Good morning ☺
Notebooks and slides are in:

https://github.com/AlineQuadros/teaching_data_science

Packages we'll need:

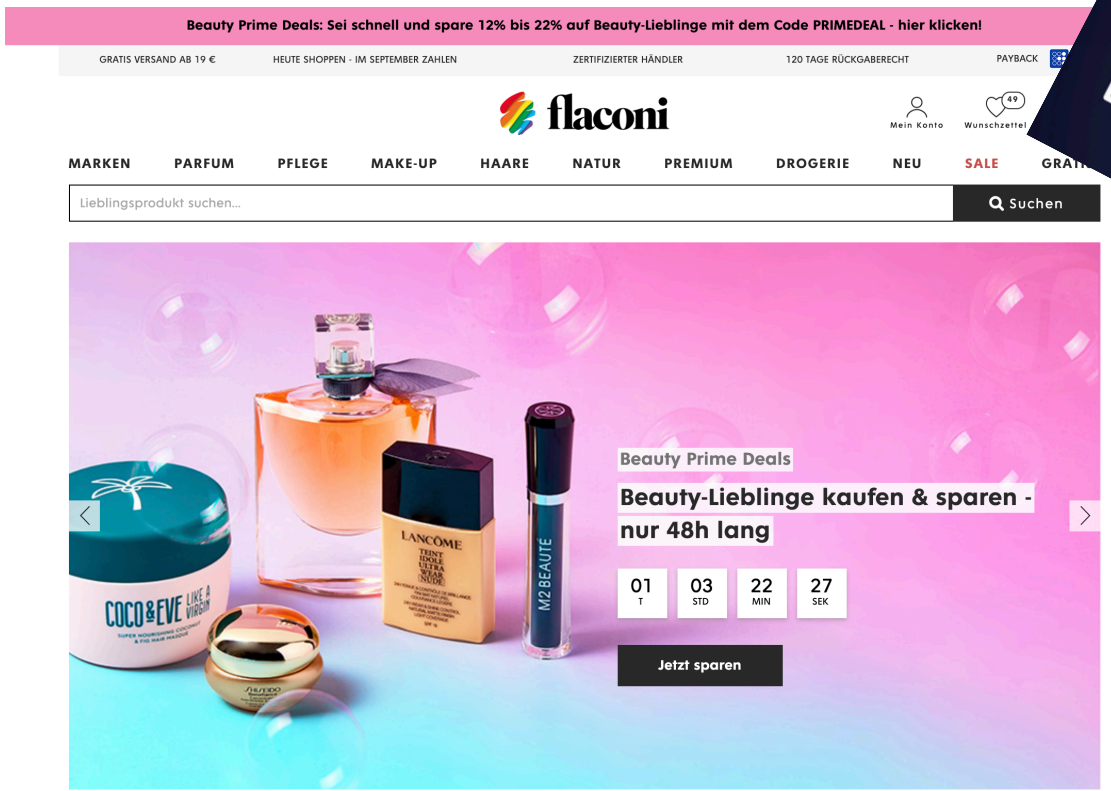
scipy==1.4.1
statsmodels==0.11.1

STATISTICS FOR DATA SCIENCE

DR. ALINE QUADROS

DSR JANUARY 2021

Data and Analytics @flaconi



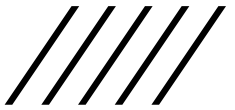
CHANEL DIOR BOSS HUGO BOSS KÉRASTASE PARIS SHISEIDO Jean Paul GAULTIER BABOR

DISCLAIMER

SORRY, I'M A FREQUENTIST

CONTENTS

- DESCRIPTIVE STATISTICS
  - PARAMETRIC VS. NON-PARAMETRIC STATISTICS
  - DISTRIBUTIONS
  - TRANSFORMATIONS
- STATISTICAL INFERENCE
  - HYPOTHESIS TESTING
  - THE P-VALUE CONUNDRUM
  - WHERE WILL WE USE STATISTICAL INFERENCE IN ML?
- EFFECT SIZES
- RESAMPLING AND BOOTSTRAPPING
- MODEL EVALUATION
- OVERVIEW OF *STATSMODELS* and *SCIPY*



DESCRIPTIVE STATISTICS

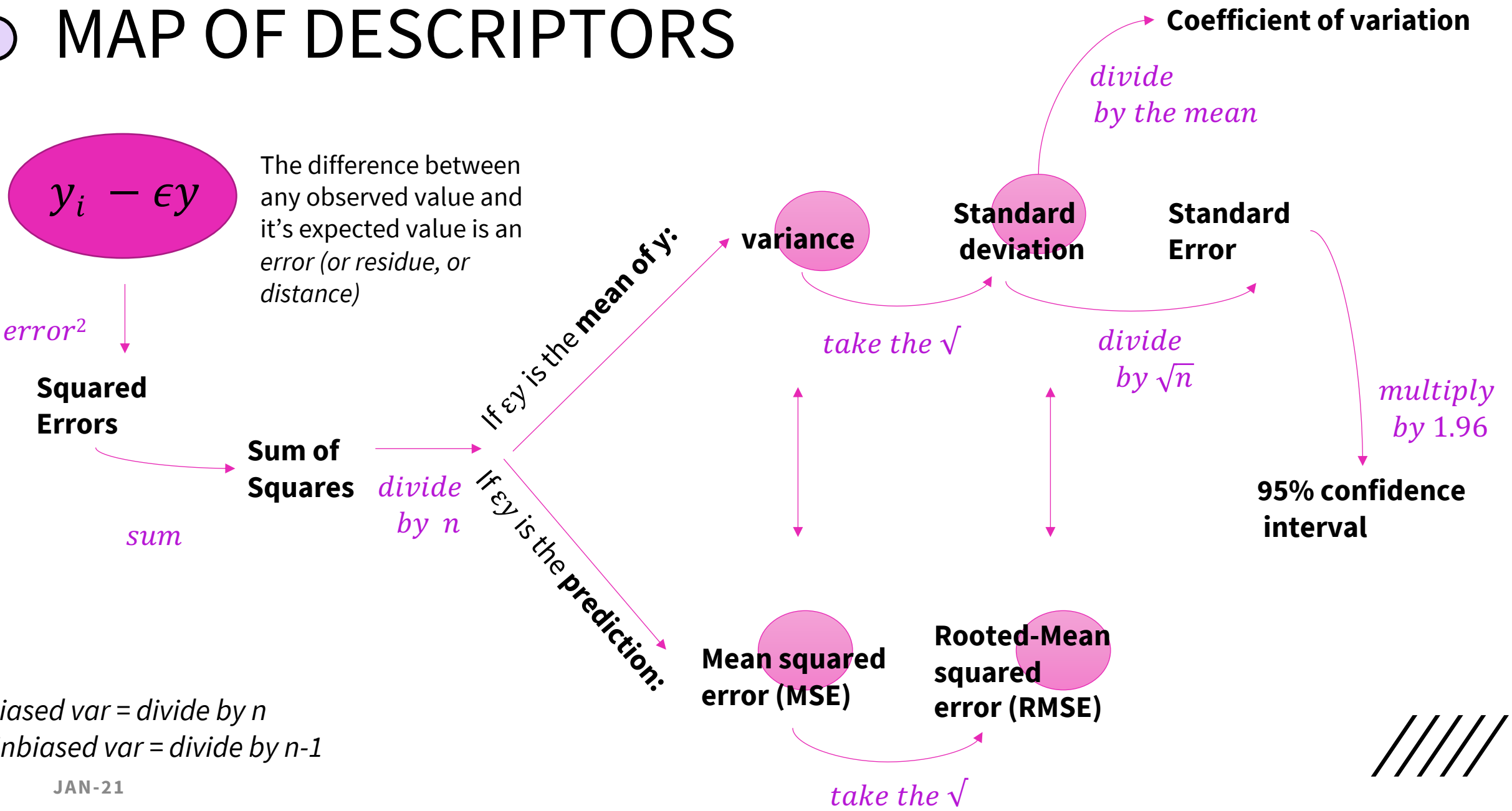
- MEAN, MEDIAN, MODE
- VARIANCE
- SD AND SE
- CONFIDENCE INTERVALS
- SS, MSE and RMSE
- COVARIANCE
- CORRELATION

Descriptive statistics and dataviz are the key components of a good POC

With descriptive statistics we learn:

- The properties of our features
  - Min-max values
  - General behavior
  - Outliers
  - Typos/data cleaning issues
- The distribution of the target variable
  - what we are trying to predict?
  - What kind of underlying process generates the data we are trying to model?

