Multivariate Verfahren

3. Multivariate Data Types and Descriptive Statistics

Hannah Schulz-Kümpel

Institut für Statistik, LMU München

Summer Semester 2024

What we will look at today:

1. Representing data mathematically
2. Working with point clouds
3. Detour: Quantiles
4. Descriptive Statistics: Why, When, and How?
5. `ggplot()`
6. From Scatterplots to Histograms: most important non-probabilistic visualization tools

The data we work with I

- As statisticians, we are often tasked with analyzing data of the following form:

	X_1	X_2	X_3	$\cdots$	X_m
1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1m}
2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2m}
3	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	x_{n1}	x_{n2}	x_{n3}	$\cdots$	x_{nm}

- To work with this data, we need to view it as (an) mathematical object(s).

The data we work with II

- One obvious option: just consider the matrix

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m}, n, m \in \mathbb{N}.$$

- This is similar to the algebraic OLS approach we dealt with before, but with two key differences
  1. In OLS we assume a linear relationship between a *target variable* and variables that influence the target variable.
  2. We needed to add a first column of ones to get the design matrix which includes the possibility of an intercept.

The data we work with III

- Often, you will also see something like "we consider a sequence of observations $\mathcal{D} = (x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ ", where, most often, $\mathcal{Y} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^p$, $p \in \mathbb{N}$.

Sidenote

Probably, you will also sometimes see $\mathcal{D}$ being referred to as "set". While "*data set*" has become an established term, $\mathcal{D}$ usually is no set in the mathematic sense, because elements of a set can neither be ordered nor be included more than once.

- The point of this notation is to immediately declare one (or more) columns of our data to be the target of interest - so the entries of which column "become" y_is depends on a given research question.

Probabilistic vs. geometric/algebraic view of data I

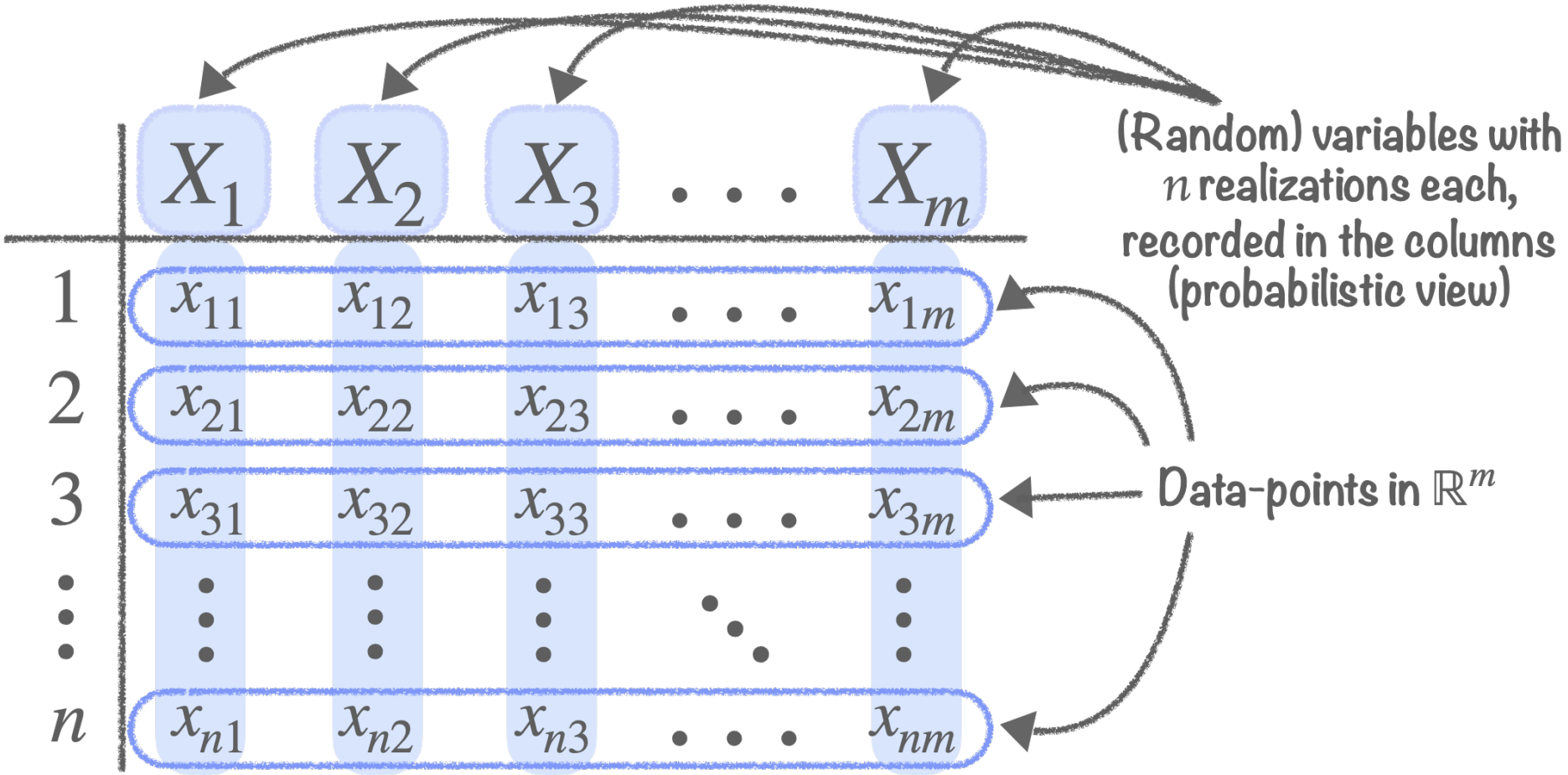
Regardless of whether we "relabel" a column of our given data as Y, we can again take a both a Probabilistic and a geometric/algebraic view of things:

	X_1	X_2	X_3	$\cdots$	X_m
1	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1m}
2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2m}
3	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	x_{n1}	x_{n2}	x_{n3}	$\cdots$	x_{nm}

Data-points in $\mathbb{R}^m$

Probabilistic vs. geometric/algebraic view of data I

Regardless of whether we "relabel" a column of our given data as Y, we can again take a both a Probabilistic and a geometric/algebraic view of things:



Probabilistic vs. geometric/algebraic view of data II

- When taking the probabilistic view, we usually model each data point in $\mathbb{R}^3$ as a(n independent) *draw from a distribution*
  and write, if one target quantity has been identified, that
  the point $(y_i, \boldsymbol{x}_i)$ is a realization of the random vector

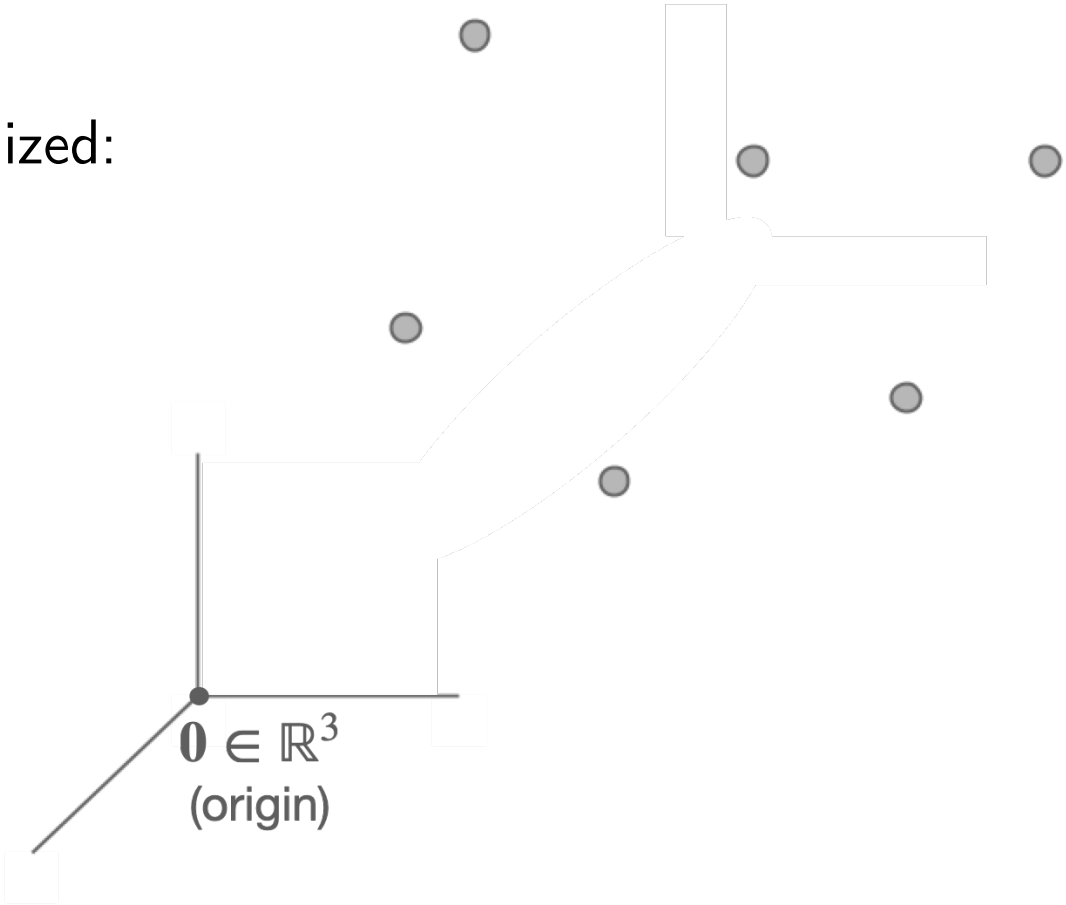
$$(Y_i, X_i) \sim P_{xy}\,.$$

  (More on random vectors later!)

- For now, let us focus on the purely algebraic/geometric view of simply considering n points in $\mathbb{R}^m$.
  $\Rightarrow$ Then, we can view our data as a "point cloud" in $\mathbb{R}^m$.

