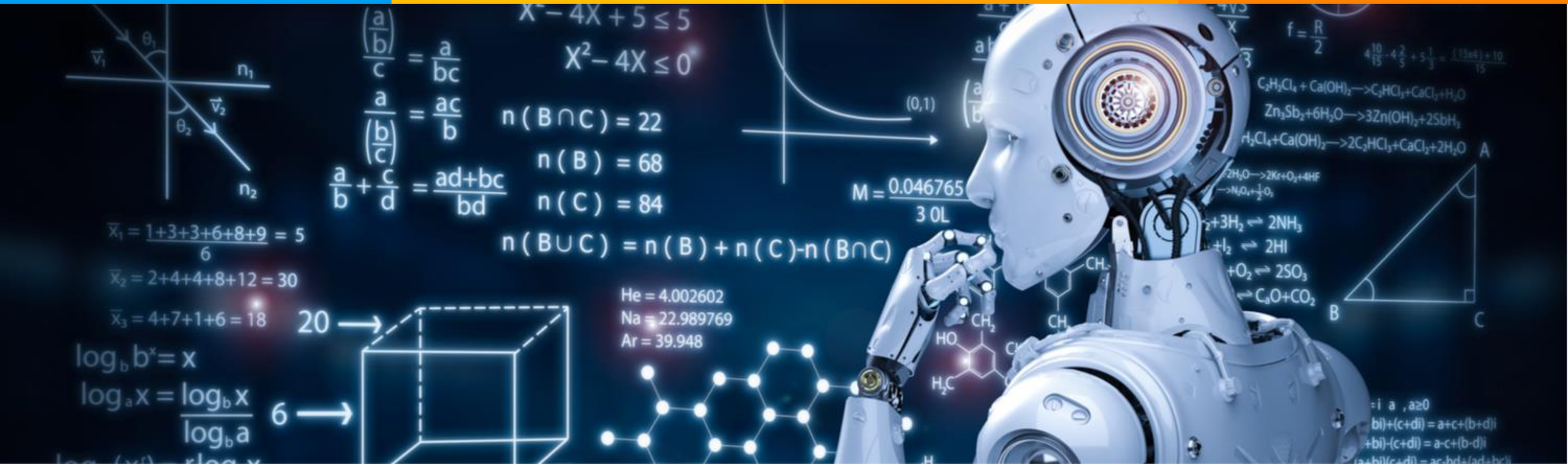
Duration: 120 min	Srikanth Shenoy	AML 5102	M.E. AI & ML

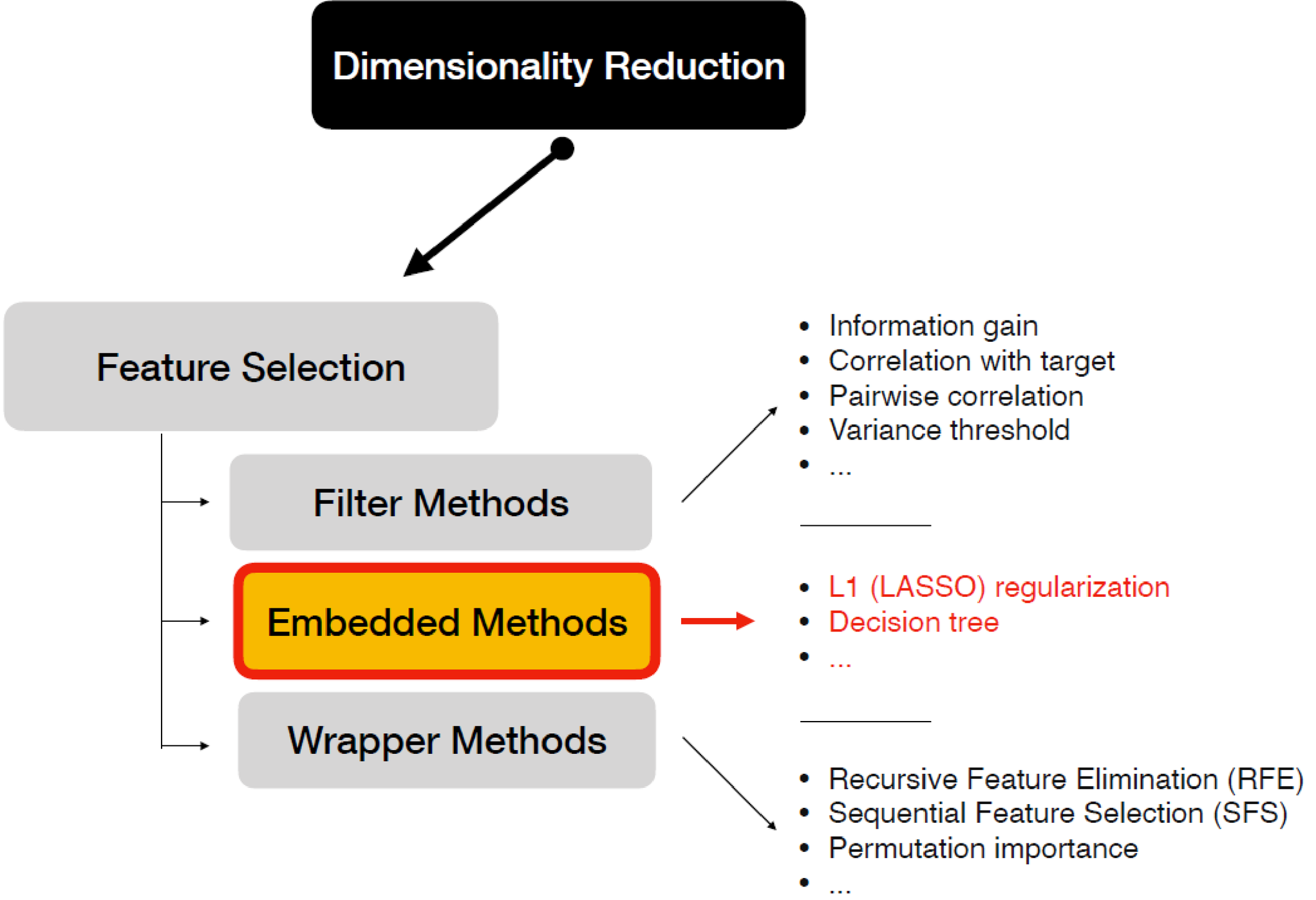


Lecture 28 & 29: Feature Selection Part 2

Recap

- Feature Selection
  - Unsupervised methods
  - Filter and wrapper methods

Feature Selection with Embedded (Intrinsic) methods

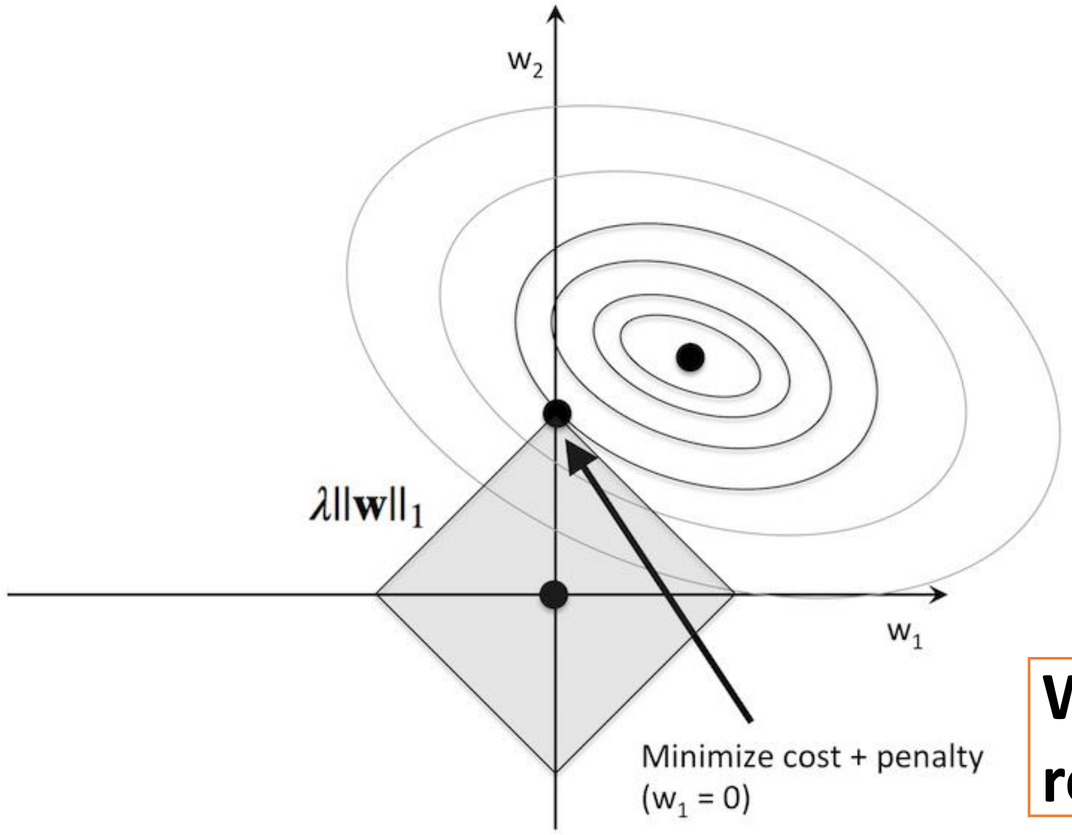

Dimensionality Reduction
Feature Selection
Filter Methods
Embedded Methods
Wrapper Methods
• Information gain
• Correlation with target
• Pairwise correlation
• Variance threshold
• ...
• L1 (LASSO) regularization
• Decision tree
• ...
• Recursive Feature Elimination (RFE)
• Sequential Feature Selection (SFS)
• Permutation importance
• ...

Embedded methods

- Always supervised
- No separate feature selection
- Feature Selection happens as part of model training
- E.g.:
  - LASSO
  - Feature Importance with Random Forest
- Returns coeff_ or feature_importances
- Other non parametric methods do not augur well

Feature Selection with LASSO

Cost function adjusted for L1 Regularization



$$\arg\min_{w} \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|_1$$

$$\nabla_w \mathcal{J} = \frac{2}{m} X^T (Xw - y) \qquad \nabla_w \|w\|_1 = \mathbf{1}$$

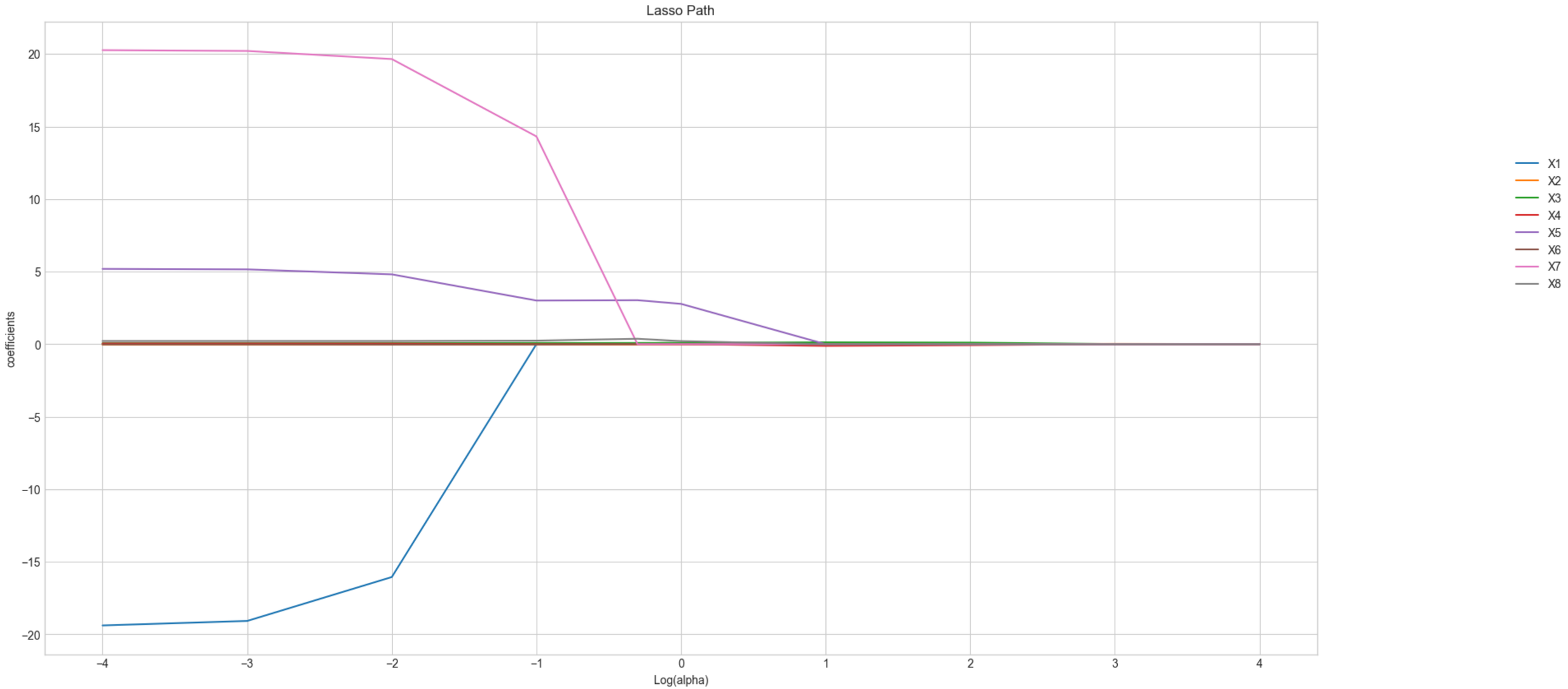
$$\mathbf{w} = \mathbf{w} - \eta \nabla_w \mathcal{J} \qquad \mathbf{w} = \mathbf{w} - \eta \nabla_w \mathcal{J} - \eta\lambda$$

