



I Microdati Istat

Susi Osti

10 novembre 2015

■ ■ ■

Microdati Istat

micro.STAT

micro.STAT sugli incidenti stradali

Metteremo le mani
in pasta con



L'informazione statistica: le forme

Macrodati:

dato aggregato, informazione statistica, indicatori ecc.

Metadati:

- descrizioni
- definizioni e informazioni

per interpretare correttamente i dati

<http://www3.istat.it/cgi-bin/glossario/indice.pl#P>



Incidenti stradali con lesioni alle persone

Personalizza Esportazioni Grafici La tua interrogazione

→ Giorno della settimana	totale
→ Mese	totale
→ Anno	2014
→ Tipo dato	incidenti stradali con lesioni alle persone
→ Territorio	
■ Veneto	13 958
Verona	3 026
Vicenza	2 230
Belluno	490
Treviso	2 195
Venezia	2 389
Padova	3 038
Rovigo	590

Open data

Sono scaricabili e rielaborabili tutti i dati dell'Istat diffusi attraverso le [banche dati e i sistemi informativi](#)



Sono rilasciati sotto licenza Creative Commons (CC BY 3.0 IT): è possibile riprodurre, distribuire, trasmettere e adattare liberamente dati e analisi dell'Istituto nazionale di statistica, anche a scopi commerciali, a condizione che venga citata la fonte



<http://www.dati.gov.it/> il portale che si occupa dello stato dell'open data in Italia

L'informazione statistica: le forme

Microdati:

dato elementare, individuale



Metadati:

- descrizioni
- definizioni e informazioni

per interpretare correttamente i dati

<http://www3.istat.it/cgi-bin/glossario/indice.pl#P>

IDQUESTIONARIO	MODELLO	TIPALL	NOCC	NFAM	TITGOD	SUPERF
1111111111L	L	1	5	1	1	200

Perché i microdati?



- . Le combinazioni personalizzate variabili → si possono creare le proprie tabelle basate su qualsiasi variabile
- . Le variabili di analisi possono essere esaminate in modo più dettagliato di quanto sia disponibile dalle tabelle di rilascio standard
- . I dati individuali forniscono un ulteriore potenziale analisi che non è possibile con i dati aggregati

Quali microdati?



Le tipologie di file messe a disposizione sono le seguenti

 [File ad uso pubblico](#), scaricabili direttamente dal sito Istat

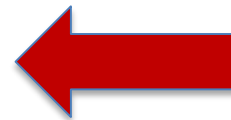
[File standard](#), rilasciabili a seguito di richiesta motivata, con finalità di studio

[File per la ricerca](#), rilasciabili a studiosi di università o enti di ricerca a seguito della presentazione di un progetto di ricerca

[File per il Sistan](#), rilasciabili esclusivamente agli uffici di statistica del Sistema Statistico Nazionale

Attraverso il [Laboratorio per l'Analisi dei Dati ELEmentari](#) (ADELE), ricercatori e studiosi possono accedere ai file di microdati di tutte le indagini dell'Istat

<http://www.istat.it/adele/ListaRilevazioni>



Quanti microdati?



Serie storica disponibile dal 1969

← → ↺ 🏠 www.istat.it/adele/ListaRilevazioni

Istat [PAGINA PRINCIPALE](#) | [PAGINA PRECEDENTE](#)

Rilevazione degli incidenti stradali con lesioni a persone

Serie storica microdati

Informazioni

Filtra anno

Periodo ☐	Tematiche ☐	Variabili ☐	N° di Records ☐	Not
Anno 2013	-		181.227	
Anno 2012	-		188.228	File
Anno 2011	-		205.638	
Anno 2010	-		212.997	Il da segi
Anno 2009	-		215.405	
Anno 2008	-		218.983	A se e fe dall
Anno 2007	-		230.871	

Quante variabili?



Le variabili rilasciate sono 134



[PAGINA PRINCIPALE](#) | [PAGINA PRECEDENTE](#)

Rilevazione degli incidenti stradali con lesioni a persone

[Informazioni](#)

Variabili disponibili nell'anno 2013

Riga ☐	Acronimo variabile ☐	Denominazione variabile ☐	Lunghezza ☐	Modalità ☐
1	anno	anno	2	
2	mese	mese	2	
3	provincia	provincia	3	
4	comune	comune	3	
5	n_d_ordine	n. d'ordine	4	
6	giorno	giorno	2	
7	organo_di_rilevazione	organo di rilevazione	1	
8	organo_coordinatore	organo coordinatore	1	
9	localizzazione_incidente	localizzazione incidente	1	
10	denominazione_della_strada	denominazione della strada	3	
11	tronco_di_strada_o_autostrada	tronco di strada o autostrada	2	
12	tipo_di_strada	tipo di strada	1	
13	pavimentazione	pavimentazione	1	



Per iniziare a conoscere i microdati...

File ad uso pubblico micro.STAT



sono **collezioni campionarie** di dati elementari relative ad alcune indagini svolte dall'Istat, per le quali siano già stati sviluppati i corrispondenti file di microdati per la ricerca (MFR)

scaricabili liberamente e gratuitamente dal sito Istat

http://www.istat.it/it/prodotti/microdati#file_microstat



I file mlcro.STAT disponibili ad oggi...

