Statistical Literacy — MINT

Lecture 2: Organize data and compute centrality indicators

Rémi Viné

The Graduate Institute | Geneva

October 2nd, 2023

Outline

Housekeeping

- By now, you should have filled the survey and submitted your first problem set
- Problem set 2 is now available

Data sets can be quite complex

	ID	IY	country	Region	income	IDA	FCY	FCM	type	stype	...	CPI2021	investment_real	realphysicalassets	realfeestogovernment	RealGDP	name
0	1334	1990	Argentina	LAC	Upper middle income	Non-IDA	1990	September	Brownfield	Rehabilitate, operate, and transfer	...	1.980824	305.046906	305.046906	0.000000	2.799942e+11	Intercity Roads - Corridor 13
1	1337	1990	Argentina	LAC	Upper middle income	Non-IDA	1990	September	Brownfield	Rehabilitate, operate, and transfer	...	1.980824	481.340240	481.340240	0.000000	2.799942e+11	Intercity Roads - Corridor 18
2	1330	1990	Argentina	LAC	Upper middle income	Non-IDA	1990	September	Brownfield	Rehabilitate, operate, and transfer	...	1.980824	350.605835	350.605835	0.000000	2.799942e+11	Intercity Roads - Corridor 7, 8 & 9
3	1346	1990	Mexico	LAC	Upper middle income	Non-IDA	1990	March	Greenfield project	Build, operate, and transfer	...	1.980824	67.744179	67.744179	0.000000	5.174974e+11	San Martin - Tlaxcala Toll Road
4	991	1990	Mexico	LAC	Upper middle income	Non-IDA	1990	November	Divestiture	Full	...	1.980824	4124.075684	0.000000	4124.075684	5.174974e+11	Telefonos de Mexico (Telmex)
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10744	10970	2021	Turkey	ECA	Upper middle income	Non-IDA	2021	March	Greenfield project	Build, operate, and transfer	...	1.000000	24.000000	24.000000	NaN	NaN	Miskevank Waste-to-Energy Power Plant
10745	11195	2021	Madagascar	AFR	Low income	IDA	2021	December	Management and lease contract	Management contract	...	1.000000	NaN	NaN	NaN	NaN	Tomasina port concession
10746	11010	2021	Vietnam	EAP	Lower middle income	Non-IDA	2021	May	Greenfield project	Build, operate, and transfer	...	1.000000	490.000000	490.000000	NaN	NaN	Dien Chau - Bai Vot section of the North - Sou...
10747	11077	2021	Burkina Faso	AFR	Low income	IDA	2021	October	Greenfield project	Build, own, and operate	...	1.000000	46.869999	46.869999	NaN	NaN	Kodeni Solar PV Plant
10748	10942	2021	China	EAP	Upper middle income	Non-IDA	2021	February	Greenfield project	Build, operate, and transfer	...	1.000000	737.179993	737.179993	NaN	NaN	Yunlong-Lanping section of Dali-Yangbi-Yunlong...

10749 rows × 59 columns

Visualization for qualitative data

- Bar chart
- Pie chart

Organize qualitative data

- Construct a frequency table from a raw data

	type
0	Brownfield
1	Brownfield
2	Brownfield
3	Greenfield project
4	Divestiture
...	...
10744	Greenfield project
10745	Management and lease contract
10746	Greenfield project
10747	Greenfield project
10748	Greenfield project

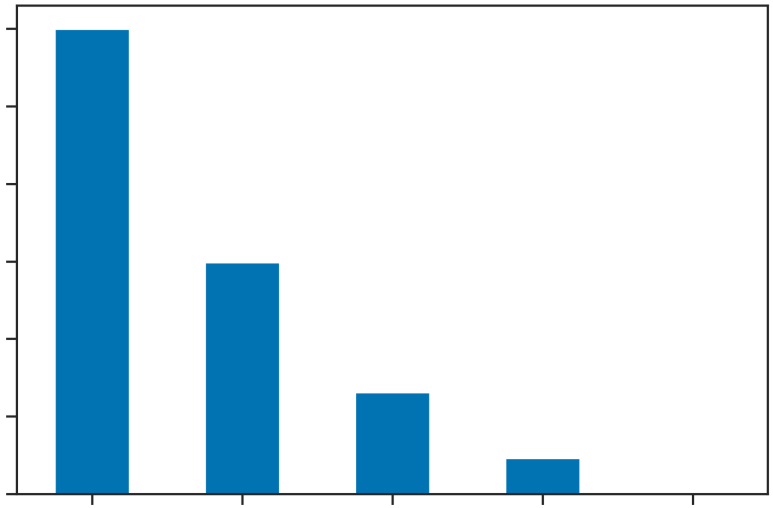
10749 rows × 1 columns

- Frequency: number of occurrence of a modality
- Relative frequency: share of a given frequency with respect to total occurrences