Without regularization

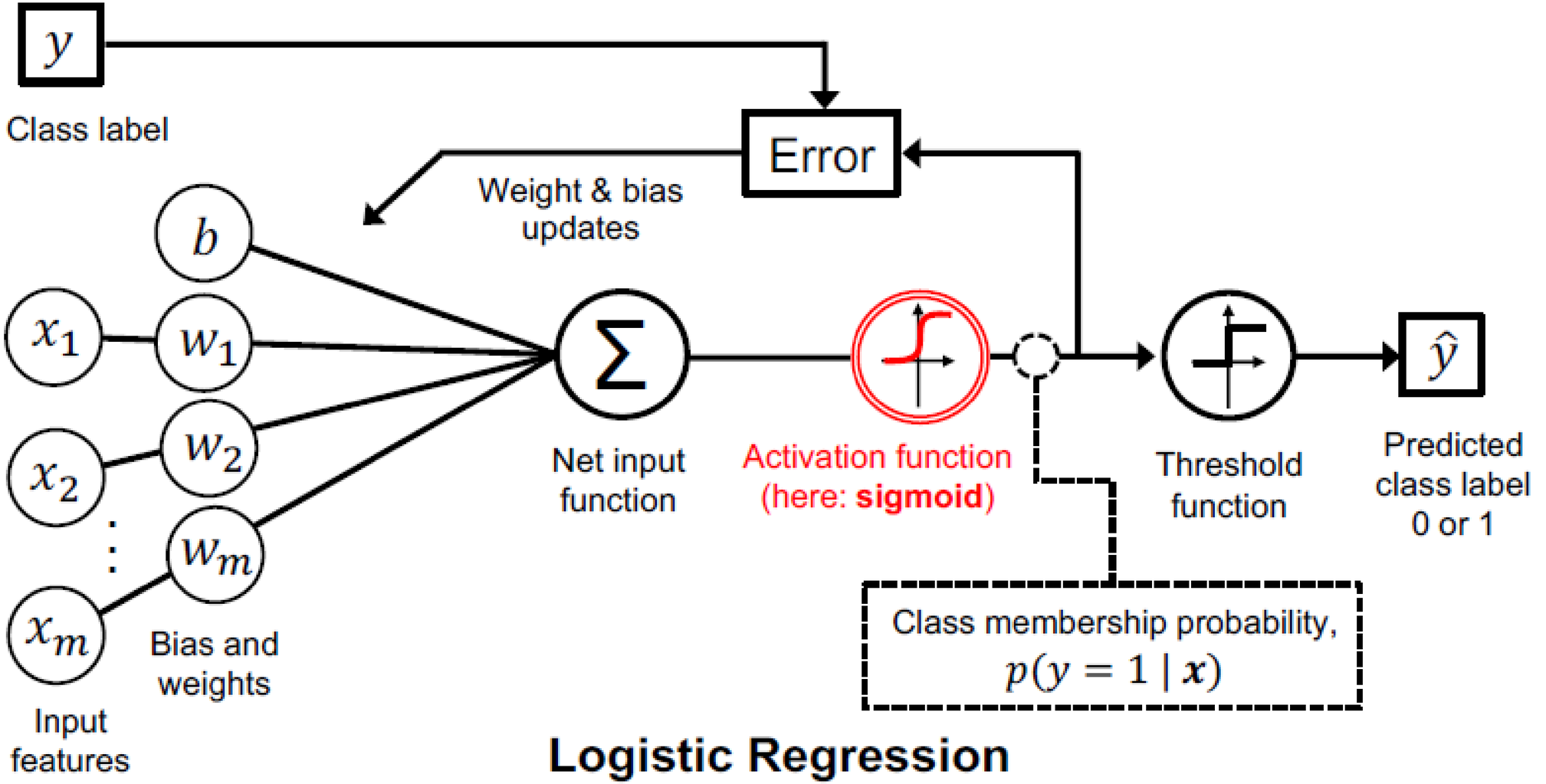
$$\mathbf{w} = (\mathbf{w} - \eta\lambda) - \eta \nabla_w \mathcal{J}$$

A FIXED small number keeps getting subtracting from a small w. Net effect w becomes 0

Lasso path



$$\arg\min_{w} \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|_1$$



Logistic Regression

$$-\sum_{i=1}\left[y^{(i)}\log\left(\sigma\left(z^{(i)}\right)\right)+\left(1-y^{(i)}\right)\log\left(1-\sigma\left(z^{(i)}\right)\right)\right]+\lambda\|w\|_1$$

Feature Selection with Decision Tree & Random Forest

Sample Tree:

- 2 Features : X1 & X2
- 2 Classes : Class1 & Class2
- Impurity Metric : Gini

- A feature is important if
  - If used many times for splitting
  - Each split on the feature is high in the tree
  - Split produces lot of decrease in impurity at each node

Sample Tree:

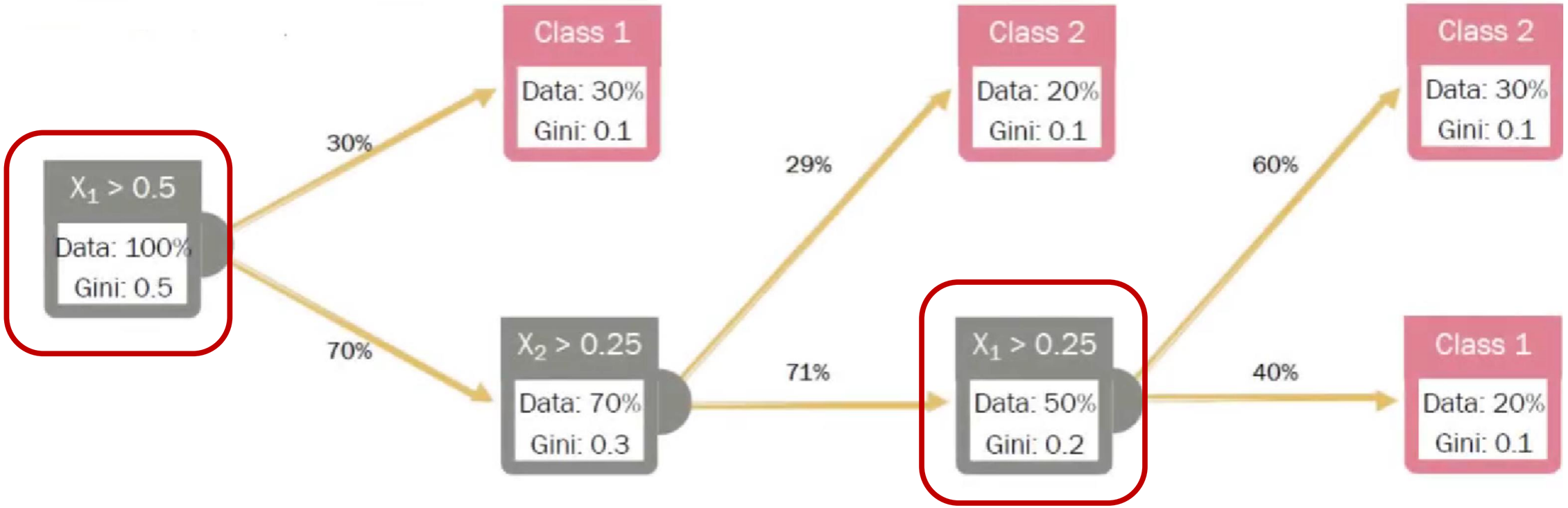
- 2 Features : X1 & X2
- 2 Classes : Class1 & Class2
- Impurity Metric : Gini

- Node Importance

$$ni_j = \frac{N_j}{N}\left(Gini_j - \mathbb{E}[Gini_{j-children}]\right)$$

- Feature Importance

$$fi_i = \frac{\sum_{j \in feature-inodes} ni_j}{\sum_{j \in allnodes} ni_j}$$



- Decrease in impurity for X1 at top:
  - Impurity in parent node – impurity in child node

$$0.5 - (0.7 \times .3 + 0.3 \times 0.1)$$

- Weighed by the ratio of data N_in/N

$$\frac{100}{100}[0.5 - (0.7 \times .3 + 0.3 \times 0.1)]$$

- How many times?

$$\frac{100}{100}[0.5 - (0.7 \times .3 + 0.3 \times 0.1)] + \frac{50}{100}[0.2 - (0.3 \times 0.1 + 0.2 \times 0.1)]$$

- Normalized Feature Importance = Sum of this feature importance divided by sum of all feature importance

Feature Importance in DT/RF summary

- Node Importance: Mean decrease in entropy/impurity from a parent node to child nodes after a feature split
  - Weighted by tree location (num examples at node)

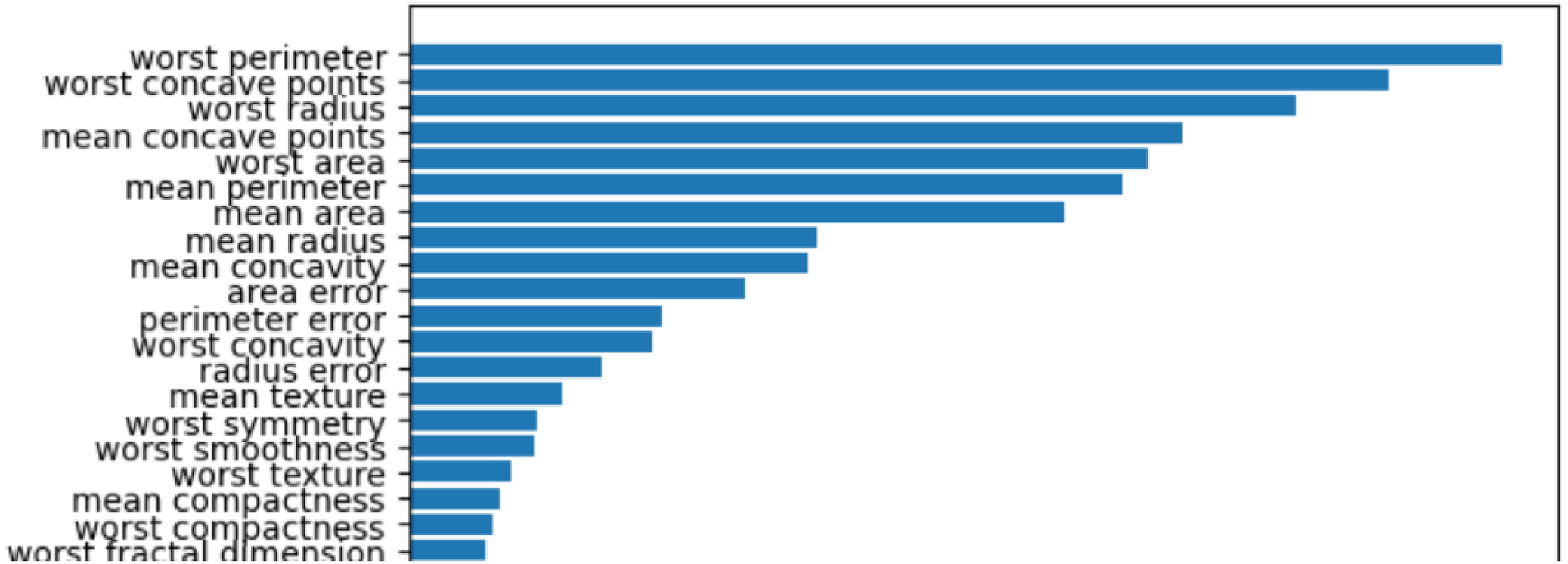
- for a given feature
  - for each tree
    - compute impurity decrease (Gini, Entropy)
    - weight by number of examples at that node
    - averaged over all trees
- normalize importances so that sum of feature importances sum to 1

Sklearn code

```
sorted_idx = model_best_rf.feature_importances_.argsort()
plt.barh(dataset.feature_names[sorted_idx], model_best_rf.feature_importances_[sorted_idx])
plt.xlabel("Random Forest Feature Importance")
```

✓ 0.3s

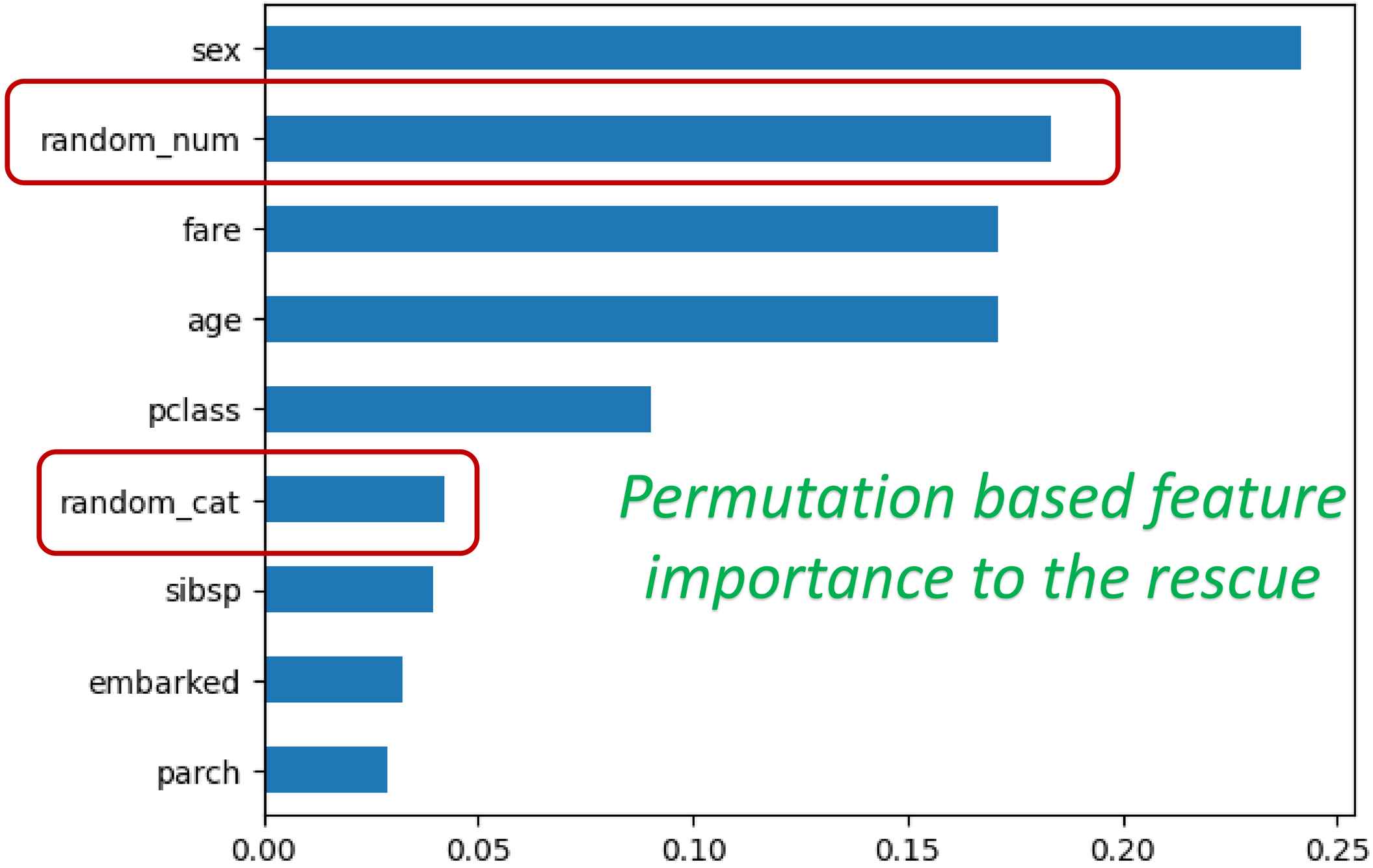
```
Text(0.5, 0, 'Random Forest Feature Importance')
```



