Data Visualization

CentraleDigitalLab
@LaPlateforme_

First: what is the problem?

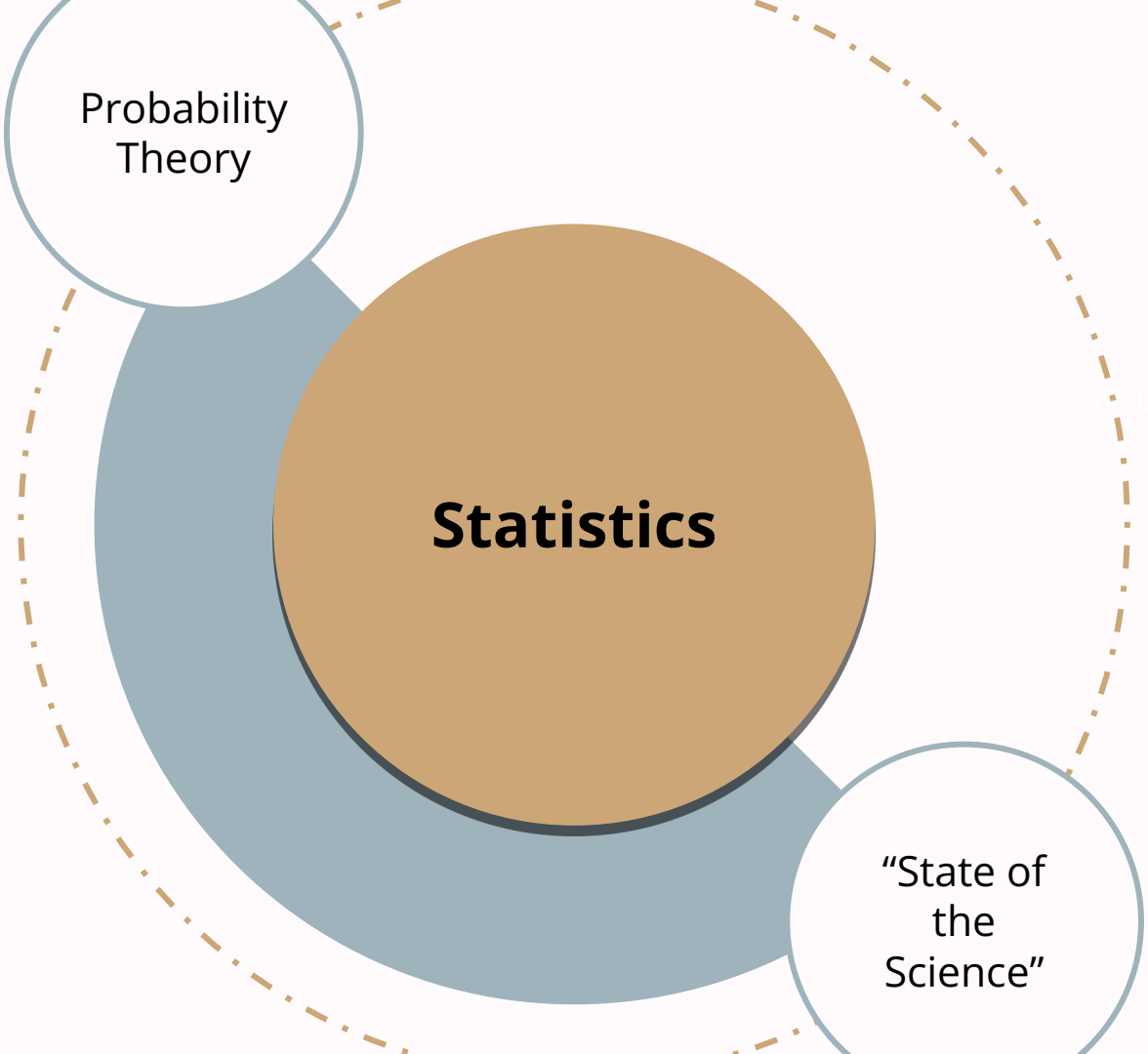
What might my salary be as a programmer?
Implement a system that, given the characteristics of a person, returns the salary that they can expect to receive as a programmer

Sysarmy Survey

- Personal and voluntary survey that seeks to collect information on salaries and working conditions of programmers, which is carried out annually.
- We will use data from Argentina
- Link to data

