ID430B: Data Analytics for Designers 디자인 특강V <디자이너를 위한 데이터 분석>

Tutorial

Descriptive Statistics on JupyterLab

Tak Yeon Lee <takyeonlee@kaist.ac.kr> (takyeonlee.com)
AI-Experience-Lab (reflect9.github.io/ael)

Goals

1. JupyterLab

Install and start working on Python Notebook

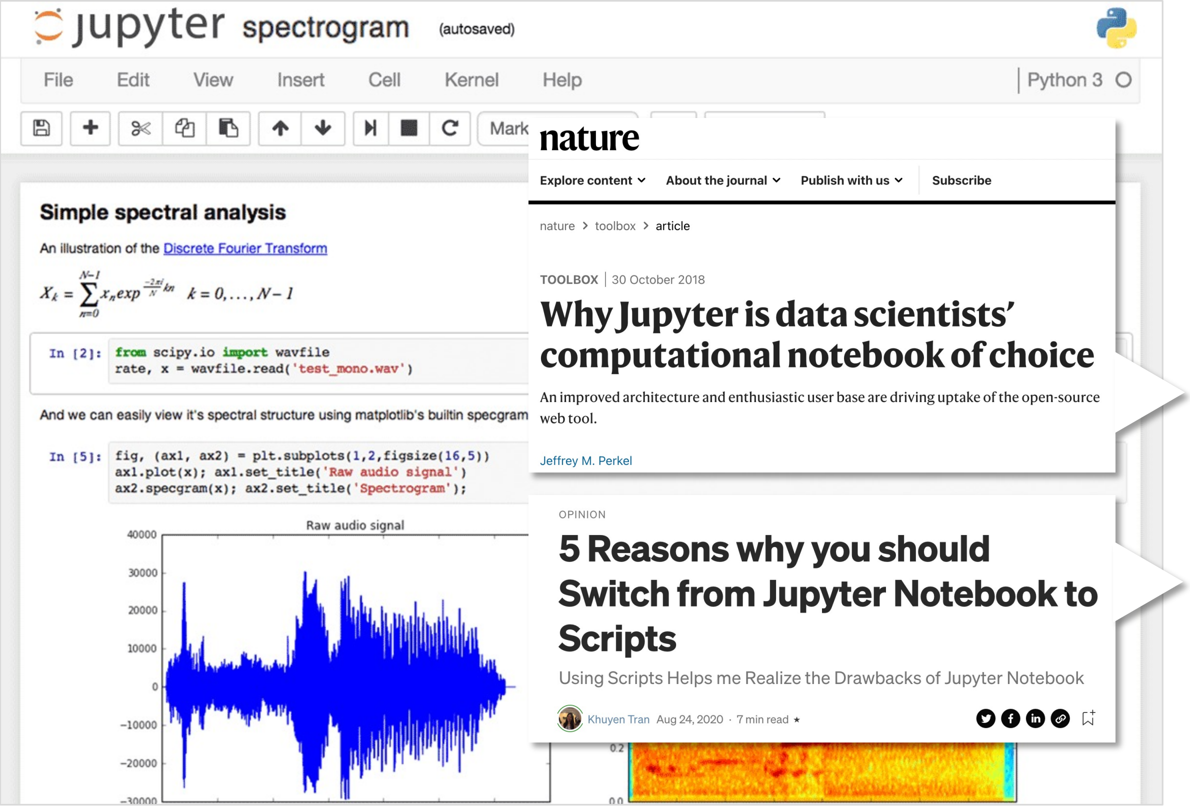
2. Descriptive Statistics

For a given dataset, understand basic characteristics and present it on Python Notebook

JupyterLab

A Popular Python Notebook Platform

What is Python Notebook?



Simple spectral analysis

An illustration of the Discrete Fourier Transform

$$X_k = \sum_{n=0}^{N-1} x_n exp^{\frac{-2\pi i}{N} kn} \quad k = 0, \ldots, N-1$$

```
In [2]: from scipy.io import wavfile
        rate, x = wavfile.read('test_mono.wav')
```

