Good morning ☺️
Notebooks and slides are in:

https://github.com/AlineQuadros/teaching_data_science

Please install these packages:

scipy
statsmodels

STATISTICS
FOR
DATA SCIENCE

DR. ALINE QUADROS
DATA SCIENCE RETREAT

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

DISCLAIMER

SORRY, I'M A FREQUENTIST

CONTENTS

DESCRIPTIVE STATISTICS

- MEAN, MEDIAN, MODE
- VARIANCE, COVARIANCE
- COEFFICIENT OF VARIATION
- Standard DEVIATION
- SUM SQUARES, MSE and RMSE
- Skewness, kurtosis, long-tails

INFERENTIAL STATISTICS

- Standard ERROR
- CONFIDENCE INTERVALS
- CORRELATION

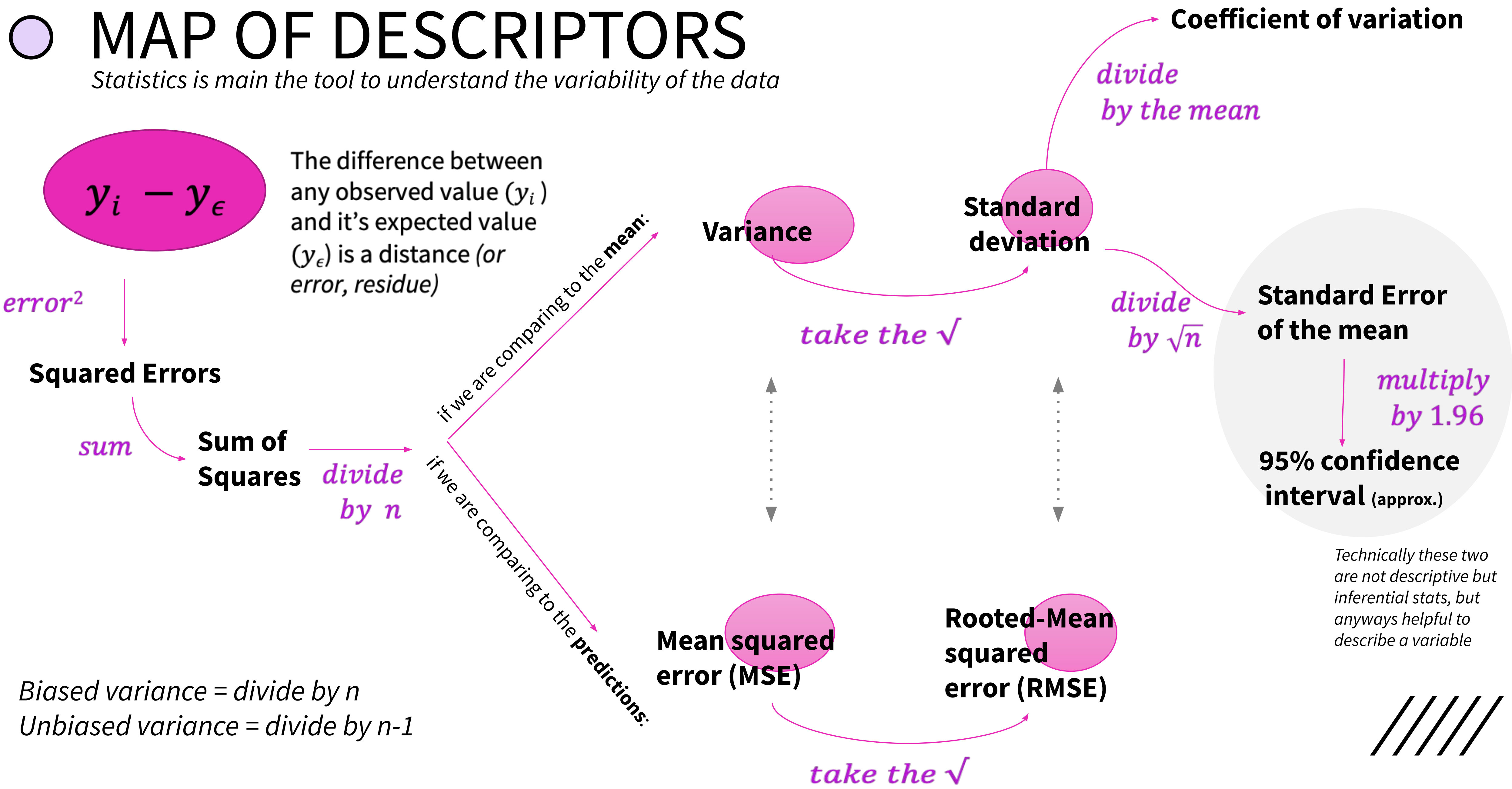
Descriptive statistics and dataviz are the key components of a good POC

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 are we trying to predict?
  - What kind of underlying process generates the data we are trying to model?