	Frequency	Relative frequency
Brownfield	2991	0.277742
Divestiture	1315	0.122110
Greenfield project	5998	0.556969
Management and lease contract	464	0.043087
Not Available	1	0.000093

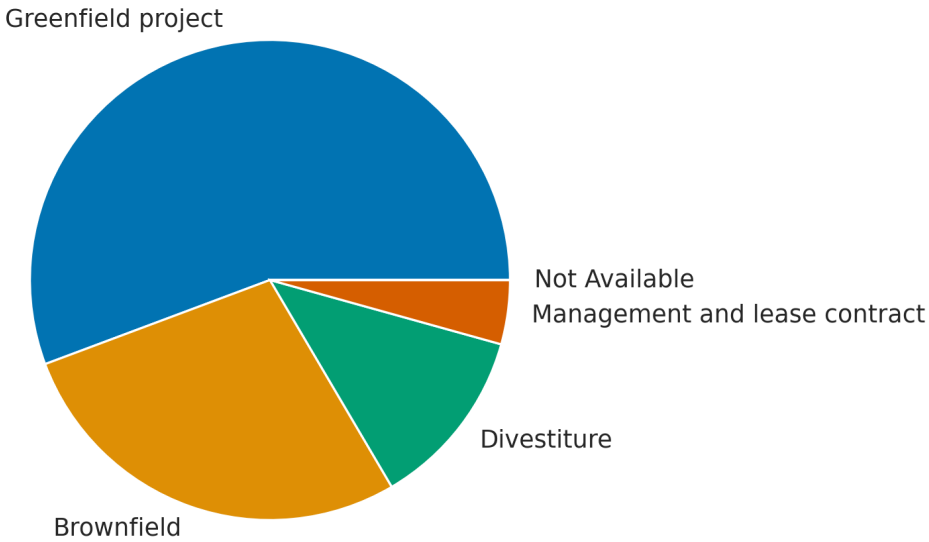
Bar chart

- Bars represent the frequency (or relative frequency) of each modality
- Bars do not touch each other
- Categories (or modalities) are on the horizontal axis



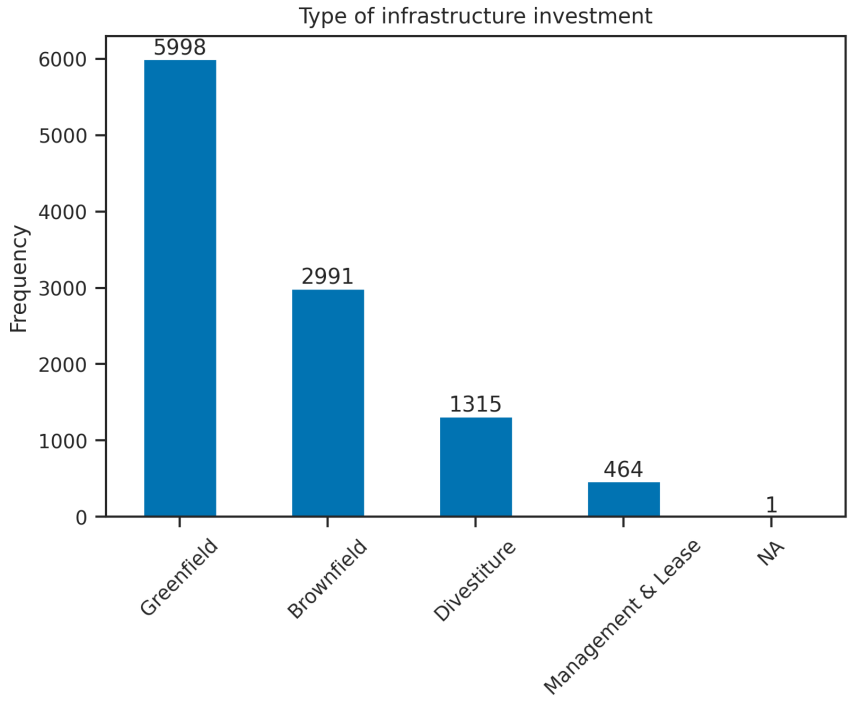
Pie chart

- Each piece (also called wedge) is proportional to the relative frequency of a modality
- Wedges add up to 100