Demo with Notebook

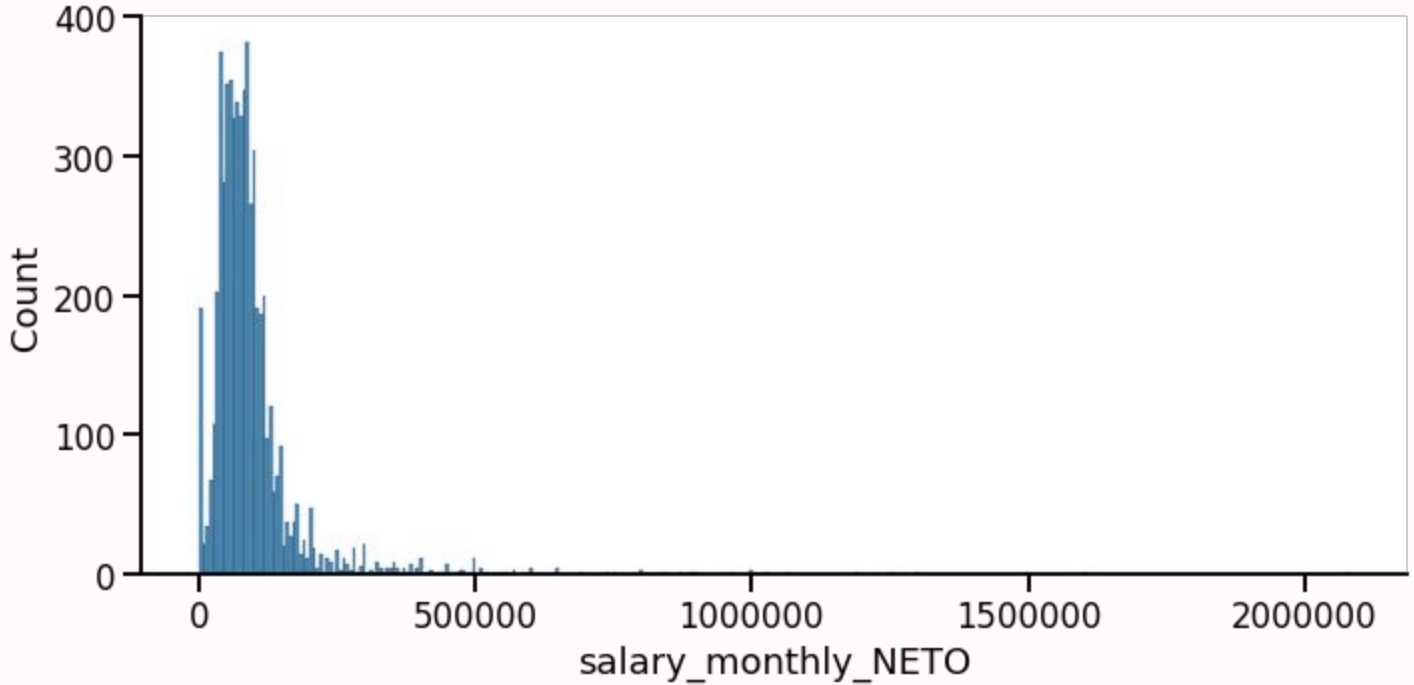
01 Probability_and_basic_plots.ipynb



Collection and use of data from a government institution

Use of Statistics

- **Data description**
- **Sample analysis**
- Hypothesis testing, decision making
- Measurement of relationships
- Inference
- Prediction



What is the mathematical concept we use to model the salary_monthly_NET column?

Random Variable

A **random variable (r.v.)** X is a function

$$X: \Omega \rightarrow R_X$$

where Ω is the **state space** and R_X is the set of values that the variable will take, called **Range.**

A **random variable** **(r.v.)** X is a function

$$X: \Omega \rightarrow R_X$$

where Ω is the **state space** and R_X is the set of values that the variable will take, called **Range.**

X is the r.v. salary_monthly_NETO, which takes a person who is a programmer in Argentina (2021) who took the survey and returns its net monthly salary.

Can we be more specific?

A **random variable** **(r.v.)** X is a function

$$X: \Omega \rightarrow R_X$$

where Ω is the **state space** and R_X is the set of values that the variable will take, called **Range.**

The state space Ω is the set of possible values (people) that we could have found in our survey.

Ω = {ω / ω is a living person who works in Argentina}

It can have more than one definition:

Ω = {ω / ω is a living person who works in Argentina as a developer(s)}

A **random variable** **(r.v.)** X is a function

$$X: \Omega \rightarrow R_X$$

where Ω is the **state space** and R_X is the set of values that the variable will take, called **Range.**

The range Rx is the set of possible values of salary_monthly_NET.