And we can easily view it's spectral structure using matplotlib's builtin specgram

```
In [5]: fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > toolbox > article

TOOLBOX | 30 October 2018

Why Jupyter is data scientists' computational notebook of choice

An improved architecture and enthusiastic user base are driving uptake of the open-source web tool.

Jeffrey M. Perkel

OPINION

5 Reasons why you should Switch from Jupyter Notebook to Scripts

Using Scripts Helps me Realize the Drawbacks of Jupyter Notebook

Khuyen Tran Aug 24, 2020 · 7 min read ★

Purposes (perfect fit for us!)

- Data Cleaning
- Statistical Modelling
- Training ML Models
- Data Visualization

Strengths (for data science)

- https://analyticsindiamag.com/why-jupyter-notebooks-are-so-popular-among-data-scientists/
- https://www.nature.com/articles/d41586-018-07196-1

Limitations

- Unorganized, Difficult to experiment, Not ideal for reproducibility, Hard to debug, Not for production
- https://towardsdatascience.com/5-reasons-why-you-should-switch-from-jupyter-notebook-to-scripts-cb3535ba9c95

When to use Jupyter Notebook?

- Code is small and not for production
- Goal is to explore data and visualize insights, and to share the process with other people
- In this course, Jupyter Notebook is a well-suited playground for students to practice data analytics.

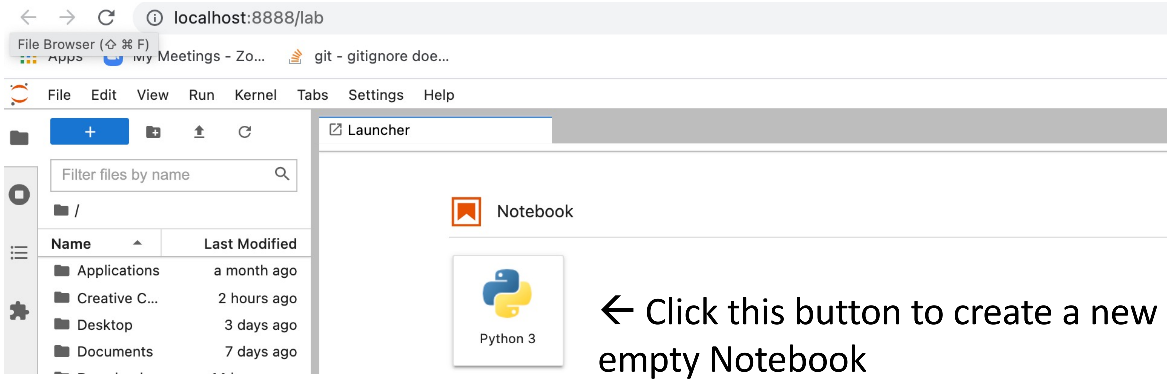
How to install and start JupyterLab

- Official Tutorial: https://jupyterlab.readthedocs.io/en/stable/getting_started/overview.html

1. Overview
2. Installation (recommend to install JupyterLab via conda)
    1. Install Anaconda (or Miniconda) for Windows or macOS
    2. Install JupterLab via conda

    ```
    conda install -c conda-forge jupyterlab
    ```
3. Start JupyterLab

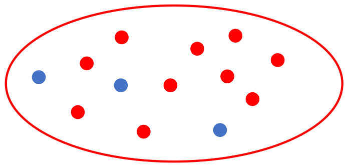
    ```
    Jupyter lab
    ```



← Click this button to create a new empty Notebook

Create a new Notebook

What is Descriptive Statistics?



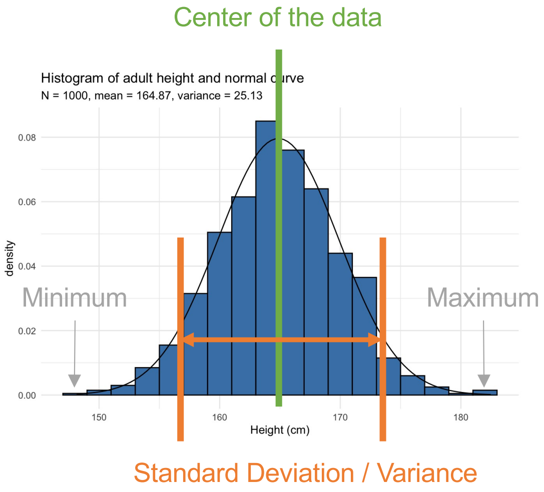