MAP OF DESCRIPTORS



Biased var = divide by n
Unbiased var = divide by n-1

You should to be able to answer these questions:

When should we use the mean? When should we not use the mean? What can we use instead?

What is the difference between standard error and standard deviation?

How much is the variance of this variable with 3 values: [2, 4, 6]

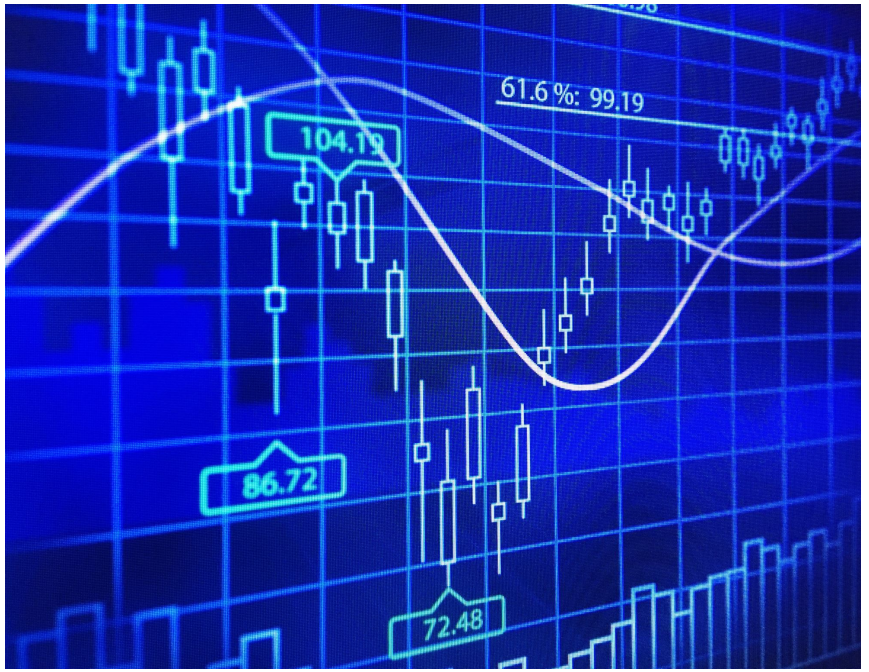
Notebook exercise 1

PARAMETRIC

VS.

NON-PARAMETRIC

STATISTICS



JAN-21

PARAMETRIC statistics

- Based on the parameters of a given probability distribution (like the Gaussian distribution has 2 parameters, the mean and s.d.).
    - T-tests, Pearson correlation, ANOVA

NON-PARAMETRIC statistics

- Doesn't make any assumptions about the underlying distribution
- But they have less power than parametric, are limited to simple experimental designs
    - Mann-Witney, Spearman, Kruskall Wallis etc.

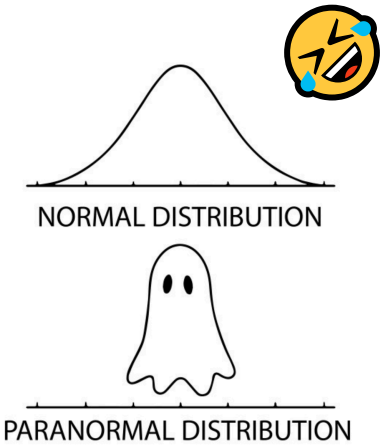
PARAMETRIC MODELS

- The number of parameters is FINITE, meaning the model's complexity does not scale with the amount of data
- Examples: Linear regression, logistic regression, perceptron, naïve Bayes

NON-PARAMETRIC models

- Parameters have infinite dimensions
- They grow in complexity as the data grows in size
- Examples: Decision trees, KNN, Kernel SVM, Gaussian processes

PROBABILITY DISTRIBUTIONS (for parametric statistics)



They describe the behavior of our variables. We must pay special attention to the distribution of our response variable.

Each distribution has one (or more) **parameters** that represent its behavior

- What is the parameter that describes the normal distribution?

- Continuous
  - Normal (Gaussian)
- Continuous non-negative
  - Gamma
- Discrete (binary)
  - Bernoulli/Binomial
- Discrete
  - Poisson
  - Negative binomial

POISSON

NUMBER OF TIMES AN EVENT IS OBSERVED IN TIME OR SPACE

Examples:

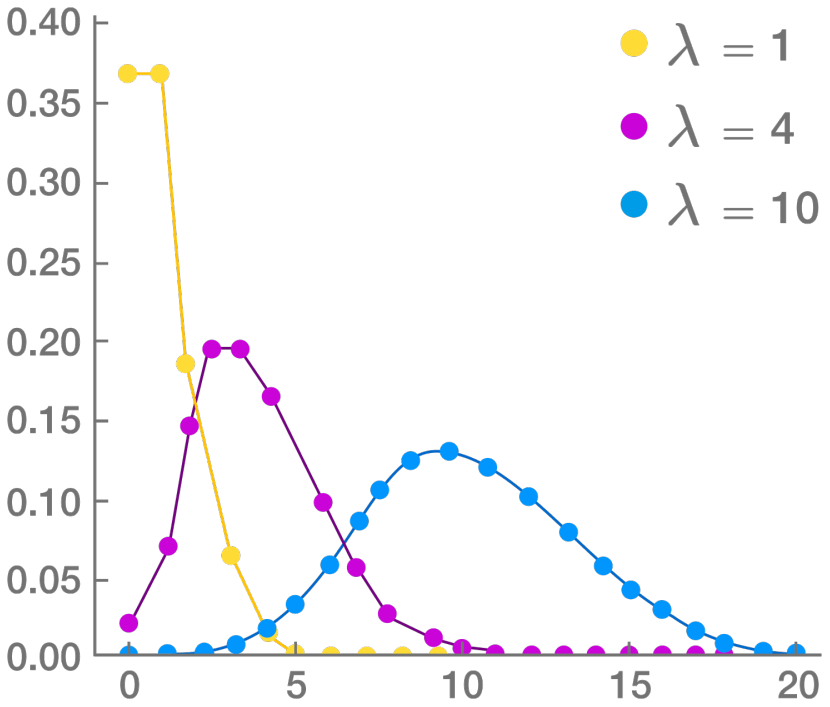
Number of sales per hour

Number of birds per tree

Number of goals per soccer match

POISSON assumes that the data is NOT overdispersed.

If it is (usually!), the alternative is the negative binomial.