Aspetti della vita quotidiana

L'indagine è condotta su un campione di circa 20.000 famiglie e 50.000 individui e rileva le informazioni sulla vita quotidiana di individui e famiglie

Forze di lavoro – dati trasversali trimestrali

L'indagine fornisce trimestralmente le stime dei principali aggregati dell'offerta di lavoro

Incidenti stradali con lesioni a persone

La rilevazione degli incidenti stradali con lesioni a persone è svolta con cadenza mensile nell'arco di un anno sul territorio nazionale

Percorsi di studio e di lavoro dei diplomati

L'indagine a cadenza triennale è condotta su una singola leva di studenti, intervistati a circa quattro anni dal diploma

I file mlcro.STAT disponibili ad oggi...

Inserimento professionale dei laureati

Indagine su un campione di 62.000 laureati sul totale dei 300.338 che avevano conseguito il titolo nel 2007

Censimento dei laureati

Nel 2009 è stata svolta, presso tutti gli atenei italiani, la rilevazione che ha permesso di censire i laureati che avevano conseguito il titolo nell'anno 2007

Rilevazione sull'innovazione nelle imprese

La rilevazione a cadenza biennale raccoglie informazioni sui processi di innovazione nelle imprese dell'industria e dei servizi

Inserimento professionale dei dottori di ricerca

L'indagine svolta tra dicembre 2009 e febbraio 2010 sui dottorati di ricerca conseguiti nel 2004 e nel 2006, aggiornata al 2014

I file mlcro.STAT disponibili ad oggi...

Indagine sui musei e le istituzioni simili

L'indagine viene effettuata su oltre 4500 istituti dall'Istat, dal Ministero dei Beni culturali e del Turismo, dalle Regioni e dalle Provincie Autonome

Indagine sulle discriminazioni

La diffusione e le forme dei fenomeni discriminatori riferiti al genere, all'orientamento sessuale e all'appartenenza etnica

Per estrarre i dati

Incidenti stradali con lesioni a persone

La rilevazione degli incidenti stradali con lesioni a persone è svolta a cadenza mensile e rileva tutti gli incidenti stradali verificatisi nell'arco di un anno solare sull'intero territorio nazionale che hanno causato lesioni alle persone (morti entro il 30° giorno e feriti).

Per l'analisi dei dati è necessario utilizzare **la variabile peso**, in quanto i mlcro.STAT costituiscono un campione dei dati originali (vedi *Descrizione del file*).

Anno 2013

Ultimo aggiornamento: 09 giu 2015 12:42:05

[Nota metodologica](#)

Lista delle variabili

Download

Per effettuare il download



Autenticazione Utente

Benvenuto nel sistema unificato di autenticazione dell'ISTAT

Utente già registrato

Inserire il nome utente e password specificati in fase di registrazione. Si ricorda che il nome utente coincide con l'indirizzo di posta elettronica indicato nel modulo di registrazione.

Utente non ancora registrato

Per accedere alle funzionalità che prevedono l'autenticazione è necessario compilare il **modulo di registrazione**, specificando un indirizzo di posta elettronica. Tale indirizzo verrà utilizzato come identificativo per l'autenticazione nel sistema.

Password dimenticata

Nel caso l'utente abbia dimenticato la propria password, potrà inserirne una nuova mediante l'apposito **modulo di richiesta**. Le istruzioni per l'inserimento della nuova password verranno inviate all'indirizzo di posta elettronica specificato.

Autenticazione

Utente (e-mail)

Password

Login

- **Nuovo Utente?**
- **Password Dimenticata?**



Bisogna registrarsi



Registrazione nuovo utente

**Compilare il seguente modulo per registrarsi.
Salvo diversa indicazione, tutte le informazioni sono obbligatorie**

Dati anagrafici e sulla professione

Nome:	
Cognome:	
Sesso:	
Età:	
Residenza:	
Titolo di studio:	
Condizione professionale/posizione:	
Ente/settore economico:	

Preferenze

Lingua preferita per le comunicazioni:	
Temi di interesse (opzionale):	☐ Statistiche territoriali
	☐ Censimenti della popolazione, delle abitazioni e delle attività produttive
	☐ Struttura e competitività del sistema delle imprese
	☐ Indicatori congiunturali

Per effettuare il download



I dati sono disponibili in diversi formati (TXT, STATA, SAS, R)

Nome	Ultima modifica	Tipo	Dimensione
 METADATI	07/11/2015 21:41	Cartella di file	
 MICRODATI	07/11/2015 21:41	Cartella di file	
 !Leggi	09/06/2015 12:42	Chrome HTML Do...	4 KB

!Leggimi.html



L'archivio compresso (.zip) contiene

1. !Leggimi.html → guida introduttiva
2. directory MICRODATI

Il file di microdati, in formato txt, è fornito con campi separati **da tabulatore** e nella prima riga riporta il nome dei campi.

