Statistical Literacy — MINT

Lecture 3: Dispersion indicators, distribution shape, Boxplots

Rémi Viné

The Graduate Institute | Geneva

October 9th, 2023

Outline

Housekeeping

Dispersion indicators
- Range
- Interquartile Range
- Variance and Standard deviation

Shape of distributions

Construct a boxplot

Housekeeping

- Problem set 3 is now available
- New room for my office hours: P1-557 (still Mondays 4.15pm-6pm)
- A few points:
  - When to use Stata?
    - No need for the problem sets (apart from a few optional points)
    - Needed for the assignments (doing these are in fact meant for you to implement Stata)
  - Decimals, fractions, *etc.*
    - The exam is closed answers, not to worry about this
    - Problem sets & assignments: just be consistent with your choice (and follow a rule whenever specified)

From centrality to dispersion

- Centrality is key to have a good guess on where data tends to be
- But it also matters to know what is the possible deviation from this centrality measure

- Assume the data (monthly income in CHF) is:
  $4900, 5000, 5000, 5400, 5450, 5550, 6500, 6600, 7000, 8600$
  - Average (monthly) income is CHF 6000 and Q2 is CHF 5500
  - What is a typical income that can be observed in the data?
    - CHF 5000
    - CHF 9000
    - CHF 1500
    - CHF 15000
    - CHF -500

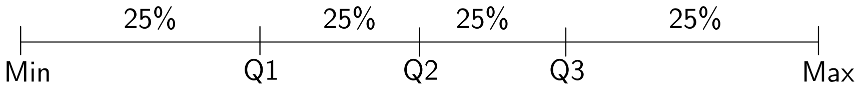
$\Rightarrow$ Dispersion indicators (Range, Interquartile Range, Variance) are needed

Range

- $Range = Maximum - Minimum$
- Within the data, it is the difference between the two *extrema*
  - In the monthly income example:
    - $Min = 4900$ (CHF)
    - $Max = 8600$ (CHF)
    - $Range = 3700$ (CHF)
- Very simple but not so informative

Interquartile range: What is a quartile?

- Recall that the median is named $Q2$
  - This is the second quartile
- There are three quartiles: $Q1, Q2, Q3$
  - $Q1$ splits the data in two so that one quarter of the data lies **beneath** it
  - $Q3$ splits the data in two so that one quarter of the data lies **above** it



A brief detour to percentiles

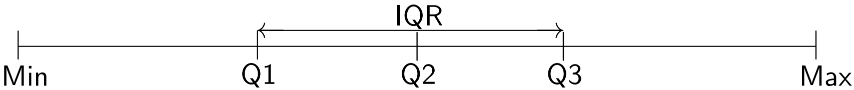
- Like quartiles, percentiles split ordered data
- Below the first percentile ($P1$) lie 1% of all observations
- Below the 50th percentile ($P50$) lie 50% of all observations
  - It is $Q2$ (the median)
- Below the 99th percentile ($P99$) lie 99% of all observations

- One can also construct **deciles** ($D1, D2, ..., D9$)

- Percentiles and deciles are useful to analyze the gap between each threshold

Interquatile range (IQR)

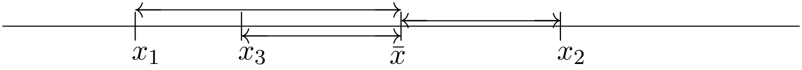
- It is the gap between the first and the third quartiles
- $IQR = Q3 - Q1$
- It gathers the 50% of the data in the center
  - It gives an idea of a typical distance around the median (spread around the median)



- In the example: $4900, 5000, 5000, 5400, 5450, 5550, 6500, 6600, 7000, 8600$
  - $Q2 = 5500$ (CHF)
  - $Q1 = 5000$ (CHF), $Q3 = \textit{something around } 6600 \textit{ and } 6700$ (CHF) (see the different definitions of quartiles)

Variance

- Assesses how much values (of a given variable, x) vary about the average
- It is the squared of the distances of each data point from the average (to have all distances positive)



- For population
  - $Var(X) = \frac{1}{N}\sum_i (x_i - \mu)^2$
- For sample
  - $Var(X) = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$
    - To understand the reason why dividing by $(n-1)$, see extra (not required to know) note on Moodle
- Units of the variance are the units of the variable *squared* (as it includes x_i^2)

Variance, example

- Compute the variance for the following data (monthly income in CHF)
- Data:
  $4900, 5000, 5000, 5400, 5450, 5550, 6500, 6600, 7000, 8600$