Geometric Mean and Variance I

- For $m = 3$ this is nicely visualized:

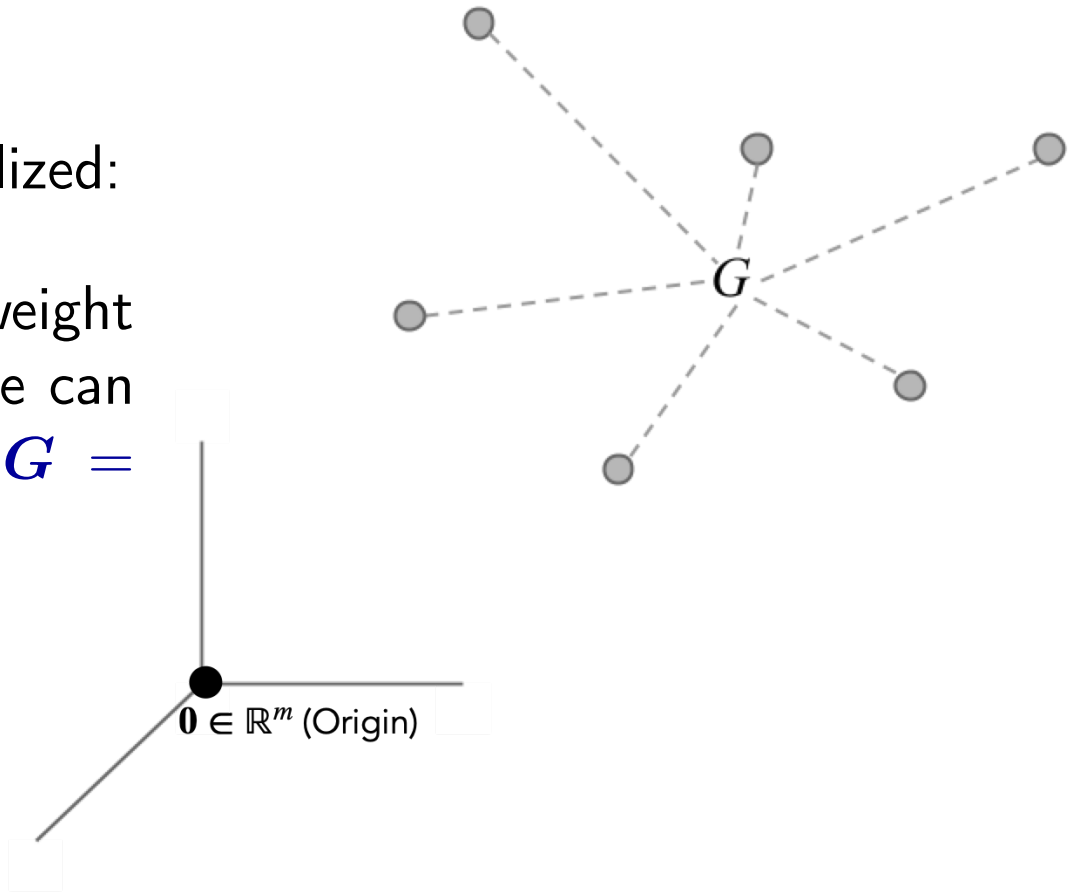


Geometric Mean and Variance I

- For $m = 3$ this is nicely visualized:

- If each point is assigned a weight $p_i \in [0, 1]$ with $\sum_{i=1}^{n} p_i = 1$, we can calculate a **center of gravity** $\boldsymbol{G} = (G_1, \ldots, G_m)^\top \in \mathbb{R}^{\boldsymbol{m}}$ with

$$G_j = \sum_{i=1}^{n} p_i \, x_{ij} \, .$$



Source: https://pca4ds.github.io/center-of-gravity.html

Geometric Mean and Variance II

- Clearly, if $p_i := \frac{1}{n}\ \forall i \in \{1, ..., n\}$, the center of gravity G is equal to the **arithmetic mean** of the sequence of data points $\{(x_{i1}, \ldots, x_{im})^\top\}_{i=1,...,n}$.

- The spread of points around a center of gravity can be measured by the concept of **Inertia**:

- Given a sequence of weights $\{p_i\}_{i\in\{1,...,n\}}$ assigned to a sequence of points $\{\boldsymbol{x}_i\}_{i\in\{1,...,n\}}$ and a corresponding center of gravity G, the inertia of our data w.r.t. some distance d can be defined as

$$I = \sum_{i=1}^{n} p_i \cdot d\Big(\boldsymbol{x}_i, G\Big)\,.$$

Spoiler: Some stuff about distances

- For some spaces $\mathcal{S}$, we call a function

$$d : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}$$

a **distance**, if it fulfils the following three requirements $\forall a, b \in \mathcal{S}$:

1. $d(a, b) = 0 \Leftrightarrow a = b$
2. $d(a, b) \geq 0$
3. $d(a, b) = d(b, a)$

