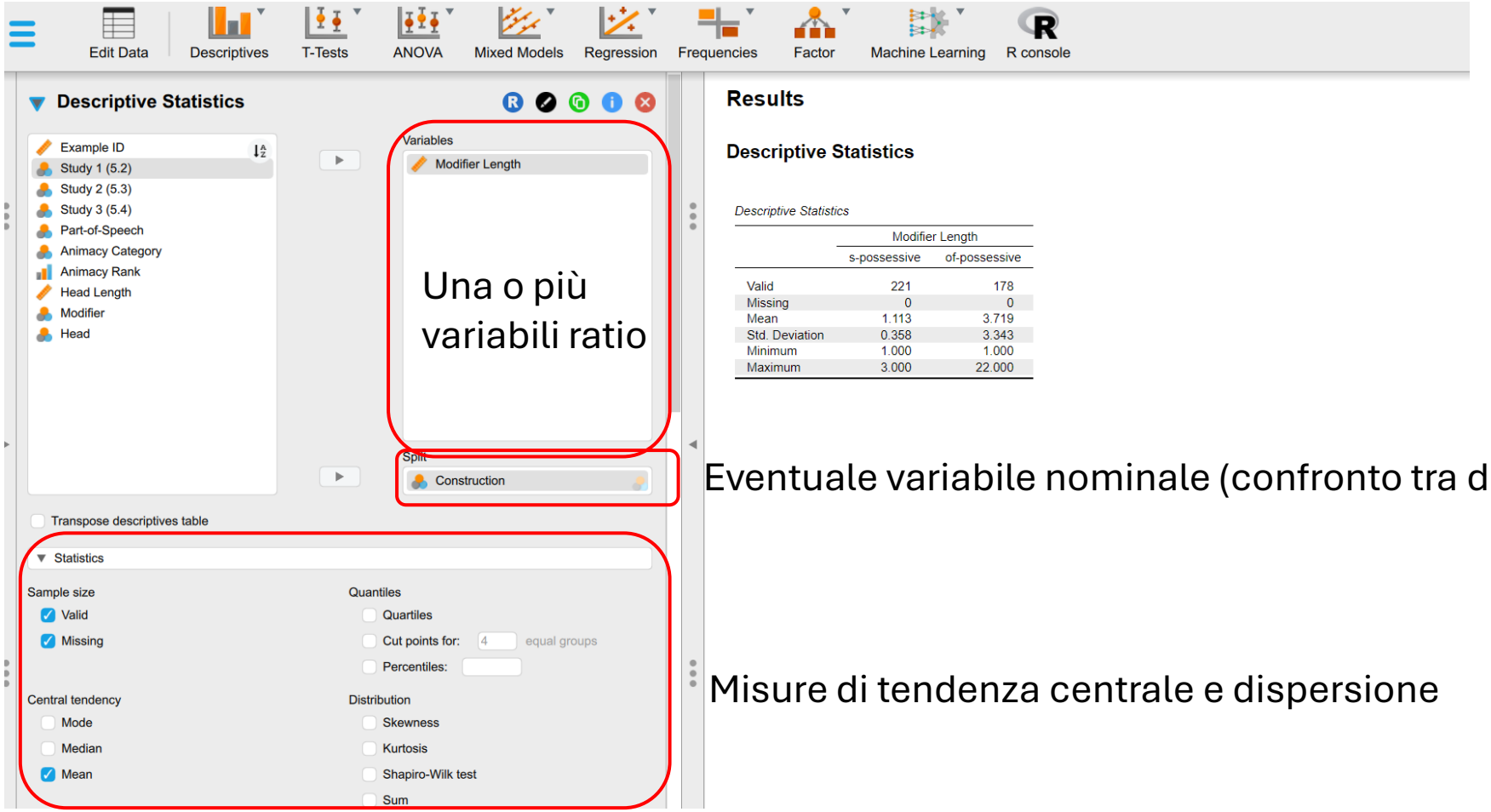


Rappresentazione dei dati in JASP

- Alcune funzioni sono sparse
- Utilizzo più intuitivo per alcuni tipi di variabili (ad es. ratio) rispetto ad altre (nominali)
- Per le variabili nominali mancano alcuni tipi di plot

  → consiglio: per le variabili nominali può essere utile Excel (grafici più intuitivi che in R)
- Vediamo come sintetizzare i dati dividendoli in ratio e nominali (le var. ordinali sono flessibili)