Riferimento temporale	Nome Microdati	Totale record	Nome T
Anno 2013	INCSTRAD_A2013.RData	79676	ISTAT

3. directory METADATI

Nome File (2)	Descrizione	Directory
ISTAT_PUF_INCSTRAD_Questionario_2013.pdf	Questionario	METADATI
ISTAT_PUF_INCSTRAD_NotaMetodologicaIndagine_2013.pdf	Nota Metodologica	METADATI
ISTAT_PUF_INCSTRAD_DescrizioneFile_2013.pdf	Descrizione	METADATI
PGM_2013_IT_DELIMITED.r	Script R	MICRODATI



RILEVAZIONE DEGLI INCIDENTI STRADALI CON LESIONI A PERSONE

DATA E LOCALITÀ DELL'INCIDENTE

ANNO MESE GIORNO

 ORA MINUTI

 PROVINCIA* Cod. Istat Comune

 COMUNE* Cod. Istat Provincia

 *L'elenco codici è disponibile sul sito www.istat.it (Strumenti/Definizioni e Classificazioni/Classificazioni adottate dall'Istat)

ORGANO DI RILEVAZIONE

Agente di Polizia Stradale ☐
 Carabiniere ☐
 Identificativo del Comando Staz. dei Carabinieri
 Agente di Pubblica Sicurezza ☐
 Agente di Polizia Municipale o Locale ☐
 Altri ☐
 Agente di Polizia Provinciale ☐

ORGANO COORDINATORE

Sezione Polizia Stradale ☐
 Gruppo Carabiniere ☐
 Off. Comunale di Statistica
 dei Capoluoghi di Provincia:
 Comune con oltre 250.000 abitanti ☐
 Altro capoluogo di Provincia ☐

1. Localizzazione dell'incidente

(Specificare la denominazione della strada, numero, eventuale n° civico in forma chiara e leggibile)

NELL'ABITATO

Strada urbana ☐
 Provinciale entro l'abitato ☐ SP N°
 Statale entro l'abitato ☐ SS N°
 Regionale entro l'abitato ☐ SR N°

FUORI ABITATO

Comunale extraurbana ☐
 Provinciale ☐ SP N°
 Statale ☐ SS N°
 Autostrada ☐ N°
 Altra strada ☐
 Regionale ☐ SR N°

Indicare il codice Istat corrispondente ai raccordi autostradali e alle tangenziali, disponibile nella seconda pagina di copertina del presente blocco.

 Progressiva chilometrica (Indicare chilometri ed ettometri)

TRONCO DI STRADA O AUTOSTRADA

diramazione; dir. A ☐
 dir. B; radd. ☐
 bis; dir. C ☐
 ter; bis dir. ☐
 quater; racc.; bis racc. ☐
 Autostrada carreggiata sinistra ☐
 Autostrada carreggiata destra ☐
 Autostrada svincolo entrata ☐
 Autostrada svincolo uscita ☐
 Autostrada svincolo tronco d.c. ☐
 Autostrada stazione ☐
 Altri casi ☐

2. Luogo dell'incidente

TIPO DI STRADA

Una carr. senso unico ☐
 Una carr. doppio senso ☐
 Due carreggiate ☐
 Più di 2 carreggiate ☐

PAVIMENTAZIONE

Strada pavimentata ☐
 Strada pavimentata
 dissestata ☐
 Strada non pavimentata ☐

INTERSEZIONE

Incrocio ☐
 Rotatoria ☐
 Intersezione segnalata ☐
 Intersezione con semaforo
 o vigile ☐
 Intersezione non segnalata ☐
 Passaggio a livello ☐

NON INTERSEZIONE

Rettilineo ☐
 Curva ☐
 Dosso, strottoia ☐
 Pendenza ☐
 Gall. illuminata ☐
 Gall. non illuminata ☐

FONDO STRADALE

Asciutto ☐
 Bagnato ☐
 Sdrucchiolato ☐
 Ghiacciato ☐
 Innevato ☐

SEGNALETICA

Assente ☐
 Verticale ☐
 Orizzontale ☐
 Verticale e
 orizzontale ☐
 Temporanea
 di cantiere ☐

CONDIZIONI
METEOROLOGICHE

Sereno ☐
 Nebbia ☐
 Pioggia ☐
 Grandine ☐
 Neve ☐
 Vento forte ☐
 Altro ☐

Ogni quesito diventa una variabile

- anno
- mese
- provincia
- comune
- giorno della settimana
- ora
- organo di rilevazione
- organo coordinatore
- localizzazione incidente
- denominazione della strada
- progressivo chilometrico
- tronco di strada o autostrada
- tipo di strada
- pavimentazione
- intersezione o non intersezione
- fondo stradale
- segnaletica
- condizioni meteorologiche.....**sono 124...10 in meno MFR**



(se estero) (veicolo a)	SOPPRESSIONE
identificazione veicoli coinvolti sigla (se estero) (veicolo b)	SOPPRESSIONE
identificazione veicoli coinvolti sigla (se estero) (veicolo c)	SOPPRESSIONE
Tipo di coordinata	SOPPRESSIONE
sistema di proiezione	SOPPRESSIONE
Minuti	SOPPRESSIONE
Codice identificativo Carabinieri	SOPPRESSIONE
Numero dei metri	SOPPRESSIONE
Altra strada	SOPPRESSIONE
Località	SOPPRESSIONE

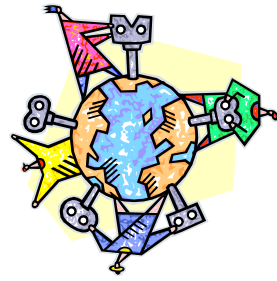
Nota metodologica



- La popolazione di riferimento
- La rilevazione e il trattamento dei dati
- La diffusione dei risultati dell'indagine
- Glossario
- Riferimenti bibliografici
- Contatti

E' obbligatorio studiarla bene!!!