- We recall the definition of the euclidean distance is, for some $p \in \mathbb{N}$,

$$d_{\text{euclid}} : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}, \quad (\boldsymbol{a}, \boldsymbol{b}) \longmapsto \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}\,.$$

Geometric Mean and Variance III

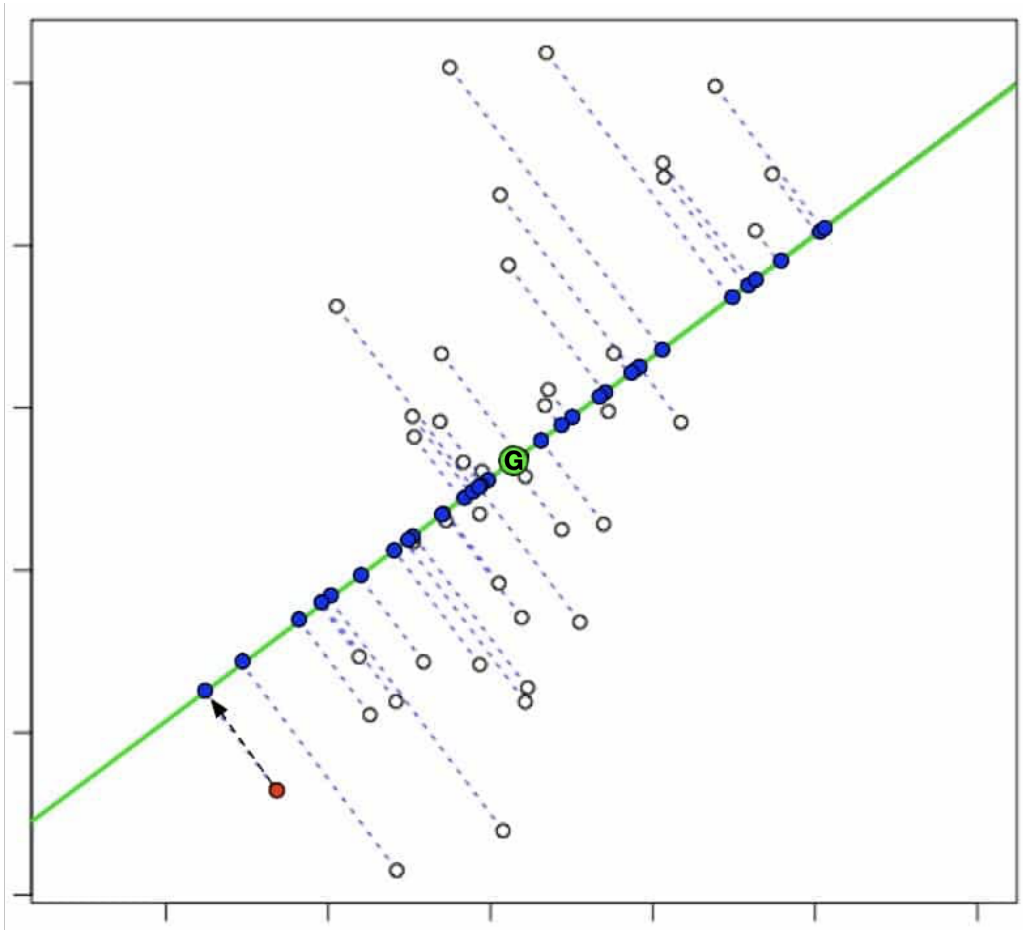
- Similar to how the arithmetic mean may be viewed as a special case of *center of gravity*, the **sample variance of data points** $\{\boldsymbol{x}_i\}_{i\in\{1,\dots,n\}}$ **with** $\boldsymbol{x}_i \in \mathbb{R}$ may be viewed as a *special case of inertia* with $p_i := \frac{1}{n-1}$ $\forall i \in \{1, ..., n\}$, and d chosen as the squared Euclidean distance.

- Clearly, the concept of sample variance **w.r.t. the Euclidean Distance** is not immediately transferable to higher dimensional data, i.e. data points $\boldsymbol{x}_i \in \mathbb{R}^m$, $m \in \mathbb{R}_{>1}$.

- However, we can of course do the following
  1. Project all points $\boldsymbol{x}_i \in \mathbb{R}^m$, $m \in \mathbb{R}_{>1}$ onto a line drawn through the origin and a point $\boldsymbol{u} \in \mathbb{R}^2$.