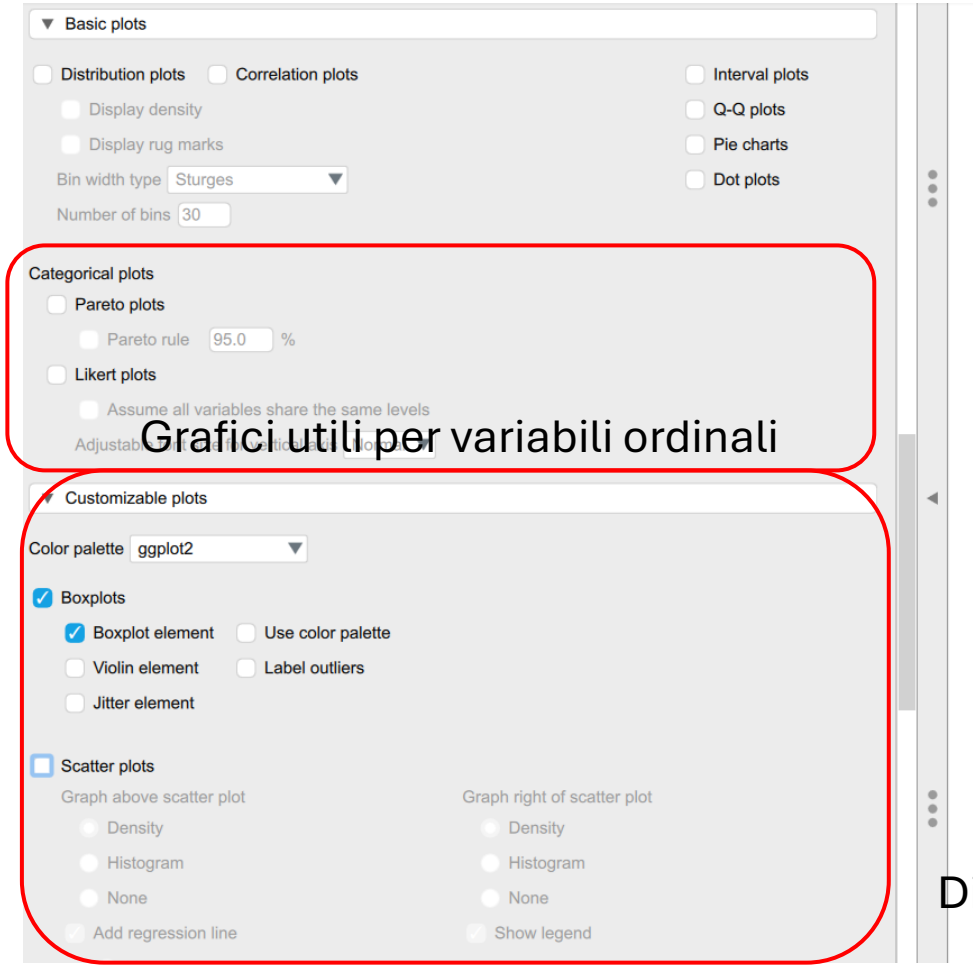
Sintesi e rappresentazione: variabili ratio



Eventuale variabile nominale (confronto tra due o più gruppi)

Misure di tendenza centrale e dispersione

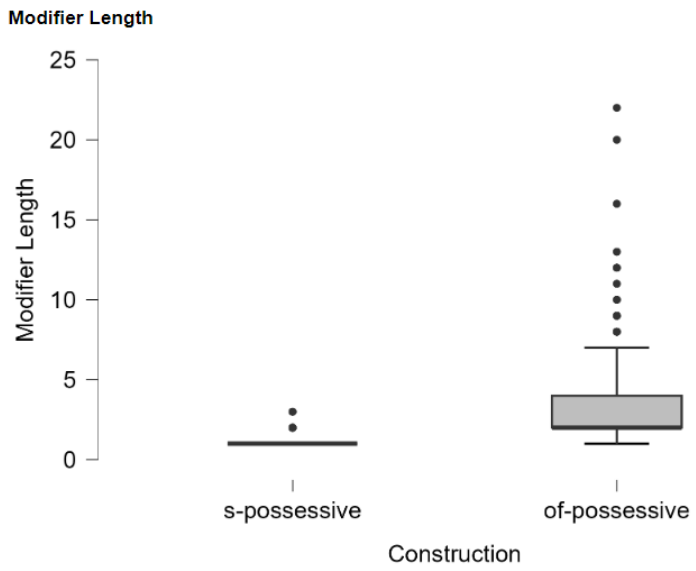
Sintesi e rappresentazione: variabili ratio



Grafici utili per variabili ordinali

Valid	221	178
Missing	0	0
Mean	1.113	3.719
Std. Deviation	0.358	3.343
Minimum	1.000	1.000
Maximum	3.000	22.000