Problems with Tree based Feature Importance

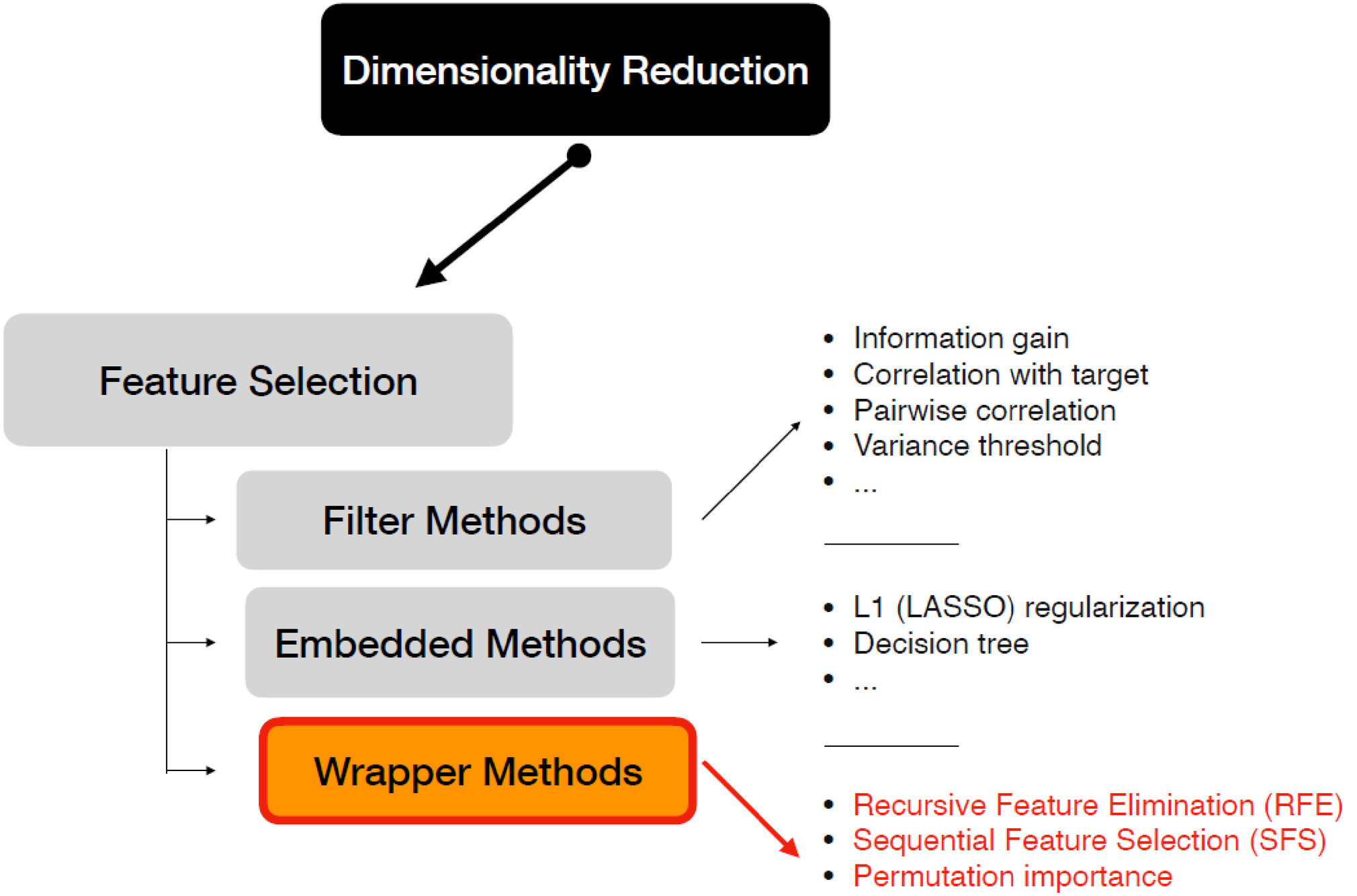
- Inflated feature importance for numerical feature
- Inflated feature importance for categorical feature with high cardinality

pclass	sex	age	sibsp	parch	fare	embarked	random_cat	random_num	survived
3.0	male	32.0	0.0	0.0	56.4958	S	0	-2.553921	1
2.0	male	27.0	0.0	0.0	26.0000	S	0	0.963879	0
3.0	male	35.0	0.0	0.0	7.8958	S	0	0.536653	0
3.0	female	26.0	1.0	1.0	22.0250	S	2	0.323079	1
3.0	male	33.0	0.0	0.0	8.6542	S	1	0.884045	0



Permutation based feature importance to the rescue

Wrapper methods


Dimensionality Reduction
Feature Selection
Filter Methods
Embedded Methods
Wrapper Methods
Information gain
Correlation with target
Pairwise correlation
Variance threshold
...
L1 (LASSO) regularization
Decision tree
...
Recursive Feature Elimination (RFE)
Sequential Feature Selection (SFS)
Permutation importance

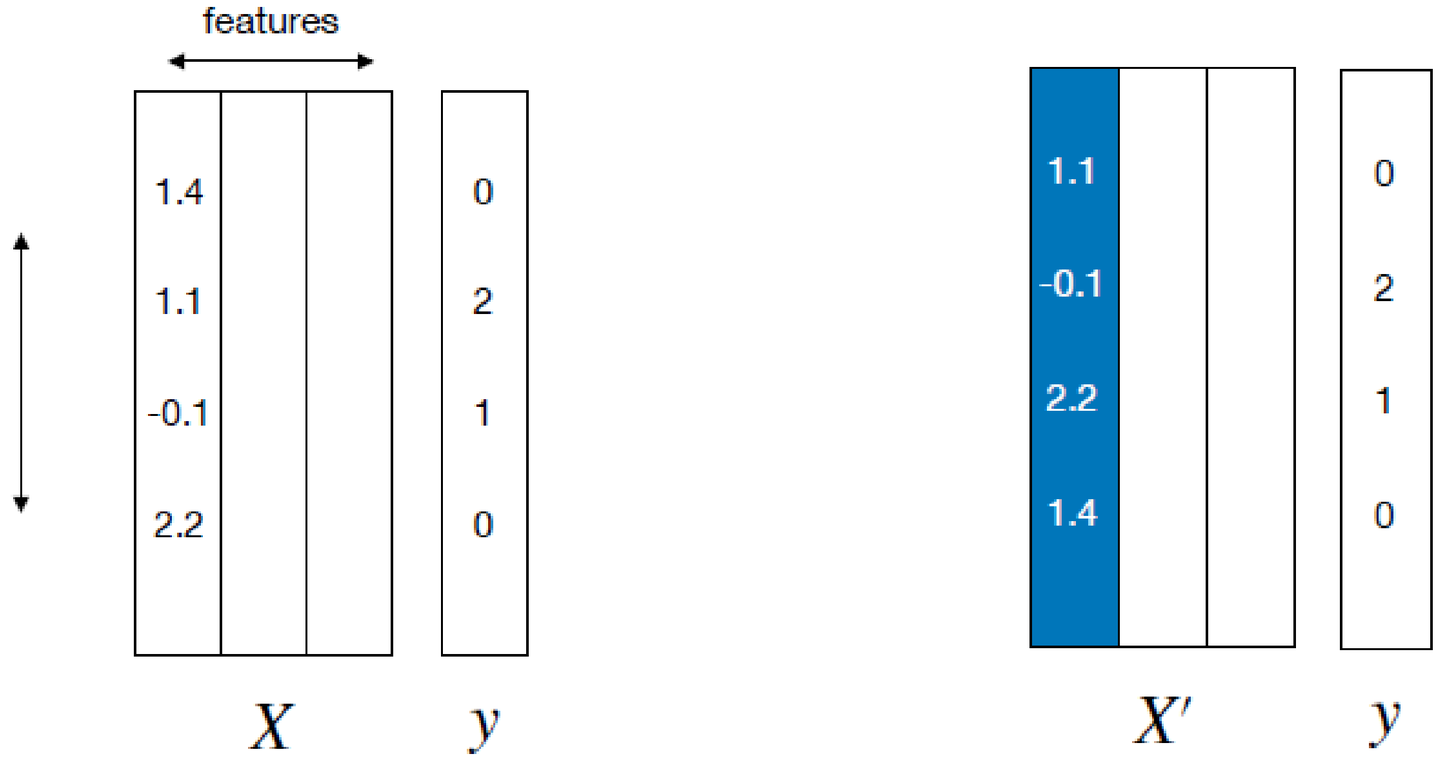
Wrapper methods

- General working
  - Wrap any algorithm to measure metrics
  - Make slight changes and run the algorithm again
  - The quantum of “slight changes” gives idea of feature importance
- E.g.:
  - RFE
  - Permutation Importance
  - Selective Feature Selection

Permutation based feature importance

Permutation importance

- model.fit(X_train, y_train), model.score(X_test)



feature importance =

baseline performance – shuffled dataset performance

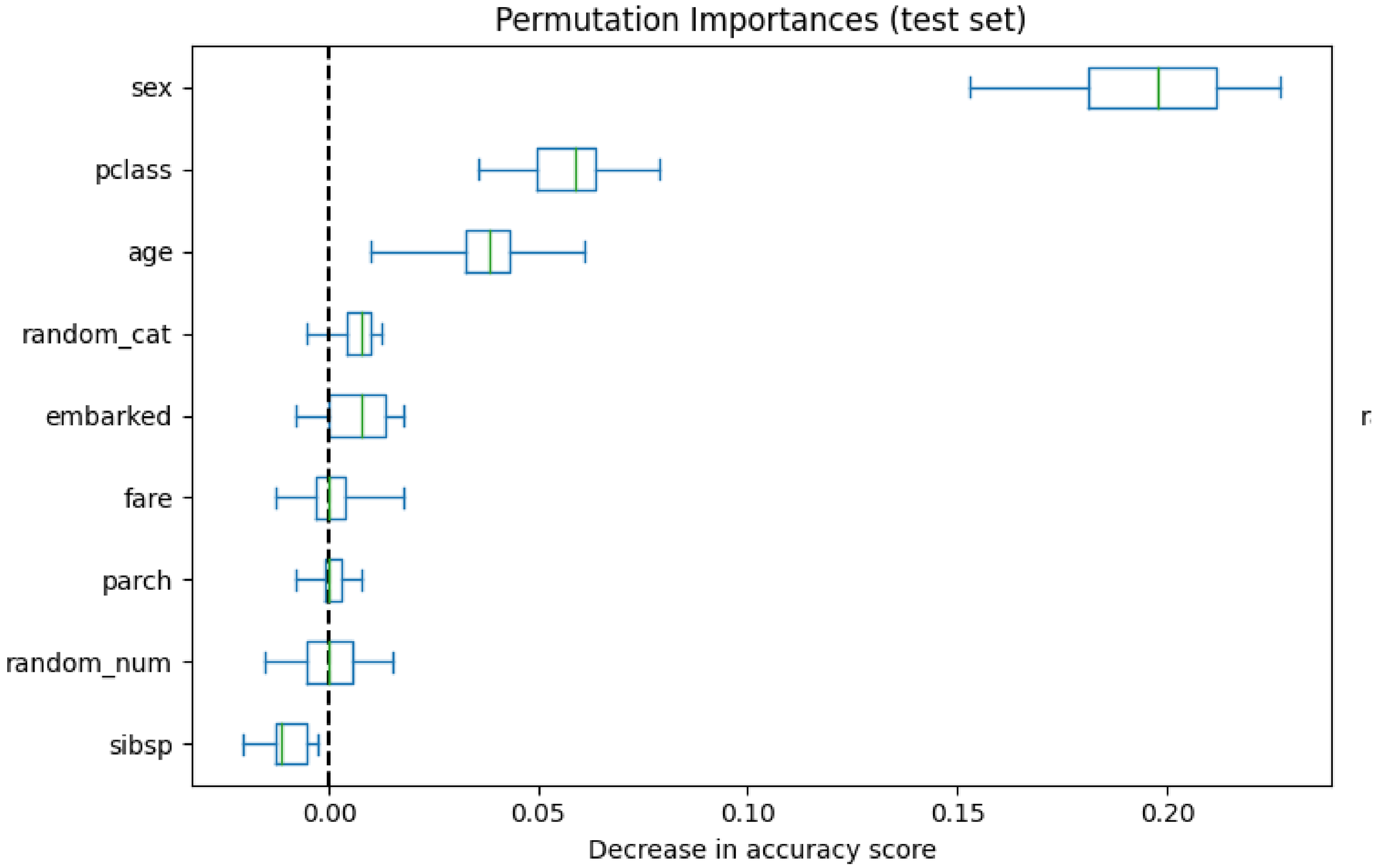
Permutation importance

- Train a baseline model
- Record model performance (any metric) on test dataset
- For each feature column in test dataset,
  - Shuffle that feature column alone, everything else unchanged
  - Observe performance and compare to original

  feature importance =

  baseline performance – shuffled dataset performance

- Do the shuffling for training dataset also & record feature importance


Permutation Importances (test set)
sex
pclass
age
random_cat
embarked
fare
parch
random_num
sibsp
0.00
0.05
0.10
0.15
0.20
Decrease in accuracy score

Column Drop Variant of permutation importance

- Train a baseline model
- Record model performance (any metric) on test dataset
- For each feature column in test dataset,
  - Drop column
  - Fit model
  - Compare test data set performance to original
- Accurate but very expensive
- Fixes the random_cat issue as well

Wrapper methods general features (except as noted)

- Permutation importance is model agnostic
  - No need for coeff_ or feature_importance implementation
    - (this is not applicable to other wrapper methods)
- Feature Importance is specific to model

- Permutation importance is flexible, can use any metric
- Feature importance is tied to impurity measure

- Permutation importance is easy to understand
- Feature importance is slightly tricky

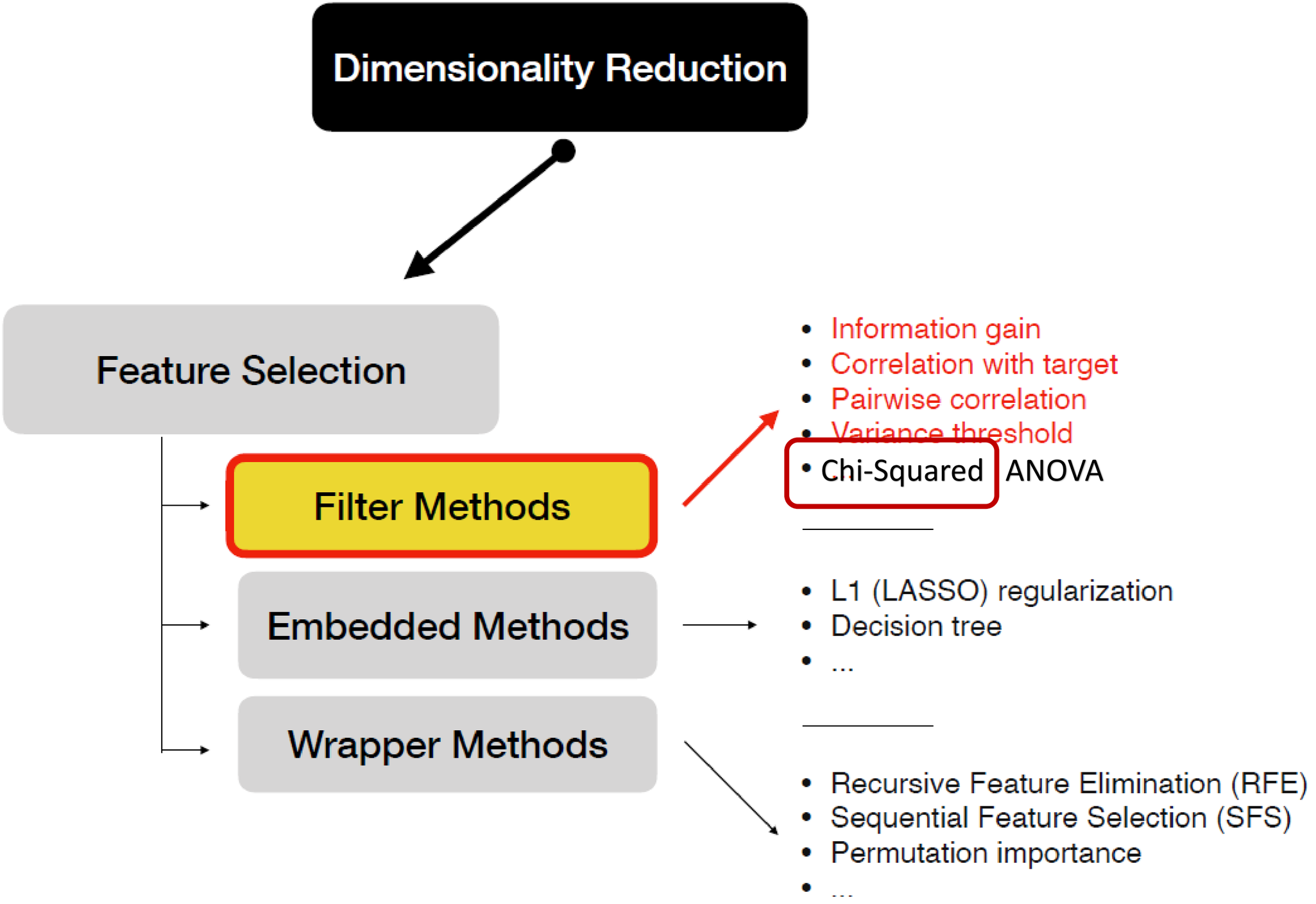
Recursive Feature Elimination

RFE

- Suppose you have number (or range) of features in mind
- Algorithm
  - Fit model to dataset
  - Eliminate feature with lowest coefficient (or lowest feature importance)
  - Repeat steps until desired features is reached
- Can be applied along with cross validation
- Comprehensive but expensive

Why repeat steps? Why not delete all low coefficient features at the outset?