Variance, example

- Compute the variance for the following data (monthly income in CHF)
- Data:
  $4900, 5000, 5000, 5400, 5450, 5550, 6500, 6600, 7000, 8600$
- First, compute the (sample) average: $\bar{x} = \frac{1}{n}\sum_i x_i = 6000$ CHF
- Second, compute the squared distances for all data points:

$$\sum(x_i - \bar{x})^2 = (4900-6000)^2 + 2(5000-6000)^2 + (5400-6000)^2 + (5450-6000)^2 + (5550-6000)^2 + (6500-6000)^2 + (6600-6000)^2 + (7000-6000)^2 + (8600-6000)^2 = 12445000$$

Variance, example

- Compute the variance for the following data (monthly income in CHF)
- Data:
  $4900, 5000, 5000, 5400, 5450, 5550, 6500, 6600, 7000, 8600$
- First, compute the (sample) average: $\bar{x} = \frac{1}{n}\sum_i x_i = 6000$ CHF
- Second, compute the squared distances for all data points:

$$\sum(x_i - \bar{x})^2 = (4900-6000)^2 + 2(5000-6000)^2 + (5400-6000)^2 + (5450-6000)^2 + (5550-6000)^2 + (6500-6000)^2 + (6600-6000)^2 + (7000-6000)^2 + (8600-6000)^2 = 12445000$$

- Third, divide by $(n-1)$ (for sample): $Var(X) = \frac{12445000}{10-1} \approx 1382777.78$
  - The variance of monthly income is 1382777.78 CHF^2 (difficult to interpret...)

Standard deviation

- Because the units of the variance make the interpretation uneasy, the square root is used
- It is the standard deviation
  - For population: $\sigma = \sqrt{Var(X)}$
  - For sample: $s = \sqrt{Var(X)}$
- In the previous example, $s = \sqrt{Var(X)} \approx 1175.92$
  - A typical monthly income deviation from the average (= CHF 6000) is CHF 1176.

Variance and standard deviation of grouped values

Frequencies (n_k)	Wage	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

- The **population** variance is:
  $\sigma^2 = \frac{1}{N} \sum_k n_k \times (x_{midpoints,k} - \mu)^2$ and $\sigma = \sqrt{\sigma^2}$
  - with k the number of intervals

Centrality & Dispersion: they both matter

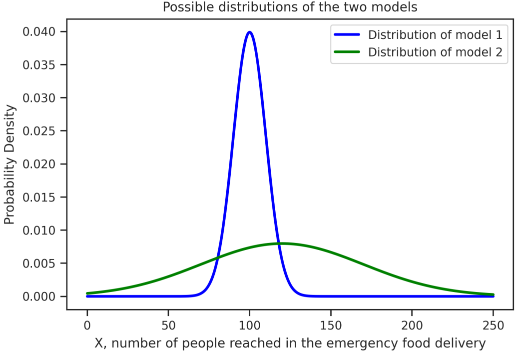
- Assume you work for an NGO delivering food in post-catastrophes areas and that there are two models to deliver food (assume costs are identical)
  - By aviation (carry less with little losses)
  - By humanitarian trucks (carry a lot but lots of stocks can be wasted)

X: Number of people reached	Model 1	Model 2
Average: $\bar{x} =$	100	120
Standard deviation: $s =$	10	50

Centrality & Dispersion: they both matter

- What is best?

  $\longrightarrow$ Not obvious...



Summary slide on indicators

Indicator	Specificity	Units of measurement
Mode	preferred for categorical variables	var. unit
Average	most common but sensitive to outliers	var. unit
Median	does not vary with values	var. unit
Quartiles	does not vary with values	var. unit
Percentiles	does not vary with values	var. unit
Range	rough idea of what's possible	var. unit
IQR	does not vary with values	var. unit
Variance	varies with values	var. unit **squared**
Standard deviation	most common (varies with values)	var. unit

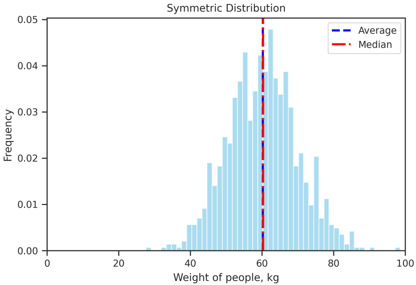
Shape of distributions

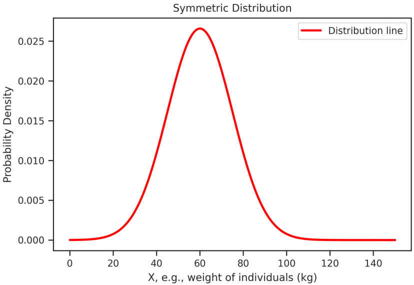
- Beyond centrality and dispersion, it can be that some observations are far from the central indicators **on one side** more than on the other side
  - Such variables' distributions are named **asymmetric**, or **skewed**
  - The absence of skewness is named **symmetry**
    - In such case, the odds of being at a specific distance from the mean (or the median) on either side is **identical**