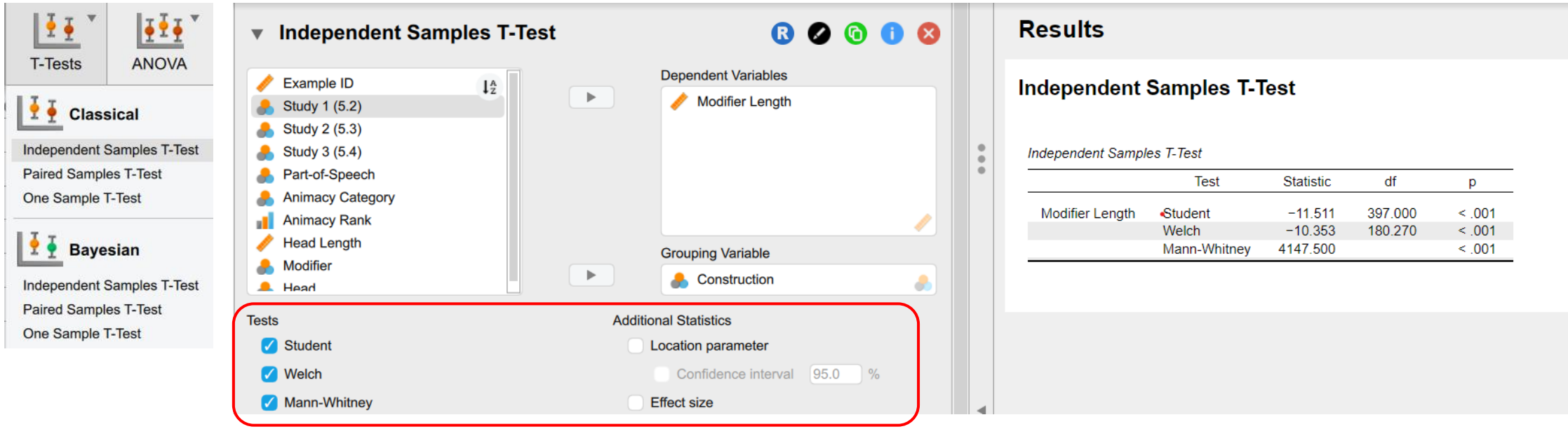
Boxplots



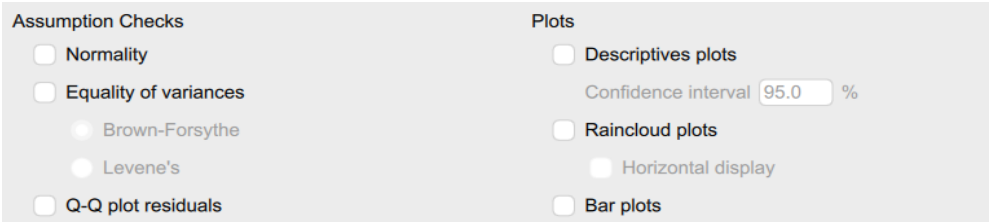
Diversi tipi di grafici, in base alla tipologia di variabili specificate sopra

Test: variabili ratio

Testare variabile ratio vs variabile nominale

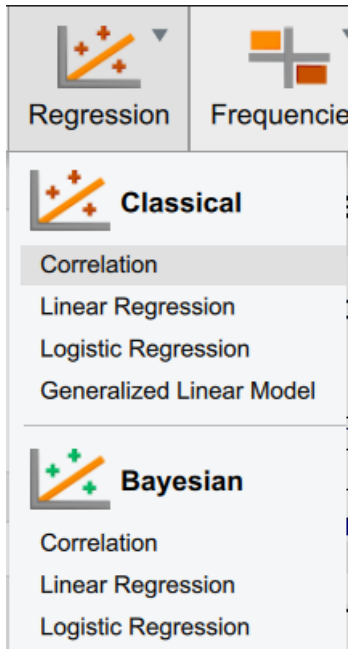


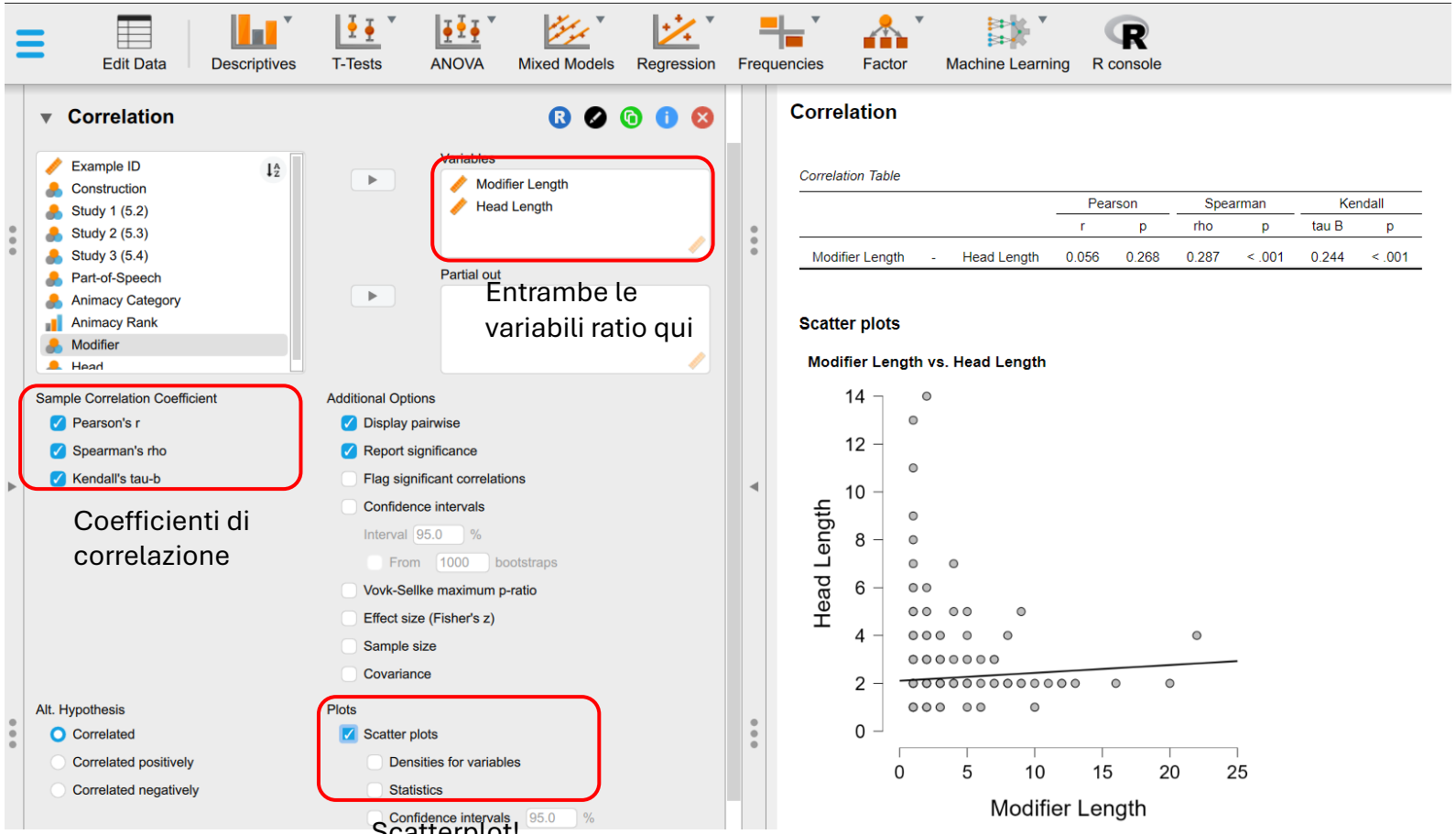
I primi due sono t-test, il terzo è l'alternativa non parametrica (wrt distribuzione normale)