Feature Selection with Chi Squared Test

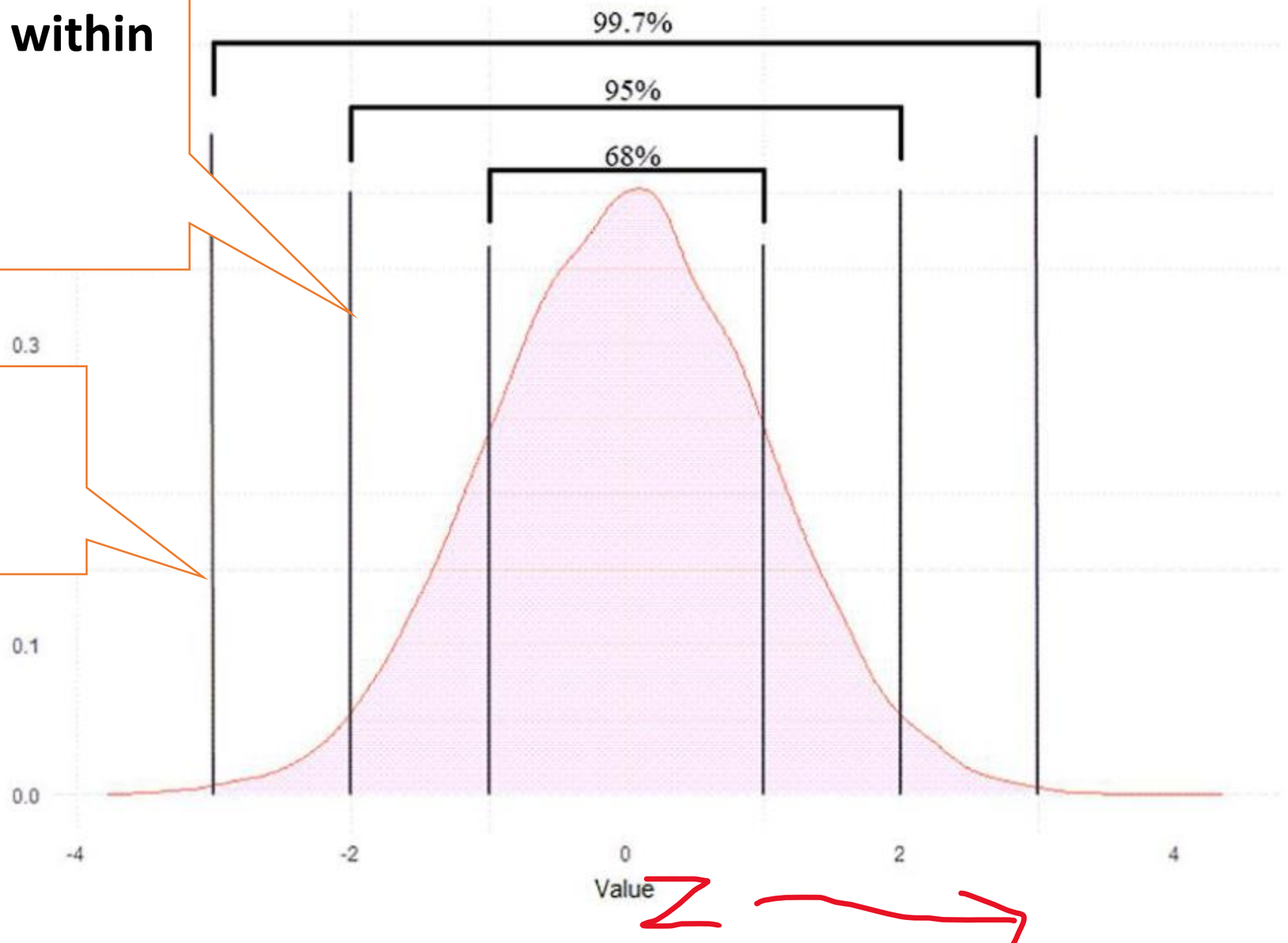

Dimensionality Reduction
Feature Selection
Filter Methods
Embedded Methods
Wrapper Methods
Information gain
Correlation with target
Pairwise correlation
Variance threshold
Chi-Squared
ANOVA
L1 (LASSO) regularization
Decision tree
...
Recursive Feature Elimination (RFE)
Sequential Feature Selection (SFS)
Permutation importance
...

Standard Normal Distribution

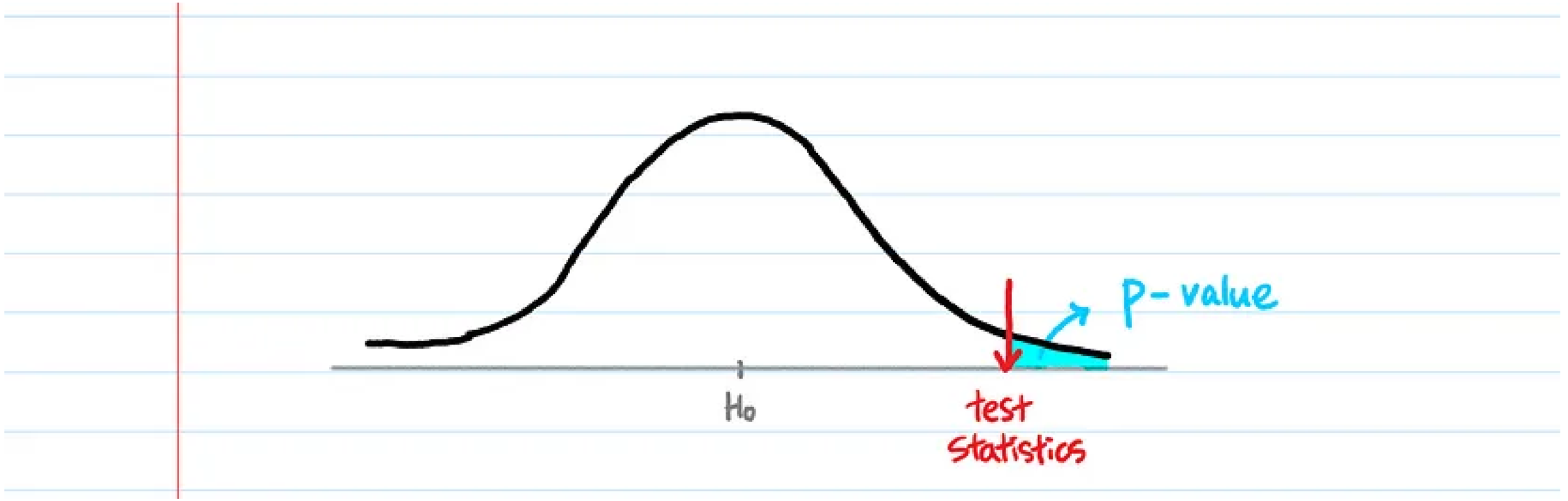
Probability of data within

1. **z=1 is 0.68**
2. **z=2 is 0.95**
3. **z=3 is 0.997**

Things are easy with normal distribution



Standard Normal distribution and p-values

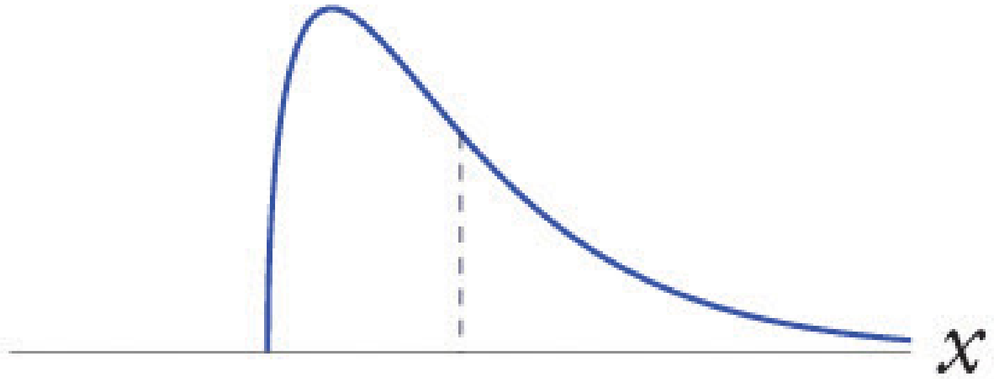


- P(X > k)

Why normal distribution?

- Not everything follows normal distribution

Some feature distribution



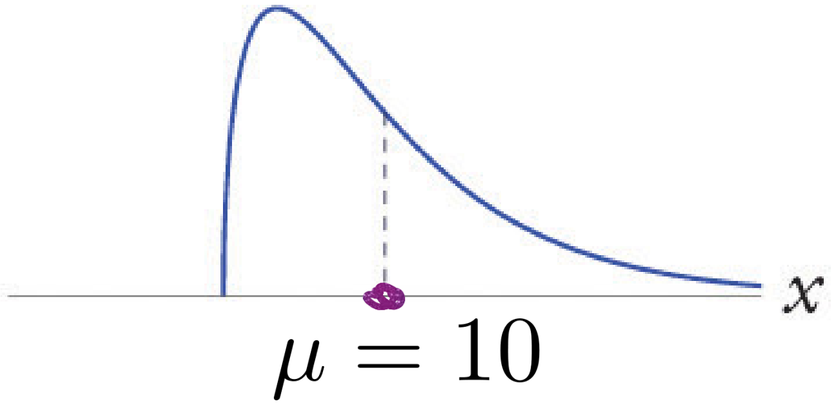
- Accidents on road follow Poisson distribution
- Volcanic eruption, asteroid strike: exponential distribution
- Why do we keep talking about normal distribution?
- The answer is in Central Limit theorem

Central Limit Theorem (CLT)

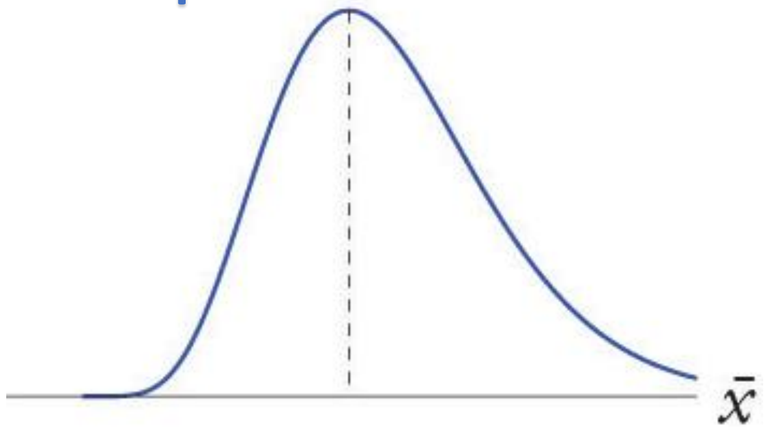
- Most fundamental to inferential statistics
- Aside: What is inferential statistics versus descriptive statistics?
- CLT provides mechanism to apply normal distribution to everything

Central Limit Theorem – Sampling distribution of mean

Population distribution



Sampling distribution of sample mean with n=5

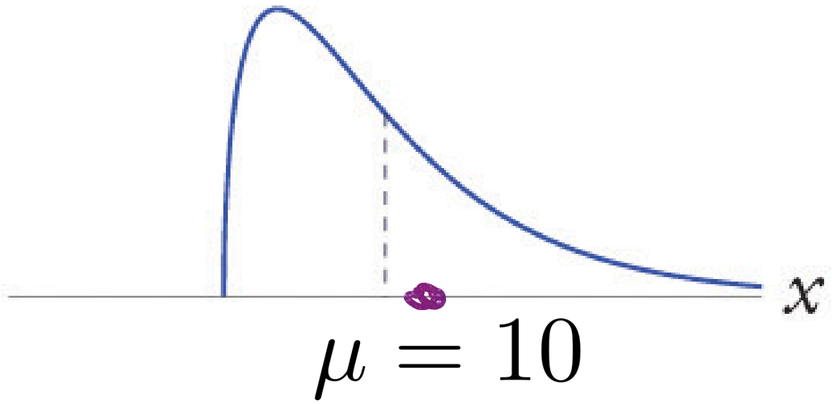


Sample size n = 5

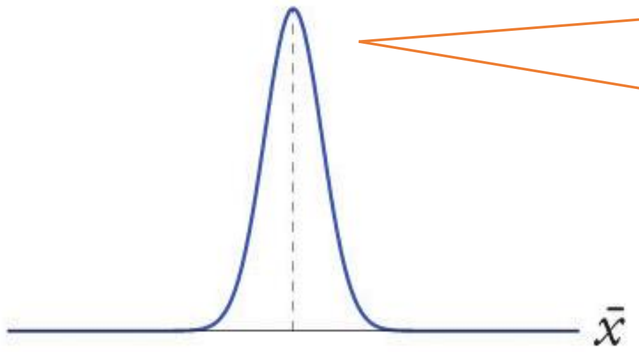
Sampling	Mean
1	7.5
2	9.5
3	11
4	9
5	11.5
6	10.5
7	9.75
8	9
9	9.25
10	9.8