MAP OF DESCRIPTORS

Statistics is main the tool to understand the variability of the data



You should to be able to answer these questions:

- When should we use the mean to describe a variable/feature?
- When should we not use the mean? What can we use instead?

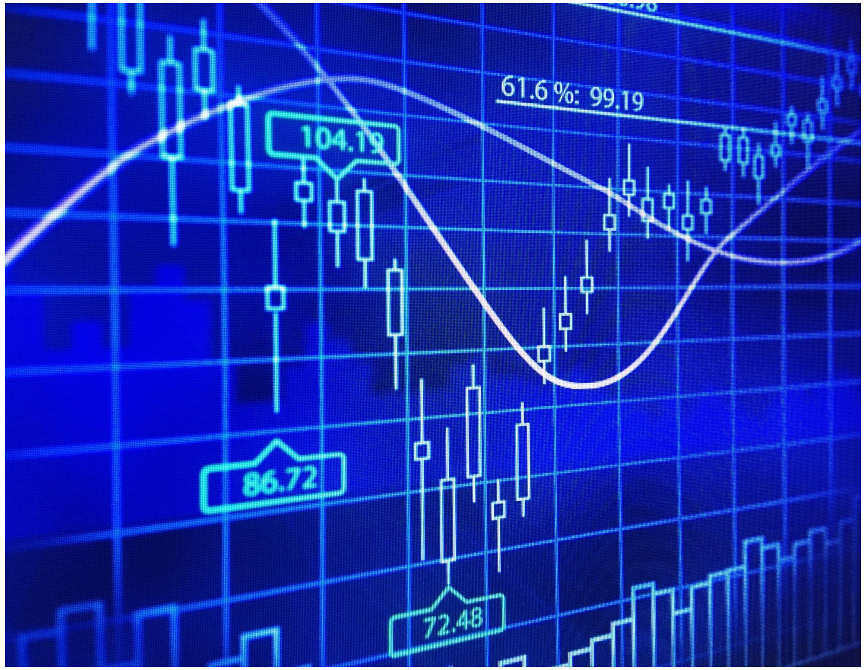
- What is the difference between standard error and standard deviation?

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

PARAMETRIC

VS.

NON-PARAMETRIC



PARAMETRIC statistics

- Based on the parameters of a given probability distribution
  - E. g. : the Gaussian distribution has 2 parameters, the mean and the standard deviation
  - 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.

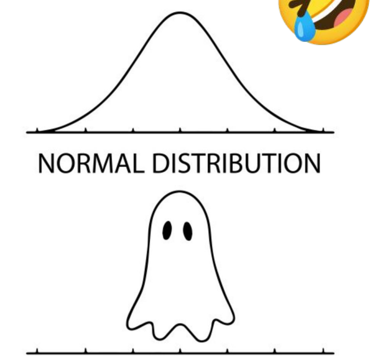
PARAMETRIC MODELS

- 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)



We can use their **parameters** describe the behavior of our variables.

Each distribution has one (or more) **parameters**

- What is the parameter that describes the normal distribution?

In regression problems/linear modelling, we must pay special attention to the distribution of our response variable.

Type of random variable	Constrains	Probability distributions
Discrete	Data is binary	Bernoulli and binomial
	Data is not overdispersed	Poisson
	Data is overdispersed	Negative Binomial
Continuous		Gaussian
	Only positive	Gamma

POISSON

NUMBER OF TIMES AN EVENT IS OBSERVED IN TIME OR SPACE

It is useful to model <u>uncommon events</u>

Examples:

Number of sales per hour

Number of complaints per customer

Number of goals per soccer match

Number of typos per book page

Lambda = the event rate; positive number.