… quantitatively **summarizes** or **describes** the characteristics of a dataset (NOT the population)

E.g. Hundreds samples

E.g. Millions of the entire users

… consists of basic categories of measures shown below

1. What is the **central tendency** of the data? (Mean, Median, Mode)
2. How **spread out** is the data? (Standard deviation, Variance)
3. What are the **extremes** of the data? (Minimum, maximum; Outliers)
4. What is the "**shape**" of the distribution? Is it symmetric or asymmetric? Are the values mostly clustered about the mean or spread at the tails?
5. How many (**unique** / **non-empty**) values are in the data?



Center of the data

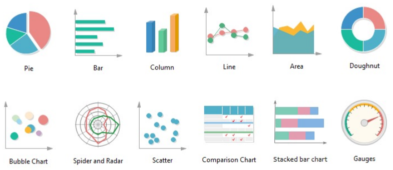
Minimum

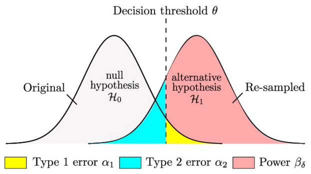
Maximum

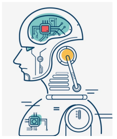
Standard Deviation / Variance

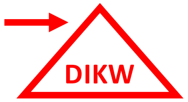
Descriptive Statistics vs Inferential Statistics

Descriptive	Inferential
Describe, summarize, and present characteristics of the known data	Inferring about the population based on the random sample of it
Organize, analyze and present	Compares, test and predicts
Describe a situation	Explain the chance of occurrence of an event
Central tendency, Variability, Distribution	Estimation of parameters, Hypothesis test
Results shown with charts and tables	Results shown with probability or model
E.g. "The log data shows that users took average 64seconds to complete the task. However, a fraction of users spent over 200 seconds. Why?"	*E.g. "Based on the latest A/B test, the upgrade would significantly increase the efficiency of our system."*





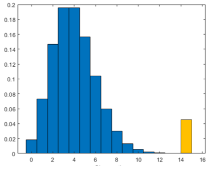




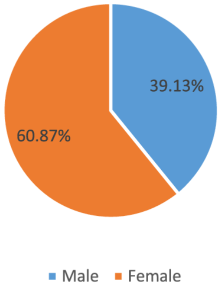
Why does Descriptive Statistics matter?

- **The first step** of understanding the dataset (i.e. the bottom layer of DIKW)

- **Sanity Check**: Use charts and descriptive statistics to spot data quality issues
  - (Completeness) *“Some students got negative scores from the exam. It might be a technical issue.”*