Central Limit Theorem – Sampling distribution of mean

Population distribution



Sampling distribution of sample mean with n=30



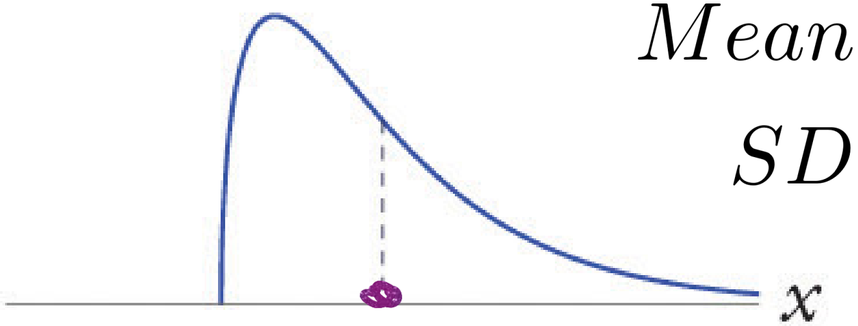
Sampling distribution approximates normal distribution when n>=30

Sample size n = 30

Sampling	Mean
1	7.5
2	9.5
3	11
4	12
5	11.5
6	10.5
7	9.75
8	10
9	10.25
10	9.8

Central Limit Theorem – Sampling distribution of mean

Population distribution

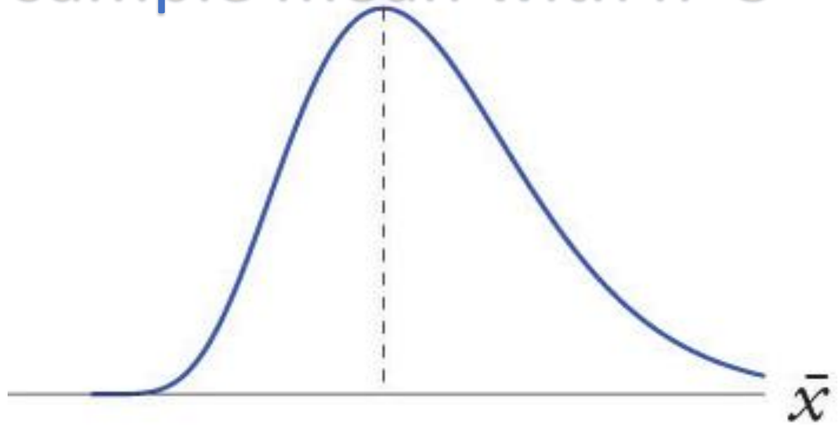


$$Mean = \mu$$

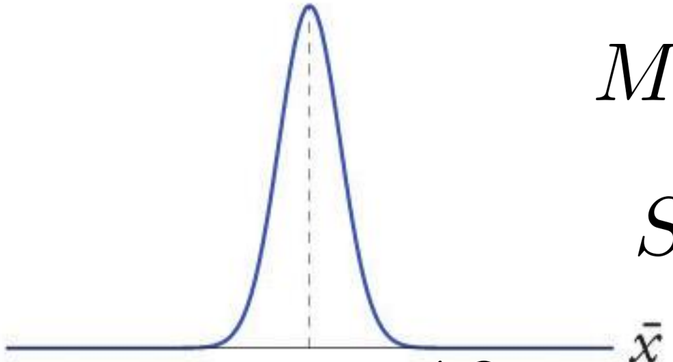
$$SD = \sigma$$

$$\mu = 10$$

Sampling distribution of sample mean with n=5



Sampling distribution of sample mean with n=30



$$Mean = \mu_s$$

$$SD = \frac{\sigma}{\sqrt{n}}$$

$$\mu_s = 10$$

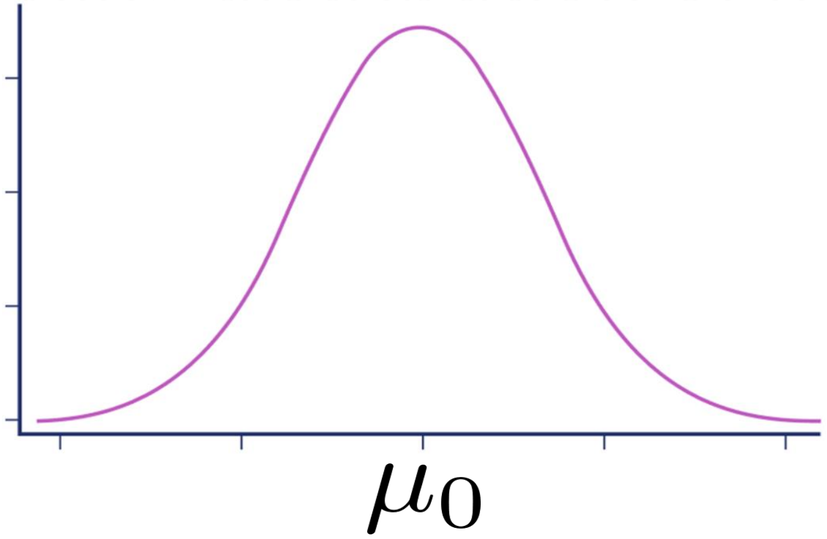
$$z \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

Made in India iPhone 15 battery life

- iPhone 15 has mean battery life μ_0 & variance σ_0
- Population has some unknown distribution

$$X \sim Unknown(\mu_0, \sigma_0)$$

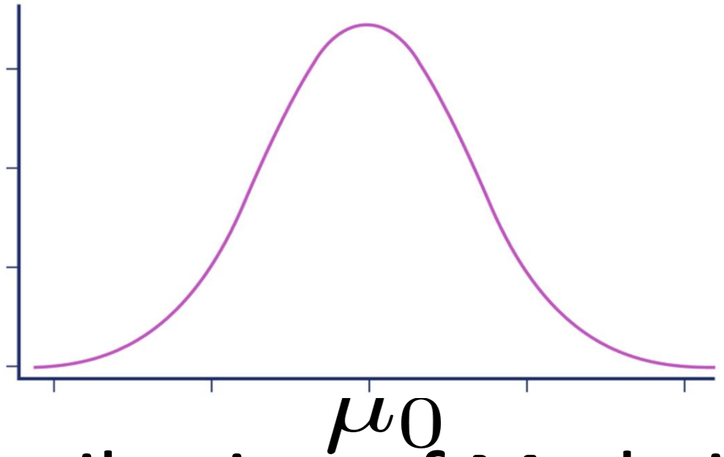
- How was population mean & variance found?
  - Draw samples (each of size >= 30) 10 times in population
  - Sampling distribution of mean battery life is gaussian

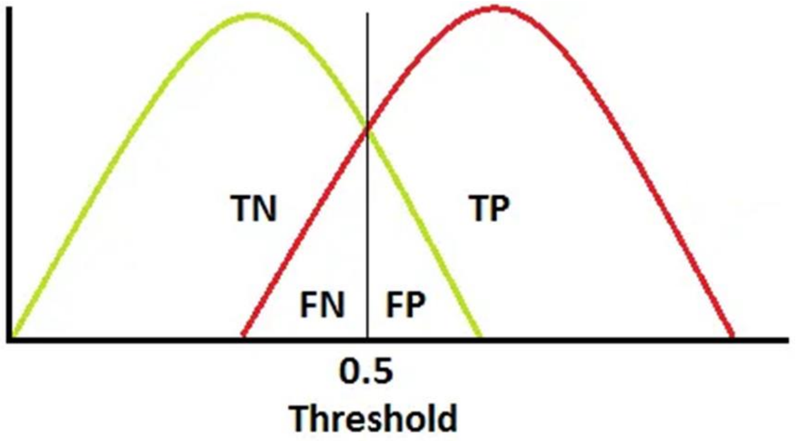


$$Y \sim \mathcal{N}(\mu_0, \frac{\sigma_0}{\sqrt{n}})$$

Made in China v/s India iPhone battery life

- Historical sampling distribution from China $Y \sim \mathcal{N}(\mu_0, \frac{\sigma_0}{\sqrt{n}})$





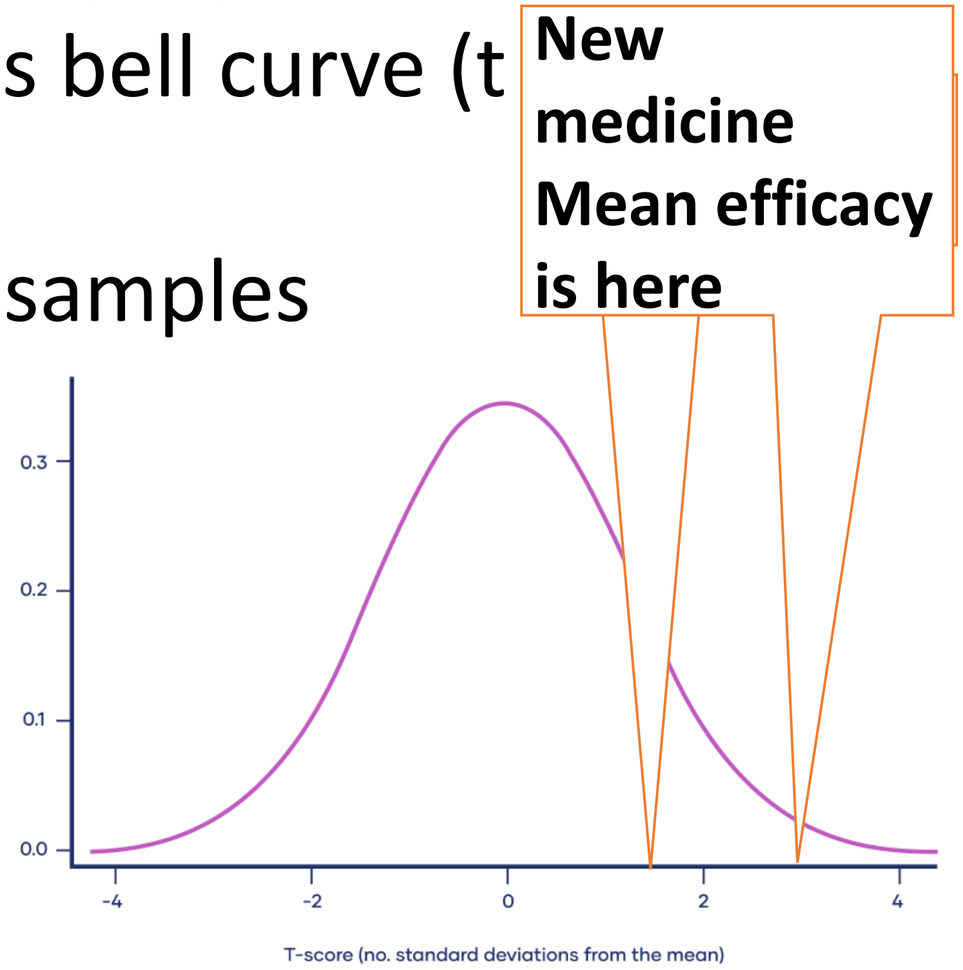
- v/s distribution of Made in India
- Where is the mean of new distribution?
- Should new distribution exactly align with historical distribution? What is confidence interval (CI)? $(\mu_0 \pm \frac{\sigma_0}{\sqrt{n}})$
- CI relation to TP/TN Type 1 Type2

Clinical Trial of a new cancer medicine

- Current medicine has some efficacy
- Efficacy of current medicine decided by equation/past data
- Central Limit Theorem “sort of” gives bell curve (t distribution)
  - New medicine efficacy based on n samples
    - (n-1 degrees of freedom)

$$z \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

$$t_{n-1} \sim \mathcal{T}(\mu, \frac{\sigma}{\sqrt{n}})$$



T Distribution

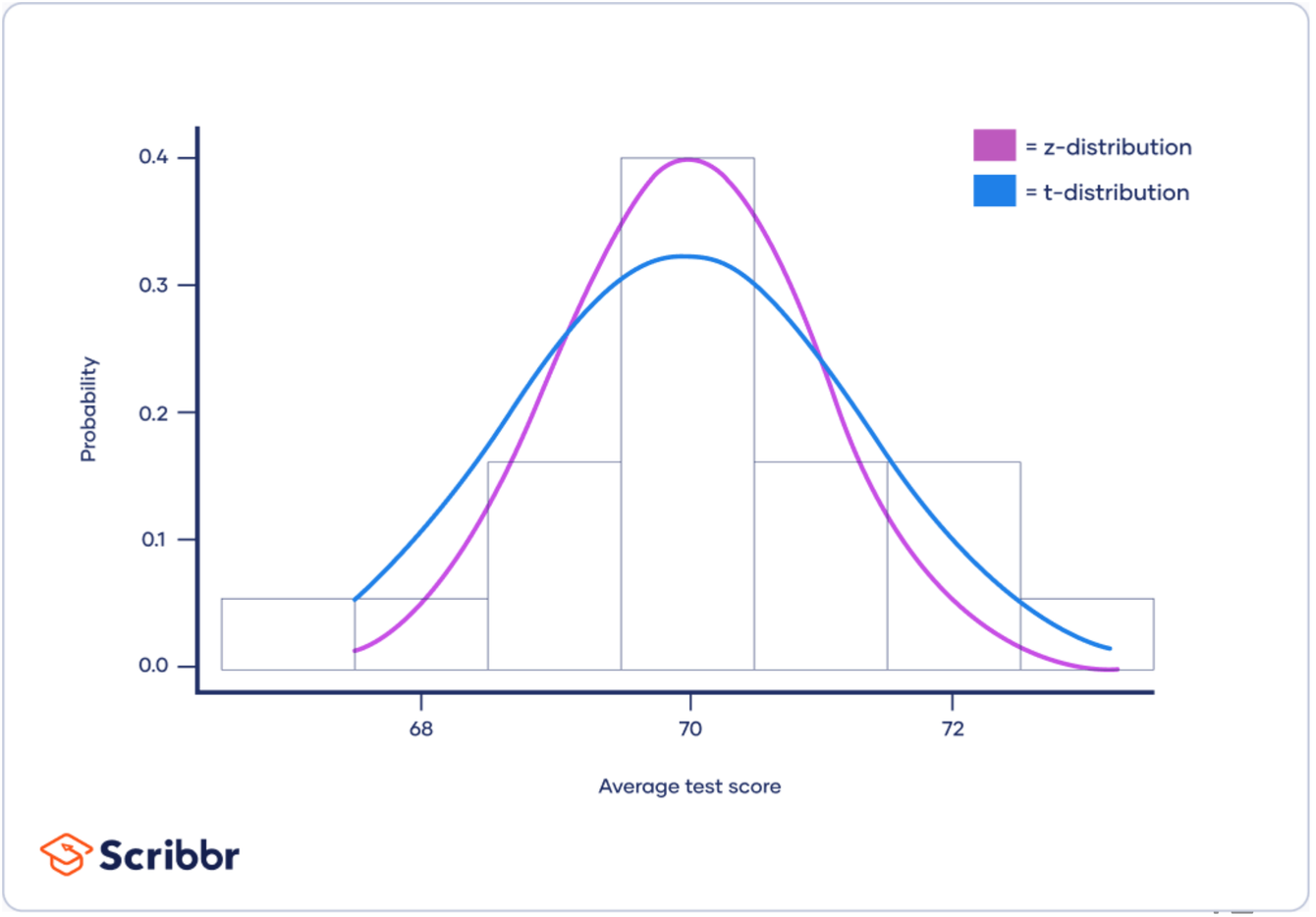
- Adjusted normal distribution for small sample sizes
- Number of samples = Degrees of freedom