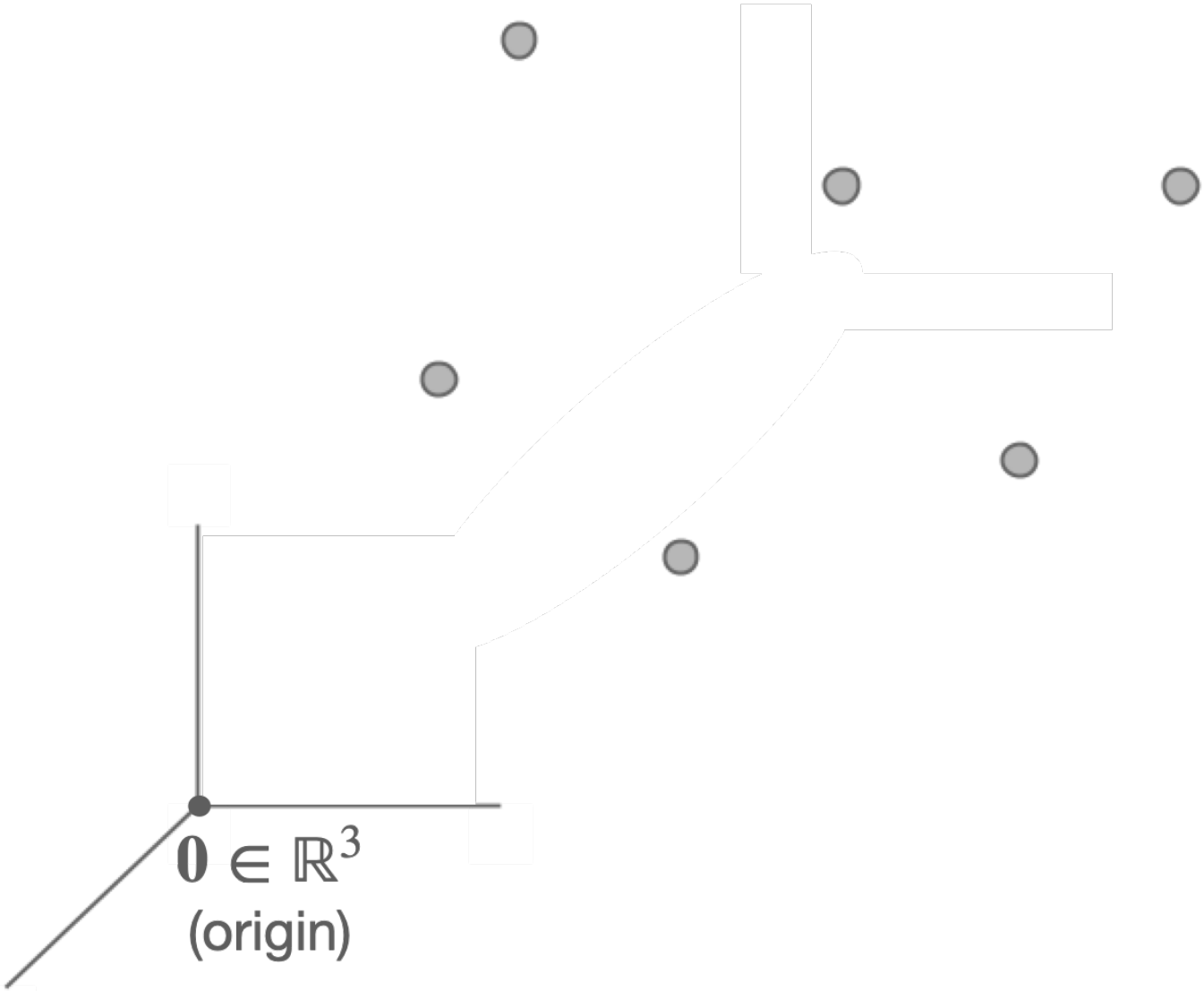
Geometric Mean and Variance IV

2. Since one of two dimensions is fixed by the choice of the point u, a sample variance may then be calculated w.r.t. the values of the other dimension.