  - (Accuracy) *“Some data points are much greater than the others (i.e. outliers). We must check our collection method.”*
  - (Redundancy) “Lots of data points are identical. We must check duplicates in our dataset.”
  - (Bias) “90% of our participants are men. Is our dataset biased?”



“There exist outliers. We need to check whether this is an accuracy issue. ”



“If 90% of our data points are male, the dataset is biased (significantly different from the real population shown left).”

Measures of Central Tendency

- **Mean (i.e. Average)** is found by adding all of the numbers together and dividing by the number of items in the set

  E.g. (20 + 10 + 70 + 40 + 10) / 5 = 30

- **Median** is found by ordering the set from lowest to highest and finding the exact middle. The median is just the middle number. If the dataset has even # values, use the average of the two median values

  E.g. Original:[20,10,70,40,10] → Sorted:[10,10,**20**,40,70] → Median: 20

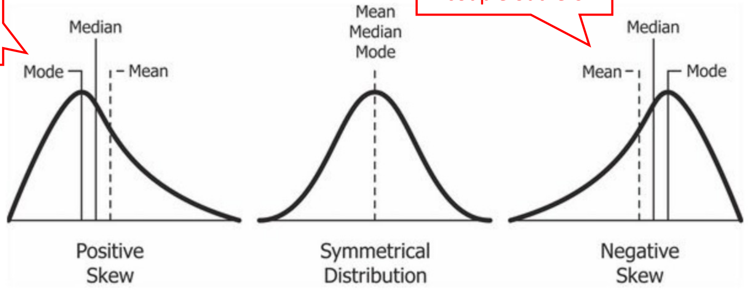
- **Mode** is the most frequently observed value

  E.g. Original:[20,**10**,70,40,**10**]

  → FreqByValue:{20:1, 10:2, 70:1, 40:1}

  → SortKeysByValue: [[10,2], [20,1], [70,1], [40,1]]

  → Mode: 10



Mode is at the peak of the distribution.

Mean can be affected by couple outliers

Measures of Variation (or Dispersion)

- Standard Deviation (SD) is calculated as below

Dataset $2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9.$

Mean
$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = \frac{40}{8} = 5$$

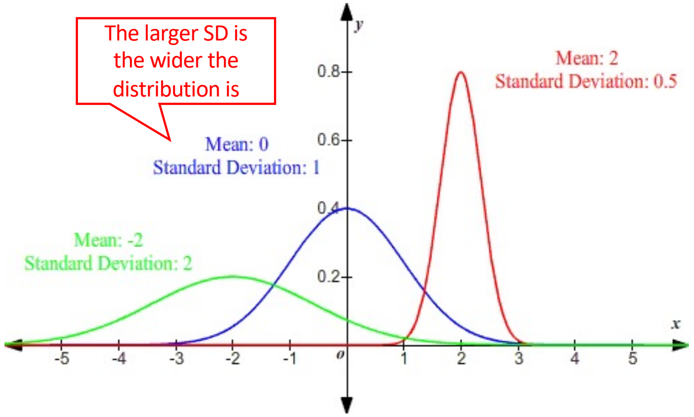
Mean

Squared Deviations
$$\begin{array}{ll} (2-5)^2 = (-3)^2 = 9 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (7-5)^2 = 2^2 = 4 \\ (4-5)^2 = (-1)^2 = 1 & (9-5)^2 = 4^2 = 16 \end{array}$$

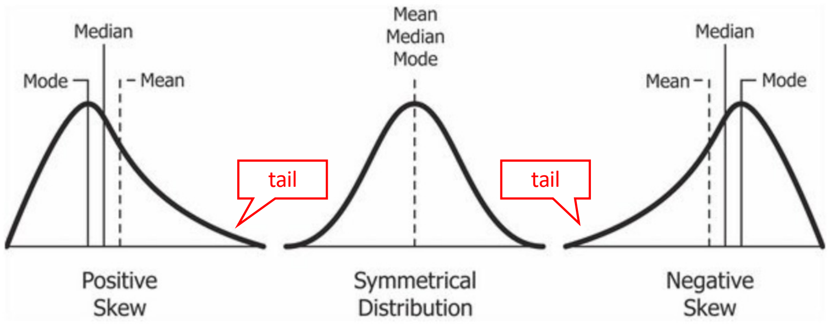
Variance
$$\sigma^2 = \frac{9+1+1+1+0+0+4+16}{8} = \frac{32}{8} = 4.$$

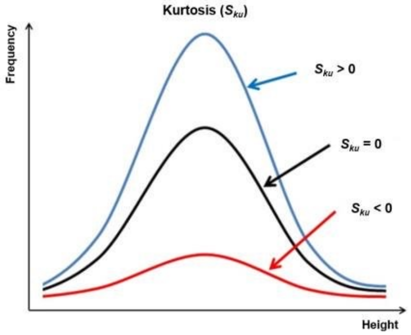
# values

Standard Deviation
$$\sigma = \sqrt{4} = 2$$


The larger SD is the wider the distribution is

