201ab Quantitative methods
L.02: Summarizing data

Tidy data

The Fellowship Of The Ring

Race	Female	Male
Elf	1229	971
Hobbit	14	3644
Man	0	1995

The Two Towers

Race	Female	Male
Elf	331	513
Hobbit	0	2463
Man	401	3589

The Return Of The King

Race	Female	Male
Elf	183	510
Hobbit	2	2673
Man	268	2459

Untidy data: separate tables, separate columns (wide format)

Tidy data: one table individuated by column, one measurement column (long format)

Film	Race	Gender	Words
The Fellowship Of The Ring	Elf	Female	1229
The Fellowship Of The Ring	Elf	Male	971
The Fellowship Of The Ring	Hobbit	Female	14
The Fellowship Of The Ring	Hobbit	Male	3644
The Fellowship Of The Ring	Man	Female	0
The Fellowship Of The Ring	Man	Male	1995
The Two Towers	Elf	Female	331
The Two Towers	Elf	Male	513
The Two Towers	Hobbit	Female	0
The Two Towers	Hobbit	Male	2463
The Two Towers	Man	Female	401
The Two Towers	Man	Male	3589
The Return Of The King	Elf	Female	183
The Return Of The King	Elf	Male	510
The Return Of The King	Hobbit	Female	2
The Return Of The King	Hobbit	Male	2673
The Return Of The King	Man	Female	268
The Return Of The King	Man	Male	2459

Useful libraries for this:

dplyr:: tidyr::

Crazy data

Mine CetinkayaRundel
@minebocek

Follow

That's some coding scheme for height

Label: Reported Height in Feet and Inches
Section Name: Demographics
Section Number: 8
Question Number: 20
Column: 182-185
Type of Variable: Num
SAS Variable Name: HEIGHT3
Question Prologue:
Question: About how tall are you without shoes? (If respondent answers in metrics, put a 9 in the first column)[Round fractions down.]

Value	Value Label	Frequency	Percentage	Weighted Percentage
200 - 711	Height (ft/inches) Notes: 0 _ / _ _ = feet / inches	467,177	97.43	96.15
7777	Don't know/Not sure	3,875	0.81	1.29
9000 - 9998	Height (meters/centimeters) Notes: The initial '9' indicates this was a metric value.	2,165	0.45	0.89
9999	Refused	6,277	1.31	1.67
BLANK	Not asked or Missing	6,809	.	.

Make sure data are tidy & variables are sensible before summarizing.

“statistics” of samples

- *Statistic:* measures some property of a sample by applying a function or algorithm to the sample.
  - *Descriptive* **statistics** summarize some features of a sample
  - *Estimators* are statistics that estimate a model parameter.
  - *Test* **statistics** are compared to their null hypothesis distribution to do significance testing.

- Difference: how we use the statistic.

- Descriptive statistics are usually (implicit) estimators
  - Most common ones are estimators of various “moments”

Rest of class in R / whiteboard