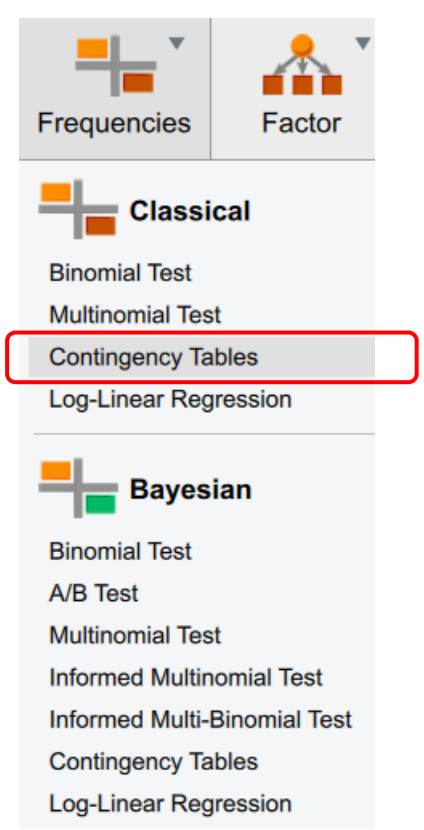
Test: variabili ratio

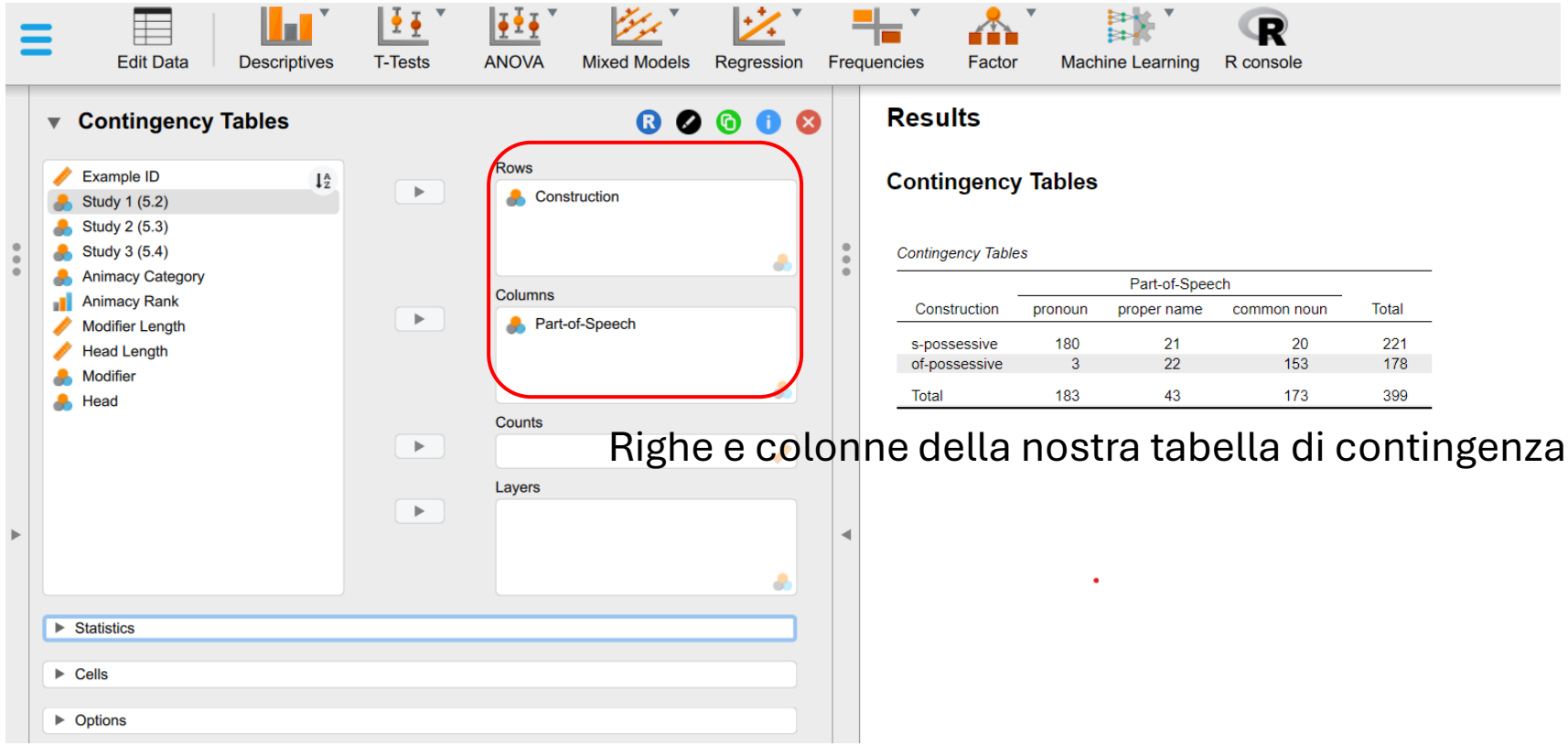
Testare variabile ratio vs variabile ratio (correlazione)