Popolazione di riferimento o campo di osservazione



E' l'insieme delle unità statistiche di interesse, omogenee rispetto a uno o più caratteri

→traccia una linea di confine tra ciò che è rilevato e ciò che non lo è

La **popolazione di riferimento** o il **campo di osservazione** è costituito dall'insieme degli incidenti stradali verbalizzati da un'autorità di polizia verificatisi sul **territorio nazionale** nell'arco di un anno solare che **hanno causato lesioni alle persone** (morti entro i 30 giorni o feriti)

Unità statistica o unità di rilevazione



è l'unità elementare su cui vengono osservati i caratteri, viene anche detta unità di rilevazione

può essere “naturale”
o convenzionale

es. un essere umano, un'auto
es. la famiglia

L'unità di rilevazione è il singolo incidente stradale

La rilevazione e il trattamento dati

E' importante capire il processo di produzione del dato per capire tutta l'informazione potenziale che ha il dato statistico

- *Chi rileva?*
- *Quando?*
- *Cosa?*
- *Come?*

Dopo aver raccolto i dati, la fase successiva riguarda la validazione finale e la correzione per la congruenza delle informazioni contenute in ciascun record

*(controlli di qualità eliminazione dei incidenti fuori campo
eliminazione duplicati
correzione valori fuori range)*

Sempre dentro nella nota metodologica...

- Diffusione *come avviene e dove trovare i dati*
- Glossario *tutte le definizioni utilizzate nelle indagini, classificazioni*
- Riferimenti bibliografici *pubblicazioni per analizzare il fenomeno nel tempo e nello spazio*
- Contatti *le persone che seguono effettivamente l'intero processo di indagine*





Adesso andiamo più nel dettaglio...

Nome File (3)	Descrizione
INCSTRAD_Questionario_Anno 2013.pdf	Questionario
INCSTRAD_NotaMetodologicaIndagine_Anno 2013.pdf	Nota Metodologica
INCSTRAD_DescrizioneFile_Anno 2013.pdf	Descrizione del file

I file possono essere scaricati liberamente
una metodologia statistica garantisce **la tutela della riservatezza dei rispondenti**

.Valutazione del rischio di violazione

Elementi o riferimenti che una volta acquisiti possano essere utilizzati per identificare l'incidente o le persone coinvolte



. Modifica delle variabili

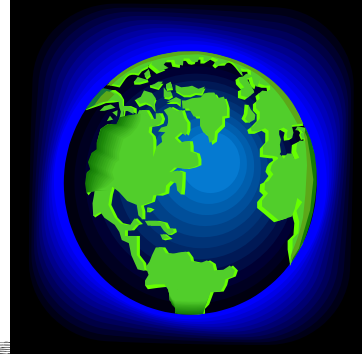
*Soppressione alcune variabili (stato psicofisico, targa, informazione casco o cintura,...)
Ricodifica globale (data incidente, anno immatricolazione, anno di rilascio patente , età...)
Arrotondamento (variabile progressiva chilometrica)
Tecniche sottocampionamento*

campionamento effettuato per i mlcro.STAT

I file di microdati mlcro.STAT sono ottenuti attraverso tecniche di sottocampionamento dal corrispondente file di microdati per la ricerca.

*Per ottenere stime dal file mlcro.STAT confrontabili con le stime ottenibili dal corrispondente file di microdati per la ricerca è opportuno utilizzare **i coefficienti di riporto all'universo***

coefficienti di riporto all'universo



ogni unità campionaria
rappresenta se stessa e altre unità del suo strato che non sono entrate a far
parte del campione

quindi

le risposte date dalle unità campionarie devono essere **RIPORTATE**
all'**UNIVERSO** in modo che
descrivano tutte le unità della popolazione

cioè si moltiplicano i risultati per i
PESI CAMPIONARI detti anche
coefficienti di riporto all'universo

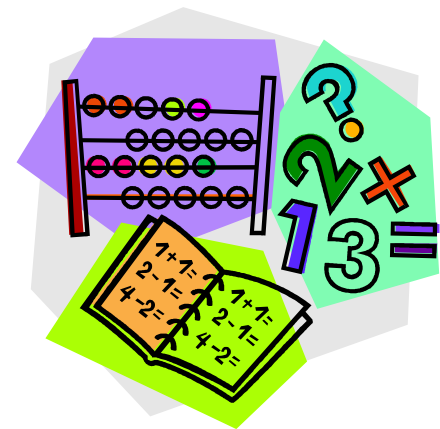
cosa significa moltiplicare per i pesi?



significa immaginare che le unità che non sono state estratte siano ben rappresentate da quelle campionarie

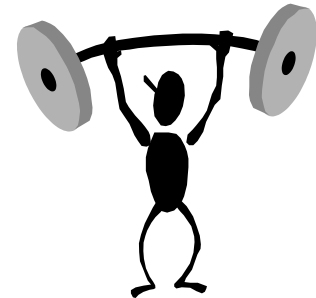
ogni unità campione PESA
tanto quante sono le unità universo
che è chiamata a rappresentare

come si calcolano i pesi?



i pesi sono pari a
l'inverso della probabilità di estrazione

Strato	Unità universo N	Unità campionarie n	Probabilità di estrazione	Coefficiente di riporto all'universo
1	1.200	140	0,117	8,571
2	1.600	160	0,100	10,000
3	2.700	280	0,104	9,643
...	...	...	...	...
...	...	...	...	...
k	800	68	0,085	11,765



perché si calcolano così?

perché devono **essere inversamente proporzionali alla numerosità campionaria:**

più sono le unità campionarie
meno devo pesarle per far loro rappresentare
le unità non intervistate

Esempio:

se da un gruppo di 1.000 persone

estraggo un campione di 100

(prob. estraz. = $100 / 1.000 = 1 / 10$)

ciascuna unità campionaria ne rappresenta 10

(se stessa + altre 9 non intervistate)

*quindi, le risposte delle 100 unità campionarie devono essere moltiplicate per un peso
pari a 10*

Come dobbiamo comportarci con la variabile peso?