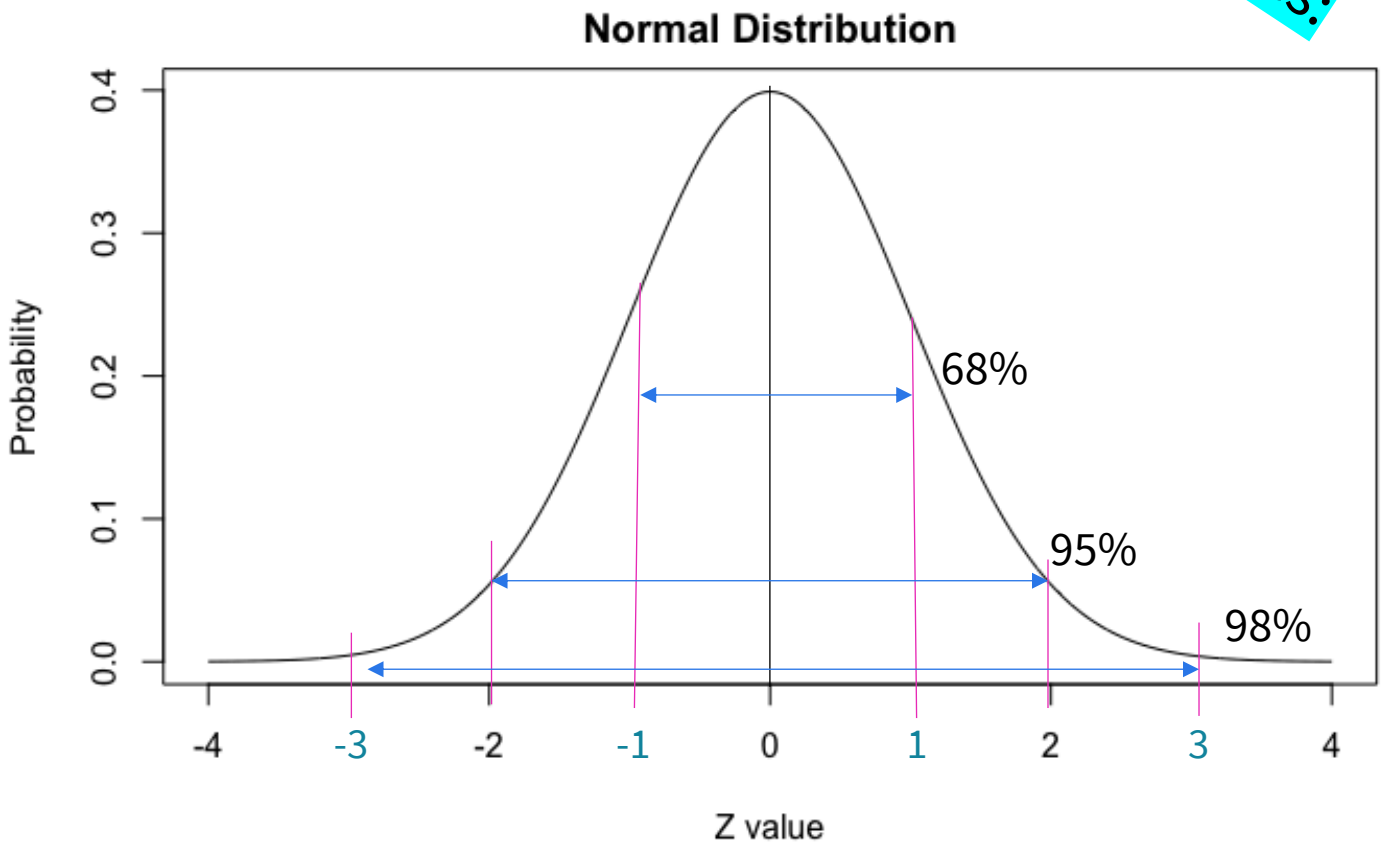
NORMAL (GAUSSIAN)

It's widely used because of interesting properties:

Properties:

Symmetry

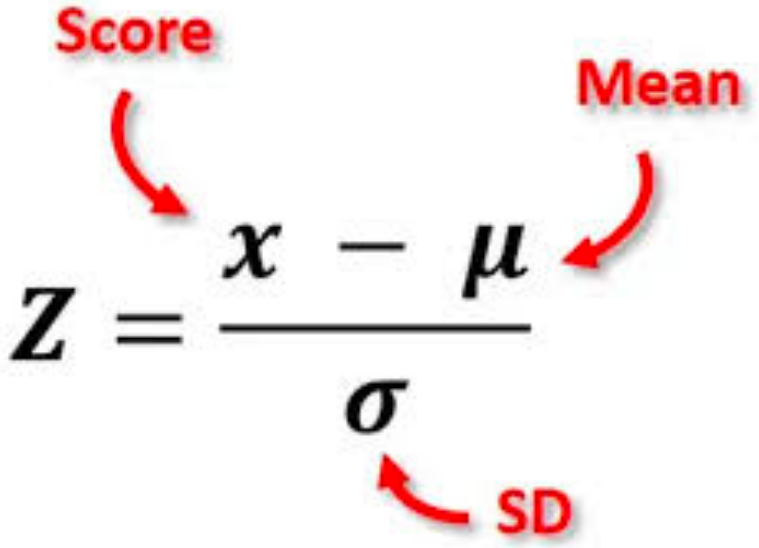
Mean = median = mode



BUT WAIT!
MY MEAN IS NOT 0 AND MY S.D. IS NOT 1!

That’s why we apply a

Z-SCORE
standardization



Notebook exercise 2

Dealing with paranormal 👻 Non-gaussian data

There's usually four ways of carrying on the analysis if you are working with regression problems and quantitative **target** variables that are not normally-distributed:

- 1. Look for models that don't need linear relationships in the data (E. g. random forests, boosted trees)
- 2. Look for models that can handle different distributions, like Poisson or Binomial (a.k.a. Generalized Linear Models)
- 3. If you are using a hypothesis test, use bootstrapping to generate to generate the null model
- 4. Apply transformations (log, sqrt, box-cox)

TRANSFORMATIONS

Very common step during model development (preprocessing of features)

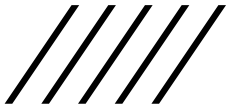
It's a trial-and-error process:

- Apply the transformation
- Inspect:
  - QQ—plot
  - Tests (Shapiro's)
- Repeat
- Choose the best

Type of Transformation	Common Applications
Log	Data with very diferent magnitudes Skewed distribution (but 0 and negatives?)
SQRT	Count data (but negatives?)
Sin/cos	Circular variables
Logit	Proportions/rates
Box-Cox	When everything else fails

Notebook exercise 3

Box, G., & Cox, D. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26*(2), 211-252. Retrieved September 29, 2020, from http://www.jstor.org/stable/2984418



STATISTICAL INFERENCE

WHAT IS IT?