Rx = $\mathbb{R}$? (set of real numbers)

R x = $\mathbb{N}$? (set of natural numbers)

How can we calculate the Rx range in the survey?

A **random variable (r.v.)** X is a function

$$X: \Omega \rightarrow R_X$$

where Ω is the **state space** and R_X is the set of values that the variable will take, called **Range.**

ω = Person who answered first

X(ω) = 43000.0

X(ω) is called the realization of the r.v. X

Random variable- Other examples

X	Ω (universe in which we will be measuring things)	R_X
Daily work hours	Developers	1 - 24
Number of red blood cells	People	Real numbers

Type of random variables

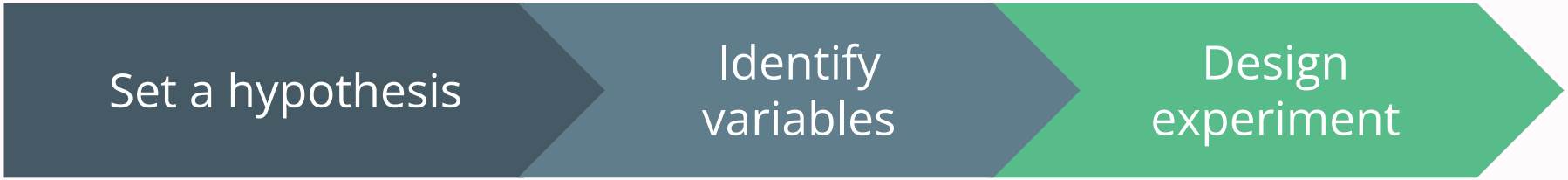
The random variables can be of different types, according to the values present in the Range and their interpretation.

- Numerical
  - Continuous
  - Discrete (Infinite or finite set of numerable values).
- Categorial
- Ordinal

Determining the types of data/variables that we are using allows us to select the appropriate tools to obtain information from them.

Let’s ask an interesting question:
Does having more years of experience mean you get paid more?

How to do this analysis?

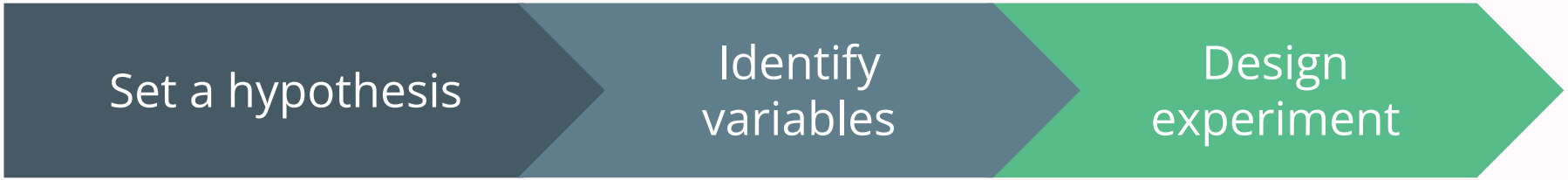


If we do not formulate a hypothesis first, it is difficult to determine what steps must be followed to be able to do the analysis

Once the hypothesis is defined, it is necessary to determine WHAT needs to be measured in order to test it.

Once what to measure is defined, the tools to measure it are selected.

How to do this analysis?



Having more years of experience means you get paid more

salary_monthly_NETO
profile_years_experience

????

Probability theory

Why can we use probability?

When we talk about doing data science, what we are looking for is being able to reason about real phenomena. ET Jaynes summarizes this need as:

1. Represent the degrees of plausibility of phenomena using numbers.
2. Qualitative correspondence with common sense.
3. Consistency.

Probability theory

Set of mathematical tools that allows us to reason about random experiments, which must comply with:

1. It can be repeated infinite times with the same experimental setup.
2. Has a fixed set of possible outcomes
3. Before doing it, you cannot predict the result that will be obtained.

Can this survey be modeled as a randomized experiment? How?

Probability? - Axiomatic interpretation

P is a **Probability** in the **state space Ω** if for each subset A of **Ω** , **P**(A) is a number such as:

- $\mathbf{0} \leq \mathbf{P}(A) \leq 1$
- $\mathbf{P}(\mathbf{\Omega}) = 1$
- $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$, for A and B disjoints
- $\mathbf{P}(\bigcup_i A_i) = \sum_i \mathbf{P}(A_i)$ for A_1, A_2,... disjoints

How is it calculated?