Skewness and Kurtosis



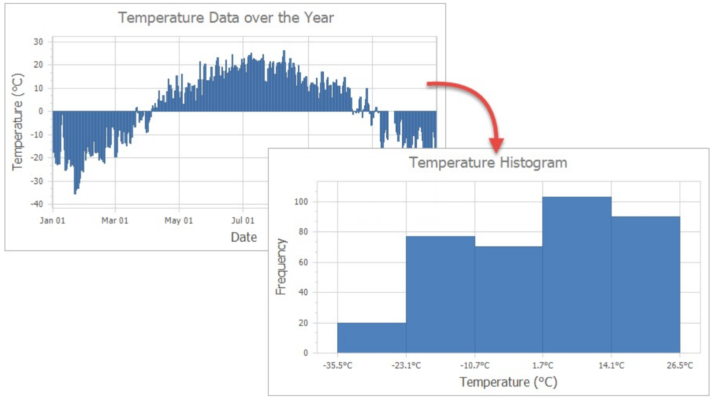


- **Skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined.
- No need to learn / memorize the formula

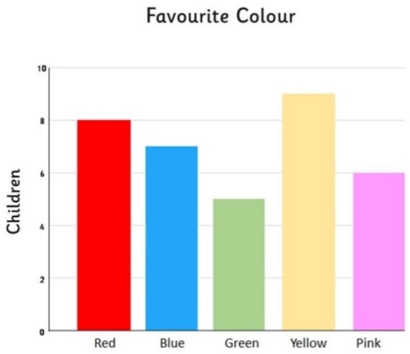
- **Kurtosis** is a measure the degree to which scores cluster in the tails or the peak of a frequency distribution.
- The smaller Kurtosis value is, the wider data points are spread out
- No need to learn / memorize the formula

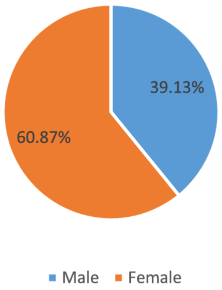
Value Counting for numeric / non-numeric data

- How many values (i.e. rows) in the dataset
  - Do we have enough data for the analysis in mind?
- How many unique (i.e. distinct) values in the dataset
  - Do they have a consistent format? Do we need to ignore / remove / fix them?
- Draw value frequency with charts
  - Histogram (for numeric data)



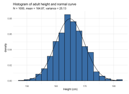
  - Bar / Pie charts (for non-numeric data)



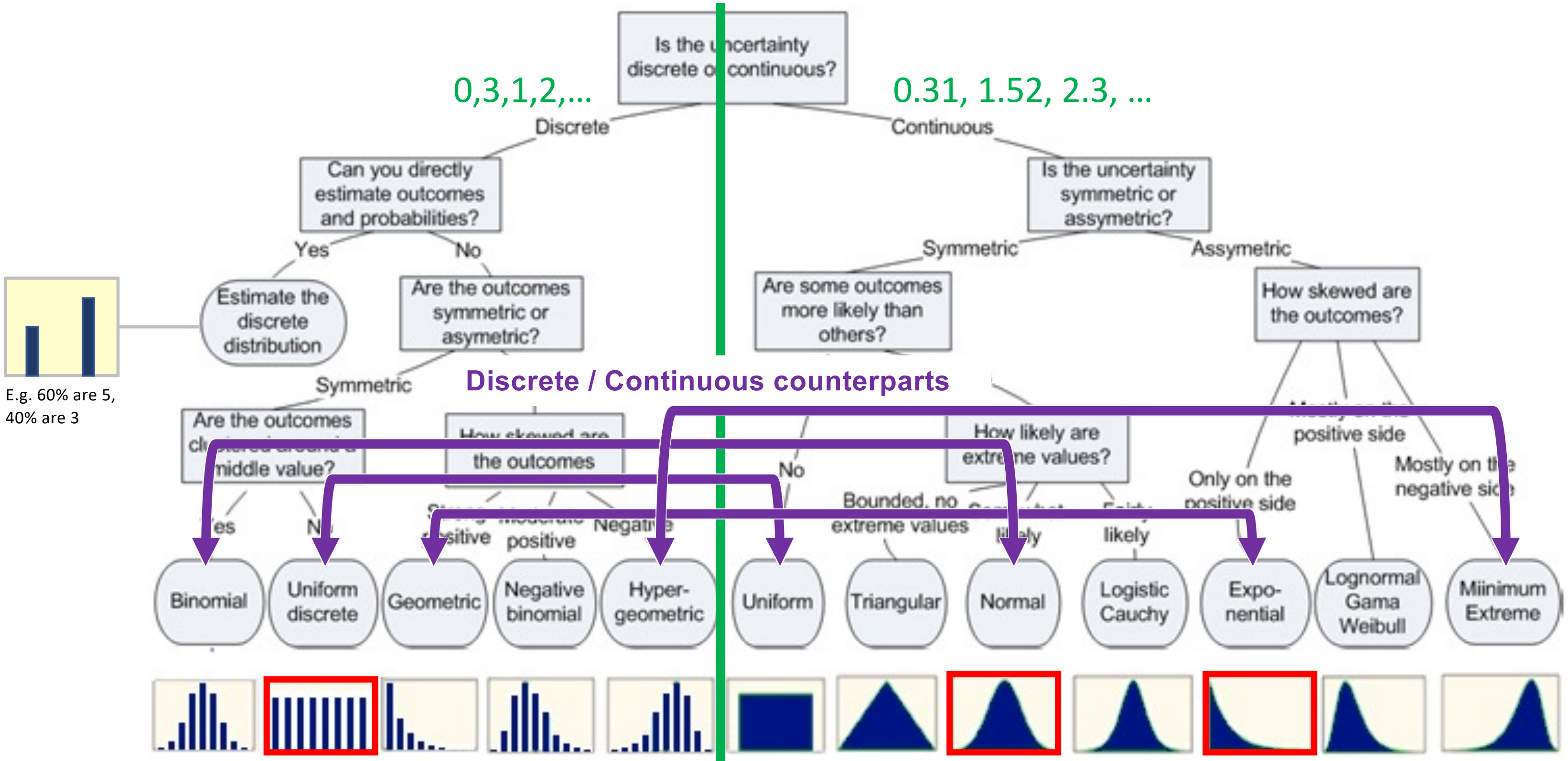


Types of Distribution

Types of Distributions



- Is it always bell-shaped? **No, there are many other types of distribution.**



Is the uncertainty discrete or continuous?

0,3,1,2,…

0.31, 1.52, 2.3, …

Discrete

Continuous

Can you directly estimate outcomes and probabilities?

Yes

No

Estimate the discrete distribution

E.g. 60% are 5, 40% are 3

Are the outcomes symmetric or asymetric?

Symmetric

Are the outcomes clustered around a middle value?

Yes

No

How skewed are the outcomes

Strong positive

Moderate positive

Negative

Is the uncertainty symmetric or assymetric?

Symmetric

Assymetric

Are some outcomes more likely than others?

No

How likely are extreme values?

Bounded, no extreme values

Somewhat likely

Fairly likely

How skewed are the outcomes?

Only on the positive side

Mostly on the positive side

Mostly on the negative side

Discrete / Continuous counterparts

Binomial

Uniform discrete

Geometric

Negative binomial

Hyper-geometric

Uniform

Triangular

Normal

Logistic Cauchy

Expo-nential

Lognormal Gama Weibull

