# Pandas

## **Scales**

**Ratio scale**: units equally spaces (+-/\* valid), e.g height

**Interval** **scale**: equally spaced but no true zero (\*/ not valid), e.g T°C

**Ordinal** **scale**: units ordered but not equally spaced, e.g A+, A-

**Nominal** **scale**: categories

grades=df['Grades'].astype('category',

categories=['C-','C','C+','B-','B','B+','A-','A','A+'],

ordered=True)

grades.head()

To categorize data: pd.cut(df['avg'], 3, labels=['Small', 'Medium', 'Large'])

## Pivot Tables

## Date and time

pd.Timestamp('9/1/2016 10:05AM')

pd.Period('1/2016')

df.index = pd.to\_datetime(df.index)

## Distributions

= set of random variables

*= shape describes the proba of a value being pulled when we sample a population*

* Binomial (0/1, discrete)
* Normal dist
* Chi square dist
* Bimodal (2 peaks) > Gaussian mixture models

discrete (binomial…) /continuous

mean value : mean of the sample

expected value : theorical mean

**Measures of central tendancy of the distribution**

mode, median, mean

**Measures of variability of the distribution**

* Standard deviation: how different each item is from the mean
* Interquartile range

kurtosis: shape of the tail of the distribution

Interpretation: >0: curve more peaky than a normal dist; <0: curve more flat

scipy.stats.kurtosis

skew : how much the peak is peaky and on the left

## Hypothesis testing

Hypothesis = a statement that we can test

Alternative hypothesis *(our idea)* vs. Null hypothesis

**Significance level**: critical value alpha (threshold *e.g 0.1-0.01)*

*p-value > H0 “we cannot reject null hypothesis” (two pop same, no statistical diff between samples)*

*p-value < H1  “we reject null hypothesis”*

*p = 0.01*

**T Test**

**Topics to dig in later**

p-hacking (video Hypothesis Testing, wk 4, Intro to DS)