$$z \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

z statistic follows from this

t statistic follows from this

$$t_{n-1} \sim \mathcal{T}(\mu, \frac{\sigma}{\sqrt{n}})$$



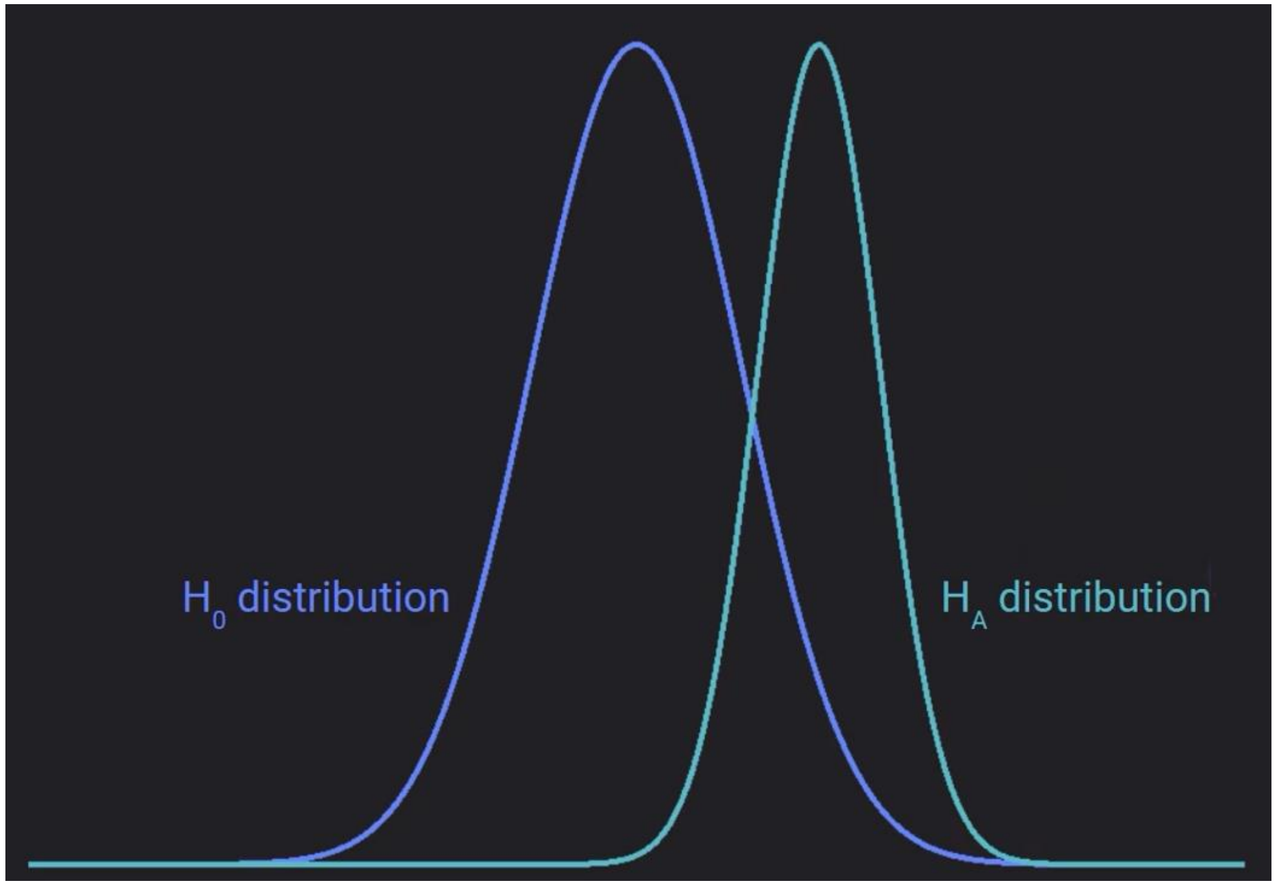
Intro to Hypothesis Test

- Null hypothesis – Ho,
- Alternate hypothesis – Ha

$$z \sim \mathcal{N}(\mu_0, \frac{\sigma_0}{\sqrt{n}})$$

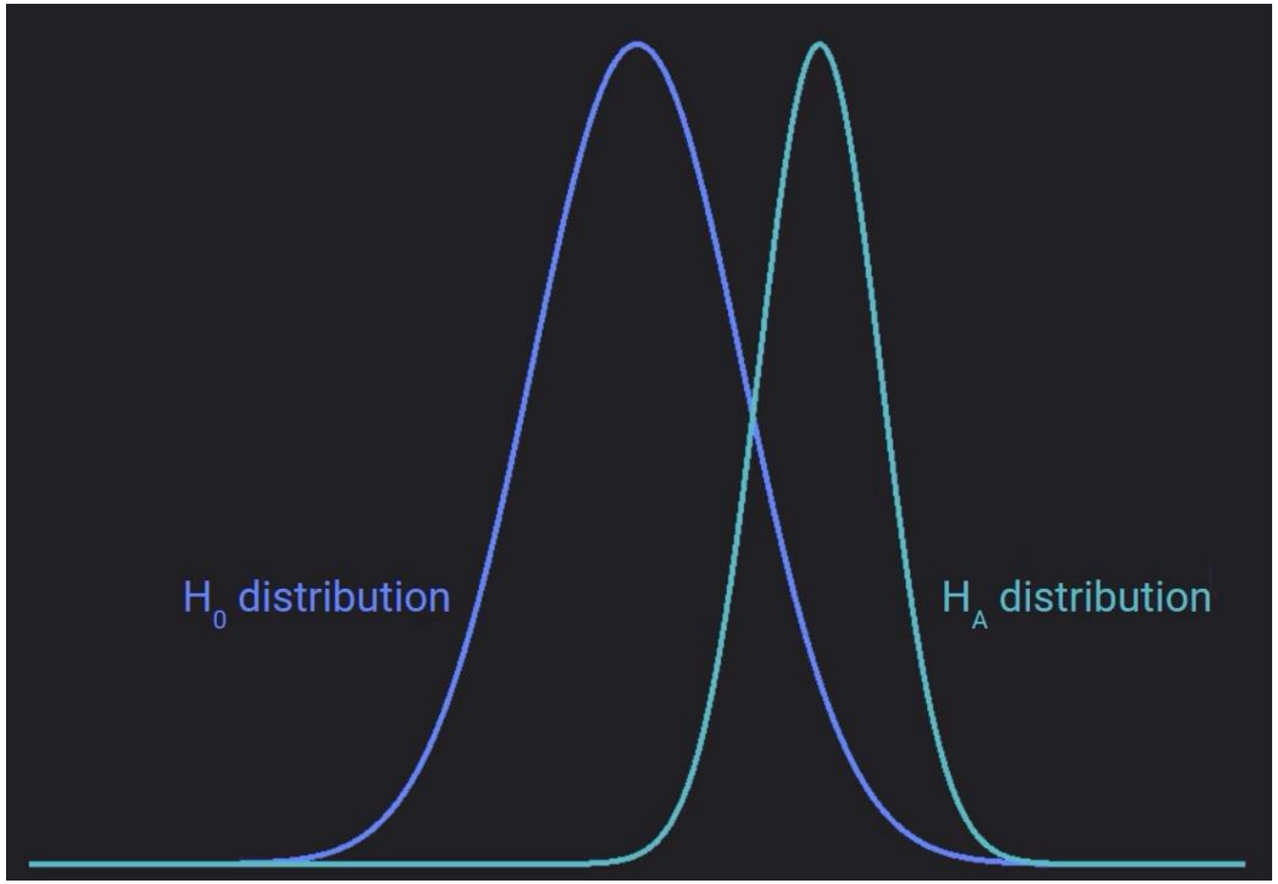
$$t_{n-1} \sim \mathcal{T}(\mu_0, \frac{\sigma_0}{\sqrt{n}})$$

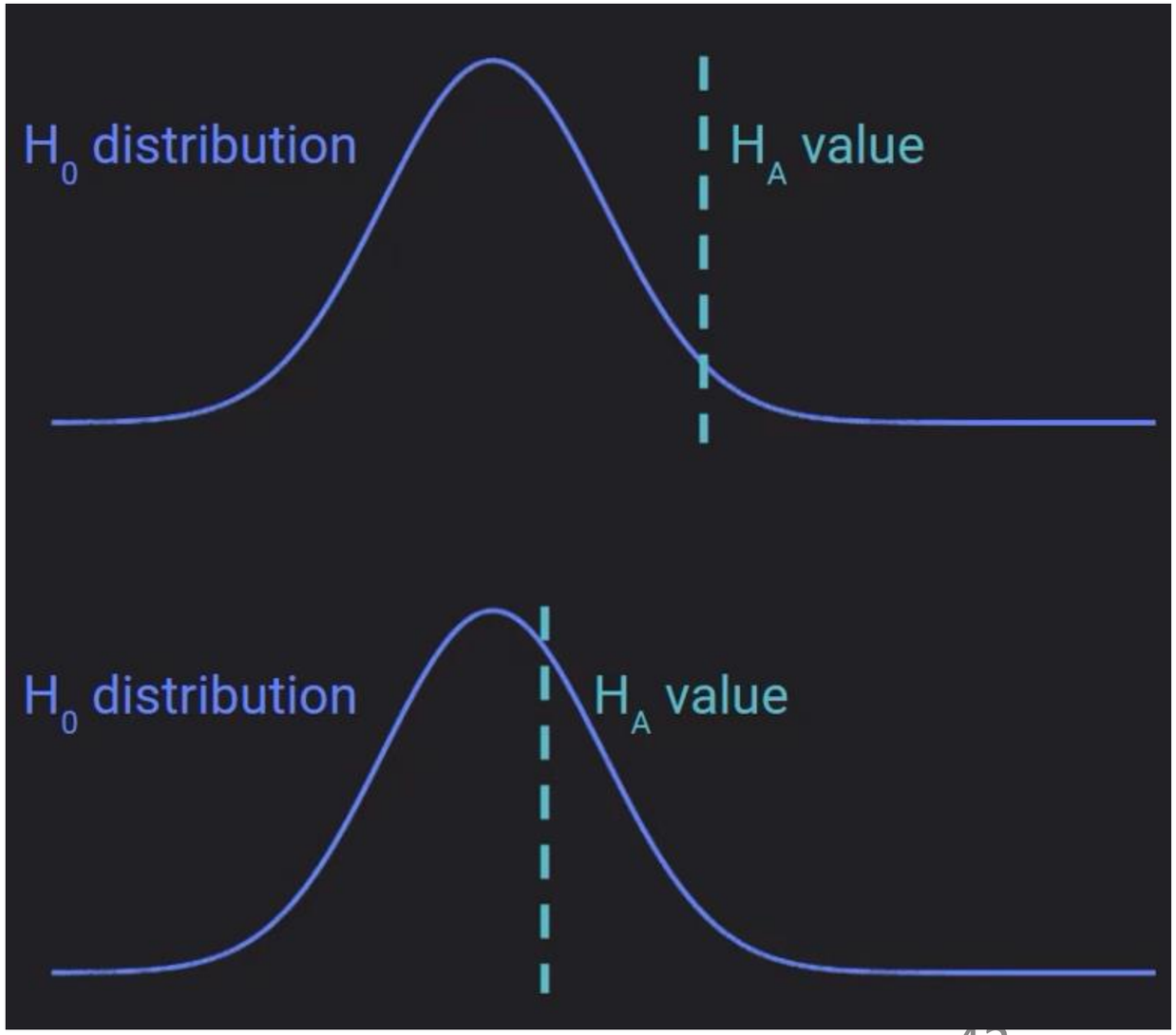
- Convert x to z or t



$$t_{n-1} = \frac{x - \mu_0}{\sigma/\sqrt{n}}$$

Intro to p-values in Hypothesis Test



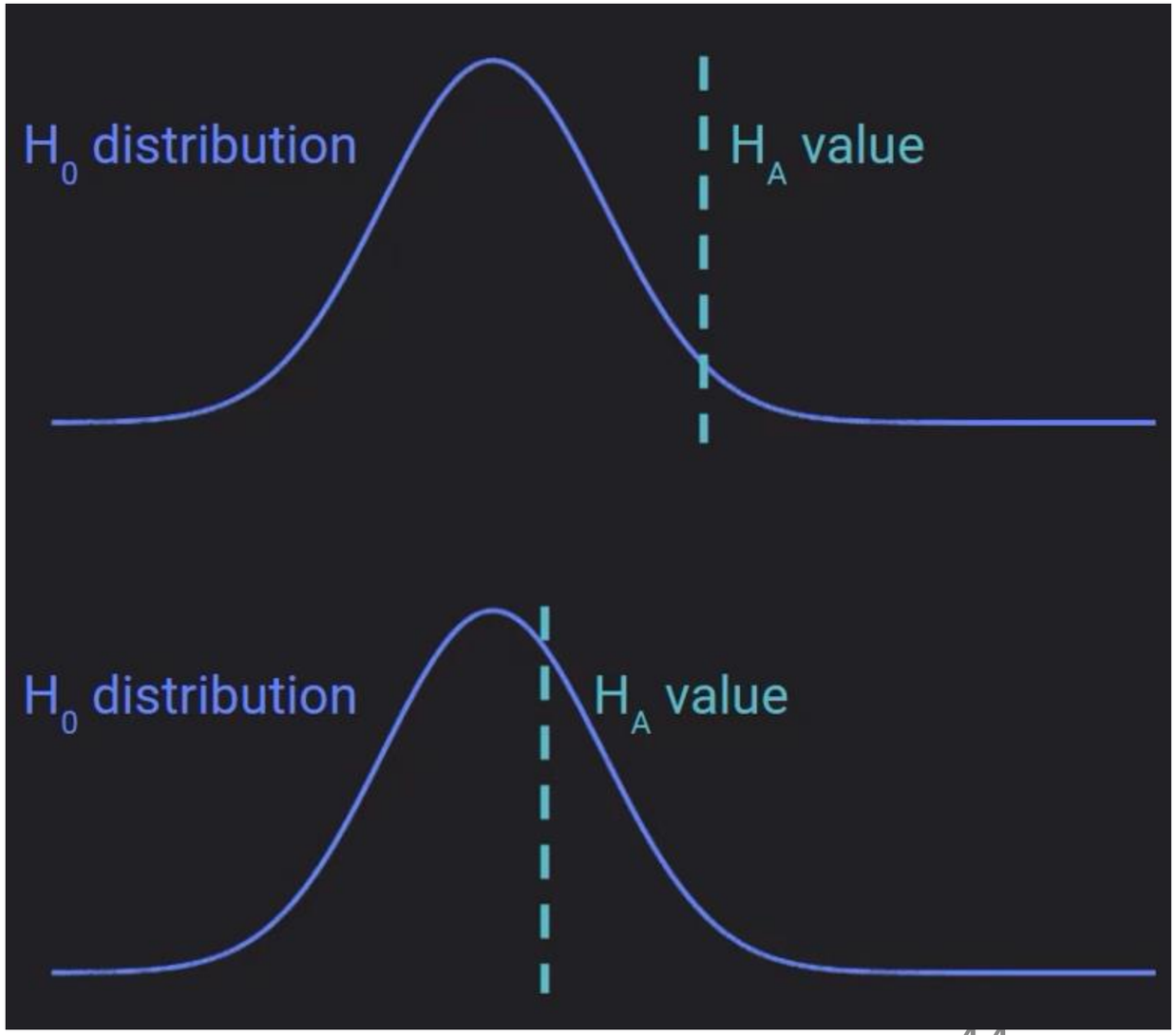


Indicates what is the probability value >= t-statistic

$$t_{n-1} = \frac{x - \mu}{\sigma / \sqrt{n}}$$

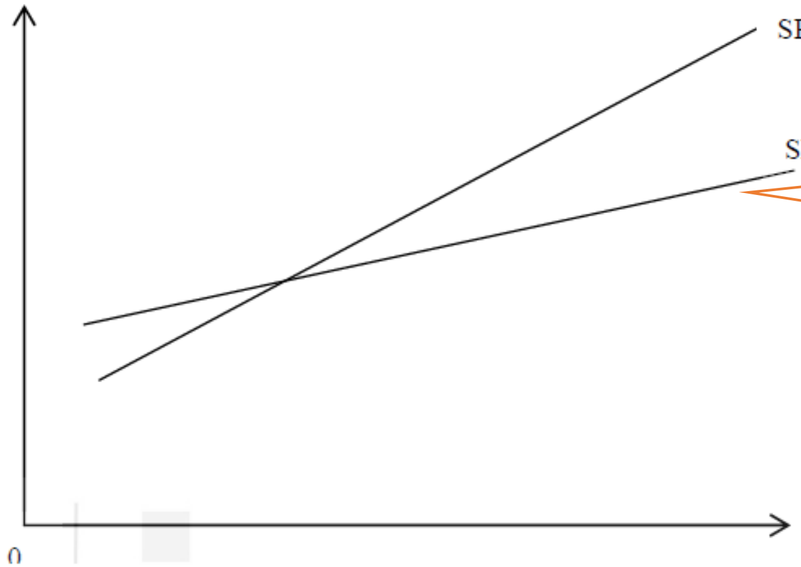
p-values in Hypothesis Test

- How likely for the Ha value to occur if Ho is true?
- What is the probability of observing a value of Ha larger than current value if there was no true effect?
- P(Ha | Ho)
- Definite Integral
  - Lookup table or coding



Regression - Population versus Sample View

- Sample Regression Functions
  - Different Regression Line/Plane/Hyperplane



Hypothesis function(s) corresponding to different null hypothesis (of what?)