Miinimum Extreme

Examples

Double dice roll

Single dice roll

# trials until the first success

Sampling without replacement

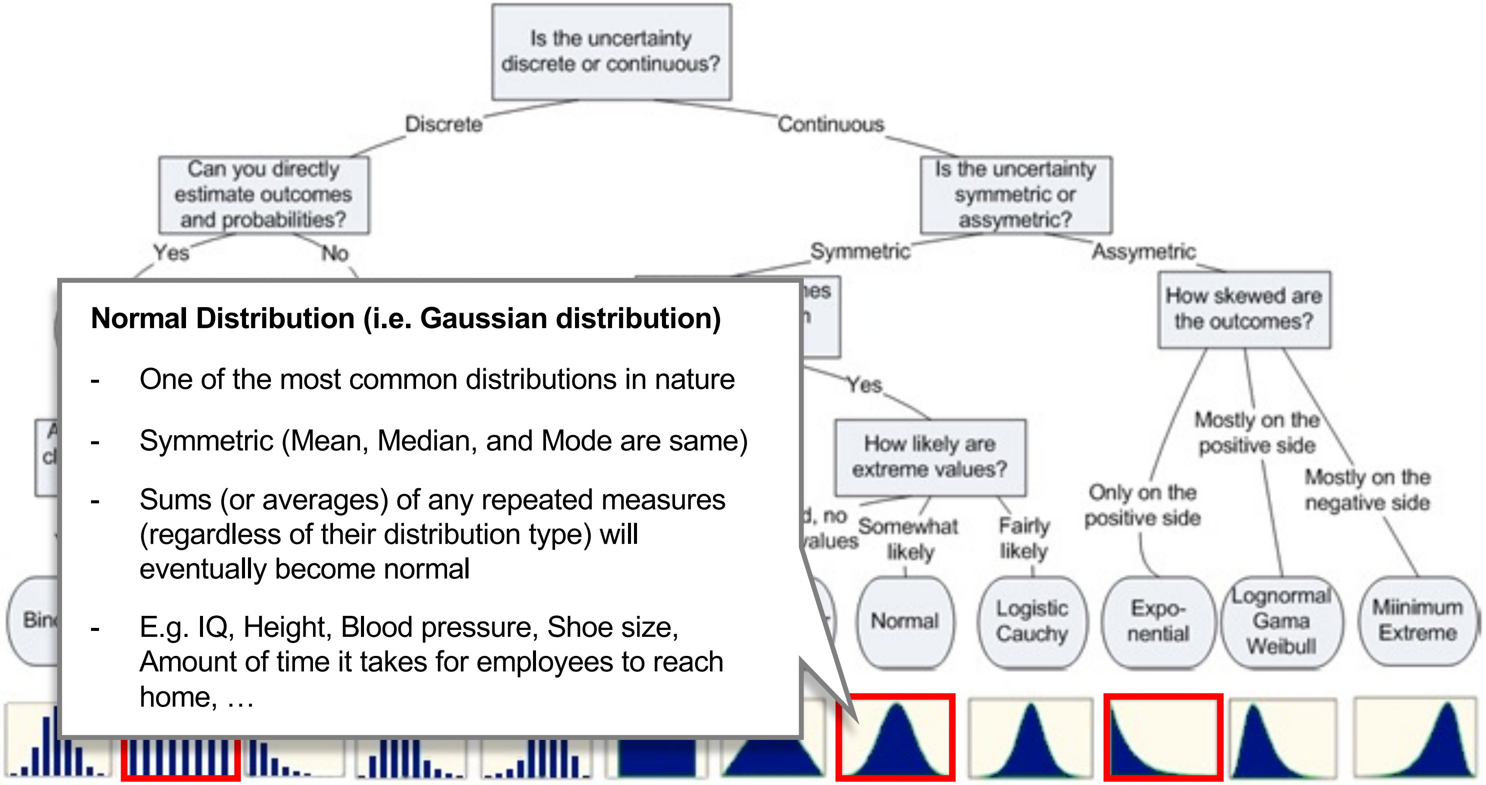
Perfect Random number generator

If you know min, max, and mode only

Most natural phenomenon

Time until the next rare event (e.g. earthquake), Money spent per grocery visit

Types of Distributions


Is the uncertainty discrete or continuous?
Discrete
Continuous
Can you directly estimate outcomes and probabilities?
Yes
No
Is the uncertainty symmetric or assymetric?
Symmetric
Assymetric
es
n
Yes
How likely are extreme values?
d, no
values
Somewhat likely
Fairly likely
How skewed are the outcomes?
Only on the positive side
Mostly on the positive side
Mostly on the negative side
A
cl
Bin
Normal
Logistic Cauchy
Expo-nential
Lognormal Gama Weibull
Miinimum Extreme

Normal Distribution (i.e. Gaussian distribution)

- One of the most common distributions in nature
- Symmetric (Mean, Median, and Mode are same)
- Sums (or averages) of any repeated measures (regardless of their distribution type) will eventually become normal
- E.g. IQ, Height, Blood pressure, Shoe size, Amount of time it takes for employees to reach home, …

Examples

Double dice roll

Single dice roll

# trials until the first success

Sampling without replacement

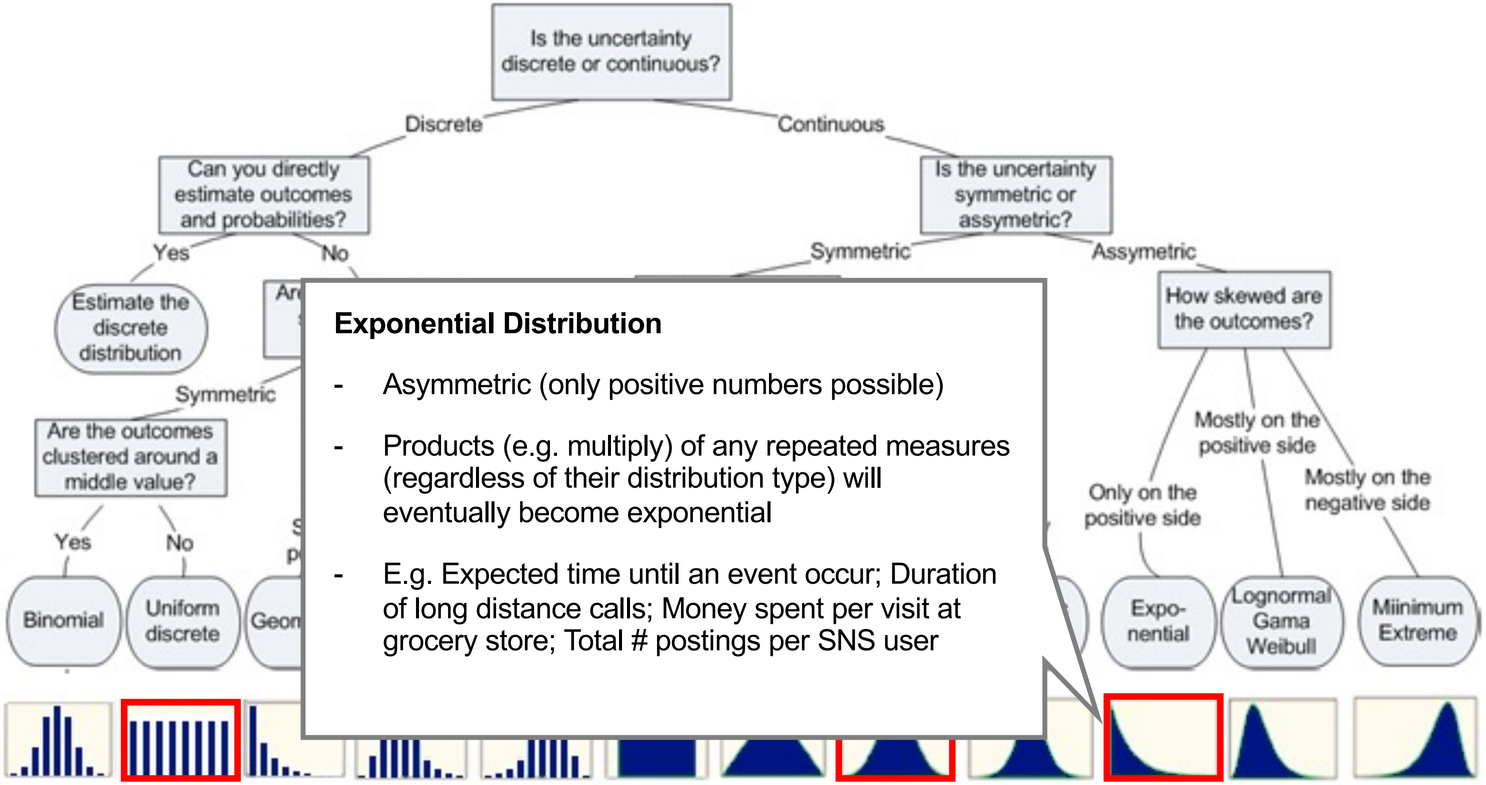
Perfect Random number generator

If you know min, max, and mode only

Most natural phenomen on

Time until the next rare event (e.g. earthquake), Money spent per grocery visit

Types of Distributions



Is the uncertainty discrete or continuous?

Discrete

Continuous

Can you directly estimate outcomes and probabilities?

Yes

No

Estimate the discrete distribution

Are

Symmetric

Are the outcomes clustered around a middle value?

Yes

No

Binomial

Uniform discrete

Geom

Is the uncertainty symmetric or assymetric?

Symmetric

Assymetric

How skewed are the outcomes?

Only on the positive side

Mostly on the positive side

Mostly on the negative side