Important property of Poisson: mean== variance

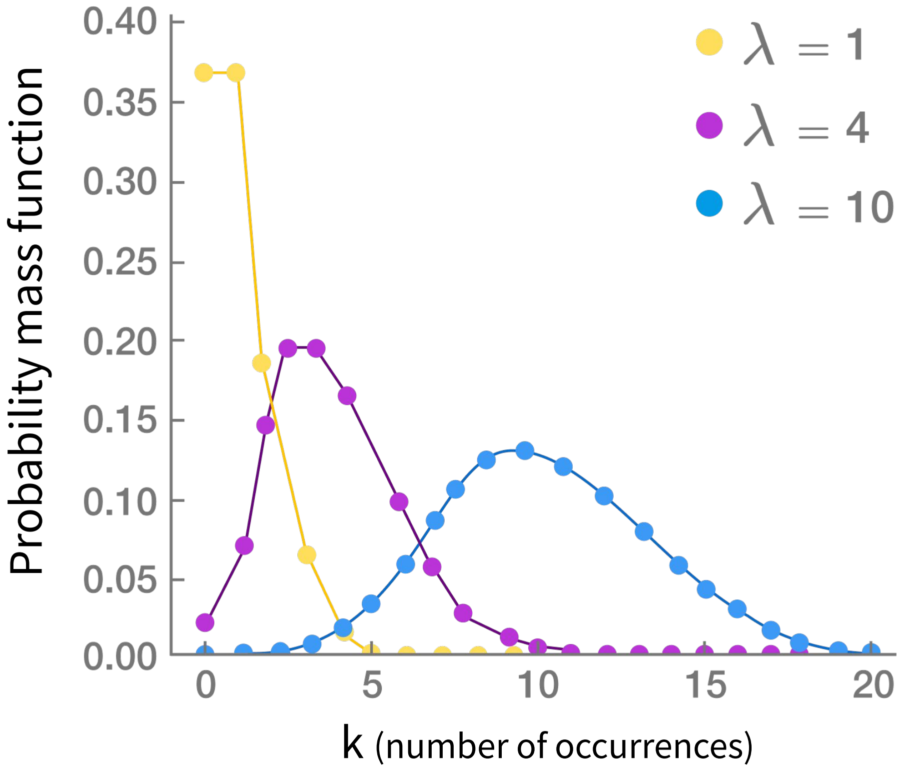
POISSON assumes that the data is NOT overdispersed.

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

The probability mass function for poisson is:

$$f(k) = \exp(-\mu)\frac{\mu^k}{k!}$$

for $k \geq 0$.



NORMAL (GAUSSIAN)

It's widely used because of interesting properties:

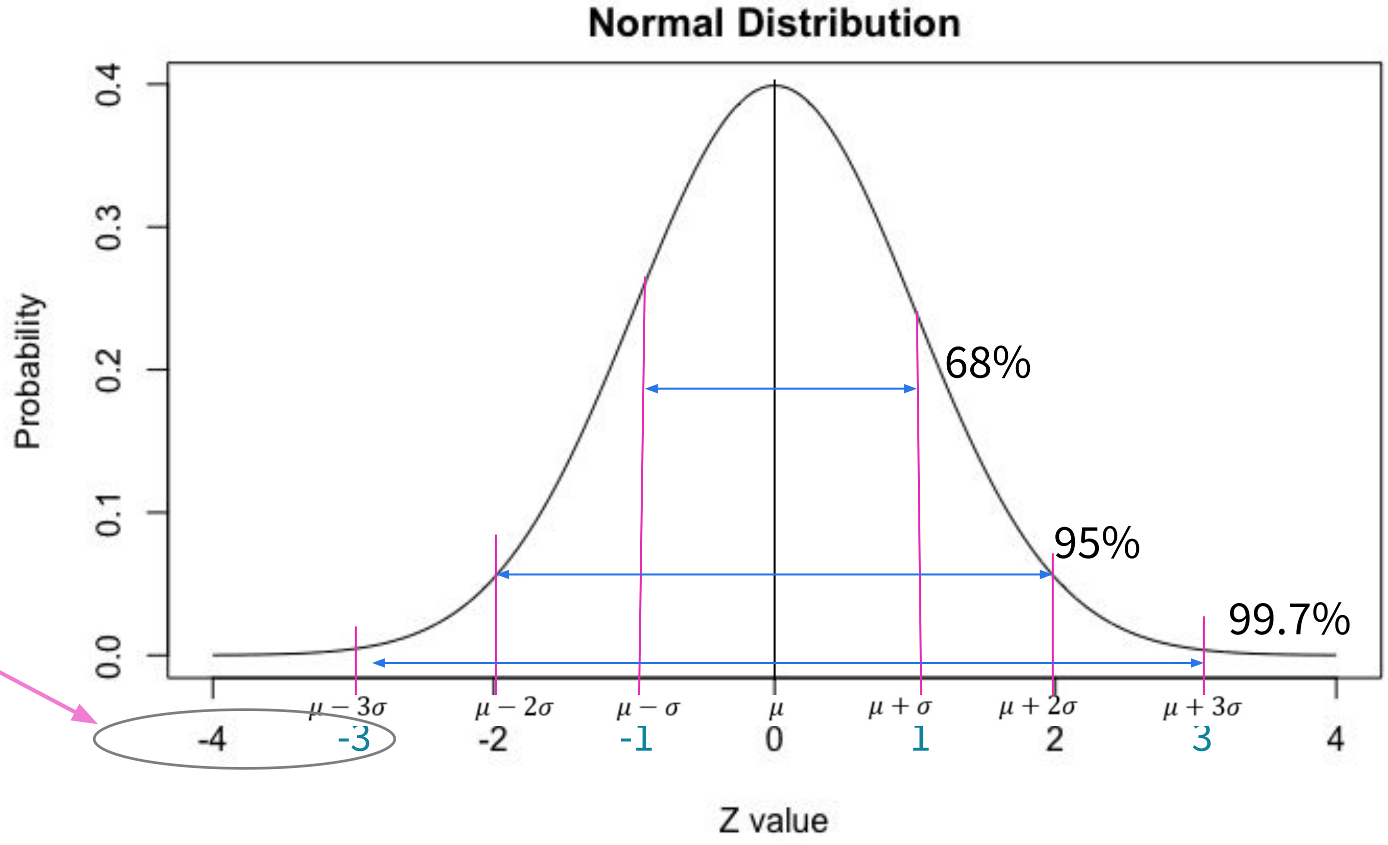
Properties:

- Symmetry
- Unimodal
- Mean = median = mode

These are standardized scores, because raw data can be in different scales. The formula is:

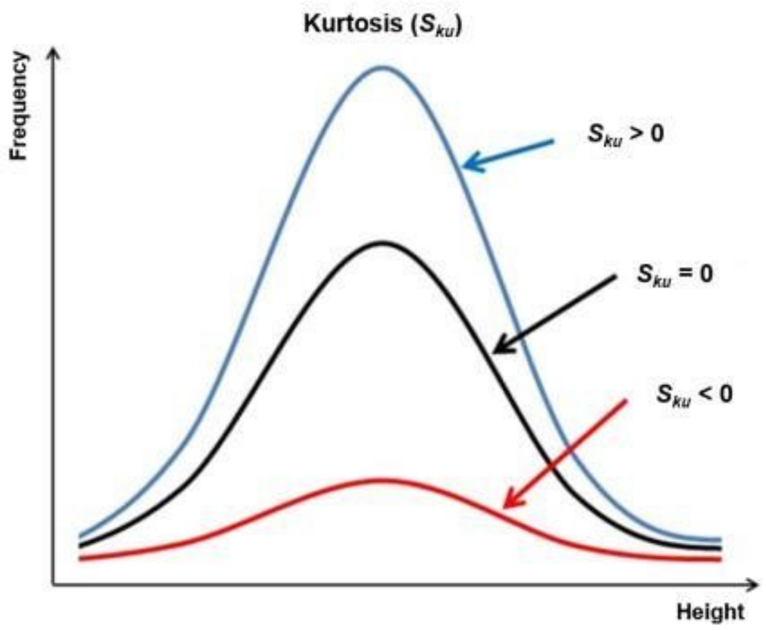
$$z = (x-\mu)/\sigma$$

Where z is the z-score we want to find, x is a data point, μ is it's mean and σ is the standard deviation

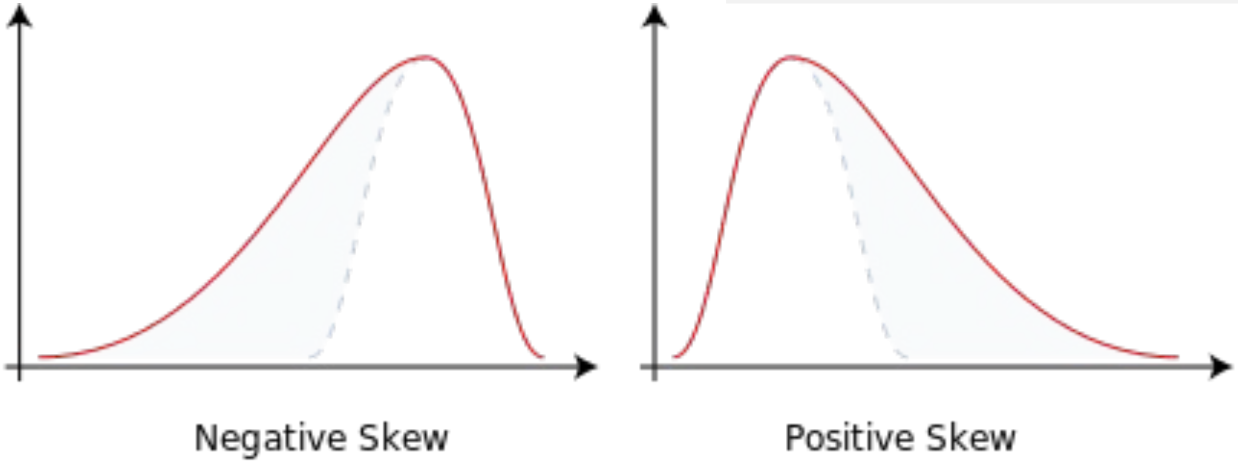


More properties of Gaussian distribution:
Zero skewness and zero excess kurtosis

Kurtosis:
How much "tail" in the data?



Skewness:
Where is the "tail"?



STATISTICAL INFERENCE

WHAT IS IT?