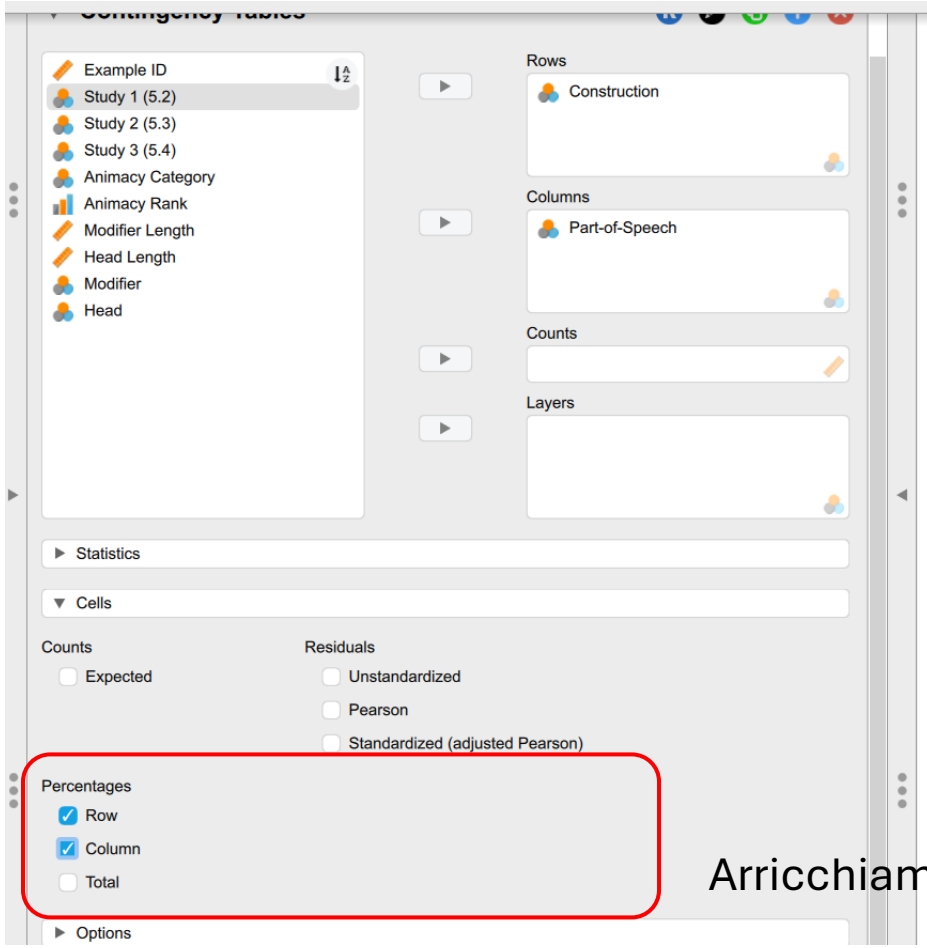
Sintesi e rappresentazione: variabili nominali