- WE WANT TO **INFER** ABOUT THE POPULATION BY ANALYSING A SAMPLE OF IT

- ALWAYS A DUAL HYPOTHESIS:

  - **NULL:** THERE'S **NO** EFFECT
  - ALTERNATIVE: THERE IS AN EFFECT

- Why? Because of OCCAM'S RAZOR: Parsimony principle
- We try to DISPROVE THE NULL

Hypothesis testing workflow

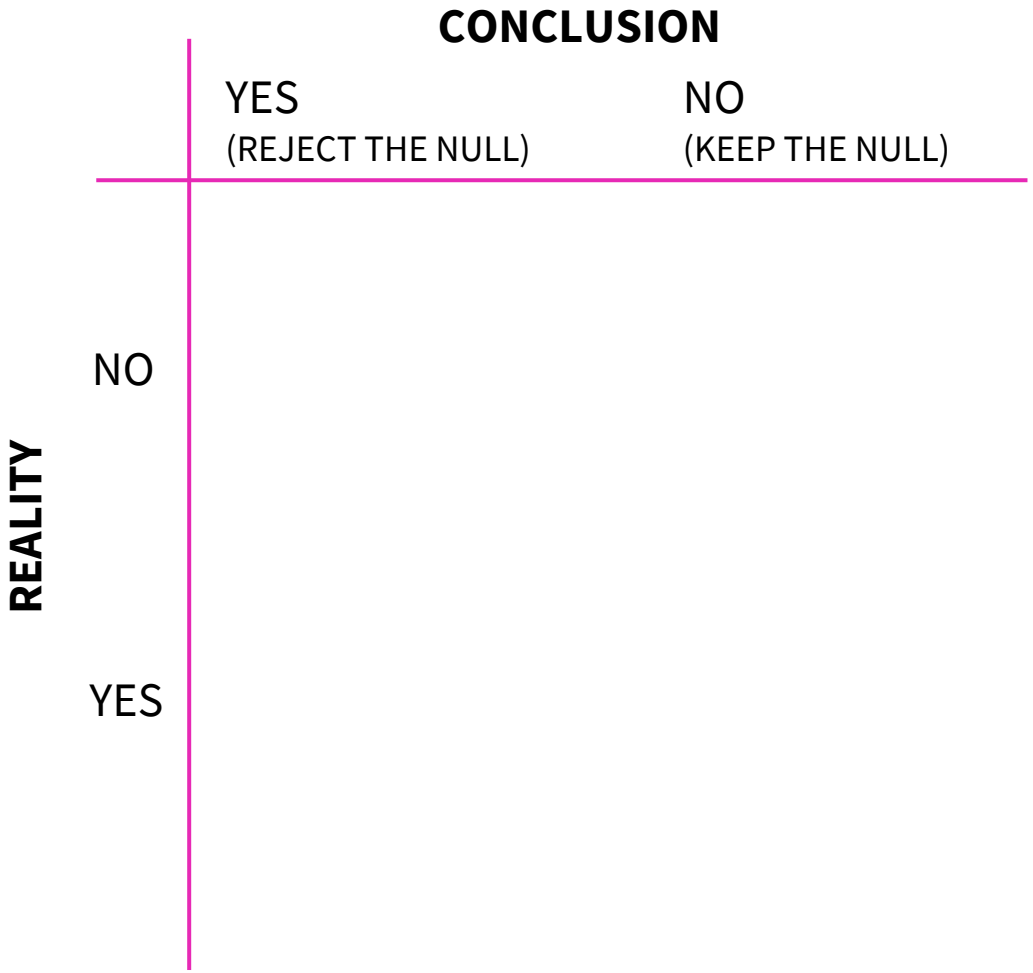
- **STATE THE QUESTION:**
  - IS THERE AN EFFECT?
  - ARE TREATMENTS DIFFERENT?
  - IS MODEL A BETTER THAN B?

- **FORMULATE THE NULL HYPOTHESIS:**
  - THERE'S NO DIFERENCE.
  - BOTH TREATMENTS HAVE SAME EFFECT
  - BOTH MODELS PERFORM THE SAME
- **COLLECT DATA**
  - DESCRIBE, TRANSFORM
- **COMPARE/TEST**
- **MAKE A DECISION**
- **ADD EFFECTS SIZE AND REPORT**

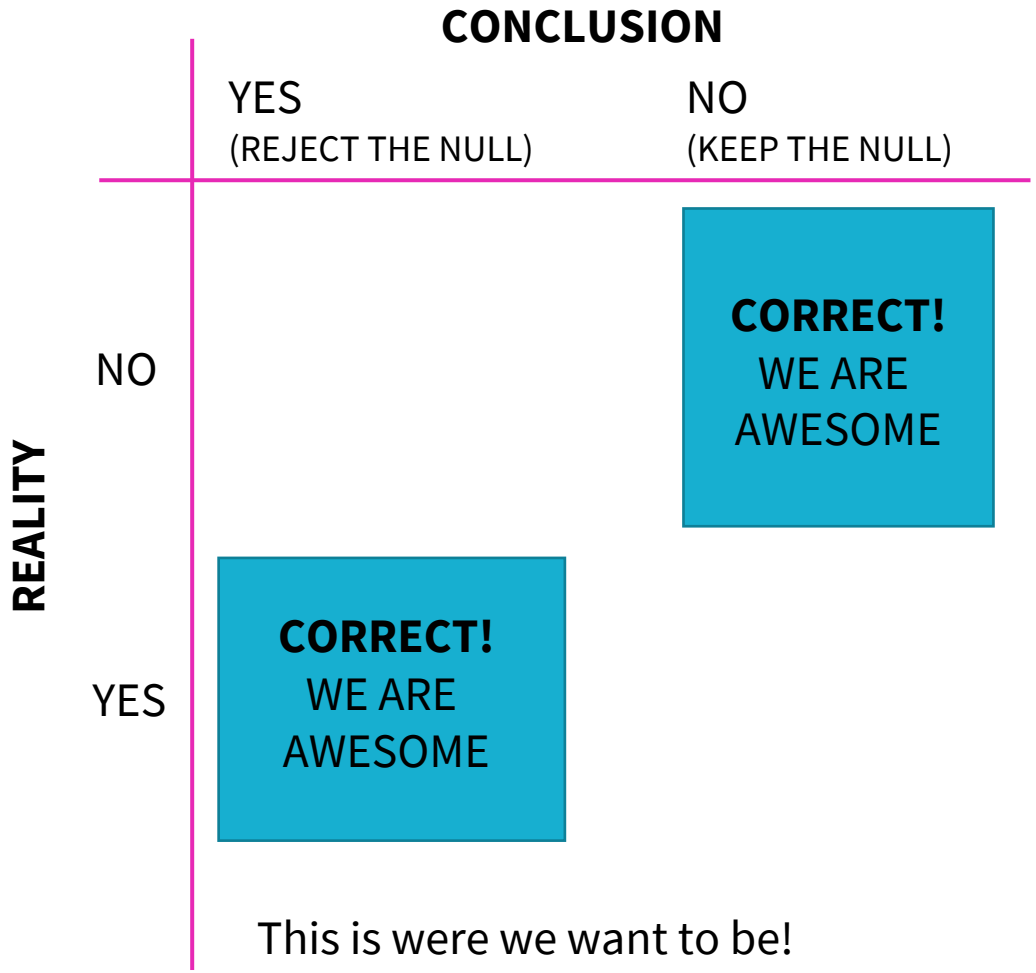
When are we using hypothesis tests in data science?

- *A/B tests*
- *Univariate outlier detection*
- *Establish the significance of a given observed pattern*
- *Model evaluation and monitoring*