Bisogna sempre ricordare il tipo di rilevazione dati (*rilevazione totale o campionaria*)

Per oggi è importante:

.capire il perché della variabile peso

.imparare a esplorare i microdati a scopo didattico

Quando saremo super esperti scopriremo che esistono procedure informatiche che permettono di lavorare con dati campionari ...



saremo in grado di elaborare tabelle o grafici
che tengono conto dei pesi e
i risultati saranno confrontabili con i dati pubblicati

**Esiste una mappa per orientarsi
nei file microdati???**



...il tracciato record

Nome campo	Descrizione
Num. ordine	Num. progressivo della variabile nel tracciato record
Lunghezza	Lunghezza in bytes della variabile
Nome campo	Nome breve o nome informatico della variabile (es: num. progressivo del quesito nel questionario)
Tipo variabile	Indica se la variabile è categorica, è possibile visualizzare la classificazione associata alla variabile cliccando su <u>Categorica</u>
Nome della variabile di aggregazione	Eventuale nome della variabile che aggrega la variabile descritta (es: la variabile 'Provincia' aggrega la variabile 'Comune')
Note	Eventuali ulteriori indicazioni sulla variabile
Tipo record	L'eventuale presenza nel tracciato di valori diversi nel campo tipo record indica l'esistenza di piu' variabili definite nelle stesse posizioni fisiche del tracciato record che hanno significati logici diversi. Pertanto le variabili appartenenti a tipi record diversi fra loro, non costituiscono variabili aggiuntive, ma sono ridefinizioni della stessa variabile al fine di poter collegare ad esempio classificazioni diverse.
Nome variabile	Descrizione estesa della variabile (es. testo del quesito del questionario)
Formato campo	A per alfabetico o N per numerico
Numero decimali	Se la variabile è numerica con decimali viene indicato il numero di decimali
Separatore decimali	Se presente il separatore tra le unità ed i decimali viene indicato il simbolo adottato
Segno aritmetico	Se presente il segno aritmetico viene indicato come '+' o '-'
Unità di misura	Unità con la quale è stata misurata la variabile

Variabili quantitative e qualitative

INDAGINE: Rilevazione degli incidenti stradali con lesioni a persone Anno 2011

num. ordine	Lunghezza	Nome Campo	TipoVariabile	Aggregazione	Note	TipoRecord	Nome Variabile	Formato Campo	Num. decimali	Separatore decimali	Segno aritmetico	Unità Misura
1	2	anno	-		ULTIME DUE CIFRE DELL'ANNO	0	anno	A				
2	2	mese	Categorica		COD. DA 01 A 12	0	mese	A				
3	3	provincia	Categorica			0	provincia	A				
4	3	comune	Categorica	provincia	COD. COMUNE NELL'AMBITO DELLA PROV.	0	comune	A				
5	1	giorno_settimana	Categorica			0	giorno della settimana	A				
6	2	ora	-		DA 01 A 24 (25=ORA IMPRECISATA)	0	ora	A				
7	1	organo_di_rilevazione	Categorica			0	organo di rilevazione	A				
8	1	organo_coordinatore	Categorica			0	organo coordinatore	A				
9	1	localizzazione_incidente	Categorica			0	localizzazione incidente	A				
10	3	denominazione_della_strada	-		NUMERO DELLA STRADA O DELL'AUTOSTRADA	0	denominazione della strada	A				
11	3	progressivo_chilometrico	-		N° KM	0	progressivo chilometrico	A				
12	2	tronco_di_strada_o_autostrada	Categorica			0	tronco di strada o autostrada	A				
13	1	tipo_di_strada	Categorica			0	tipo di strada	A				
14	1	pavimentazione	Categorica			0	pavimentazione	A				
15	2	intersezione o non interse	Categorica			0	intersezione o non	A				

Come fare per elaborare i microdati?

INCSTRAD_Microdati_Anno_2013.txt ×											
anno	mese	provincia		comune	giorno_settimana			ora	organo_di_rilevazione		
13	1	007	043	6	7	1	1	7	A05	007	
13	1	007	057	4	20	1	1	9	028	000	12
13	2	007	066	7	18	1	1	3	026	010	12
13	3	007	074	1	1	1	1	7	A05	011	
13	4	007	058	3	19	1	1	6	026	009	12
13	4	007	011	4	4	1	1	7	A05	009	
13	5	007	003	1	18	1	1	9	038	000	12
13	5	007	043	3	9	1	1	7	A05	007	
13	5	007	003	1	16	1	1	7	A05	000	
13	6	007	073	7	16	1	1	7	A05	006	
13	6	007	027	4	14	1	1	7	A05	009	
13	6	007	066	6	12	1	1	1	000	000	12
13	7	007	008	6	14	1	1	9	507	000	12
13	7	007	054	2	11	1	1	7	A05	010	

Prima di mettere le mani finalmente in pasta
e utilizzare il programma R

Vediamo in anteprima alcuni facili comandi di R per scoprire cosa c'è
dentro un file di microdati
concentrandoci sui risultati



esempio file degli incidenti stradali

Matrice dei dati → Un data.frame

Il data.frame è una lista in cui ogni componente è un vettore di una prefissata lunghezza

Il data.frame è organizzato e stampato a schermo come una matrice di dati di n righe e tante colonne quante sono le componenti della lista.