Righe e colonne della nostra tabella di contingenza

Sintesi e rappresentazione: variabili nominali



Results

Contingency Tables

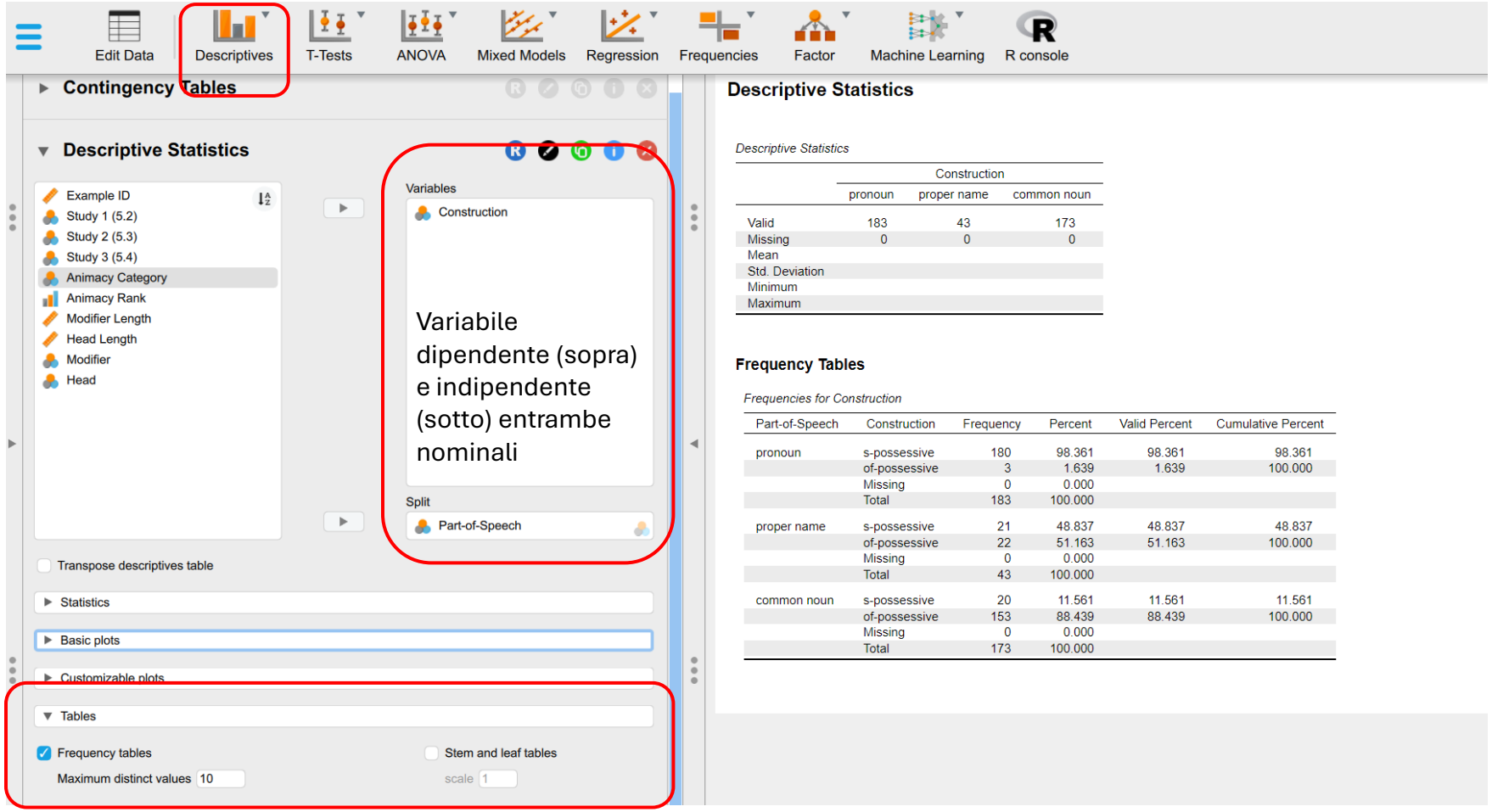
Contingency Tables

Construction		Part-of-Speech			Total
		pronoun	proper name	common noun	
s-possessive	Count	180.000	21.000	20.000	221.000
	% within row	81.448 %	9.502 %	9.050 %	100.000 %
	% within column	98.361 %	48.837 %	11.561 %	55.388 %
of-possessive	Count	3.000	22.000	153.000	178.000
	% within row	1.685 %	12.360 %	85.955 %	100.000 %
	% within column	1.639 %	51.163 %	88.439 %	44.612 %
Total	Count	183.000	43.000	173.000	399.000
	% within row	45.865 %	10.777 %	43.358 %	100.000 %
	% within column	100.000 %	100.000 %	100.000 %	100.000 %

Arricchiamo la nostra tabella con le freq. percentuali (su righe e o su colonne)

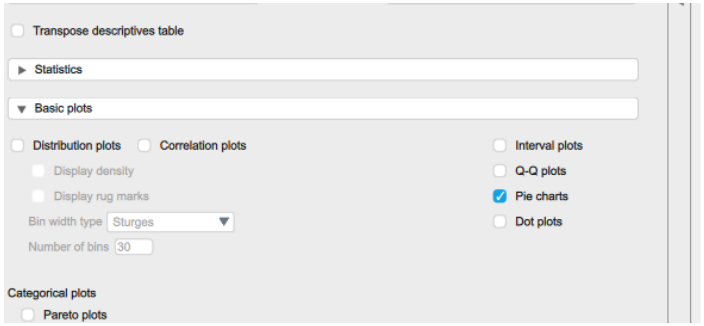
Sintesi e rappresentazione: variabili nominali