WE WANT TO **INFER** SOMETHING ABOUT THE POPULATION BY ANALYSING SAMPLES 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 DIFFERENCE.*
  - *BOTH TREATMENTS HAVE SAME EFFECT*
  - *BOTH MODELS HAVE SAME PERFORMANCE*
- **COLLECT DATA**
  - DESCRIBE, TRANSFORM
- **CHOOSE and APPLY an INFERENCE TEST**
- **MAKE A DECISION**
- **ADD EFFECTS SIZE AND MAKE YOUR 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

Dealing with paranormal 👻 Non-gaussian data

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

- 1. Look for models that don't assume 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,** library `statsmodels` is very good for that)
- 3. For a simple hypothesis test, use bootstrapping to generate to generate the null model
- 4. Remove outliers and 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 and apply to the data. Don't forget to mention the transformation used in any documentation or report

Type of Transformation	Common Applications
Log	Data with very different 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

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

THE MOST IMPORTANT TABLE EVER

		CONCLUSION	
		YES (REJECT THE NULL)	NO (KEEP THE NULL)
REALITY	NO		
	YES		

THE MOST IMPORTANT TABLE EVER

		CONCLUSION	
		YES (REJECT THE NULL)	NO (KEEP THE NULL)
REALITY	NO		**CORRECT!** WE ARE AWESOME *true negative*
	YES	**CORRECT!** WE ARE AWESOME *true positive*	

This is where we want to be!

THE MOST IMPORTANT TABLE EVER

The infamous p-value!!!

REALITY \ CONCLUSION	YES (REJECT THE NULL)	NO (KEEP THE NULL)
NO	**ERROR A** ALPHA α (TYPE I) *false positive*	**CORRECT!** WE ARE AWESOME *true negative*
YES	**CORRECT!** WE ARE AWESOME *true positive*	

THE MOST IMPORTANT TABLE EVER

This is α, or p-value
The **probability of making a TYPE 1 error**

		CONCLUSION	
		YES (REJECT THE NULL)	NO (KEEP THE NULL)
REALITY	NO	**ERROR A** ALPHA α (TYPE I) *false positive*	**CORRECT!** WE ARE AWESOME *true negative*
	YES	**CORRECT!** WE ARE AWESOME (1-β) *true positive*	**ERROR B** BETA B (TYPE II) *false negative*

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

A T-test compares the means of 2 independent samples

The effect we are testing is the difference between the 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, i.e. it's greater or lesser than

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

LIMITATIONS

What if we have more than 2 groups?

- **Analysis of Variance** (least-squares)
- Not a comparison of means anymore
- Tests based on **Analysis of Variance:** How many times is the variance **between** groups larger than the variance **within** groups? (https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/)

Writing your own T-Test

(example for 2 independent samples)

The values in this table are known as “T critical”. They represent thresholds under the assumption that the null hypothesis is TRUE

Calculate the t statistic for your sample:

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

Calculate the t-critical for your sample size and alpha:

What is the t-critical for a sample size of 10, and alpha of 95%?

```
stats.t.ppf(0.95, df=10)
```

```
1.8124611228107335
```

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	$\boxed{1.812}$	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300

Applications of inference tests: A/B Tests

A/B test is the name given to an experiment where a company or a team wants to compare **2 alternatives** or **before-after** situations (hence the name A/B test) **in a controlled, structured way** (like in an experiment)

Business use-cases:

- Compare 2 versions of a website to see each one leads to a higher conversion rate
- Compare bounce rates before and after increasing the speed of the homepage
- Compare 2 strategies to increase sales *(1 or 2 mini-samples example)*
- Check more resources on that in *https://vwo.com/blog/*

ML use-cases:

- Compare the performance of 2 or more models

Applications of inference tests: A/B Tests

A/B test logic

Should we send 1 or 2 free mini-samples with every order?

Implementation:
- Formulate the null hypothesis:
  - There's no differences between the purchases of customer who receive 1 or 2 mini-samples
- Formulate the alternative hypothesis:
  - Customers who receive 2 mini-samples purchase more (this is a one-tail test)
- Define target groups (gender, age), sample size, duration of experiment
  - Example:
    - Send 1 mini-sample to 100 customers (female; age 20 to 40)
    - Send 2 mini-samples to another 100 customers
    - Check how many products each customer buys in the next 3 months
- Apply statistics: T-test comparing group 1 and 2
- Check the results: statistical significance, power of test, effect sizes (are the differences meaningful)?
- Give your recomendation: Should we send 1 or 2 mini-samples?