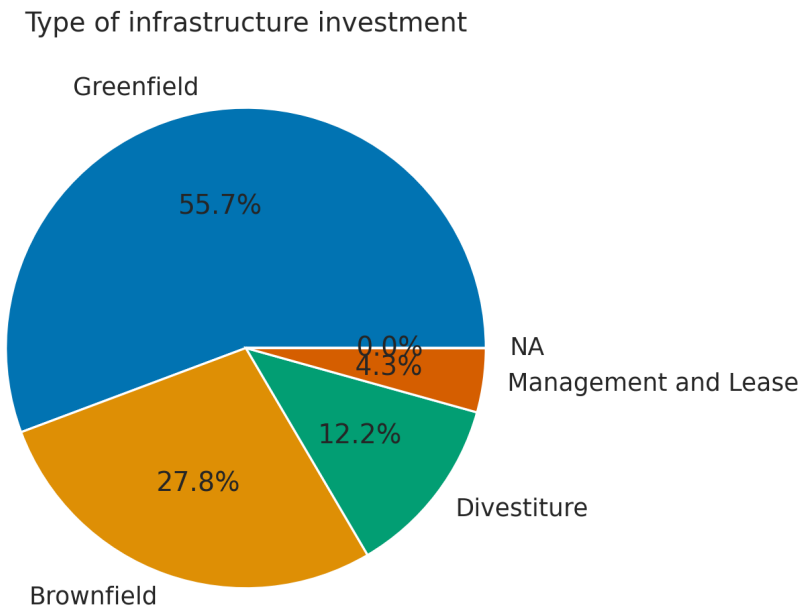


Missing pieces on the former visualizations

- Titles
- Labels
- Units of measurement
- Data source



Data source: Private Participation in Infrastructure, World Bank



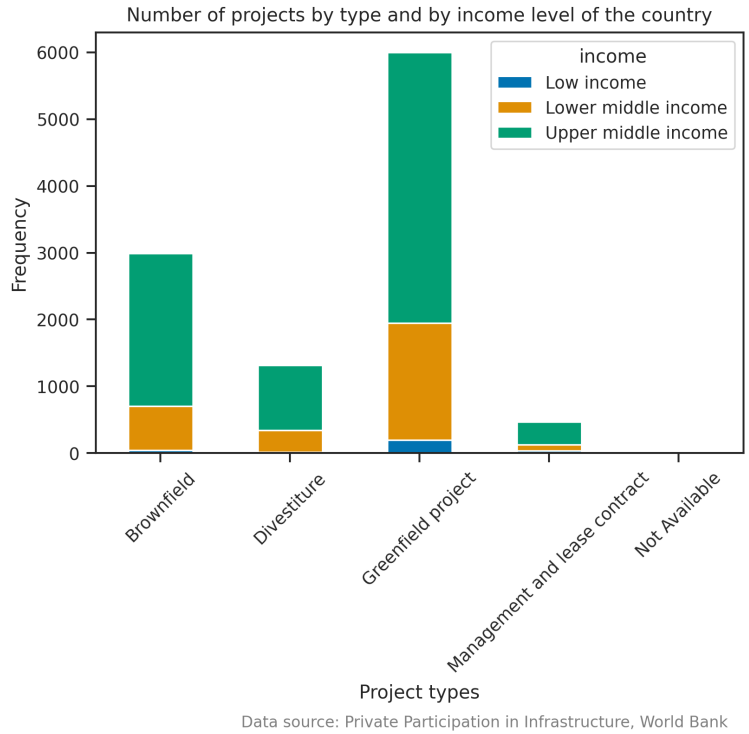
Data source: Private Participation in Infrastructure, World Bank

Representing two categorical variables together

- Contingency table
  - One variable in rows
  - One variable in columns

income type	Low income	Lower middle income	Upper middle income
Brownfield	41	661	2289
Divestiture	15	327	973
Greenfield project	198	1750	4050
Management and lease contract	36	88	340
Not Available	0	0	1

- Visualization: stacked bars



- Stacked bars can be *normalized* to all be 100%, depending on the message to convey

Quantitative variables

- Discrete
  - Countable values
- Continuous
  - Any possible values are possible (fractions, *etc.*)
- Beware of the unit of measurement!
  - *period* is expressed in months
  - *investment_real* is expressed in million USD

