Exploratory data analysis (EDA) - describing variables

5 reason to do EDA!

1. To check data cleaning (part of iterative process)

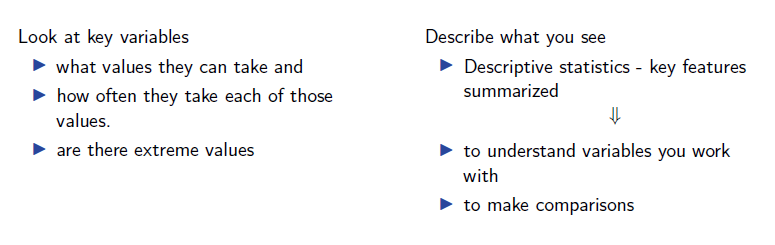
2. To guide subsequent analysis (for further analysis)

3. To give context of the results of subsequent analysis (for interpretation)

4. To ask additional questions (for specifying the (research) question)

5. Offer simple, but possibly important answers to questions.

03



Distribution - theory

A random variable (RV) can be fully characterized by its probability distribution function - PDF - also called density function.

* Probability is a measure of the likelihood of an event.
* Probabilities are always between zero and one.
* PDF - connects all the possible event to a certain probability.

Theoretical vs Empirical distribution

In theory RV has a nice theoretical distribution - e.g. normal distribution. However in real life, we do not have infinitely many observations to have perfect theoretical distributions. We are only ‘converging’ to these distributions (see later - LLN, CLT).

In practice we have empirical distribution!

* I The frequency or more precisely, absolute frequency or count, of a value of a variable is simply the number of observations with that particular value.
* I The relative frequency is the frequency expressed in relative, or percentage, terms: the proportion of observations with that particular value among all observations.
* I Practical note: With missing values – proportion can be relative to all observations OR only observations with non-missing values (usual choice).

Probabilities and distributions are much wider notions, than we will use here. They are important if you want to prove some new data analysis tool to work (e.g. create a new estimator).

The distribution and the histogram

A key part of EDA is to look at (empirical) distribution of most important variables.

* I All variables have a distribution.
* I The distribution of a variable tells the frequency of each value of the variable in the data.
* I May be expressed in terms of absolute frequencies (number of observations) or relative frequencies (percent of observations).
* I The distribution of a variable completely describes the variable as it occurs in the data.
* I independent from values the other variables may show.

Histograms

Histogram reveals important properties of a distribution.

* I Number and location of modes: these are the peaks in the distribution that stand out from their immediate neighborhood.
* I Approximate regions for center and tails
* I Symmetric or not - asymmetric distributions have a long left tail or a long right tail
* I Extreme values: values that are very different from the rest. Extreme values are at the far end of the tails of histograms.

Extreme values

Some variables have extreme values: substantially larger or smaller values for one or a handful of observations than the values for the rest of the observations.

Need conscious decision.

* I Is this an error? (drop or replace)
* I Is this not an error but not part of what we want to talk about? (drop)
* I Is this an integral feature of the data? (keep)

The Expected value

I The expected value is the value that one can expect for a randomly chosen observation

I The notation for the expected value is E[x].

I For a quantitative variable, the expected value is the mean

I For a qualitative variable, it can only be determined if transformed to a number

* I Male/Female binary variable. Expected value could be probability / relative frequency of females.
* I Quality of hotel: 1 to 5 stars, mean can be calculated, but its meaning is less straightforward.
  + I What is the assumption for getting the mean as number?
* 03

Summary statistics: The median and other quantiles

I Quantiles: a quantile is the value that divides the observations in the dataset to two parts in specific proportions.

I The median is the middle value of the distribution - half the observations have lower value and the other half have higher value.

I Percentiles divide the data into two parts along a certain percentage.

* I The first percentile is the value below which one percent of the observations are and 99 percent above.

I Quartiles divide the data into two parts along fourths.

* I 1st quartile has one quarter of the observations below and three quarters above; it is the 25th percentile.
* I 2nd quartile has two quarters of the observations below and two quarters above; this is the median, and also the 50th percentile.

Summary statistics: The mode

* The mode is the value with the highest frequency in the data.
* Some distributions are unimodal, others have multiple modes.
* Multiple modes are apart from each other, each standing out in its "neighborhood", but they may have different frequencies.

Summary statistics: central tendency

* I The mean, median and mode are different statistics for the central value of the distribution
* I Central tendency.
  + I The mode is the most frequent value
  + I The median is the middle value
  + I The mean is the value that one can expect for a randomly chosen observation.

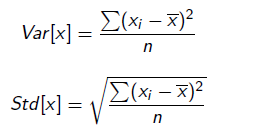
Summary statistics: spread of distributions

* I spread of distributions is also often used in analysis.
* I Statistics that measure the spread of distributions are the range, inter-quantile ranges, the standard deviation and the variance.
* I The range is the difference between the highest value (the maximum) and the lowest value (the minimum) of a variable.
* I The inter-quantile ranges is the difference between two quantiles- the third quartile (the 75th percentile) and the first quartile (the 25th percentile).
* I The 90- 10 percentile range gives the difference between the 90th percentile and the 10th percentile.

03

Summary statistics: standard deviation

* I The most widely used measure of spread is the standard deviation. Its square is the variance.
* I Variance is the average squared difference of each observed value from the mean.
* I The standard deviation captures the typical difference between a randomly chosen observation and the mean.
* I The variance is a less intuitive measure. At the same time, the variance is easier to
* work with, because it is a mean value itself.