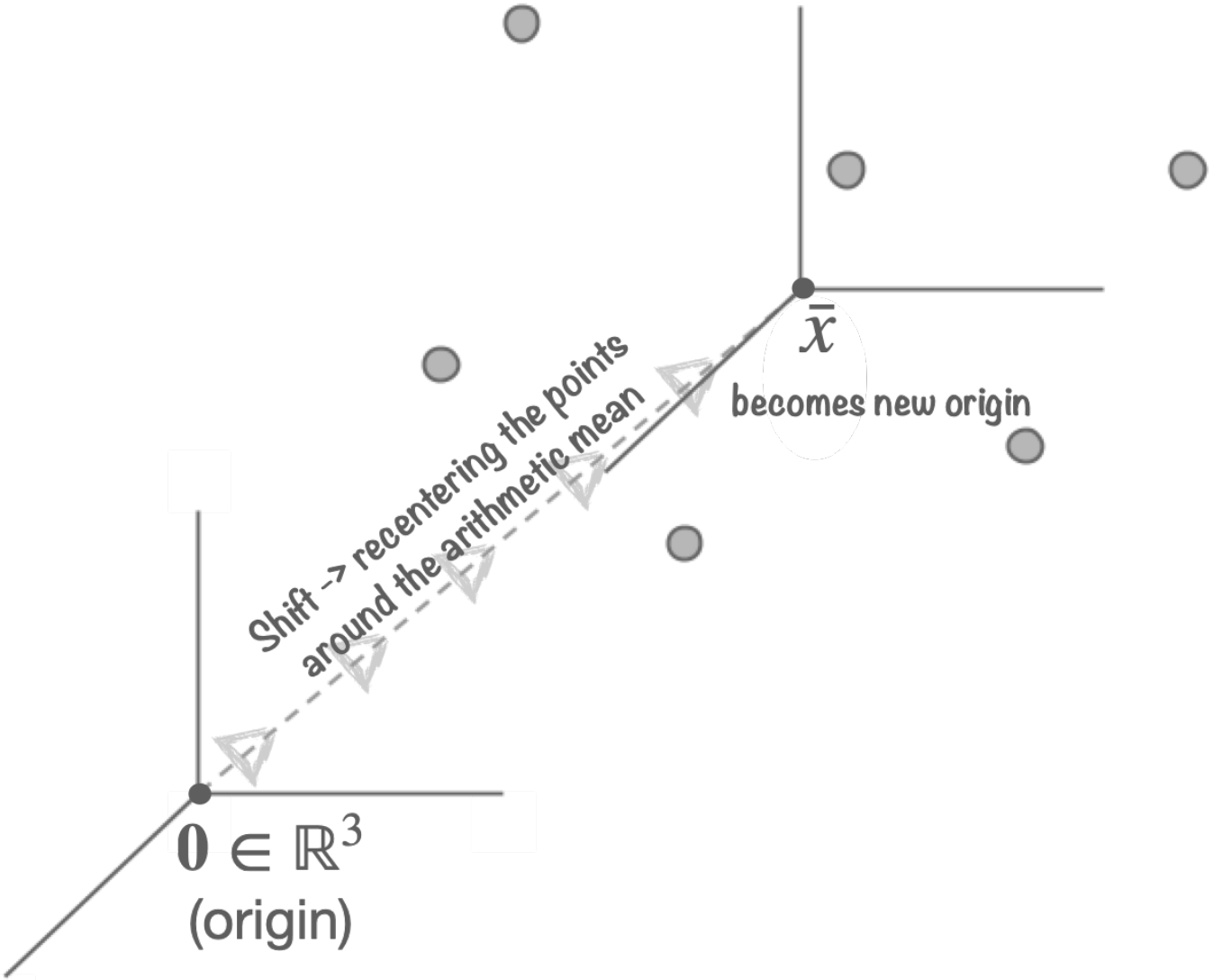


Source: https://dimensionless.in/principal-component-analysis-in-r/

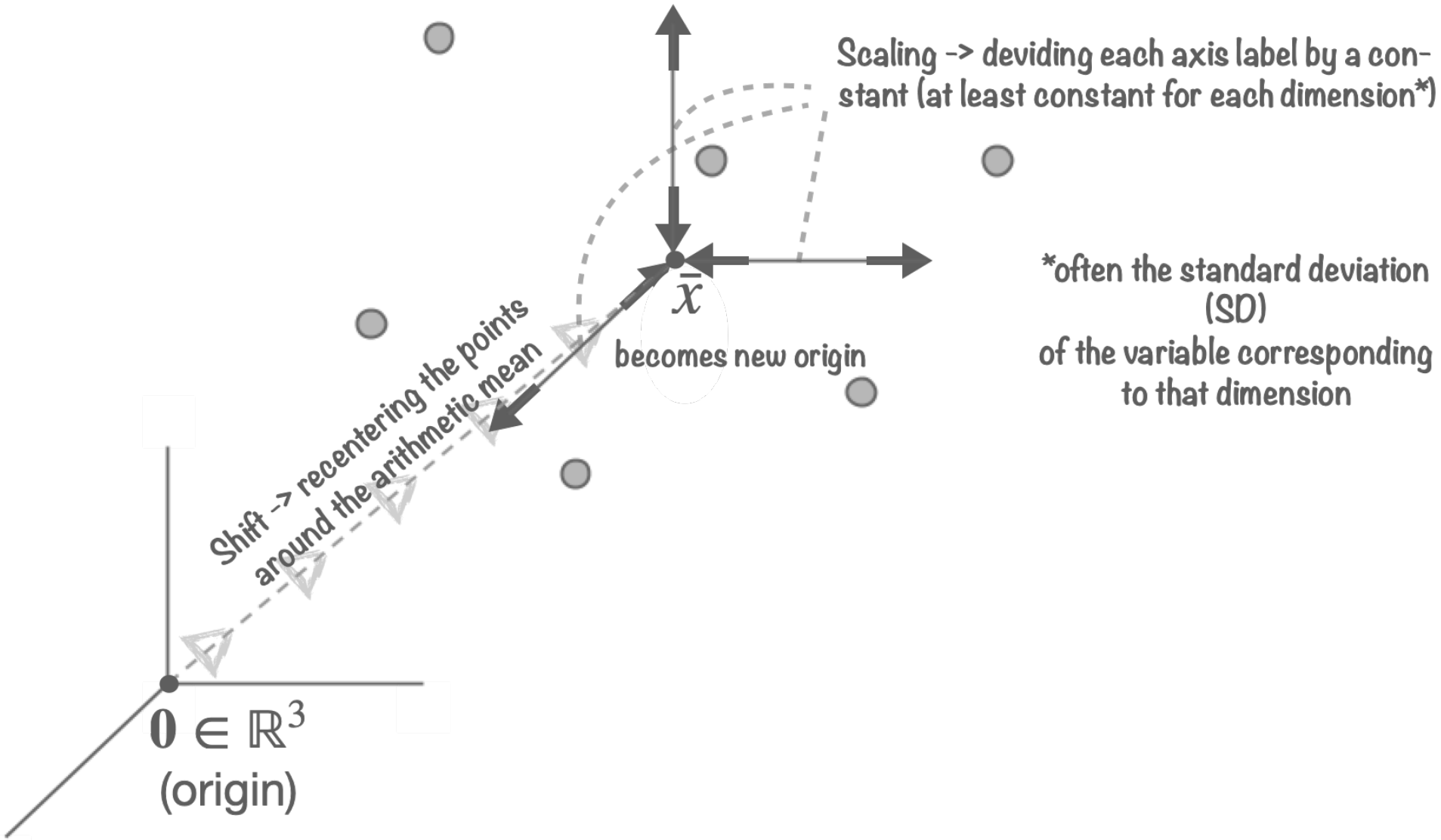
Some other transformations



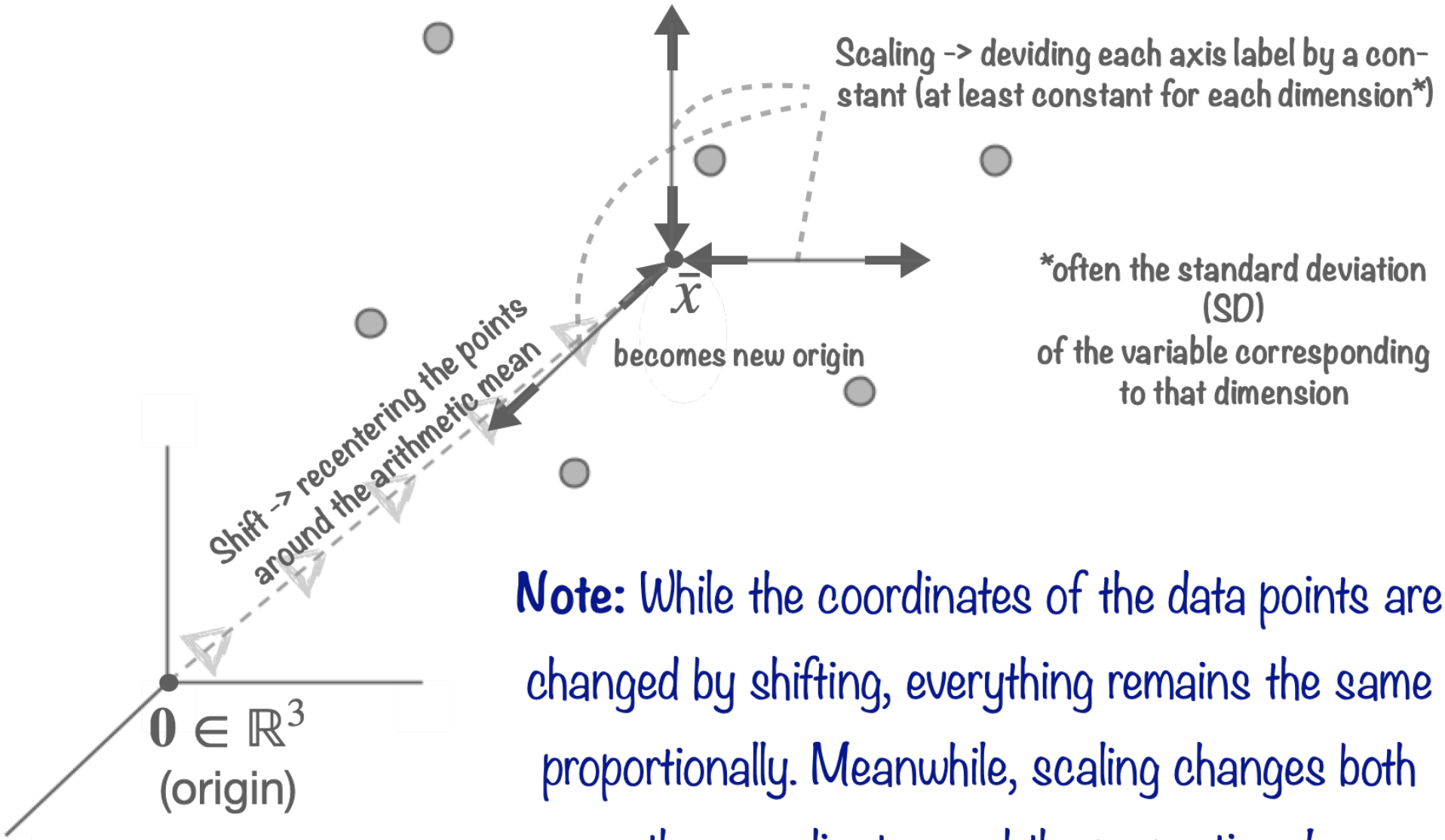
Some other transformations



Some other transformations



Some other transformations



Note: While the coordinates of the data points are changed by shifting, everything remains the same proportionally. Meanwhile, scaling changes both the coordinates and the proportions!