Expo-nential

Lognormal Gama Weibull

Miinimum Extreme

Exponential Distribution

- Asymmetric (only positive numbers possible)
- Products (e.g. multiply) of any repeated measures (regardless of their distribution type) will eventually become exponential
- E.g. Expected time until an event occur; Duration of long distance calls; Money spent per visit at grocery store; Total # postings per SNS user

Examples

Double dice roll

Single dice roll

# trials until the first success

Sampling without replacement

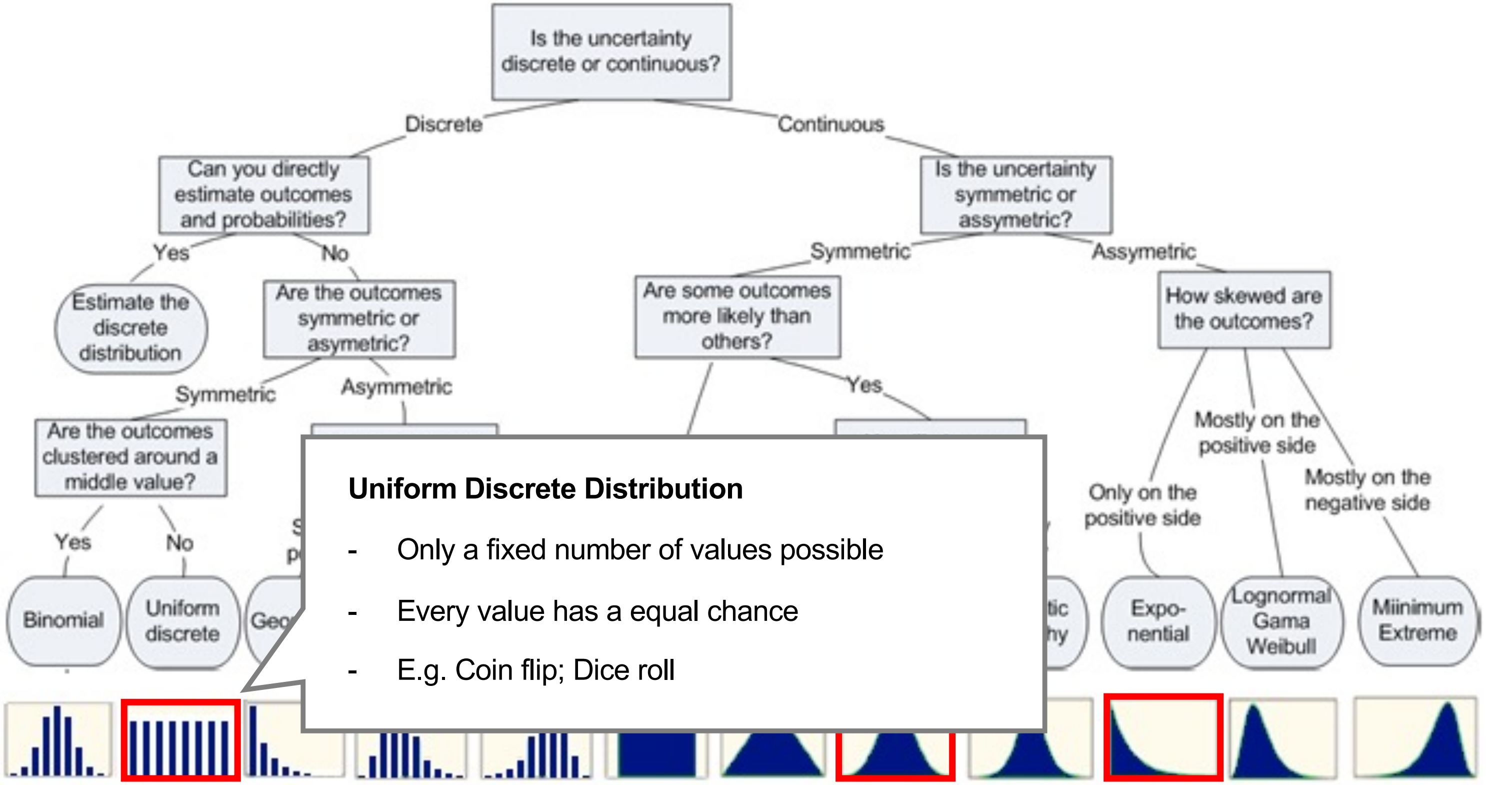
Perfect Random number generator

If you know min, max, and mode only

Most natural phenomenon

Time until the next rare event (e.g. earthquake), Money spent per grocery visit

Types of Distributions


Is the uncertainty discrete or continuous?
Discrete
Continuous
Can you directly estimate outcomes and probabilities?
Yes
No
Estimate the discrete distribution
Are the outcomes symmetric or asymetric?
Symmetric
Asymmetric
Are the outcomes clustered around a middle value?
Yes
No
Binomial
Uniform discrete
Geo
Is the uncertainty symmetric or assymetric?
Symmetric
Assymetric
Are some outcomes more likely than others?
Yes
How skewed are the outcomes?
Only on the positive side
Mostly on the positive side
Mostly on the negative side
Expo-nential
Lognormal Gama Weibull
Miinimum Extreme

Uniform Discrete Distribution

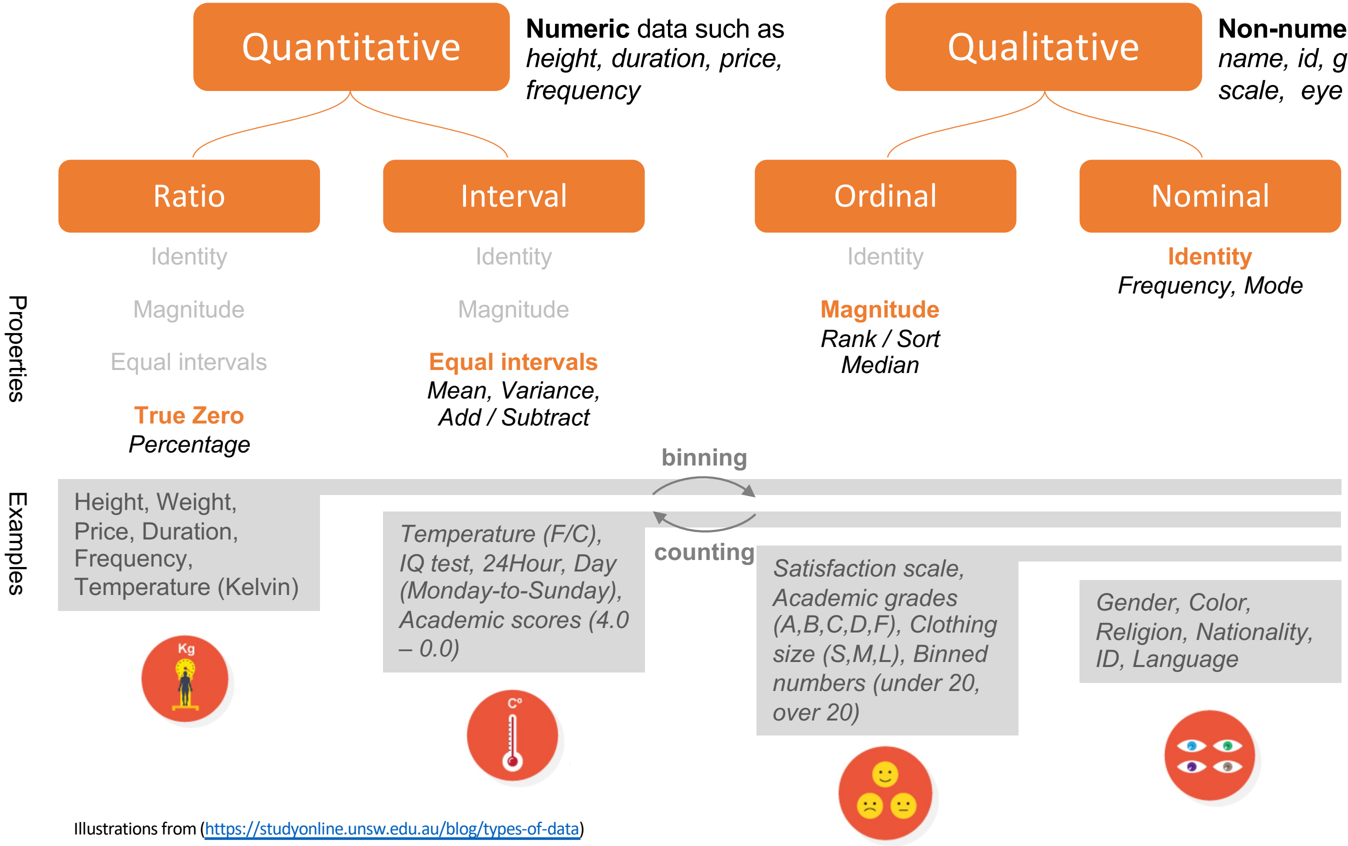
- Only a fixed number of values possible
- Every value has a equal chance
- E.g. Coin flip; Dice roll

Examples	Double dice roll	Single dice roll	**# trials** until the first success		Sampling without replacement	Perfect Random number generator	If you know min, max, and mode only	Most natural phenomen on		**Time** until the next rare event (e.g. earthquake), Money spent per grocery visit		

Levels of Measurement

Four Levels of Measurement

Whether a value is numeric or non-numeric is decided by its semantic meaning (not character itself). For instance, “5” is numeric (quantitative) if it means “5 pieces of cake”; or non-numeric (qualitative) if it means “5 out of 10 satisfaction scale on survey data”