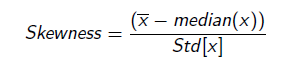


Summary statistics: skewness

* I A distribution is skewed if it isn’t symmetric.
* I It may be skewed in two ways, having a long left tail or having a long right tail.
* I Example: hotel price distributions having a long right tail - such as in hotel price distribution.
* I Skewness and the prevalence of extreme values are related. With distributions with long tails, values far away from all other values are more likely.
* I When extreme values are important for the analysis, skewness of distributions is important, too.
* 03

Summary statistics: skewness measure

Simplest measure is mean–median measure of skewness.



* I When the distribution is symmetric its mean and median are the same.
* I When it is skewed with a long right tail the mean is larger than the median: the few very large values in the right tail tilt the mean further to the right.
* I When a distribution is skewed with a long left tail the mean is smaller than the median
* I To make this measure comparable across various distributions use a standardized measure
* I If multiplied by 3, and then it’s called Pearson’s second measure of skewness.

Density plots

* I Density plots - also called kernel density estimates
* I alternative to histograms - instead of bars density plots show continuous curves.
* I Instead of bars, density plots show continuous curves. We may think of them as curves that wrap around the corresponding histograms.
* I density plots complementing histograms - some believe density plots allow for easier comparison of distributions across groups in the data.

03

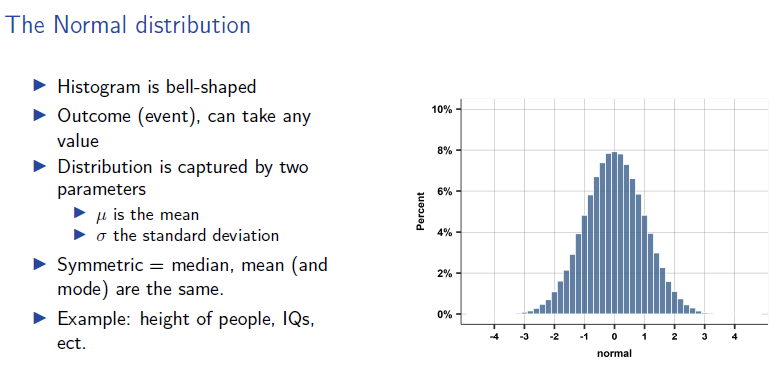
Theoretical distributions

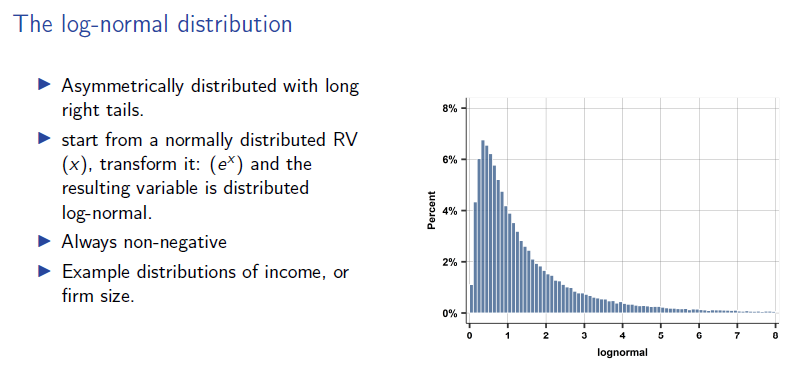
Theoretical distributions are distributions of variables with idealized properties.

* I Show frequencies for theoretical distributions and not for empirical’s.
* I The likelihood of each value in a more abstract setting - hypothetical "dataset" or "population," or the abstract space of the possible realizations of events.
* I Theoretical distributions are fully captured by few parameters: these are statistics determine the whole distributions

Theoretical distributions can be helpful

* I Have well-known properties!
* I If variable in our data well approximated by a theoretical distribution –> attribute properties to the variable
* I Real life, many variables surprisingly close to theoretical distributions.
* I Will be useful when generalizing from data - Class 05

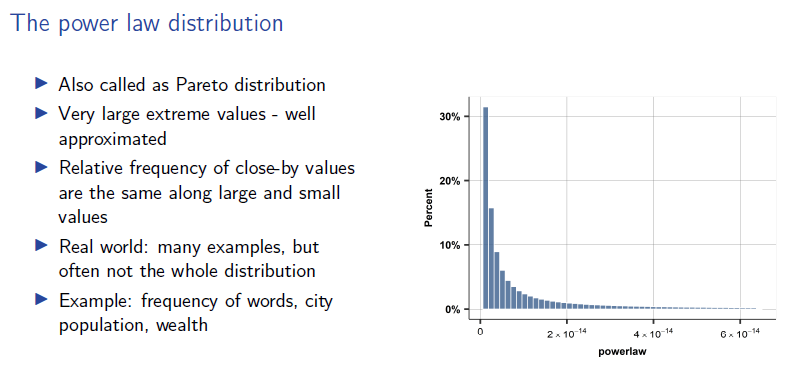




A few more points on the Normal and log-normal

* I Many many variables in real life are close to normal
* I Especially when based on elementary things which are added up
* I Not good approximation when
  + I some reasons for non-symmetry
  + I extreme values are important
* I Variables are well approximated by the log-normal if they are the result of many things multiplied (the natural log of them is thus a sum).

0



(good for extreme values) -> also to use it for a particular part of the dataset (to present extreme)