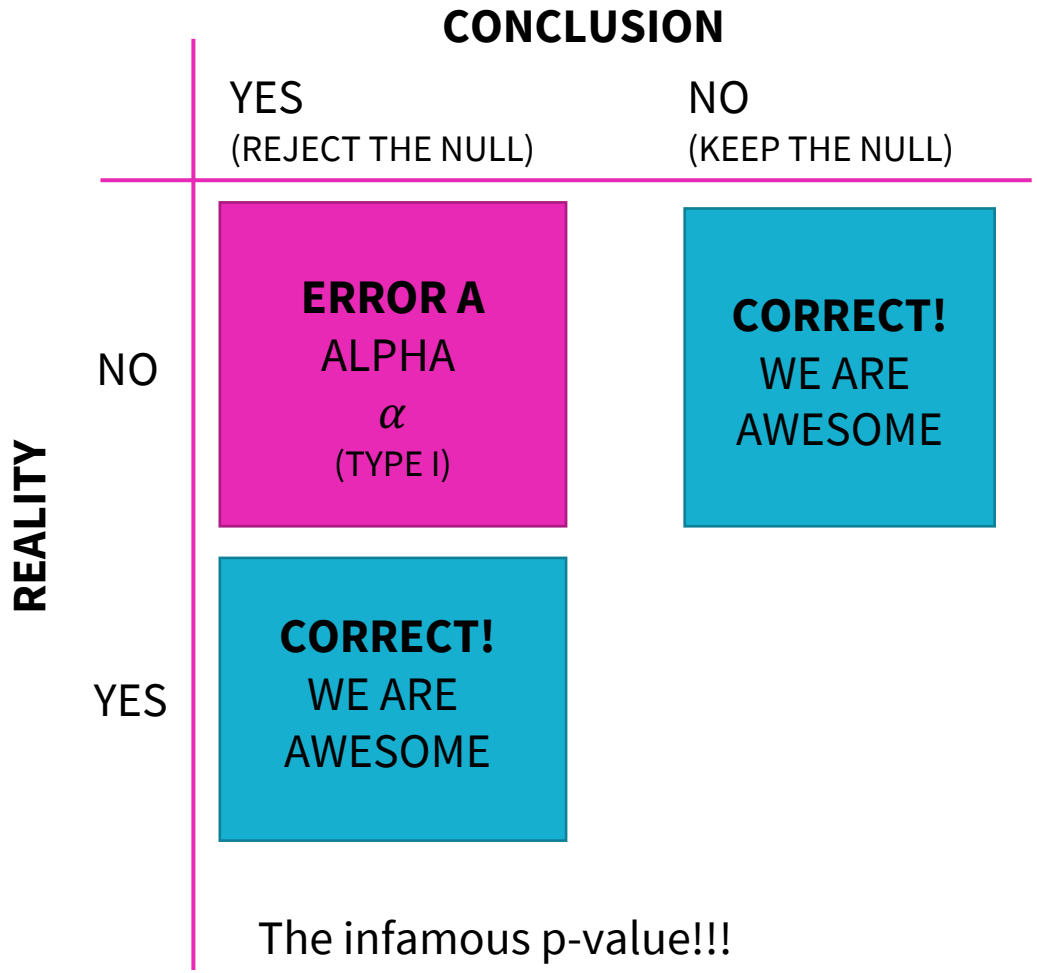
THE MOST IMPORTANT TABLE EVER



THE MOST IMPORTANT TABLE EVER

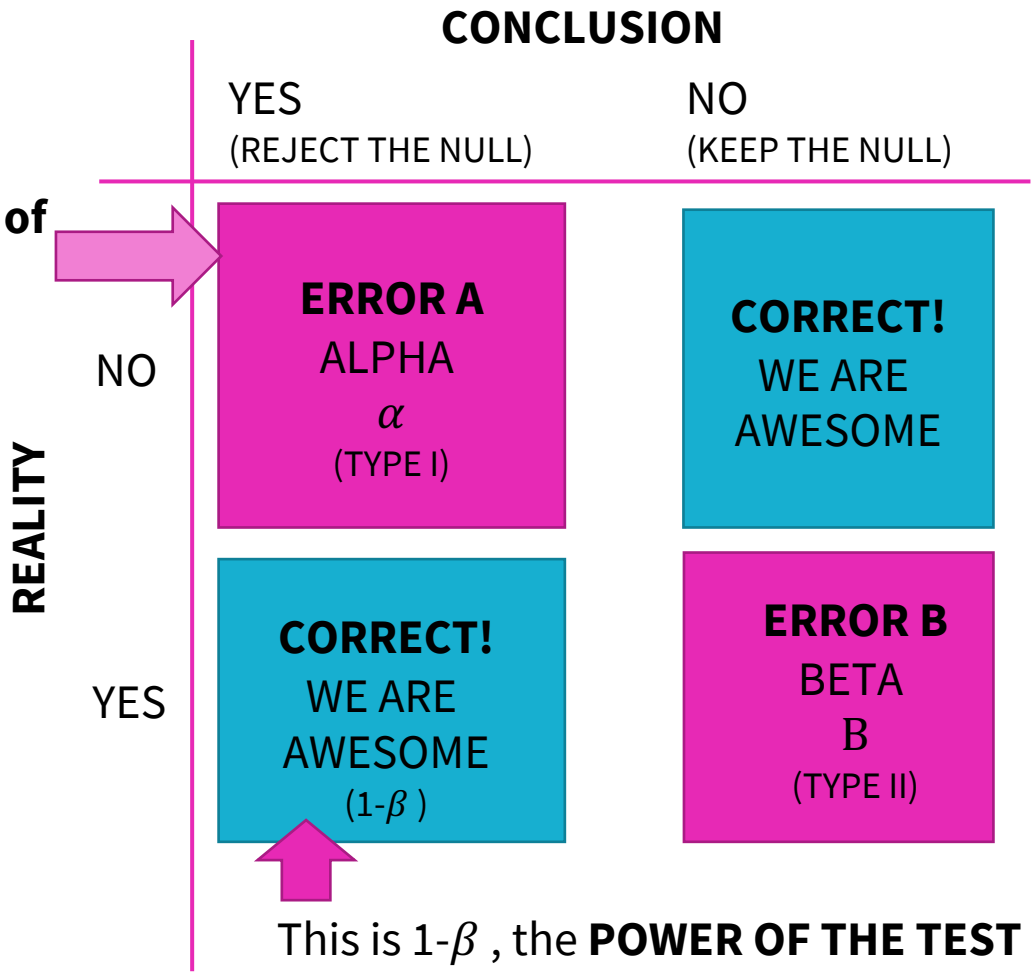


THE MOST IMPORTANT TABLE EVER



THE MOST IMPORTANT TABLE EVER

This is α, the **probability of making a TYPE 1 error**



This is 1-β , the **POWER OF THE TEST**
Probability of detecting an effect when it really exists

WHY ARE P-VALUES SO MESSY?

- Lack of statistical education and knowledge about the scientific method

- People tried to simplify the concept and then they twisted its meaning

- P(observation | reality) ***is not the same as*** P(reality | observation)

The p-value is a property of YOUR DATA,
not of the reality

Want to avoid the misuse of p-values? Just don't try to say it with your own words

Simplest hypothesis test: T-Test

If
The effect we are testing is the difference between means of 2 groups

Then
The null hypothesis is…

And
The alternative hypothesis is…

Simplest hypothesis test: T-Test (for 2 independent samples)

The effect we are testing is the difference between means of 2 groups, A and B
The null hypothesis is that the difference is 0
The alternative hypothesis is that the difference is not zero

1. One-side hypothesis: the difference is GREATER than 0 (or smaller)

What if we have more than 2 groups?

- Analysis of Variance (least-squares)
- Not a comparison of means anymore
- How many times is the variance **between** groups larger than the variance **within** groups?

Notebook exercise 4

EFFECT SIZES AND POWER TESTS

JAN-21

Effect sizes and Power tests

You got a p-value < 0.05. So what?