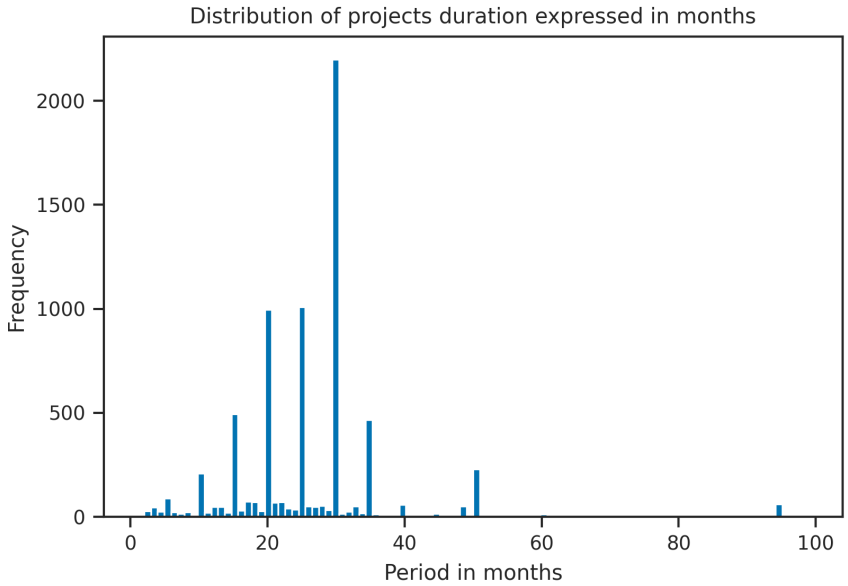
	period	investment_real
0	13.0	305.046906
1	28.0	481.340240
2	13.0	350.605835
3	30.0	67.744179
4	NaN	4124.075684
...	...	...
10744	NaN	24.000000
10745	15.0	NaN
10746	16.0	490.000000
10747	25.0	46.869999
10748	34.0	737.179993