Applications of inference: model selection

Example of an experiment setup to understand if adding certain features significantly decreased the model error

the error metric

	Mean Absolute Error (Summer)			
	Control (current 10 features)	**All features (group 1 and 2)**	**Only group 1 (Weather effect)**	**Only group 2**
Category A	2.03	1.95*	1.85*	2.0
Category A	2.10	2.11	1.96*	2.16*worst than control
Category A	0.80	0.77	0.70*	0.79
Category A	1.58	1.32*	1.25*	1.33*
Category A	1.28	1.19*	1.08*	1.17*

These are the 4 levels of the experiment (control, all, only 1 and only 2)

*The * indicates that the error is significantly lower than in the control*

Highlighted numbers indicate likely source of effect

EFFECT SIZES AND POWER TESTS

Effect sizes and Power tests

Your test shows 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?**

Effect sizes give another perspective to help make a decision

We should always report effect sizes, in addition to any statistical test

COMMON EFFECT SIZES

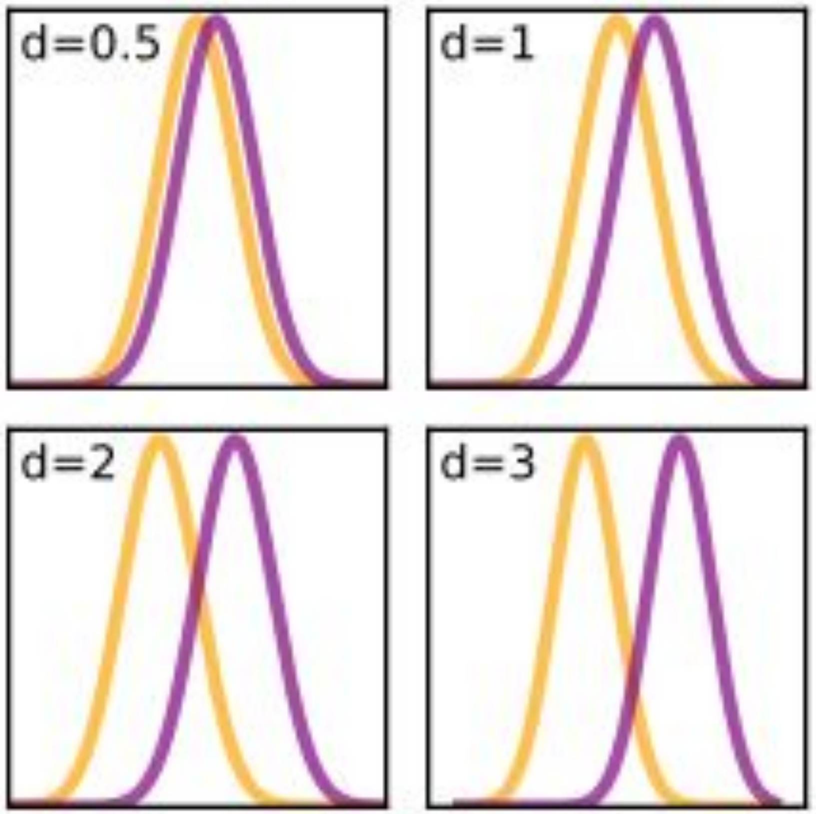
STRENGHT OF A RELATIONSHIP: **CORRELATION** COEFFICIENTS

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

Cohen’s D (Standardized EFFECT SIZE)

Cohen’s D is the most common measure of effect size:

- It is the difference 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

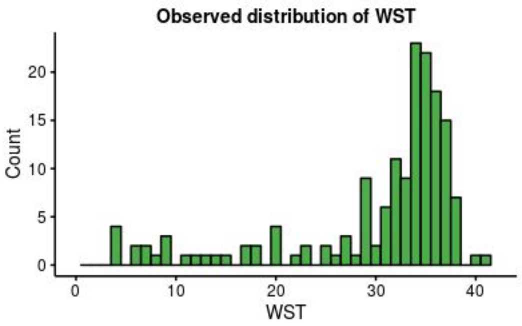
Additional reading:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/