Quantitative

Numeric data such as *height, duration, price, frequency*

Qualitative

Non-numeric data such as *name, id, gender, yes/no, Likert scale, eye color*

Properties

Ratio

Identity

Magnitude

Equal intervals

True Zero

Percentage

Interval

Identity

Magnitude

Equal intervals

Mean, Variance, Add / Subtract

Ordinal

Identity

Magnitude

Rank / Sort Median

Nominal

Identity

Frequency, Mode

binning

counting

Examples

Height, Weight, Price, Duration, Frequency, Temperature (Kelvin)

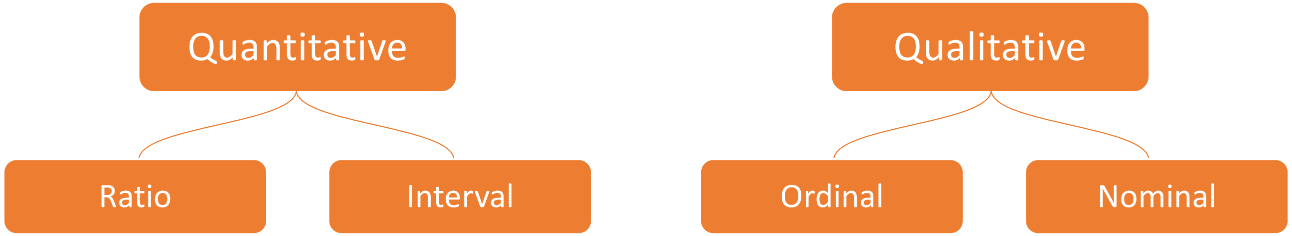
Temperature (F/C), IQ test, 24Hour, Day (Monday-to-Sunday), Academic scores (4.0 – 0.0)

Satisfaction scale, Academic grades (A,B,C,D,F), Clothing size (S,M,L), Binned numbers (under 20, over 20)

Gender, Color, Religion, Nationality, ID, Language

Illustrations from (https://studyonline.unsw.edu.au/blog/types-of-data)

E.g. Academic Grading is … weird



Quantitative

Ratio

Interval

Qualitative

Ordinal

Nominal

(1)

40% Assignments
30% Final exam
30% Midterm exam
Total: 100%

Total score earned throughout the course **is ratio.**
If a student didn’t submit anything he/she may get true zero point.

(4)

Total GPA: 420
# classes = 120
Average GPA:
420/120 = 3.5

Average GPA is ratio again.

(3)

4.0
3.0
2.0
1.0
0.0

GPA is interval because they are added to calculate the total GPA. However GPA is not ratio since A is not twice of C by any means. Also F does not mean true zero.

(2)

A: 90% - 100%
B: 80% - 90%
C: 70% - 80%
D: 60% - 70%
F: - 60%

Letter grade is ordinal because they can be ranked (i.e. A is better than B). Differences are not precisely meaningful, for example, if one student scores an A and another a B on an assignment, we cannot say precisely the difference in their scores, only that an A is larger than a B.

Note.
Letter grade also qualifies as nominal values since we can check the equality (e.g. *two students got the same grade A*)