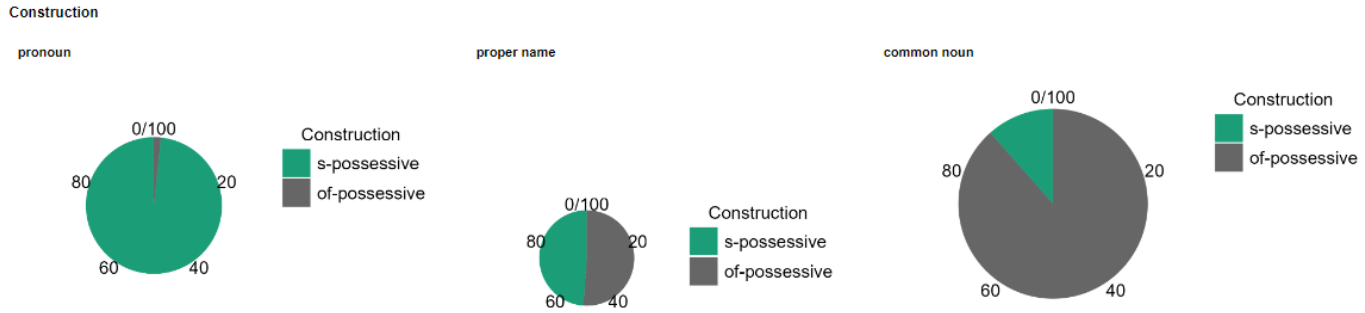
Torniamo su Descriptives > Descriptive Statistics



Sintesi e rappresentazione: variabili nominali

Purtroppo però i grafici per le variabili nominali non sono disponibili quando abbiamo un incrocio di due variabili, ma solo valore per valore delle celle (no grafico a barre stacked, no mosaicplot…)

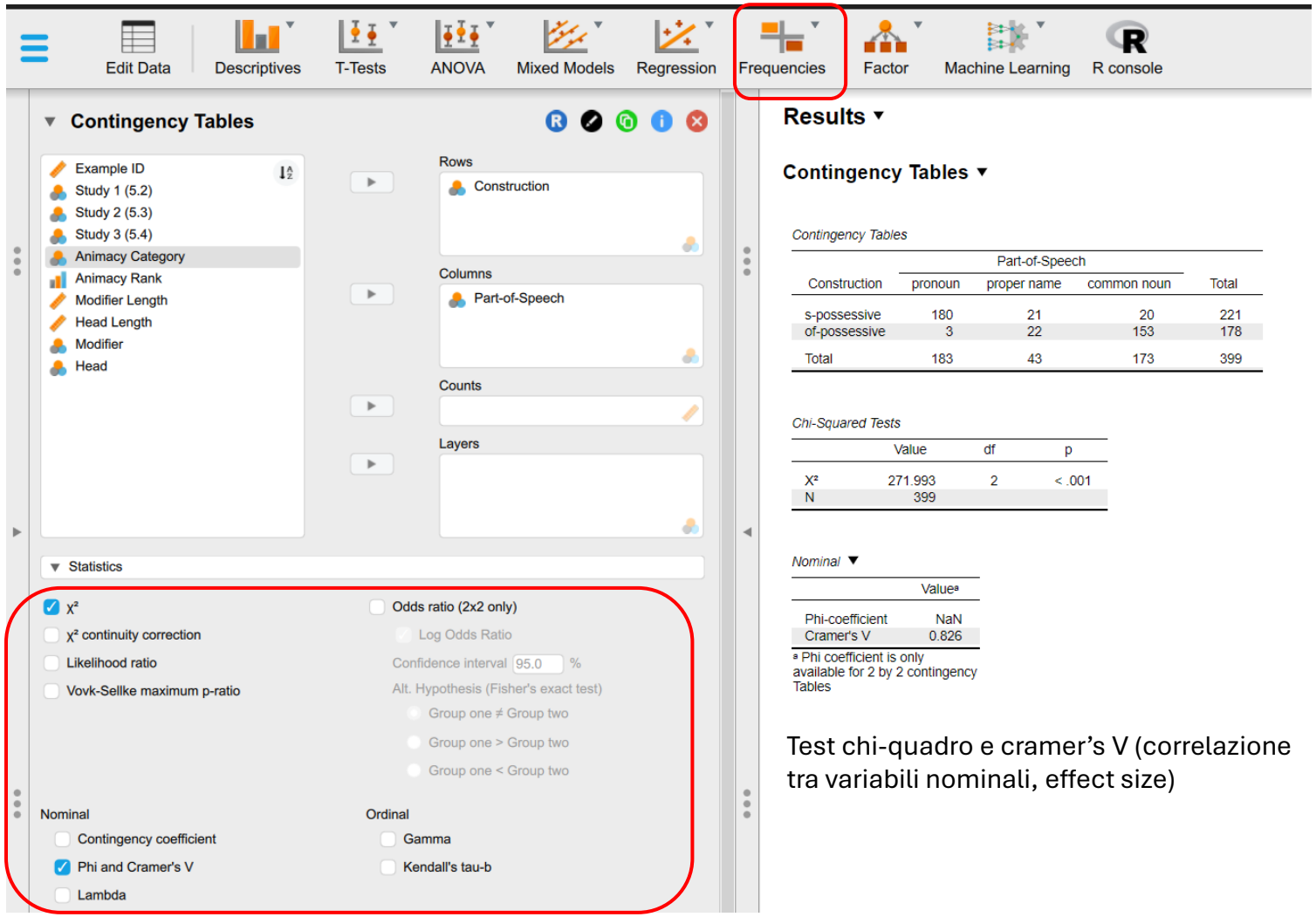




MA vedremo un plot carino relativo al *chi-square*…

Test: variabili nominali

Torniamo su Frequencies > Contingency tables



Results

Contingency Tables

Contingency Tables

Construction	Part-of-Speech			
	pronoun	proper name	common noun	Total
s-possessive	180	21	20	221
of-possessive	3	22	153	178
Total	183	43	173	399