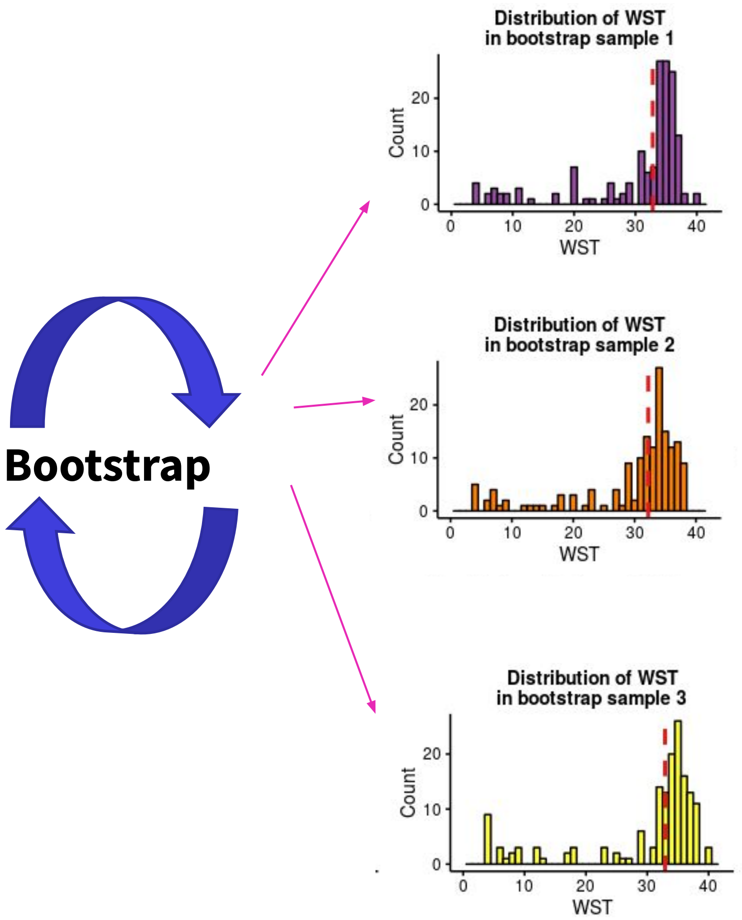
RESAMPLING AND BOOTSTRAPPING

- Resampling is the process of repeatedly taking samples from the same data
- Bootstrapping is a specific kind of resampling with replacement
  - In bootstrapping, the probabilities of each value don't change at each iteration, because the sampled examples are "returned" to the pool
  - 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

BOOTSTRAP

Original observations
Number of students and their scores



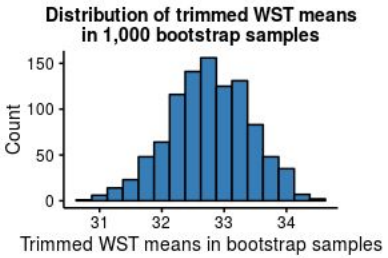


Logic:

For n in n_iterations:

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

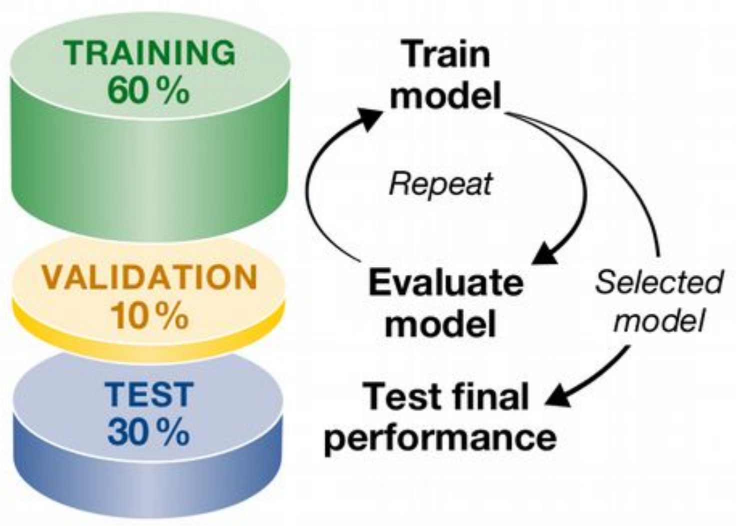
After n_iterations, you'll have a population of means, which will be normally distributed, allowing you to calculate confidence intervals…



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

MODEL EVALUATION

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 perfume 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 algorithm with different 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	
KAPPA		

ERROR METRICS in regression models

MAE = simplest interpretation

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

- what do we do with zeros?
- asymmetric measure

USE the median instead of the mean if the distribution of your errors is very skewed (Median APE)

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

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

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

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

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

AIC and BIC

Metrics of fit + complexity:

AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) express the distance between the (unknown) true likelihood function of the data and the fitted likelihood function (i. e. the model's performance), taking into account the amount of parameters used to obtain the fit

Both AIC and BIC represent a distance, thus, the smaller the better!

Which one should you use?
BIC penalizes model complexity more heavily; If keeping the model simple is important for your case, choose based on **BIC**

The simplest answer is often the right one.

Occam's Razor

$$AIC=-2*\ln L+2*k$$
$$BIC=-2*\ln L+2*\ln N*k$$

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

Example – model evaluation

Comparison of sales of shops with different assortments

```
X = X[[ 'Promo', 'type_a', 'type_c', 'type_d',
'assortment_a', 'assortment_c']]
y = sales.Sales

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)

reg = LinearRegression().fit(X_train, y_train)
y_pred = reg.predict(X_test)
```

Overall MAE in the test set = 1803

But when you look at data in detail and break it down, you'll see it varies from 1305 to 2213, depending of the store's features. This is helpful because it gives hints on what you can look to try improve the model.