Our Ω are all the responses in the survey, each ωi is a response, and the set A are the responses in which the phenomenon occurs.

If each of our events is independent and identically distributed, that is, P({ωi}) = 1/k, then the probability of a set A ⊂ Ω is the proportion of events in A.

$$P(\{\omega_i\}) = 1/k \implies |A|/k$$

More complex situations

If there are two situations to study, then the problem is modeled using the salary_monthly_NET and profile_years_experience columns to create sets of events and check if there is a relationship between them.

The sets that are chosen are those that determine the experiment.

- $A = \{ \omega_i : salary\_monthly\_NETO > avg(salary\_monthly\_NETO) \}$
- $B = \{ \omega_i : profile\_years\_experience > 5 \}$

More complex situations

A = { ω_i : salary_monthly_NETO > avg }

B = { ω_i : profile_years_experience > 5 }

The **joint probability** that two events occur at the same time is modeled as the probability of the set intersection.

$$P(A \cap B)$$

More complex situations

A = { ω_i : salary_monthly_NETO > avg }

B = { ω_i : profile_years_experience > 5 }

The conditional probability that the salary is above the average, assuming that the event of having more than 5 years of experience occurs, is calculated as:

$$P(B) \neq 0 \implies P(A|B) = \frac{P(A \cap B)}{P(B)}$$

More complex situations

A = { ω_i : salary_monthly_NETO > avg }

B = { ω_i : profile_years_experience > 5 }

A and B are said to be independent if

$$P(A \cap B) = P(A)P(B)$$

$$P(B) \neq 0 \implies P(A|B) = P(A)$$

If one has more than 5 years of experience, does the probability of earning more than the average increase? Are these events independent?

Descriptive statistics

Measures of central tendency

Given X a numerical r.v

and a set of realizations

$x = \{x_1, x_2, ...\}$

where $x_i = X(\omega)$ for some

$\omega \in \Omega$, y $N = |x|$

The sample mean (arithmetic) or average is calculated as:

$$\bar{x} = \frac{1}{N} \sum_{i}^{N} x_i$$

There are other types of media, but this is the most used.

Measures of central tendency

Given X a numerical r.v

and a set of realizations

$x = \{x_1, x_2, ...\}$

where $x_i = X(\omega)$ for some

$\omega \in \Omega$, y $N = |x|$

The median is calculated as:

1. Order the realizations from least to greatest
2. If N is odd, the median is the middle value:

$$median = x_{N/2}$$

3. If N is even, the median is the average of the two middle values:

$$median = \frac{1}{2}(x_{\lfloor N/2 \rfloor} + x_{\lfloor N/2 \rfloor + 1})$$

Measures of central tendency

Given X a numerical r.v

and a set of realizations

$x = \{x_1, x_2, ...\}$

where $x_i = X(\omega)$ for some

$\omega \in \Omega$, y $N = |x|$

The mode is the values with the greatest frequency, that is, the ones that are repeated the most.

There is only more than one mode when the count of two values is equal.

Positioning

Given X a numerical r.v

and a set of realizations

$x = \{x_1, x_2, \ldots\}$

where $x_i = X(\omega)$ for some

$\omega \in \Omega$, y $N = |x|$

The **percentil-k** of a set x is the value x_i such that k% of the values of the sample are less than x_i.

There's no an unique formula to calculate the percentil-k.

1. Order the realizations such that $x_j \leq x_{j+1}$
2. Select the element of the series at the position

$$n = \left\lceil \frac{P}{100} \times N \right\rceil .$$

Measures of dispersion

Given X a numerical r.v

and a set of realizations

$x = \{x_1, x_2, ...\}$

where $x_i = X(\omega)$ for some

$\omega \in \Omega$, y $N = |x|$

The sample variance measures the variation of the data across the squared distance to the sample mean.

$$v = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

The standard deviation is the square root of the variance. It is on the same drive as the data.

The coefficient of variation is the standard deviation divided by the sample mean. It is comparable between different v.a.

Measures of dispersion

Given X a numerical r.v

and a set of realizations

$x = \{x_1, x_2, ...\}$

where $x_i = X(\omega)$ for some

$\omega \in \Omega$, y $N = |x|$

The range and the interquartile range measure in which interval a certain percentage of the data falls.

Rank:
percentile-100 - percentile-0

Interquartile range:
percentile-75 - percentile-25