Results can be statistically significant but **trivial. Would you spend 200 work hours to implement a model that is 1% more accurate?**

We should always report effect sizes, in addition to any stats tests

Effect sizes that you probably already known:

- Correlation (r)
- R^2
- Odds-ratio

COMMON EFFECT SIZES

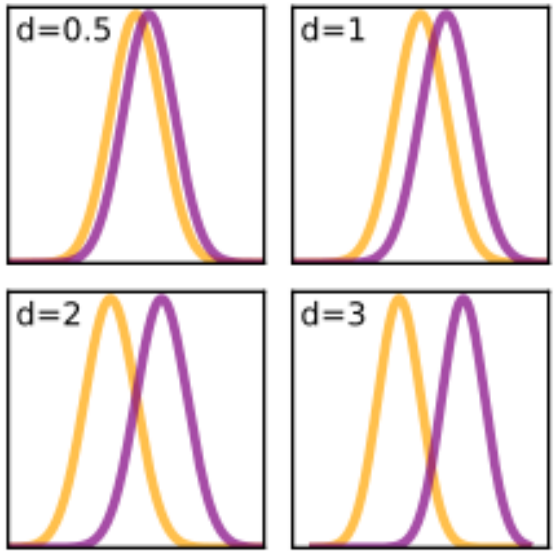
STRENGHT OF A RELATIONSHIP: **CORRELATION** COEFFICIENTS

EXPLANABILITY OF A GIVEN MODEL: **R^2**

Cohen's D (Standardized EFFECT SIZE)

It's a standardized coeficient:

- The diference between the means relative to the pooled standard deviation



- Convention:

Small Effect Size: d=0.20
Medium Effect Size: d=0.50
Large Effect Size: d=0.80

> After I learned what noncentral distributions were and figured out that it was important to decompose noncentrality parameters into their constituents of effect size and sample size, I realized that I had a framework for hypothesis testing that had four parameters: the alpha significance criterion, the sample size, the population effect size, and the power of the test. For any statistical test, any one of these was a function of the other three.

Cohen, 1990. Things I Have Learned (So Far)

Aditional reading:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/

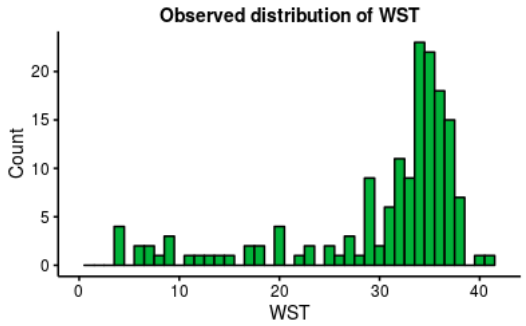
Notebook exercise 5

RESAMPLING AND BOOTSTRAPPING

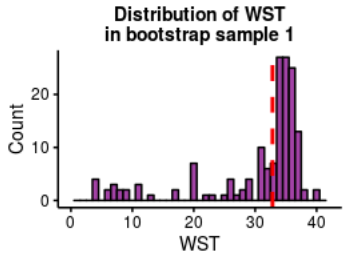
- It’s another way to estimate a confidence interval for an estimate
- It’s a more robust alternative when the data is not normally distributed and can’t be transformed
- It’s relatively “new” because it’s only possible with computers
- When are they used in ML?
  - Coefficient estimation with confidence intervals
  - K-fold cross-validation (resampling)
  - Random forests and boosted trees

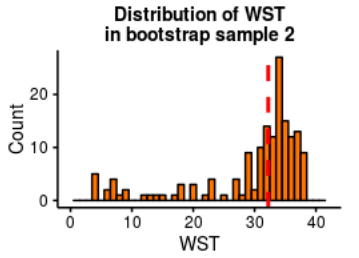
RESAMPLING AND BOOTSTRAPPING

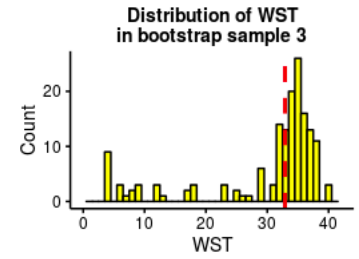
Original observations
Number of students and their scores



Bootstrap





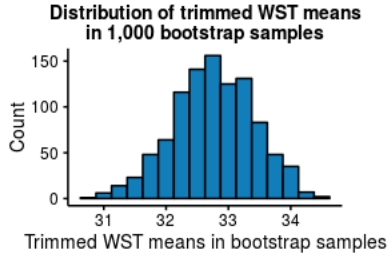


Logic:

For n in n_iterations:

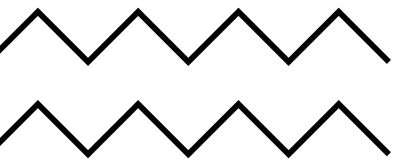
1. sample with replacement
2. calculate the mean
3. add the mean to a list

After n_iterations, you'll have a populations of means, which will be normally distributed, allowing you to calculate confidence intervals…

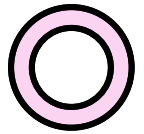


Notebook exercise 6

Source: https://janhove.github.io/teaching/2016/12/20/bootstrapping

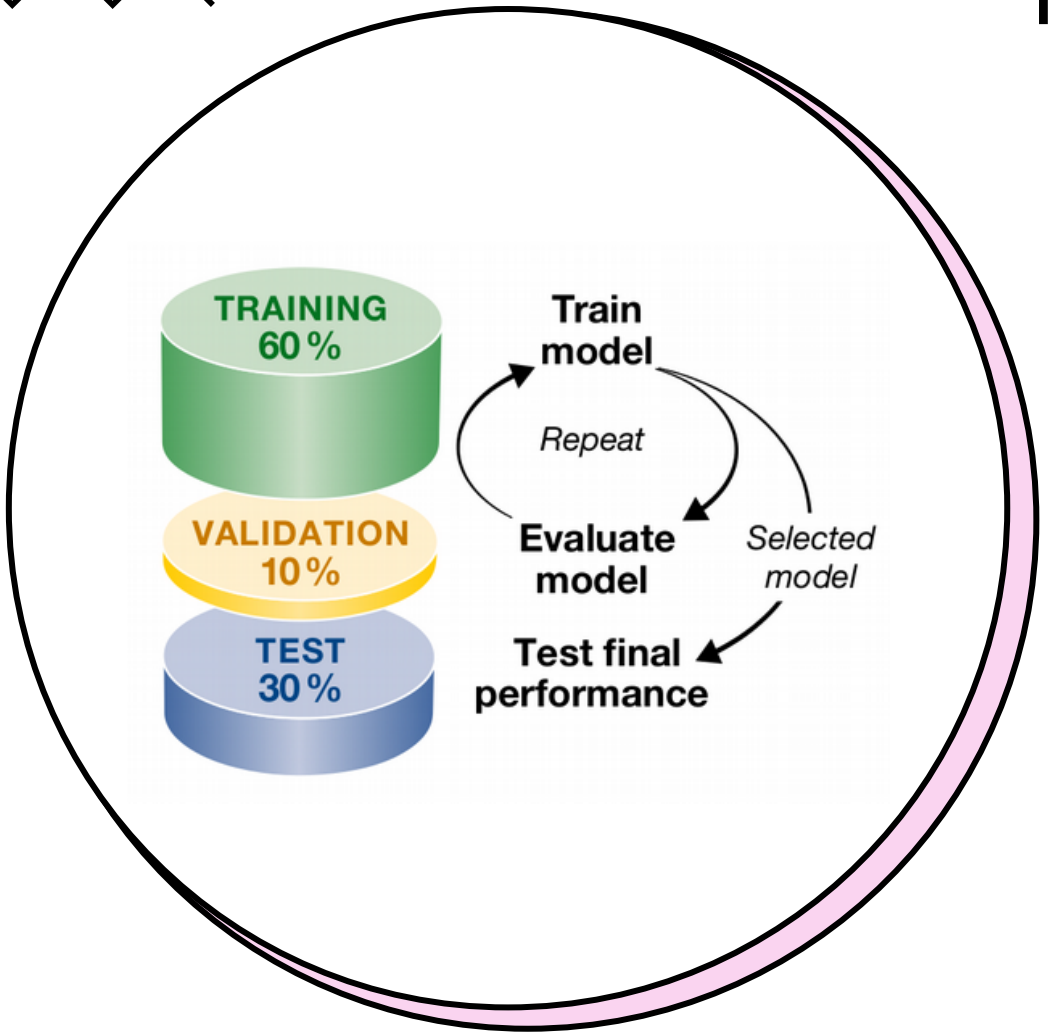


MODEL EVALUATION



JAN-21

MODEL EVALUATION



- TRAIN-TEST-VALIDATION SPLIT
- MODEL ERROR METRICS
- COMPARISON OF METRICS
- PROS AND CONS OF EACH METRIC
- USE IN PRACTICE