We can also check if observed differences are statistically significant (with a T-Test). Alternatively you can compare all categories with a Anova test)

	Promo	type_a	type_c	type_d	assortment_a	assortment_c	Abs_error	
							mean	std
0	0	0	0	1	0	1	1438.16	1242.91
1	0	0	0	1	1	0	1305.88	1132.77
2	0	0	1	0	0	1	1668.18	1323.35
3	0	0	1	0	1	0	1480.64	1514.00
4	0	1	0	0	0	1	1654.64	1420.64
5	0	1	0	0	1	0	1671.97	1491.18
6	1	0	0	1	0	1	2130.43	1865.51
7	1	0	0	1	1	0	1760.98	1355.54
8	1	0	1	0	0	1	2176.41	1513.92
9	1	0	1	0	1	0	1646.28	1497.24
10	1	1	0	0	0	1	2176.87	1960.96
11	1	1	0	0	1	0	2213.95	1926.83

```
groupa = errors_df[(errors_df.assortment_a==1)&(errors_df.type_d==1)&(errors_df.Promo==0)].Abs_error
groupb = errors_df[(errors_df.assortment_a==1)&(errors_df.type_a==1)&(errors_df.Promo==1)].Abs_error

stats.ttest_ind(groupa, groupb, equal_var=False)
```

```
Ttest_indResult(statistic=-42.60189646790727, pvalue=0.0)
```

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

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

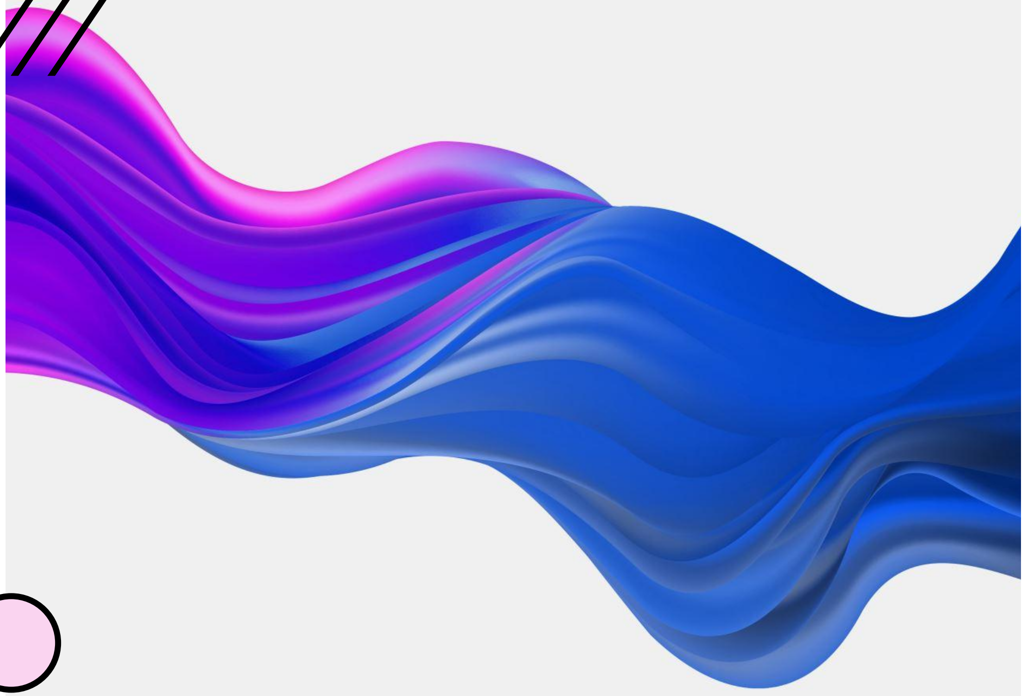
$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Information about the model's features
Hypothesis test for each coefficient
(T-test for one sample)

Description of the data

THANKS

QUESTIONS? REACH OUT TO ME ANYTIME



EXTRAS

Resources for job interviews:

- This is a very nice, curated list of job interview questions more on the theoretical side, and they include many of the concepts we discussed here: https://github.com/alexeygrigorev/data-science-interviews/blob/master/theory.md
- There's a section on questions here: https://end-to-end-machine-learning.teachable.com/p/navigating-a-data-science-career
- I really like some of the questions here: https://towardsdatascience.com/40-statistics-interview-problems-and-answers-for-data-scientists-6971a02b7eee
- Another cool list: https://www.nicksingh.com/posts/40-probability-statistics-data-science-interview-questions-asked-by-fang-wall-street
- Try the quiz (p.s. I do not agree with question #1):
  - http://interview-questions-247.appspot.com/data-science-probability-statistics-14
  - http://interview-questions-247.appspot.com/data-science-probability-statistics-17

Resources for job interviews:

Examples of statistical questions:

- **What are MSE and RMSE?**
- **What is a p-value? Where is it used?**
- **Which metrics for evaluating regression models do you know?**
- **How do we check if a variable is normally distributed?**
- **What is the normal distribution? Why do we care about it?**
- **What are the main assumptions of linear regression?**