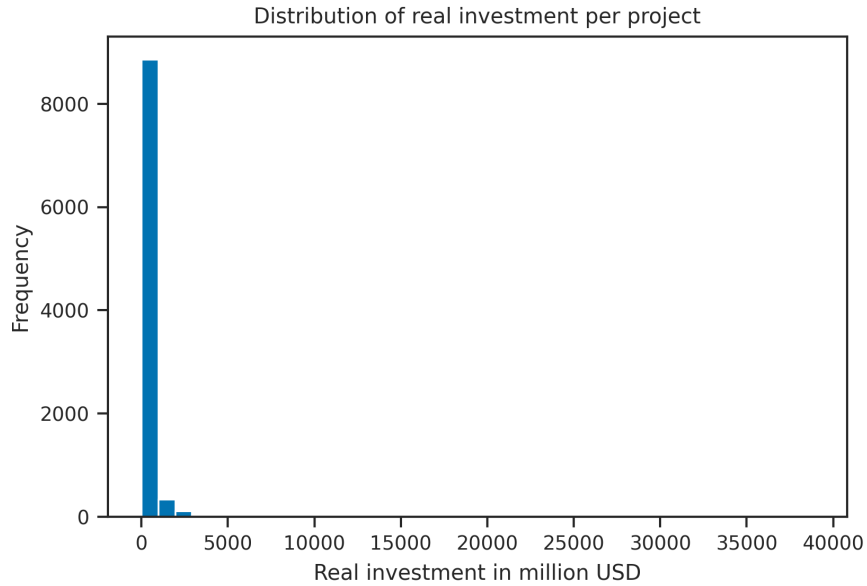
10749 rows × 2 columns

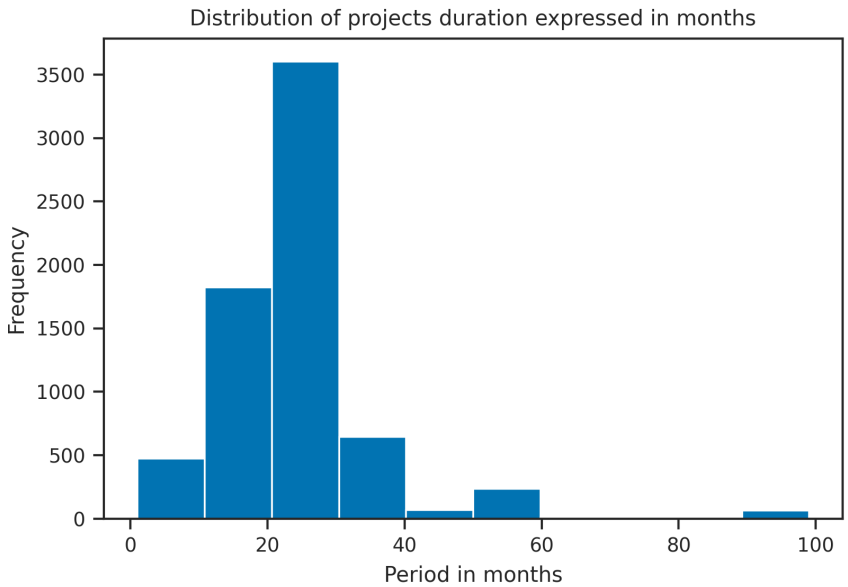
Visualization for quantitative data

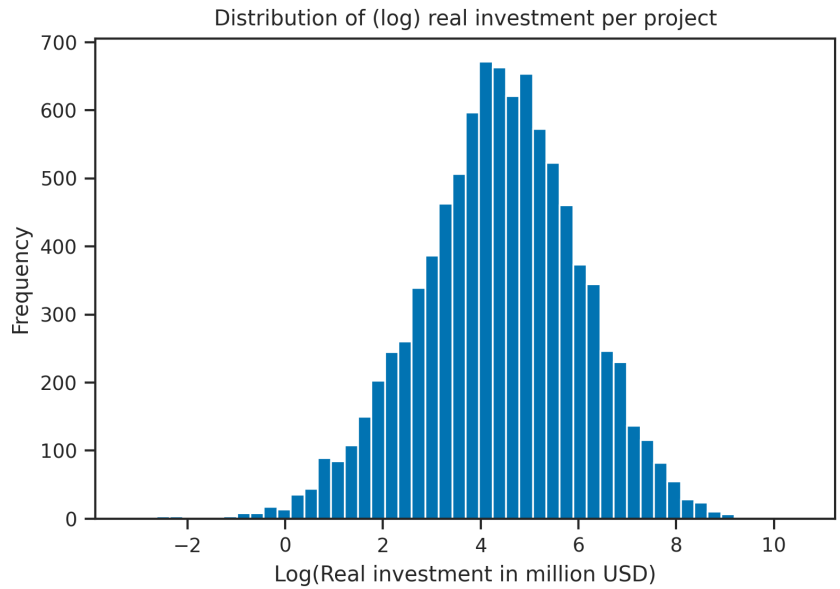
- Histograms (discrete and continuous)
  - Data is displayed in non-overlapping but adjacent bins
  - Heights of bars give the relative frequency of the bin
  - Shape gives an idea of the **distribution**
- Dot plots (discrete)
  - Each observation is represented by a dot on the horizontal axis
- Stem and leaf (discrete)
  - Each observation's value is split in two parts (the stem and the leaf)
  - Here, the tens are the stem and the units are the leaf

Visualization for quantitative data: histogram



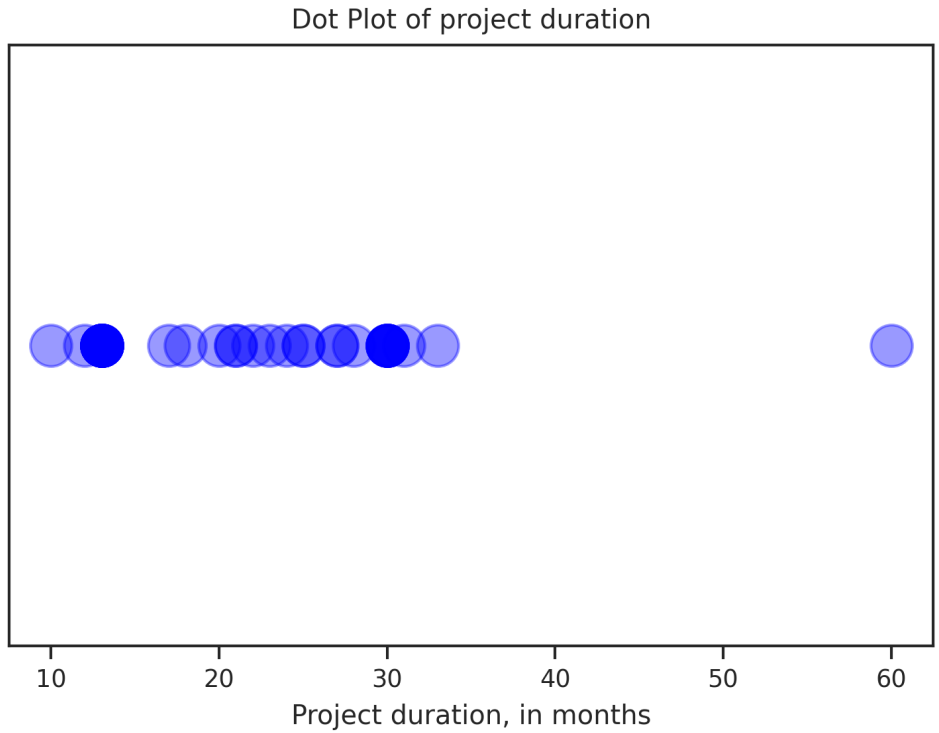






Visualization for quantitative data: dot plot

- Each observation is a dot on the x-axis
- More identical dots $\rightarrow$ more opaque color