- BASIC STATISTICAL KNOWLEDGE IS CRUCIAL

MODEL EVALUATION IS A BIG PART OF A DATA SCIENTIST'S JOB

- Compare models with different hyperparameters (same model and same data)

But we also need to **design experiments**:

- Compare the model's performance across clusters of data *(my model predicts well the sales of perfum but not of make-up…)*
- Compare different train-test splits (same model and hyperparameters) *(this is crucial for unbalanced classification datasets and for time-series predictions)*
- Compare different algorithms *(same data but XGBoost ot Catboost?)*
- Compare the same model with diferent data *(e.g. feature selection)*

In the POC stage of the development of your data product, you'll very likely have to do ALL these evaluations

ERROR METRICS

In general they indicate the **model's performance**
They are specific for the type of problem that you have:

https://scikit-learn.org/stable/modules/model_evaluation.html

CLASSIFICATION	REGRESSION	TIME-SERIES
F-SCORE	MAE	MASE
AUC	MSE	
ACCURACY	MAPE	
PRECISION/RECALL	RMSE	

ERROR METRICS

CLASSIFICATION
(based on the confusion matrix)

F-SCORE
ACCURACY
PRECISION/RECALL

CONFUSION MATRIX

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-score} = \frac{2 * precision * recall}{precision + recall}$$

ERROR METRICS

REGRESSION

MAE = simplest interpretation

MSE = sensitive to outliers
RMSE = sensitive to outliers but easier to interprate; most used in practice
MAPE = very used for business and analytics but has a lot of pitfalls…

- what do we do with zeros?
- asymetric measure

USE the Median if the distribution of your errors is very skewed

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of y*
$\bar{y}$ – *mean value of y*

Likelihood, AIC, BIC

The simplest answer is often the right one.

Occam's Razor

AIC and BIC express the distance between the (unknown) true likelihood function of the data and the fitted likelihood function (i. e. the model's performance)

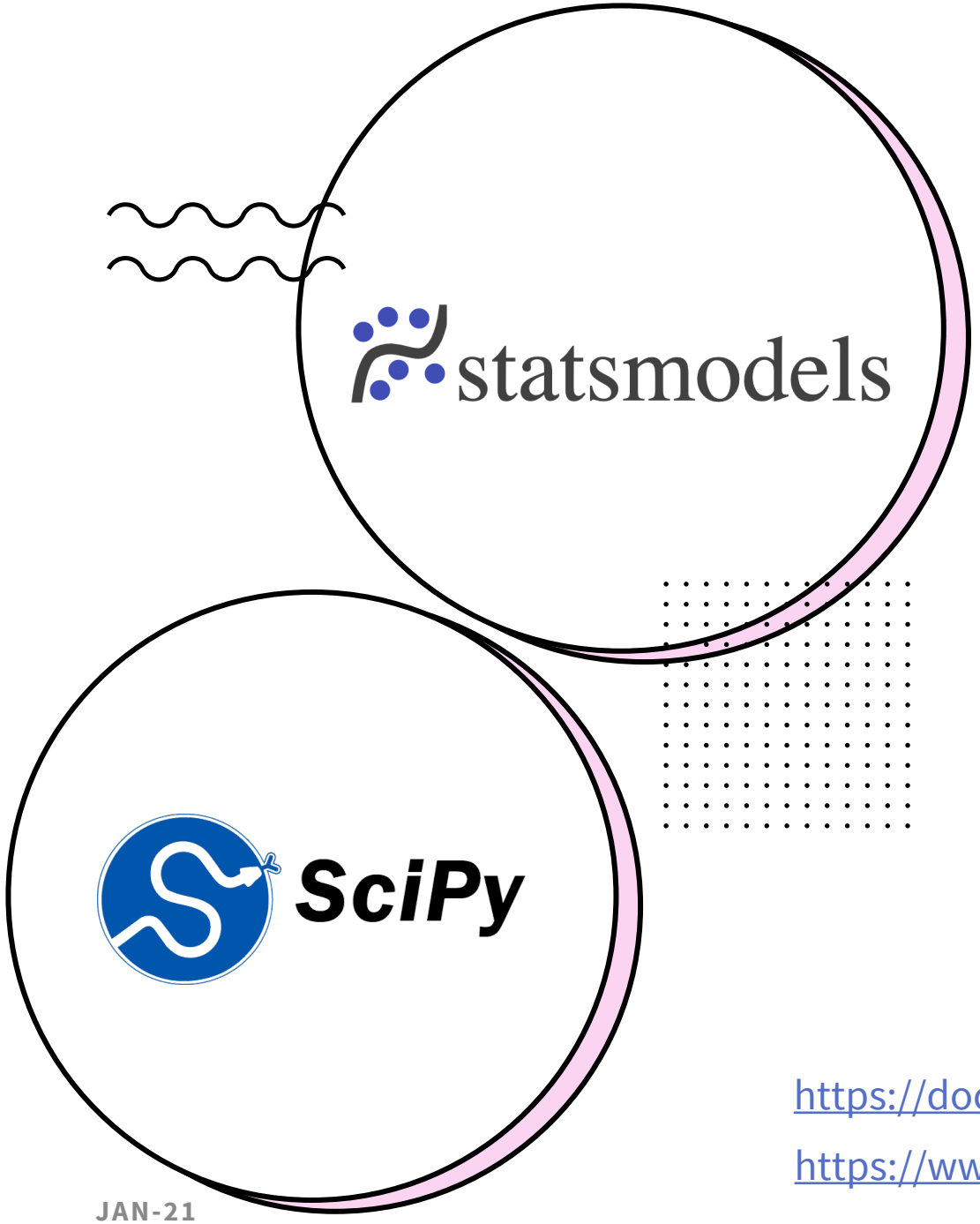
Because AIC and BIC represent a distance, the lower the better!

- BIC penalizes model complexity more heavily;
- If keeping the model simple is important for your case, choose base on **BIC**

$$AIC=-2*lnL+2*k$$
$$BIC=-2*lnL+2*lnN*k$$

where L is the value of the likelihood,
N is the number of obs,
and k is the number of estimated parameters.