Un data frame può essere considerato come una matrice le cui colonne rappresentano **dati eterogenei**

Data frame

Dati:	val.num	val.car.	val. log.	val.num	val.car.
Dato1	3.45	ATTA	TRUE	0.45	CCAT
Dato2	5.67	TGAT	FALSE	0.91	TATT
Dato3	1.45	TATA	FALSE	3.78	CCCC
Dato4	4.56	TGAG	TRUE	8.03	GAGA
Dato5	8.09	CCCG	TRUE	8.09	AGAG
Dato6	3.11	CATG	TRUE	4.56	ATAT
Dato7	1.40	TGAG	FALSE	1.80	GCTA
Dato8	7.73	GGAC	TRUE	5.90	TGAT

Matrice (array bidimensionale)

Dati:	val.num	val.num	val.num	val.num	val.num
Dato1	3.45	1.20	2.54	0.45	1.45
Dato2	5.67	2.95	5.44	0.91	0.95
Dato3	1.45	5.21	3.60	3.78	6.78
Dato4	4.56	8.00	2.12	8.03	8.73
Dato5	8.09	1.65	6.00	8.09	6.65
Dato6	3.11	2.58	3.98	4.56	3.56
Dato7	1.40	0.95	2.72	1.80	5.21
Dato8	7.73	8.82	4.43	5.90	3.18

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

Project: (None)

Environment History

To Console To Source

dati<-DF_INCSTRAD_A2013
dati
names(dati)
var(dati)

Demos in package 'base':

error.catching More examples on catching and handling errors

is.things Expl and new

recursion Using inte An i

scoping

Demos in package 'grap

Hershey Table Hers

Japanese Table the

graphics A sho capa

Console ~/ ↩

Platform: i386-w64-ming

R is free software and
You are welcome to redi
Type 'license()' or 'li

R is a collaborative pr
Type 'contributors()' or
'citation()' on how to

Type 'demo()' for some
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Open File

Look in: C:\Users\Administrator\Documents

Name	Size	Type	Date Modified
Cartella Scambio Bluetooth		Fil...der	06/07/... 15:17
Modelli di Office personalizzati		Fil...der	17/08/... 11:36
My SAS Files		Fil...der	15/10/... 11:18
PDF Architect		Fil...der	16/07/... 23:00
Schemi		Fil...der	17/07/... 12:23
.Rhistory	1 KB	Rhi...ile	10/11/... 11:38
DF_INCSTRAD_A2013.RData	42 bytes	RDa...ile	04/11/... 23:33

File name: DF_INCSTRAD_A2013.RData

Files of type: All Files (*)

Open Cancel

R demos * DF_INCSTRAD_A2013 *

Filter

	anno	mese	provincia	comune	giorno_settimana	ora	organo_di_rilevazione	organo_coordinato
1	13	1	7	43	6	7	1	
2	13	1	7	57	4	20	1	
3	13	2	7	66	7	18	1	
4	13	3	7	74	1	1	1	
5	13	4	7	58	3	19	1	
6	13	4	7	11	4	4	1	
7	13	5	7	3	1	18	1	
8	13	5	7	43	3	9	1	
9	13	5	7	3	1	16	1	
10	13	6	7	73	7	16	1	
11	13	6	7	27	4	14	1	

Showing 1 to 12 of 79,676 entries

Console ~/

R is a collaborative project with many contributors.
 Type 'contributors()' for more information and
 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
 'help.start()' for an HTML browser interface to help.
 Type 'q()' to quit R.

```
> load("C:/Users/Administrator/Dropbox/2015 11 10 Micro.Stat Treviso/INCSTRAD_2013_IT_R/MICRODATI/DF_INCSTRAD_A2013.RData")
> load("C:/Users/Administrator/Dropbox/2015 11 10 Micro.Stat Treviso/INCSTRAD_2013_IT_R/MICRODATI/DF_INCSTRAD_A2013.RData")
>
> view(DF_INCSTRAD_A2013)
```

Environment History

Import Dataset

Global Environment

Data

DF_INCSTRAD_A2013 79676 obs. of 124 variables

values

comune_Fact Large factor (79676 elements, 667
 condizioni_meteor... Factor w/ 7 levels "sereno","Nebb
 direttorio "G:/A_MICROSTAT_DATASET/INCSTRAD_
 fileInput "G:/A_MICROSTAT_DATASET/INCSTRAD_
 fondo_stradale_Fa... Factor w/ 5 levels "Asciutto","Ba

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

	Name	Size
	.Rhistory	5.1 KB
	Cartella Scambio Bluetooth	
	DF_INCSTRAD_A2013.RData	42 B
	Modelli di Office personalizzati	
	My SAS Files	
	PDF Architect	
	Schemi	

Come è fatto il data.frame degli incidenti?



```
dati <- DF_INCSTRAD_A2013
```

dati è il nome che ho dato al data.frame

Possiamo sapere da quante colonne è formato → variabili

```
> ncol(dati)
```

```
[1] 124
```



quante righe → unità statistiche

```
> nrow(dati)
```

```
[1] 79676
```



e le sue dimensioni

```
> dim(dati)
```

```
[1] 79676 124
```

Qual è il contenuto del data.frame?