- In practice, rarely used

Visualization for quantitative data: stem and leaf

- Helps identify concentration of the data
- Unlike the two other charts, it keeps all data information

```
1 | 3 3 0 3 3 8 3 3 3 2 3 3 3 3 7 3 3
2 | 8 5 4 1 0 2 7 1 3 5 7
3 | 0 0 1 0 0 0 0 0 3 0
6 | 0
```

- In practice, rarely used
- Only useful when the data is not too large
  - For the current data set used (beyond 10000 observations), the Stem and Leaf is not readable

Centrality indicators

- This is the first and perhaps most common type of **descriptive measures**
  - Three common indicators: **average** (arithmetic mean), **median**, **mode**
  - These are specific to quantitative variables (except for the mode)

- Helps understand what is common, what is typical
- Also helps as starting points for future inference (later in the semester)

- Assume you have one more observation but know nothing about it, what would be the best guess of the value/category of this new observation?

The average (arithmetic mean)

- This is the sum of all values over the number of observations
- In sample: $\bar{x} = \frac{x_1+x_2+...+x_n}{n} = \sum_i \frac{x_i}{n}$
- In population: $\mu = \frac{x_1+x_2+...+x_N}{N} = \sum_i \frac{x_i}{N}$

- Some notation:
  - $\sum_i$ is the sum over all observations "i"
  - Population size $= N$ and sample size $= n$
  - Population average $= \mu$ and sample average $= \bar{x}$

A slight refinement: the weighted average

- This is the **weighted** sum of all values over the sum of weights
- In sample: $\bar{x}_{weighted} = \frac{w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n}{w_1 + w_2 + ... + w_n} = \frac{\sum_i w_i * x_i}{\sum_i w_i}$

- Example: you obtained 5 in attendance, 5.5 in problem sets, 3.75 in assignments, and 4.5 in the final exam. What is the average grade?
  - Recall that attendance accounts for 10% of the grade, problem sets 15%, assignments 30%, and the final 45%

A slight refinement: the weighted average

- Example: you obtained 5 in attendance, 5.5 in problem sets, 3.75 in assignments, and 4.5 in the final exam. What is the average grade?
  - Recall that attendance accounts for 10% of the grade, problem sets 15%, assignments 30%, and the final 45%

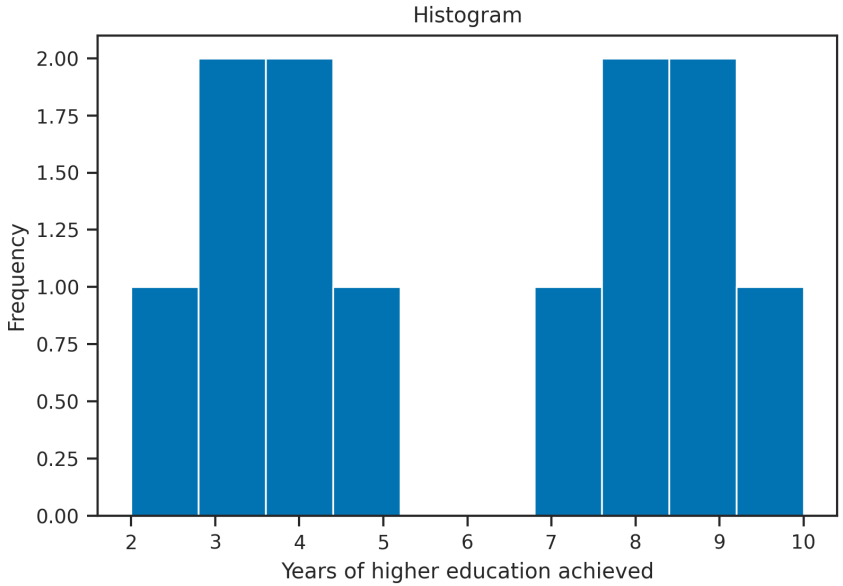
- $Grade\ Average = \frac{0.1*5+0.15*5.5+0.3*3.75+0.45*4.5}{0.1+0.15+0.3+0.45} = 4.475$

An example of average

- Assume you survey 10 people ($= n$) in a street and ask for the number of years of higher education
  - The result is: 2, 3, 3, 4, 4, 5, 7, 8, 8, 9, 9, 10
  - What is the average?