Shape of distributions: symmetry

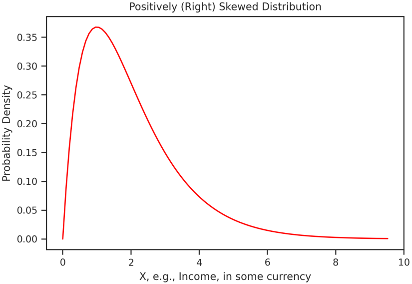
- The data is spread out on either side of the median similarly: same distances from Q2
  - By construction: $average = median$
  - Is it also always equal to the mode?

    $\rightarrow$ No! *(think of a bimodal and symmetric distribution)*





Shape of distributions: asymmetry



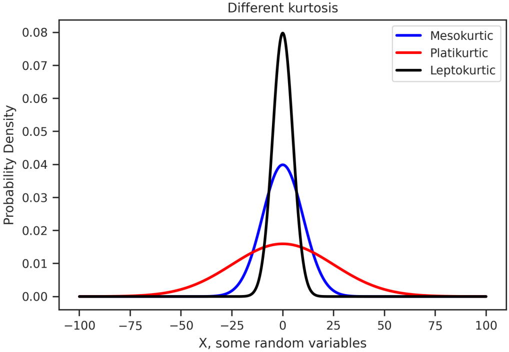
- Asymmetry can be problematic because the tail may capture *outliers*
- One can compute the skewness coefficient *(average = 1st centred moment, variance = 2nd, skewness = 3rd)*
  - $Skewness > 0 \Leftrightarrow Positive\ (Right)\ skewness$
  - $Skewness < 0 \Leftrightarrow Negative\ (Left)\ skewness$

A brief detour to outliers

- Outliers are values that are distinct because they are far away from others
  - Very large values
    - *E.g.*, a monthly income of CHF $5'500'000$
  - Very small values
    - *E.g.*, a height (of human being) of 1.7 cm
- It can be an extremely rare event
  - The monthly income Stellantis CEO in 2021 was 5'500'000 euros (!)
- It can be an error
  - Measurement unit, here the height might be 1.7m instead of 1.7cm
  - Measurement error in coding: the data might have been wrongly compiled (*typo*)

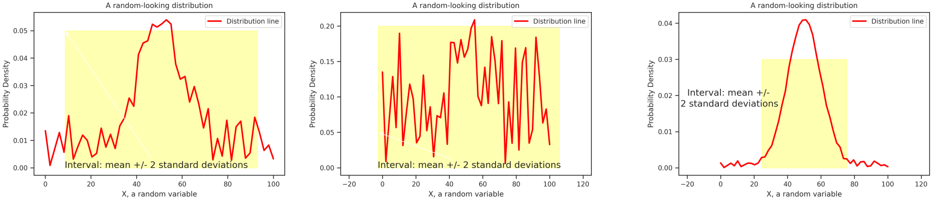
Shapes of distributions: Kurtosis

- The kurtosis (4th central moment) is about the thickness of the tails of the distributions
  - How likely are rare events?
  - Much less used than other measures, but might be informative on the "Black swan" issue
    - A Black Swan (Nassim Nicholas Taleb) is (i) unpredictable, (ii) massive, (iii) explained *ex post* by a fully fledged narrative



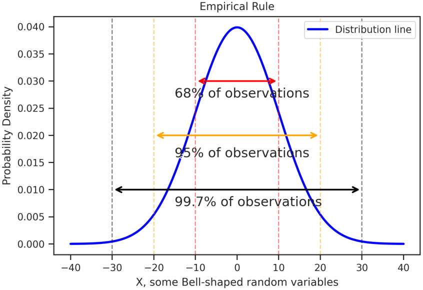
Shapes of distributions: how much of the data to expect in an interval?

- **Chebyschev's rule:** *regardless of the distribution*
  - 75% of all observations are located within 2 standard deviations around the mean
  - $8/9 \approx 89\%$ of observations are within 3 standard deviations



Shapes of distributions: how much of the data to expect in an interval?