Chi-Squared Tests

	Value	df	p
X^2	271.993	2	< .001
N	399		

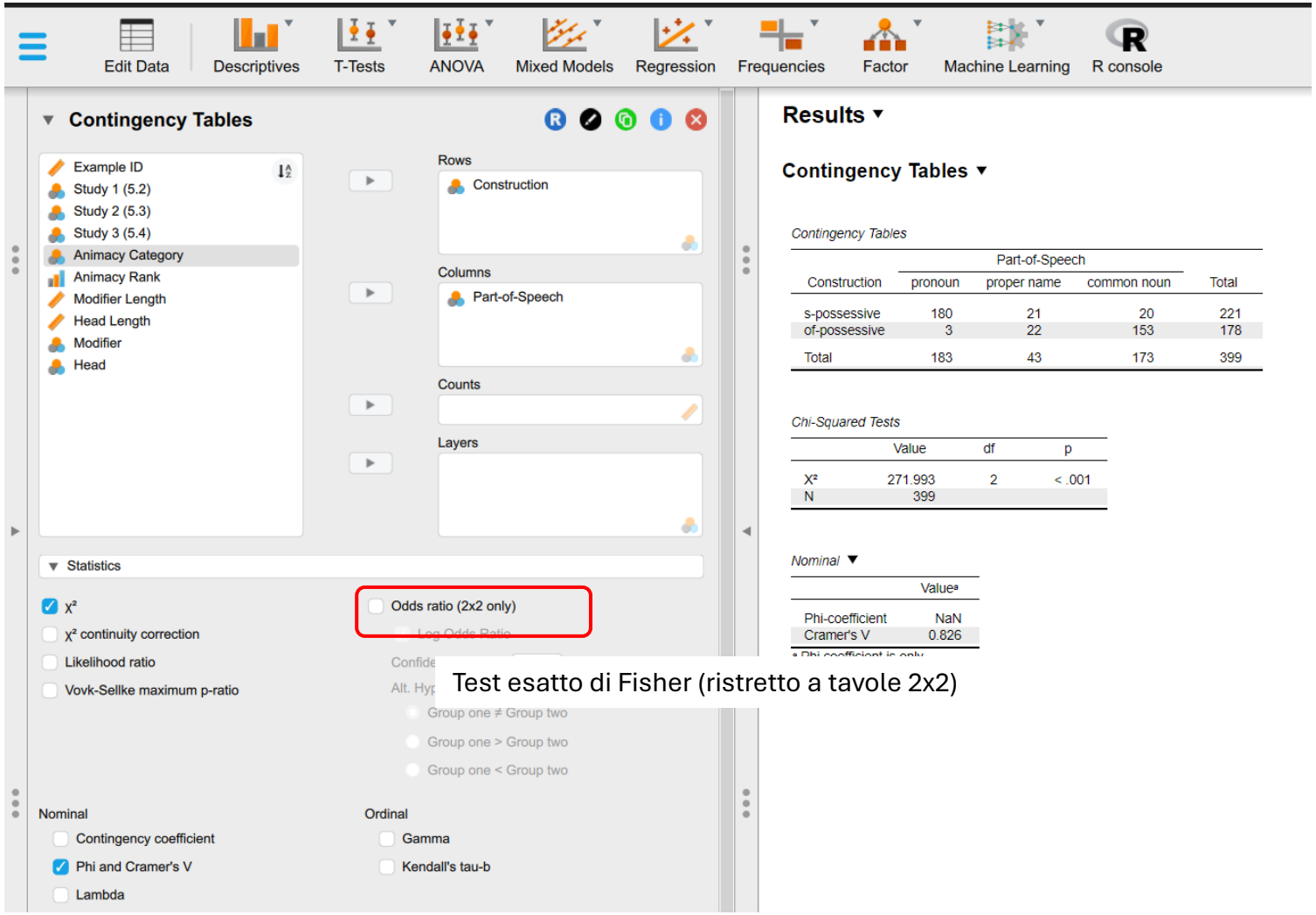
Nominal

	Value[a]
Phi-coefficient	NaN
Cramer's V	0.826

[a] Phi coefficient is only available for 2 by 2 contingency Tables

Test chi-quadro e cramer's V (correlazione tra variabili nominali, effect size)

Test: variabili nominali



Test esatto di Fisher (ristretto a tavole 2x2)

Deviation plot per variabili nominali

Deviazione dei conti osservati da quelli attesi (in positivo e in negativo)

Piccola esercitazione (se ce la facciamo)

- caricate il dataset *suffix_competition* → competizione tra i suffissi *-ic* e *-ical* in inglese
- Che cosa analizzare?
  Domanda di ricerca: Analizziamo l'influenza della lunghezza della base sulla formazione di aggettivi tramite uno o l'altro suffisso
- Come prima: partiamo con il riportare i dati, poi un grafico, e successivamente testiamo il tutto!