`str(dati)`

- . Numero delle osservazioni
- . Numero delle variabili caricate
- . Il nome e il tipo delle variabili

```
> str(dati)
'data.frame': 79676 obs. of 124 variables:
 $ anno          : int  13 13 13 13 13 13 13 13 13 13 ...
 $ mese          : num  1 1 2 3 4 4 5 5 5 6 ...
 $ provincia     : int  7 7 7 7 7 7 7 7 7 7 ...
 $ comune       : int  43 57 66 74 58 11 3 43 3 73 ...
 $ giorno_settimana : int  6 4 7 1 3 4 1 3 1 7 ...
 $ ora          : num  7 20 18 1 19 4 18 9 16 16 ...
 $ organo_di_rilevazione : int  1 1 1 1 1 1 1 1 1 1 ...
 $ organo_coordinatore : int  1 1 1 1 1 1 1 1 1 1 ...
 $ localizzazione_incidente : int  7 9 3 7 6 7 9 7 7 7 ...
 $ denominazione_della_strada : Factor w/ 627 levels " ", "000", "001", ...: 537 30 28
37 28 537 40 537 537 537 ...
 $ progressivo_chilometrico : num  7 0 10 11 9 9 0 7 0 6 ...
 $ tronco_di_strada_o_autostrada : num  NA 12 12 NA 12 NA 12 NA NA NA ...
 $ tipo_di_strada : int  3 2 2 1 2 3 2 3 2 3 ...
 $ pavimentazione : int  1 1 1 1 1 1 1 1 1 1 ...
 $ intersezione_o_non_interse1 : num  7 3 4 8 4 7 4 7 8 8 ...
 $ fondo_stradale : int  1 1 1 5 1 1 1 2 1 1 ...
 $ segnaletica : int  4 4 4 4 4 4 4 4 4 3 ...
```

`names(dati)` → vettore dei nomi usati per identificare le colonne

Ho 79676 osservazioni, posso fare un riassunto?

Si ma con molta ATTENZIONE!!!

La funzione è `summary (dati)`

Si può fare la media della variabile provincia??

```
> summary(dati)
```

anno	mese	provincia	comune	giorno_settimana
Min. :13	Min. : 1.000	Min. : 1.00	Min. : 1.00	Min. :1.000
1st Qu.:13	1st Qu.: 4.000	1st Qu.: 16.00	1st Qu.: 15.00	1st Qu.:2.000
Median :13	Median : 7.000	Median : 48.00	Median : 33.00	Median :4.000
Mean :13	Mean : 6.664	Mean : 46.39	Mean : 60.18	Mean :3.708
3rd Qu.:13	3rd Qu.:10.000	3rd Qu.: 63.00	3rd Qu.: 91.00	3rd Qu.:5.000
Max. :13	Max. :12.000	Max. :110.00	Max. :315.00	Max. :7.000

ora	organo_di_rilevazione	organo_coordinatore	localizzazione_incidente
Min. : 0.00	Min. :1.00	Min. :0.000	Min. :0.000
1st Qu.:10.00	1st Qu.:4.00	1st Qu.:3.000	1st Qu.:1.000
Median :13.00	Median :4.00	Median :4.000	Median :1.000
Mean :13.42	Mean :3.63	Mean :3.202	Mean :1.381
3rd Qu.:17.00	3rd Qu.:4.00	3rd Qu.:4.000	3rd Qu.:1.000
Max. :25.00	Max. :6.00	Max. :5.000	Max. :9.000
		NA's :2231	

numero di
osservazioni mancanti



Se voglio lavorare con una sola variabile?

Per lavorare con una sola variabile posso richiamarla con la seguente notazione

nome data.frame \$nome variabile

Se voglio la variabile mese *dati\$mese*
 dati\$organo_di_rilevazione

può troppo laborioso!

essere eccessivamente verbosa e poco conveniente.

La funzione `attach` “rende disponibili” nel cammino di ricerca corrente i nomi delle componenti come se fossero variabili “stand alone”

`attach(dati)` posso lavorare con le variabili solo con il loro nome

Attenzione ai Not Available



Bisogna fare attenzione ai dati mancanti o non validi
Bisogna utilizzare una opportuna codifica per riconoscerli,
→ caratteri particolari (":", ".", "999", "NA", "NAN", ecc.)

NA è la variabile che controlla i dati mancanti o non validi
(NA sta per *not available*),

NA è una **variabile logica** se vera segnala l'assenza di un dato valido....

```
mean(dati$cilindrata__veicolo__a_)
```

```
[1] NA → R non riesce a calcolarla...
```

La variabile `cilindrata_veicolo_a_` è una variabile
QUANTITATIVA con molti valori mancanti

NOT AVAILABLE

```
> summary(cilindrata_veicolo_a_)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
40	1000	1299	1447	1800	9999

NA's

26307

```
> cilindrata=na.omit(cilindrata_veicolo_a_)
```

```
> summary(cilindrata)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
40	1000	1299	1447	1800	9999

```
> length(cilindrata_veicolo_a_)
```

```
[1] 63243
```

```
> length(cilindrata)
```

```
[1] 36936
```

```
> |
```

La variabile `cilindrata` è una nuova variabile ottenuta eliminando
i valori mancanti