- **Empirical rule:** *for Bell-shaped distributions* (unimodal, $Mode = Median = Average$)
  - 68% of the data are within the interval $Average \pm 1\ Standard\ Deviation$
  - 95% of the data are within the interval $Average \pm 2\ Standard\ Deviation$
  - 99.7% of the data are within the interval $Average \pm 3\ Standard\ Deviation$



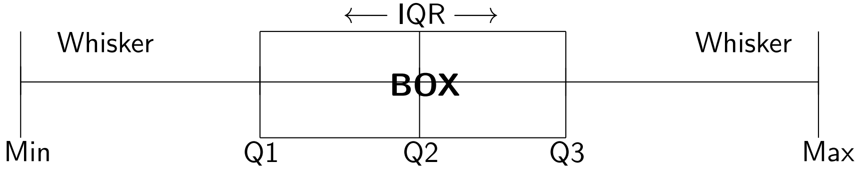
Construct a Boxplot

- A boxplot - also named a 5 number summary - requires

  1. The median (Q2)
  2. The first quartile (Q1) $\rightarrow$ Different methods for quartiles
  3. The second quartile (Q3)
  4. The minimum
  5. The maximum

- A boxplot is only based on *ranked* data

  - When no suspected **outliers**:



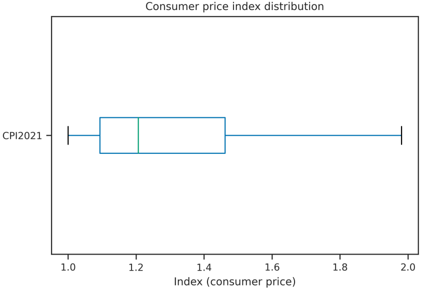
Boxplot: construct the whiskers

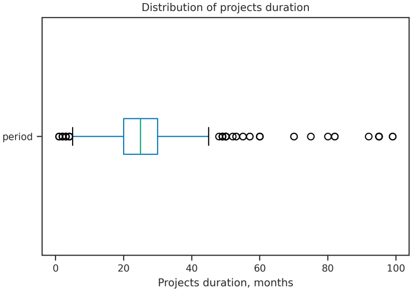
- Whiskers are located according to lower and upper limits
  - Lower limit: $Q1 - 1.5 * IQR$
  - Upper limit: $Q3 + 1.5 * IQR$
- If no data point beyond these limits: the whiskers depict the Min and the Max
- If some data points go further than these bounds:
  - The data points beyond the limits are represented as dots
    - One would suspect them to be **outliers**
    - Spotting potential outliers is an important added-value of boxplot
  - The new whisker is the data point right below (or right above for the left whisker) the theoretical whisker

Boxplot: construct the whiskers, example

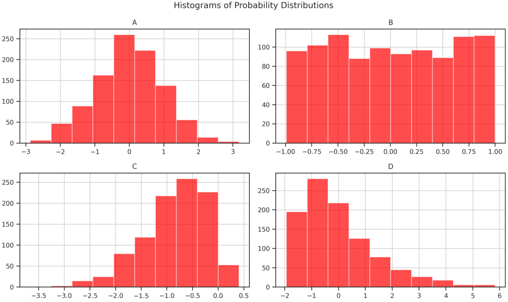
- No suspected outliers
- Hence, whiskers are within $[Q1 - 1.5IQR; Q3 + 1.5IQR]$
- Positive skewness suspected
  - $|Q2 - Q1| < |Q2 - Q3|$
  - $|Q1 - Min| < |Q3 - Max|$

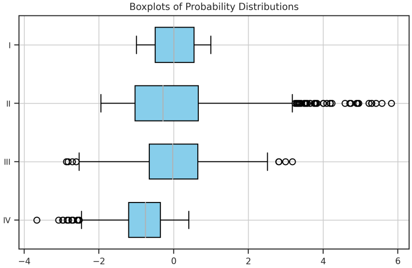
- Many suspected outliers (low and large values)
- Apart from (suspected) outliers, some symmetry



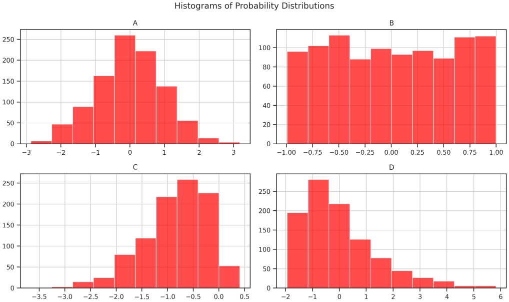


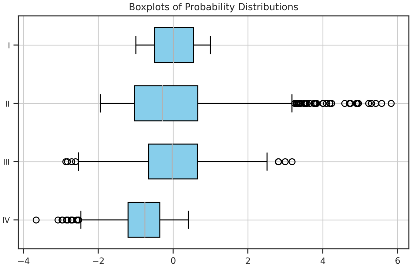
Boxplot & Histogram: match them





Boxplot & Histogram: match them





- A-III
- B-I
- C-IV
- D-II

Next session

- Next session is on probability and random variables in the discrete case