An example of average

- Assume you survey 12 people $(= n)$ in a street and ask for the number of years of higher education
  - The result is: 2, 3, 3, 4, 4, 5, 7, 8, 8, 9, 9, 10
  - What is the average of years of education in the sample?
  - $\bar{x} = \frac{2+3+3+4+4+5+7+8+8+9+9+10}{12} = 6$
    - The average years of higher education in the sample is 6 years



- The average captures a central value even if here no one studied for 6 years

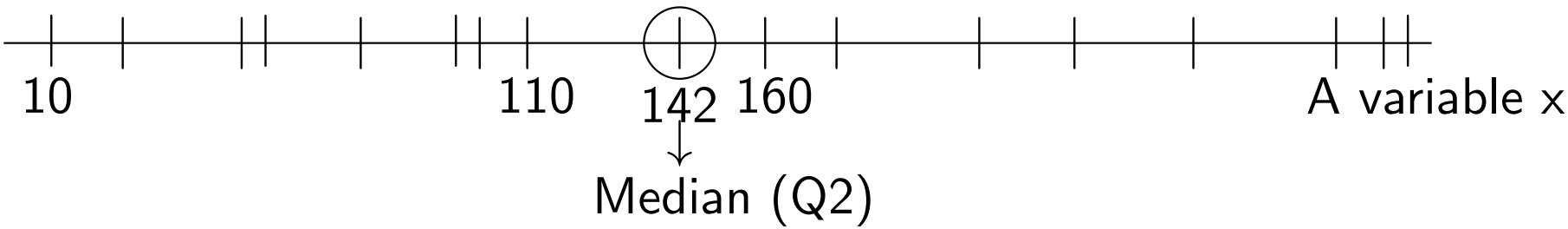
Average of grouped values

Frequencies (n_k)	Wage (CHF)	Wage midpoints
24	$[0 - 1000)$	500
61	$[1000 - 2000)$	1500
44	$[2000 - 3000)$	2500
48	$[3000 - 5000)$	4000
103	$[5000 - 8000)$	6500
95	$[8000 - 13000)$	10500
16	$[13000 - 20000)$	16500

- Grouped values are **continuous** variables because all values are possible within the range
- Here, the average (population) is: $\bar{x} = \frac{1}{n} \sum_k n_k \times x_{midpoints,k}$
  - with k the number of intervals (or classes)

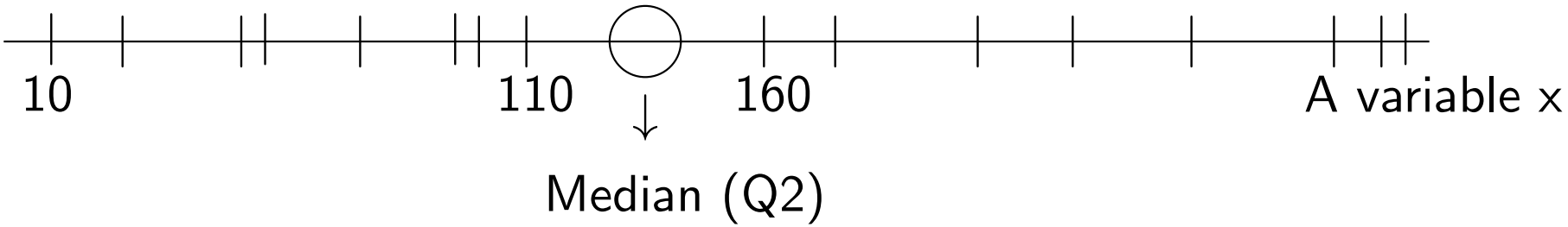
Another centrality indicator: the median

- The median splits the ordered variable in two subsets of equal number of observations
  - With n odd, one observation splits the data in two identical subsets
  - The $\frac{n+1}{2}$ observation is the median
  - Example below
    - Data:
      $10, 25, 50, 55, 75, 95, 100, 110, 142, 160, 175, 205, 225, 250, 280, 290, 295$
    - $n = 17$ so that $Q2 = Value\ of\ 9th\ observation = 142$



Another centrality indicator: the median (cont')

- With even number of observations, no observation splits the data in two equal subsets
  - The $\frac{n+1}{2}th$ observation would be the median
    - $\rightarrow$ This is between two observations in the data set, so take the average of these
  - Example below (dropping the 9th observation from previous example)
    - Data:
      $10, 25, 50, 55, 75, 95, 100, 110, 160, 175, 205, 225, 250, 280, 290, 295$
    - $n = 16$ so that
      $Q2 = Value\ of\ 8.5th\ observation = \frac{110+160}{2} = 135$



Median of grouped values

Frequencies (n_k)	Cumulative Frequencies (N_k)	Wage (CHF)	Wage midpoints
24	24	$[0-1000)$	500
61	85	$[1000-2000)$	1500
44	129	$[2000-3000)$	2500
48	177	$[3000-5000)$	4000
103	280	$[5000-8000)$	6500
95	375	$[8000-13000)$	10500
16	391	$[13000-20000)$	16500