Descriptive Quantiles

- Just as with the mean as *center of gravity*, we can define the quantiles of a **one dimensional data sequence** $\mathcal{D} = (x_i)_{i=1}^n$, $x_i \in \mathbb{R}\ \forall i = 1, ..., n$, without any probabilistic assumptions, simply as follows:

  1. order the data points in $\mathbb{R}$ so that

  $$x^{(1)} \leq x^{(2)} \leq \ldots \leq x^{(n)}$$

  2. then, we define, for $\alpha \in (0, 1)$, the α-quantile as:

  $$q_\alpha(\mathcal{D}) := \begin{cases} x^{(\alpha \cdot (n+1))}, & \text{if } \alpha \cdot (n+1) \in \mathbb{N} \\ \frac{1}{2}\left(x^{(\alpha \cdot (n+1))} + x^{(\alpha \cdot (n+1)+1)}\right), & \text{otherwise.} \end{cases}$$

Multivariate descriptive quantiles?

- The definition of descriptive quantiles for higher dimensional data points is not quite as straightforward, because $\mathbb{R}^p$, $p \in \mathbb{N}_{\geq 2}$, *is not a totally ordered vector space*.

- More intuitively: For $\mathbb{R}^p$, $p \in \mathbb{N}_{\geq 2}$, we do not immediately have comparison operators as nice as $\leq$ and $\geq$.

- Of course, we can work around this fact! See, for example this interesting paper.

- But, for higher dimensions, it is often much easier to try estimating quantiles via the data's distribution - the **probabilistic view!**
  More on this later; for now: What other advantages does the probabilistic view have in this context, and which assumptions do we need to make?

What are descriptive statistics? I

- Well, the The Oxford Dictionary of Statistical Terms says:

"**descriptive statistics :=** *A term used to denote statistical data of a descriptive kind or the methods of handling such data; more broadly, methods of analysis, graphical or tabular, without any probabilistic formulation. When the data are determined by national authorities they are referred to as official statistics.*"

- What do you think about this definition?

- What definitely makes sense is that we do not take a probabilistic view for descriptive statistics in any way.

What are descriptive statistics? II

- That being said, what does that make, for example, the OLS estimate of linear regression coefficients, which we can calculate without any probabilistic thinking? *(Hint: there is no right answer)*

- Going further in the other direction, there is the field of *causal inference*, where most anything that is not defined under specific probabilistic assumptions is often called descriptive.

- For the purposes of this course we will say that **descriptive statistics** is any plot or numerical output that

  1. was derived without any probabilistic assumptions

  2. is meant to visualize/summarize given data.

Why would anyone use that?

There are many very situations where descriptive statistics can be very helpful:

Why would anyone use that?

There are many very situations where descriptive statistics can be very helpful:

- Quickly getting an overview and an intuition for the data.
- In an **exploratory analysis** setting, **and only then**, descriptive statistics can be very useful to generate hypotheses.
- Effectively communicating certain elements/concerns/take-aways; especially to people less familiar with statistical concepts.
- Visually checking whether certain assumptions for probabilistic modelling are met (arguably not purely descriptive anymore).
- Sometimes, descriptive statistics can be the only unarguably valid analysis path. *Why? When? Discuss.*

How does one use it

- Of course, we can compute all sorts of descriptive summary statistics, such as *mean (center of gravity)*, *intertia*, and *quantiles*.
- However, visualizations are often more powerful; especially since we can add the above values in them as points/lines.

ggplot()

When you are working in R, you will usually use the `ggplot2` package. Some of the advantages include:

- Use of a consistent grammar
- Plot specification at a high level of abstraction complex graphics easier than with base R functions
- High flexibility
- Appearance of plots easily customisable via themes
- Especially in comparison to base R graphics: ggplot2 follows a grammar Commands follow a more logical syntax Commands are usually slightly longer for simple graphics Commands are sometimes significantly shorter for complex graphics Commands are always applied to a data.frame There are no class-specific methods (`plot.lm`, `plot.gam`, ...)

Visualizing data

- Generally, the "perfect" visualization for certain data does not exist. It really depends on the research question and, even, to a certain extend the viewer.

- Some standard options, examples of which you can view here, include
  - Scatterplots
  - Lineplots
  - Bar graphs
  - Box plots
  - Histograms

Are there "rules"?

- Generally, any descriptive plot that is correctly interpreted is not "wrong"; and it takes some time to get an intuition for helpful plotting of data.

- However, there are some general guidelines:
  - If you want to visualize the "effect" of one variable on another, plot the first on the x and the second on the y axis.
  - Try not to avoid large areas containing no information (such as a HUGE boxes in boxplots or including an outlier in a scatterplot due to which all other points occupy a small space).
  - Mind your axis limits. Especially when you are comparing plots, you should make sure that the axis are not on completely different scales, so that the plot elements are somewhat comparable.
  - Do not overload the viewer.