Notebook exercise 7



Scipy and Statsmodels

Scipy and Statsmodels are your tools for statistics in Python. Sklearn also has a lot of stuff

- Scipy is more user-friendly
- Statsmodels is easier if you come from R and are used to its formula-like modelling (I like their summary reports)
- You'll need Statsmodels for general and generalized linear models

https://docs.scipy.org/doc/scipy/reference/stats.html

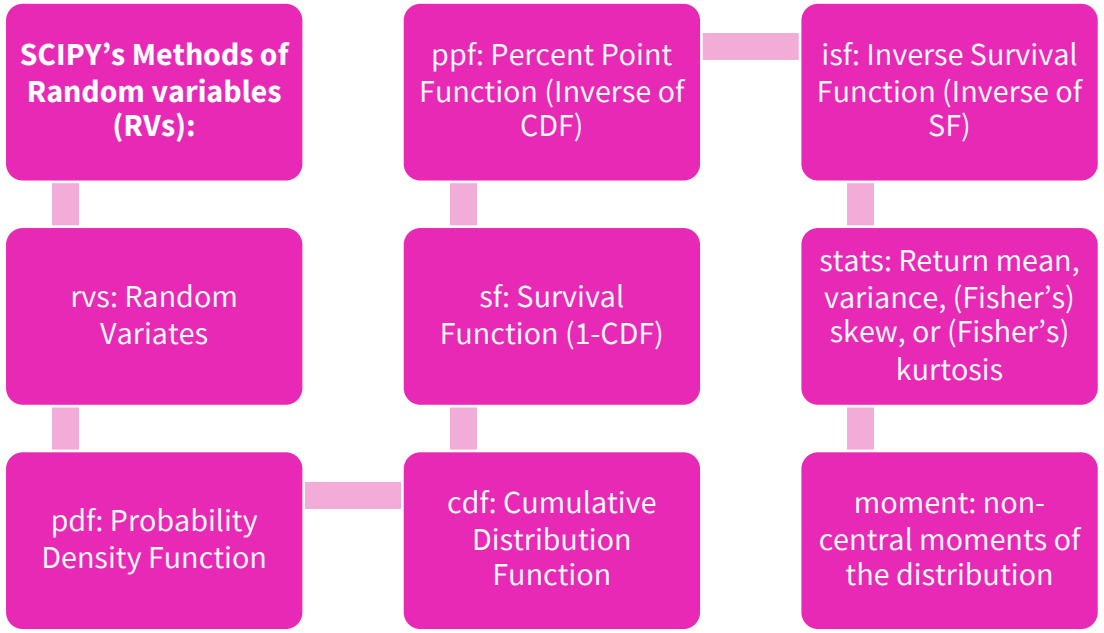
https://www.statsmodels.org/devel/gettingstarted.html

A little dictionary:
(to help using Scipy and Statsmodels)

Statsmodels:

Exogenous variables = x;
a.k.a. features, predictors, independent variables

Endogenous variables = y;
a.k.a. target, response, dependant variable;



INTERPRETATION OF A LINEAR REGRESSION IN STATSMODELS:

```
In [22]: res.summary()
Out[22]:
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 TOTEMP   R-squared (uncentered):                   1.000
Model:                            OLS   Adj. R-squared (uncentered):              1.000
Method:                 Least Squares   F-statistic:                          5.052e+04
Date:                Wed, 15 Jul 2020   Prob (F-statistic):                    8.20e-22
Time:                        12:58:48   Log-Likelihood:                         -117.56
No. Observations:                  16   AIC:                                      247.1
Df Residuals:                      10   BIC:                                      251.8
Df Model:                           6
Covariance Type:            nonrobust
------------------------------------------------------------------------------
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
GNPDEFL      -52.9936    129.545     -0.409      0.691    -341.638     235.650
GNP            0.0711      0.030      2.356      0.040       0.004       0.138
UNEMP         -0.4235      0.418     -1.014      0.335      -1.354       0.507
ARMED         -0.5726      0.279     -2.052      0.067      -1.194       0.049
POP           -0.4142      0.321     -1.289      0.226      -1.130       0.302
YEAR          48.4179     17.689      2.737      0.021       9.003      87.832
==============================================================================
Omnibus:                        1.443   Durbin-Watson:                   1.277
Prob(Omnibus):                  0.486   Jarque-Bera (JB):                0.605
Skew:                           0.476   Prob(JB):                        0.739
Kurtosis:                       3.031   Cond. No.                     4.56e+05
==============================================================================
```

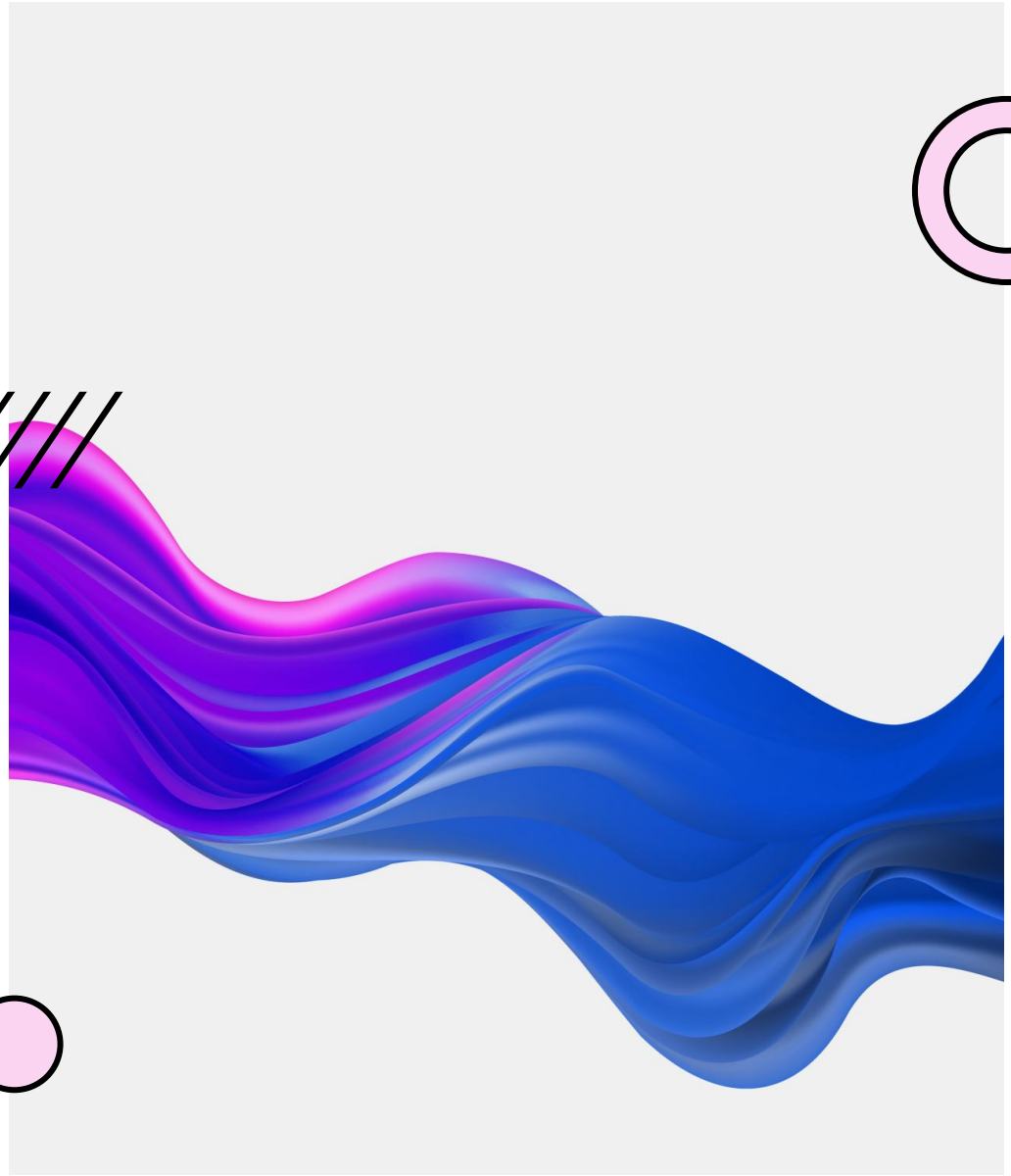
General Information about the model

Information about the model's features

Description of the data

THANKS

QUESTIONS? REACH OUT TO ME ANYTIME



Central Limit Theorem
given a sufficiently large sample size from a population with a finite level of variance, **the mean of all samples from the same population will be approximately equal to the mean of the population**. Furthermore, all of the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size.