Tabella di frequenza univariata

Modalità	Frequenza assoluta	Frequenza relativa	Frequenza percentuale	Frequenza relativa
X_1	n_1	f_1	p_1	$f_j = \frac{n_j}{n}$
X_2	n_2	f_2	p_2	
.....				
X_j	n_j	f_j	p_j	Frequenza percentuale
.....				
X_k	n_k	f_k	p_k	$p_j = f_j * 100$
Totale	n	1	100	

$$\sum_{j=1}^k n_j = n \quad \sum_{j=1}^k f_j = 1 \quad \sum_{j=1}^k p_j = 100$$

Calcoliamo le frequenze con un clic!

```
> table(mese)
```

```
mese  
 1    2    3    4    5    6    7    8    9   10   11   12  
5862 5097 5901 6261 7655 8030 7820 5675 7225 7257 6751 6142
```

Tabella di frequenze
assolute

```
> n=sum(table(mese))
```

```
> n
```

```
[1] 79676
```

Il totale n

```
> table(mese)/n
```

```
mese  
      1      2      3      4      5      6      7      8      9     10     11     12  
0.07357297 0.06397158 0.07406245 0.07858075 0.09607661 0.10078317 0.09814750  
0.07122597 0.09067975 0.09108138 0.08473066 0.07708720
```

Tabella di frequenze
relative

```
> table(mese)/n*100
```

```
mese  
      1      2      3      4      5      6      7      8      9     10     11     12  
7.357297 6.397158 7.406245 7.858075 9.607661 10.078317 9.814750 7.122597  
9.067975 9.108138 8.473066 7.708720
```

Tabella di frequenze
percentuali

```
> round(table(mese)/n*100,2)
```

```
mese  
 1    2    3    4    5    6    7    8    9   10   11   12  
7.36 6.40 7.41 7.86 9.61 10.08 9.81 7.12 9.07 9.11 8.47 7.71  
> |
```

Mettiamo le etichette!

La variabile mese è una variabile QUALITATIVA

```
mese_Fact <- factor(DF_INCSTRAD_A2013$mese,  
levels=c("1", "2", "3", "4", "5", "6", "7", "8", "9",  
"10", "11", "12"),  
labels=c("Gennaio", "Febbraio", "Marzo", "Aprile",  
"Maggio", "Giugno", "Luglio", "Agosto",  
"Settembre", "Ottobre", "Novembre", "Dicembre"))
```

Scriviamo la tabella in un file chiamato tabella.txt, il carattere separatore è il tabulatore e la virgola è il separatore dei decimali e facciamo il grafico a barre!
...usando le parole di R

```
freqperc=round(table(mese_Fact)/n*100,2)  
write.table(x=freqperc,file="tabella.txt",sep="\t",dec=",")  
barplot(freqperc,main="Incidenti per mese",col="red")
```

Ecco la tabella e il suo grafico!

Scriviamo la tabella in un file chiamato `tabella.txt`, il carattere separatore è il tabulatore e la virgola è il separatore dei decimali e facciamo il grafico a barre!

...usando le parole di R

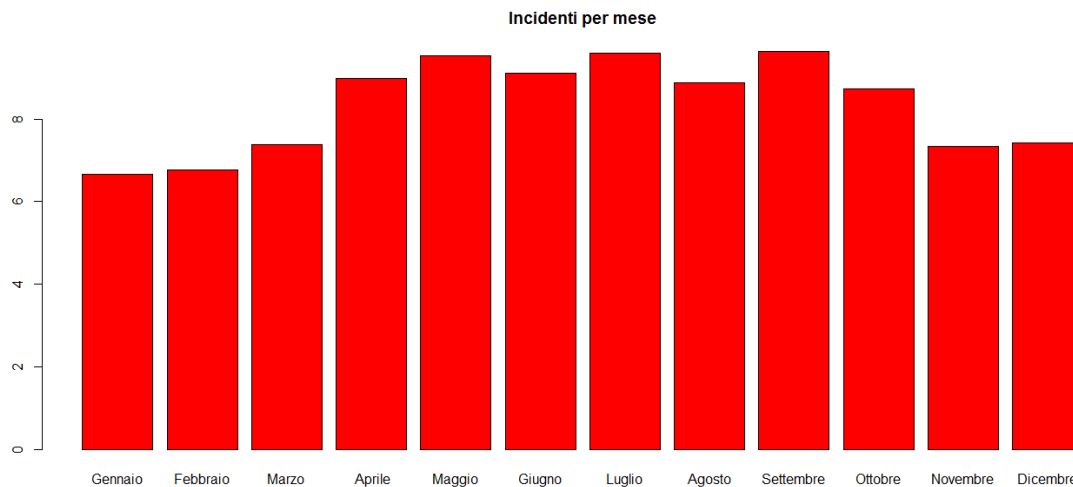
```
freqperc=round(table(mese_Fact)/n*100,2)
```

```
write.table(x=freqperc, file="tabella.txt", sep="\t", dec = ",", "
```

```
barplot(freqperc, main="Incidenti per mese", col="red")
```

Ecco fatto!

	mese1	Freq
1	Gennaio	6,66
2	Febbraio	6,76
3	Marzo	7,38
4	Aprile	8,99
5	Maggio	9,53
6	Giugno	9,1
7	Luglio	9,59
8	Agosto	8,88
9	Settembre	9,63
10	Ottobre	8,72
11	Novembre	7,33
12	Dicembre	7,43
		100



```
barplot(freqperc, main="Incidenti per mese", col="red")
```

Grazie!

Susi Osti

osti@istat.it