- Difference between lines – values of coefficients
- Distribution of coefficients

$$\hat{y} = h(x) = w_1 TV + w_2 radio + w_3 newspaper$$

P value tells us total probability of given value under null hypothesis

Null Hypothesis is coefficient = 0

Alternate Hypothesis is coefficient not 0

OLS Regression Results

Dep. Variable:	sales	**R-squared (uncentered):**	0.982
Model:	OLS	**Adj. R-squared (uncentered):**	0.982
Method:	Least Squares	**F-statistic:**	3566.
Date:	Sun, 28 Mar 2021	**Prob (F-statistic):**	2.43e-171
Time:	13:42:33	**Log-Likelihood:**	-423.54
No. Observations:	200	**AIC:**	853.1
Df Residuals:	197	**BIC:**	863.0
Df Model:	3		
Covariance Type:	nonrobust		

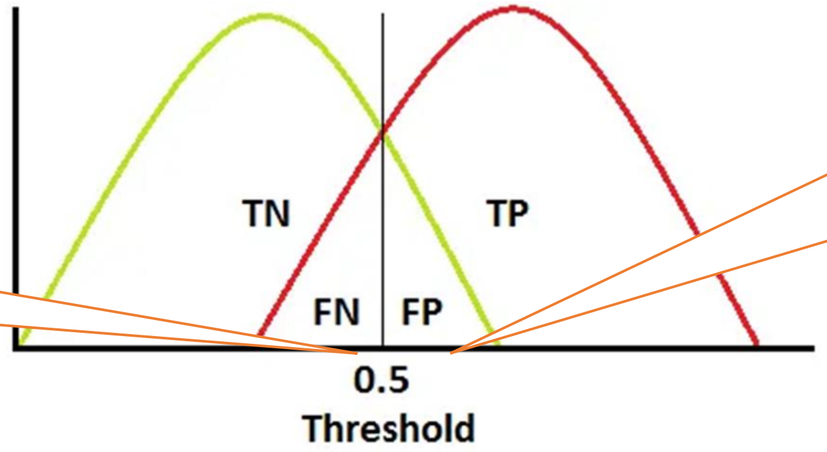
	coef	std err	t	P>|t|	[0.025	0.975]
TV	0.0538	0.001	40.507	0.000	0.051	0.056
radio	0.2222	0.009	23.595	0.000	0.204	0.241
newspaper	0.0168	0.007	2.517	0.013	0.004	0.030

Omnibus:	5.982	**Durbin-Watson:**	2.038
Prob(Omnibus):	0.050	**Jarque-Bera (JB):**	7.039
Skew:	-0.232	**Prob(JB):**	0.0296
Kurtosis:	3.794	**Cond. No.**	12.6

Viewing hypothesis test from generative ML perspective

- p values & statistical significance
- Machine Learning
  - Training is not a strict Ho, but a foundation
  - Each X_test record is sample from different dist (RV)
  - Each X_test record is the mean of the RV
  - Prediction on each X_test is different Ha

ML looks for optimal threshold

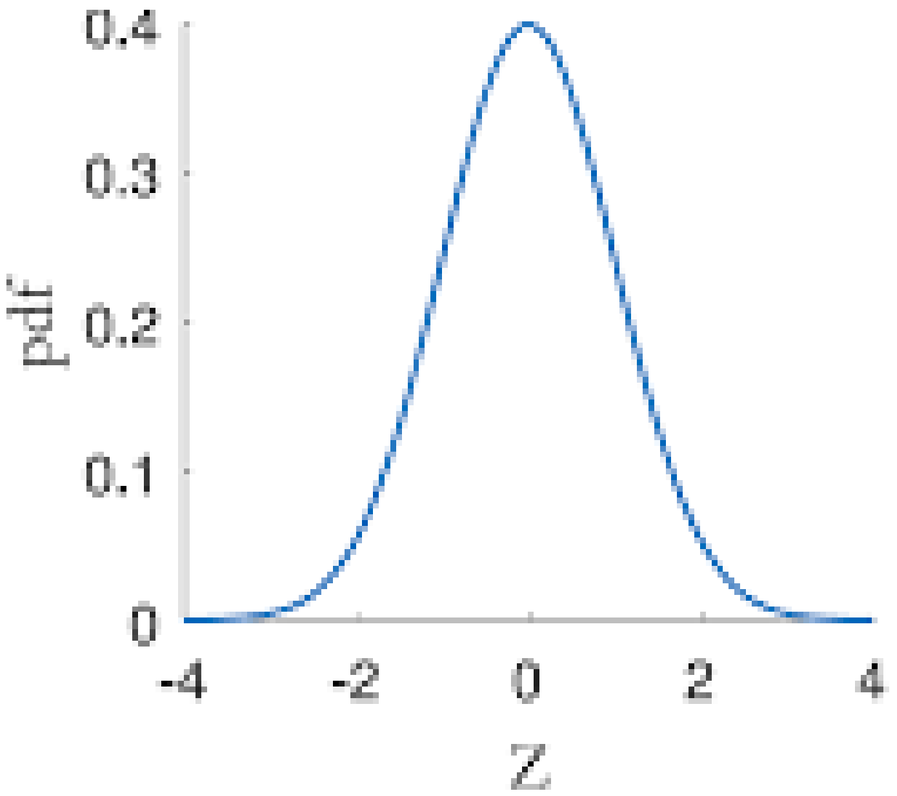


Hypothesis testing looks for conservative threshold for a given p-value

Chi-squared Distribution

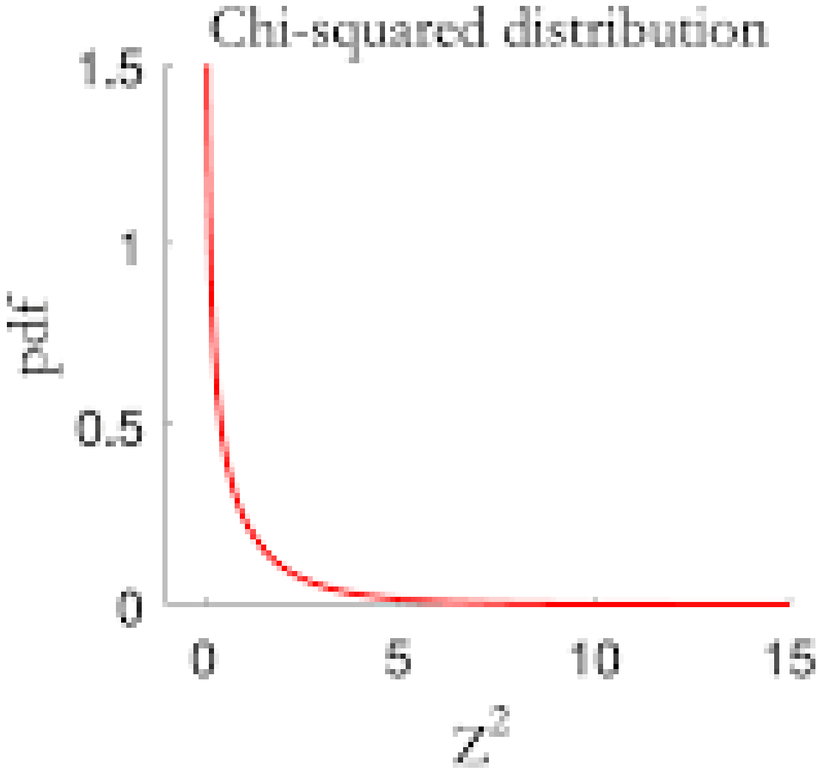
$$Z \sim N(0,1)$$

Standard Normal Distribution



$$Q = Z^2 \sim \chi^2$$

Chi-Squared Distribution

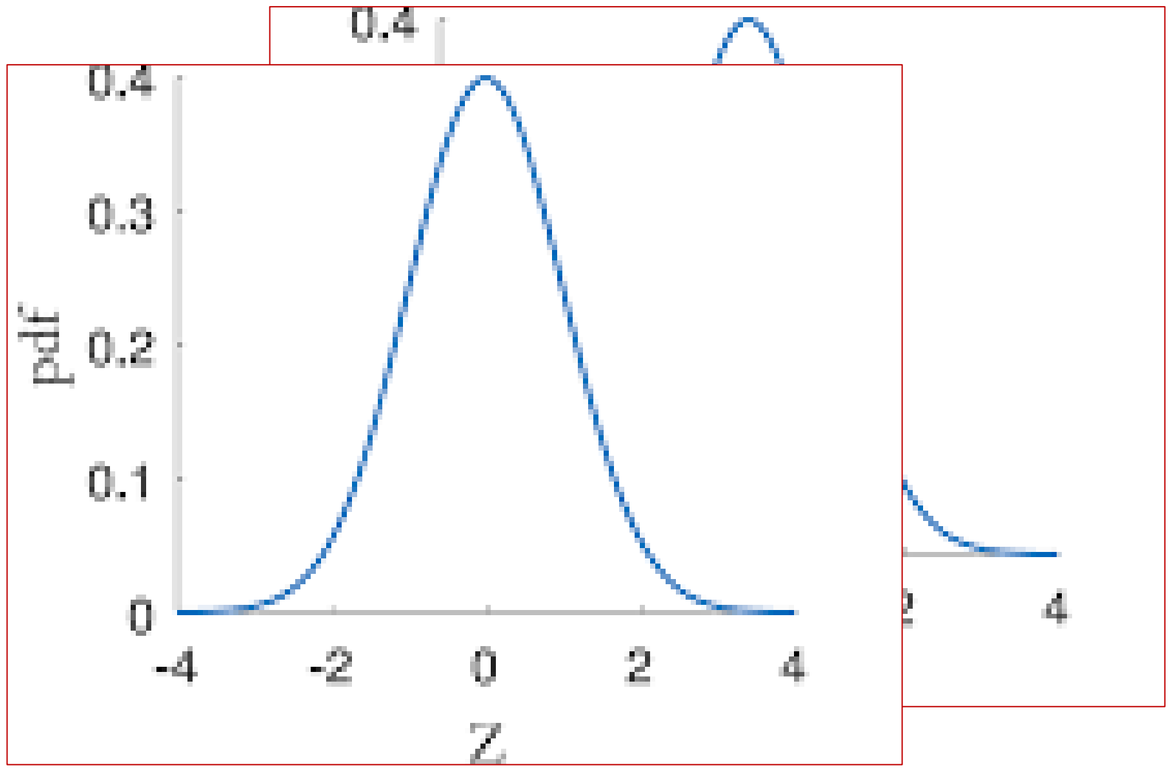


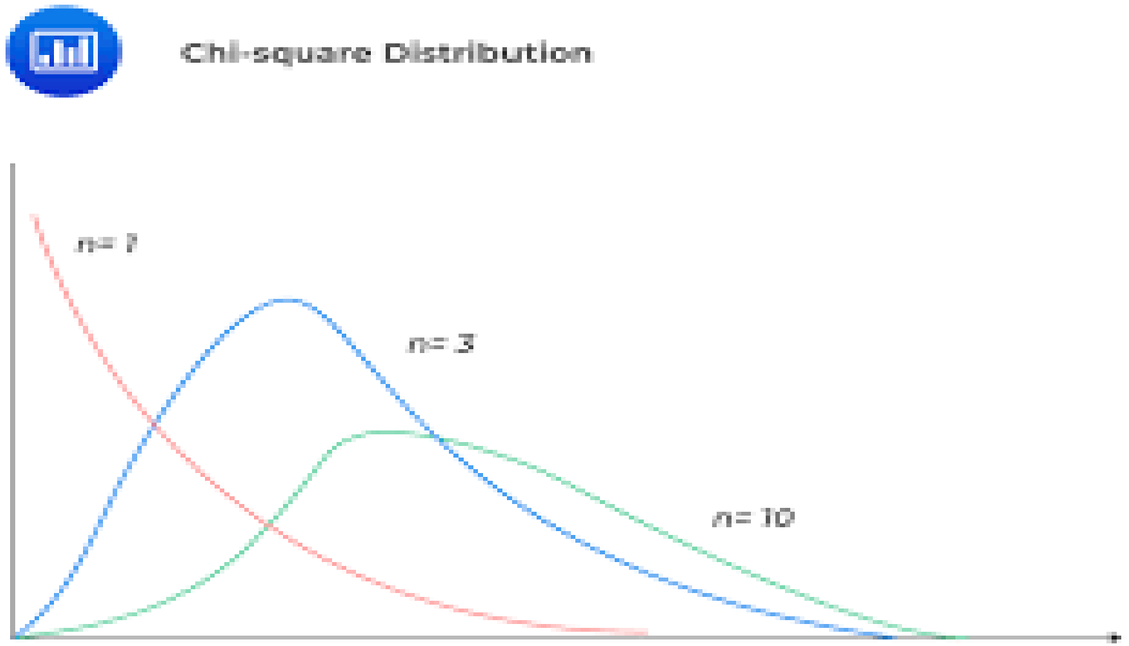
Chi-squared Distribution with two degrees of freedom