- $n = 391$ so that the median would be the $\frac{n+1}{2}th$ value $= 196th$ value
  - Which is in the $[5000-8000)$ interval
  - But it is closer to the interval below than the interval above
- Use **linear interpolation** to find Q2's linear approximation
  - Here: $Q2 = 5000 + \frac{196-177}{280-177}(8000-5000) \approx 5553$ CHF

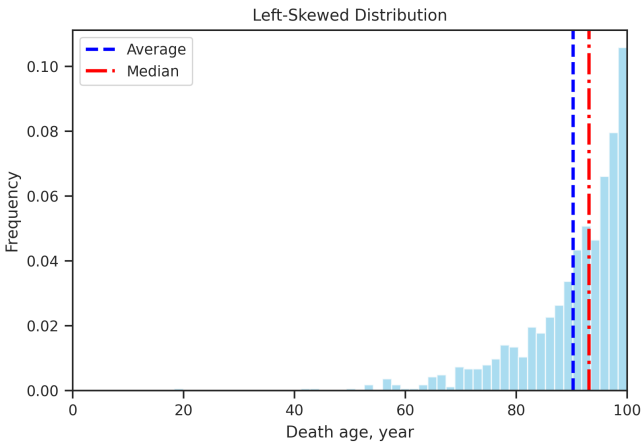
Comparison Average *versus* Median

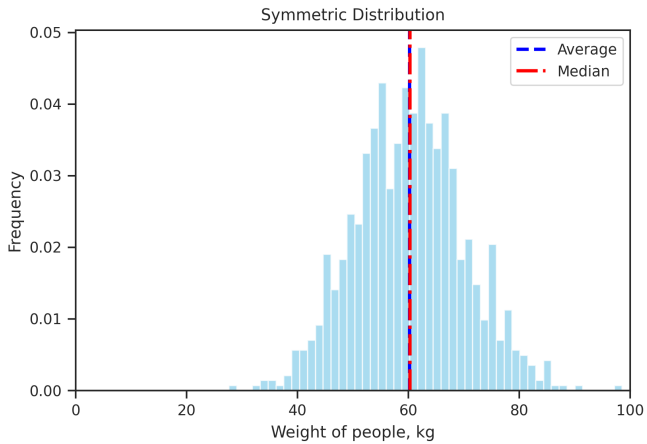
- Data: $1, 2, 3, 4, 5$
  - $\bar{x} = 3$, $Q2 = 3$
- New Data: $1, 2, 3, 4, 1000000$
  - $\bar{x} = 200002$, $Q2 = 3$

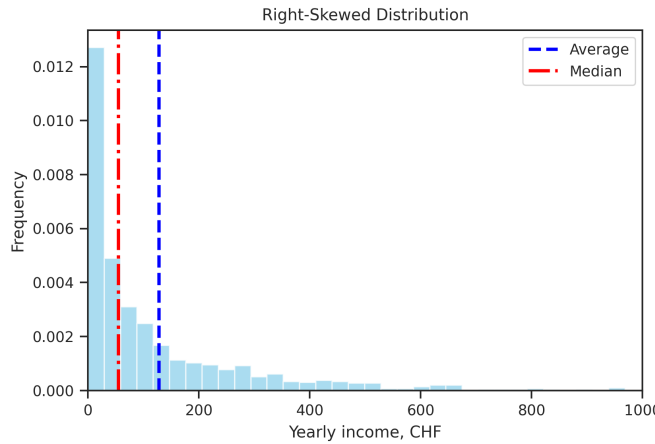
$\Rightarrow$ The average is sensitive to *outliers*, while the median is not

A quick detour to distribution (a)symmetry

- Because the average depends on values, it is more prone to variation than the median
- Symmetric distribution: $Median = Average$
- Right-skewed (or positively skewed): $Median < Average$ (in general)
- Left-skewed (or negatively skewed): $Median > Average$ (in general)







A last centrality indicator: the mode

- It is the most frequent value/modality in a variable
- May not exist; may not be unique (bimodal, multimodal distributions)
  - What is/are the mode(s) of the following data?
  - Data: $1, 3, 3, 3, 4, 5, 5, 6, 7, 9, 9, 9, 10, 11, 15$
- Can be used with qualitative variables (where in fact it is most useful)
  - What is the mode of the categorical variable "type" (displayed below)?

	Frequency	Relative frequency
Brownfield	2991	0.277742
Divestiture	1315	0.122110
Greenfield project	5998	0.556969
Management and lease contract	464	0.043087
Not Available	1	0.000093

Next session

- Next session is on calculating most usual dispersion indicators, constructing and understanding a Box plot, and encountering some distribution-specific rules