- Two random var Z_1, Z_2 with std normal distribution

$$Z_1 \sim N(0,1) \quad Z_2 \sim N(0,1)$$

$$Q = Z_1^2 + Z_2^2 \sim \chi_2^2$$



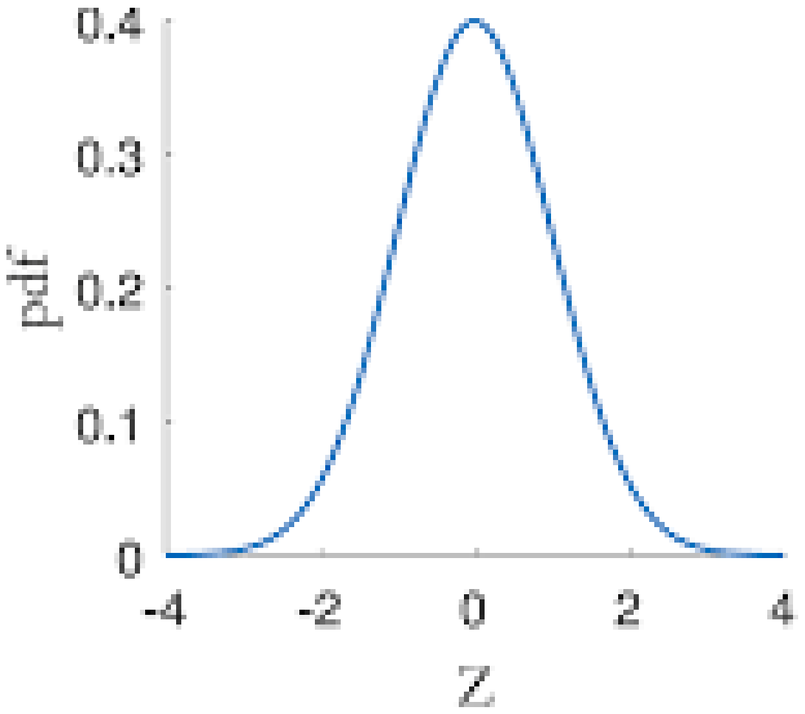


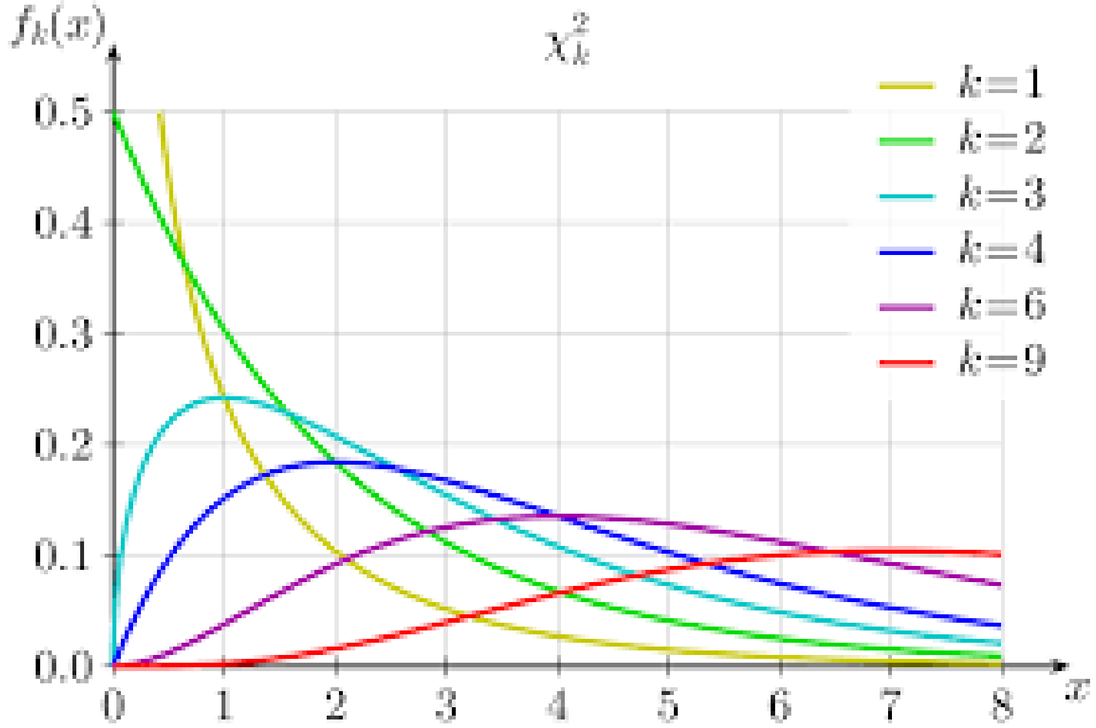
Chi-squared Distribution with n degrees of freedom

- n random var $Z_1, ..Z_n$ with std normal distribution

$$Z_1 \sim N(0,1) \quad Z_n \sim N(0,1)$$

$$Q = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$





Revisiting the contingency table for independence

- Titanic dataset
  - Sex & Survived (x & y)

sex	female	male	
survived			
0	91	479	570
1	235	109	344
Total	326	588	914

Observed

	Female	Male	Total
Survived	0.1	0.52	0.62
Not Survived	0.26	0.12	0.38
Total	0.36	0.64	1

Expected

	Female	Male	Total
Survived	.22	0.4	0.62
Not Survived	0.14	0.25	0.38
Total	0.36	0.64	1

Observed counts to expected counts

sex	female	male	
survived			
0	91	479	570
1	235	109	344
Total	326	588	914

Observed

	Female	Male	Total
Survived	0.1	0.52	0.62
Not Survived	0.26	0.12	0.38
Total	0.36	0.64	1

P(X,Y) = P(X)P(Y)

Expected

	Female	Male	Total
Survived	0.22 x 570	0.4 x 570	570
Not Survived	0.14 x 344	0.25 x 344	344
Total	0.36	0.64	1

	Female	Male	Total
Survived	.22	0.4	0.62
Not Survived	0.14	0.25	0.38
Total	0.36	0.64	1

Distribution of a single column in contingency table

sex	female
survived	
0	91
1	235
Total	326

Observed

Binomial distribution with large n

- n = 326
- p=0.1
- X =91

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim \mathbb{N}(0,1)$$

$$Z^2 = \frac{(X - np)^2}{np(1-p)}$$

	Female
Survived	0.1
Not Survived	0.26
Total	0.36

$$Z^2 = \frac{(X - np)^2}{np} + \frac{(n - X - np(1-p))^2}{n(1-p)}$$

$$Z^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

Chi-Squared distribution

sex	female	male	
survived			
0	91	479	570
1	235	109	344
Total	326	588	914

$$\chi^2 = Z^2 = \sum_{all-cells} \frac{(o_i - e_i)^2}{e_i}$$

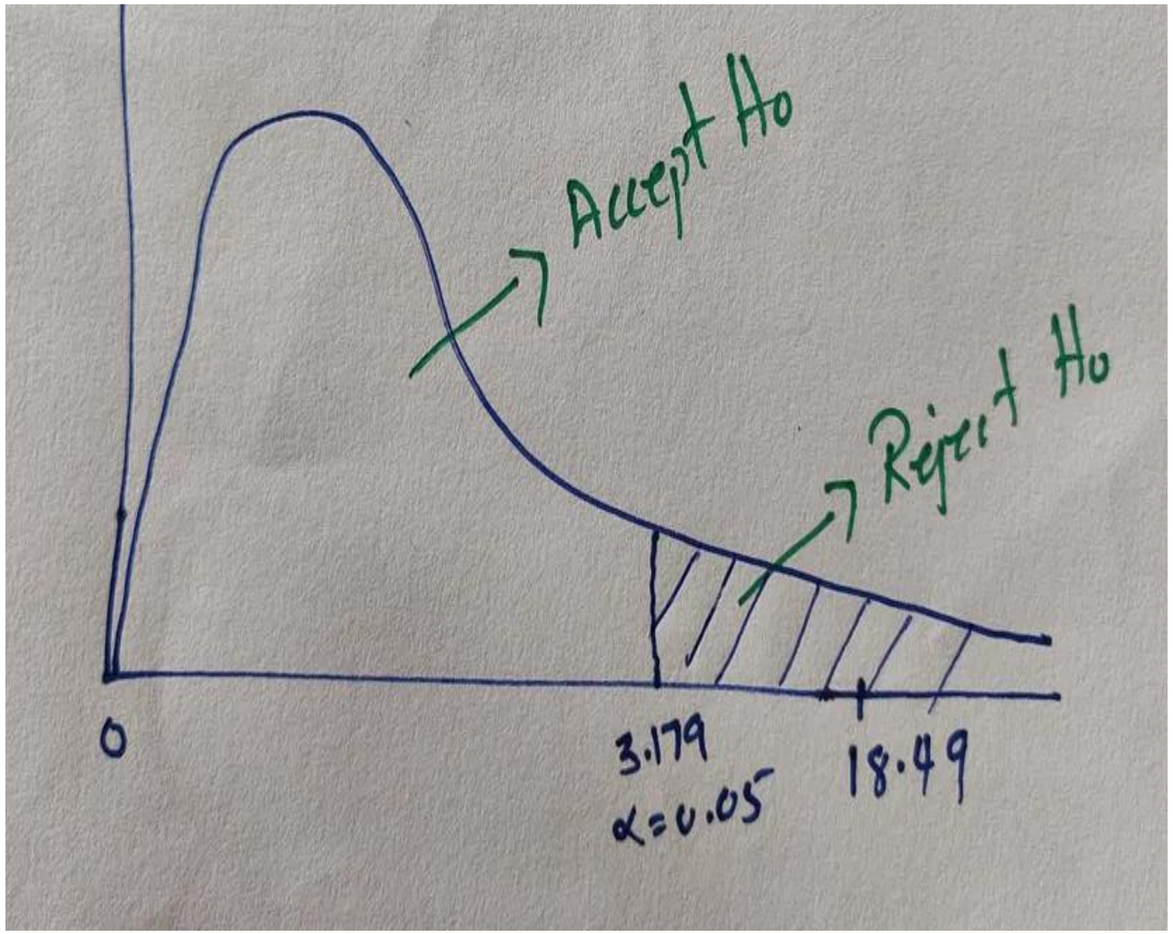
- How many degrees of freedom
- Ho Two features or feature-target are independent
- Ha Features not independent

	Female	Male	Total
Survived	0.22 x 570	0.4 x 570	570
Not Survived	0.14 x 344	0.25 x 344	344
Total	0.36	0.64	1

$$Z_1^2 = \sum_{i \in k} \frac{(o_i - e_i)^2}{e_i} \qquad Z_2^2 = \sum_{i \in k} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = Z_1^2 + Z_2^2 = \sum_{all-cells} \frac{(o_i - e_i)^2}{e_i}$$

$$\chi^2 = Z_1^2 + Z_2^2 + ... = \sum_{all-cells} \frac{(o_i - e_i)^2}{e_i}$$



- Ho Two features (or feature-target) are independent
- Ha Features (or feature-target) not independent

- Observed diverges A LOT from expected
- Chi-Squared value increases


$X^2-4X+5\le 5$
$X^2-4X\le 0$
$n(B\cap C)=22$
$n(B)=68$
$n(C)=84$
$n(B\cup C)=n(B)+n(C)-n(B\cap C)$
$\frac{\left(\frac{a}{b}\right)}{c}=\frac{a}{bc}$
$\frac{a}{\left(\frac{b}{c}\right)}=\frac{ac}{b}$
$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}$
$\log_b b^x = x$
$\log_a x=\frac{\log_b x}{\log_b a}$
$\bar{x}_1=\frac{1+3+3+6+8+9}{6}=5$
$\bar{x}_2=2+4+4+8+12=30$
$\bar{x}_3=4+7+1+6=18$
20
6
(0,1)
$M=\frac{0.046765}{3\,0L}$
He = 4.002602
Na = 22.989769
Ar = 39.948
$f=\frac{R}{2}$
$C_2H_3Cl_4+Ca(OH)_2\rightarrow C_2HCl_3+CaCl_2+H_2O$
$Zn_3Sb_2+6H_2O\rightarrow 3Zn(OH)_2+2SbH_3$
$+3H_2\rightleftharpoons 2NH_3$
$+I_2\rightleftharpoons 2HI$
$+O_2\rightleftharpoons 2SO_3$
A
B
C

ANOVA

ANOVA intuitively from ML perspective

- Categorical Feature: Crime rate = High, Medium, Low
- Numerical target: House price
- Does crime rate have impact on house price?

- Is the price difference between groups a mere chance (noise) or significant enough to be good predictor?
- Quantifying the difference as not significant/significant
- Think of it as lining up all combinations of Ho-Ha between categorical values

$$SSB = \sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2$$

$$SSW = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{i,j} - \bar{x}_i)^2$$

$$F = \frac{SSB/n_1}{SSW/n_2}$$

F-statistic from F distribution Has its own p-values

- n1, n2 degrees of freedom
- Ho Categorical feature has no variance between groups
- Ha Categorical feature has significant variance between groups

- Ho Categorical target cannot be predicted by a numerical feature
- Ha ??

Code example in sklearn

```python
from sklearn.feature_selection import f_classif

from sklearn.datasets import load_iris
data = load_iris()
X = data.data
y = data.target
```

```python
# Perform ANOVA
F_scores, p_values = f_classif(X, y)

for feature, F, p in zip(data.feature_names, F_scores, p_values):
    print(f"Feature: {feature}, F-score: {F}, p-value: {p}")
```

```
Feature: sepal length (cm), F-score: 119.26450218449871, p-value: 1.6696691907731823e-31
Feature: sepal width (cm), F-score: 49.16004008961098, p-value: 4.4920171333119 86e-17
Feature: petal length (cm), F-score: 1180.1611822529776, p-value: 2.856776610962102e-91
Feature: petal width (cm), F-score: 960.0071468018025, p-value: 4.169445839445031e-85
```

F-statistic in OLS

$$\hat{y} = h(x) = w_1 TV + w_2 radio + w_3 newspaper$$

P value tells us total probability of given value under null hypothesis

Null Hypothesis is coefficient = 0

Alternate Hypothesis is coefficient not 0

OLS Regression Results

Dep. Variable:	sales	**R-squared (uncentered):**	0.982
Model:	OLS	**Adj. R-squared (uncentered):**	0.982
Method:	Least Squares	**F-statistic:**	3566.
Date:	Sun, 28 Mar 2021	**Prob (F-statistic):**	2.43e-171
Time:	13:42:33	**Log-Likelihood:**	-423.54
No. Observations:	200	**AIC:**	853.1
Df Residuals:	197	**BIC:**	863.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TV	0.0538	0.001	40.507	0.000	0.051	0.056
radio	0.2222	0.009	23.595	0.000	0.204	0.241
newspaper	0.0168	0.007	2.517	0.013	0.004	0.030

Omnibus:	5.982	**Durbin-Watson:**	2.038
Prob(Omnibus):	0.050	**Jarque-Bera (JB):**	7.039
Skew:	-0.232	**Prob(JB):**	0.0296
Kurtosis:	3.794	**Cond. No.**	12.6

F-statistic in OLS

ANOVA intuitively from ML perspective

- Categorical Feature: Crime rate = High, Medium, Low
- Numerical target: House price
- Does crime rate have impact on house price?

- Is the price difference between groups a mere chance (noise) or significant enough to be good predictor?
- Quantifying the difference as not significant/significant
- Think of it as lining up all combinations of Ho-Ha between categorical values


QUESTIONS