



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

- Information gain
- Correlation with target
- Pairwise correlation
- Variance threshold
- ...

Embedded Methods

- L1 (LASSO) regularization
- Decision tree
- ...

Wrapper Methods

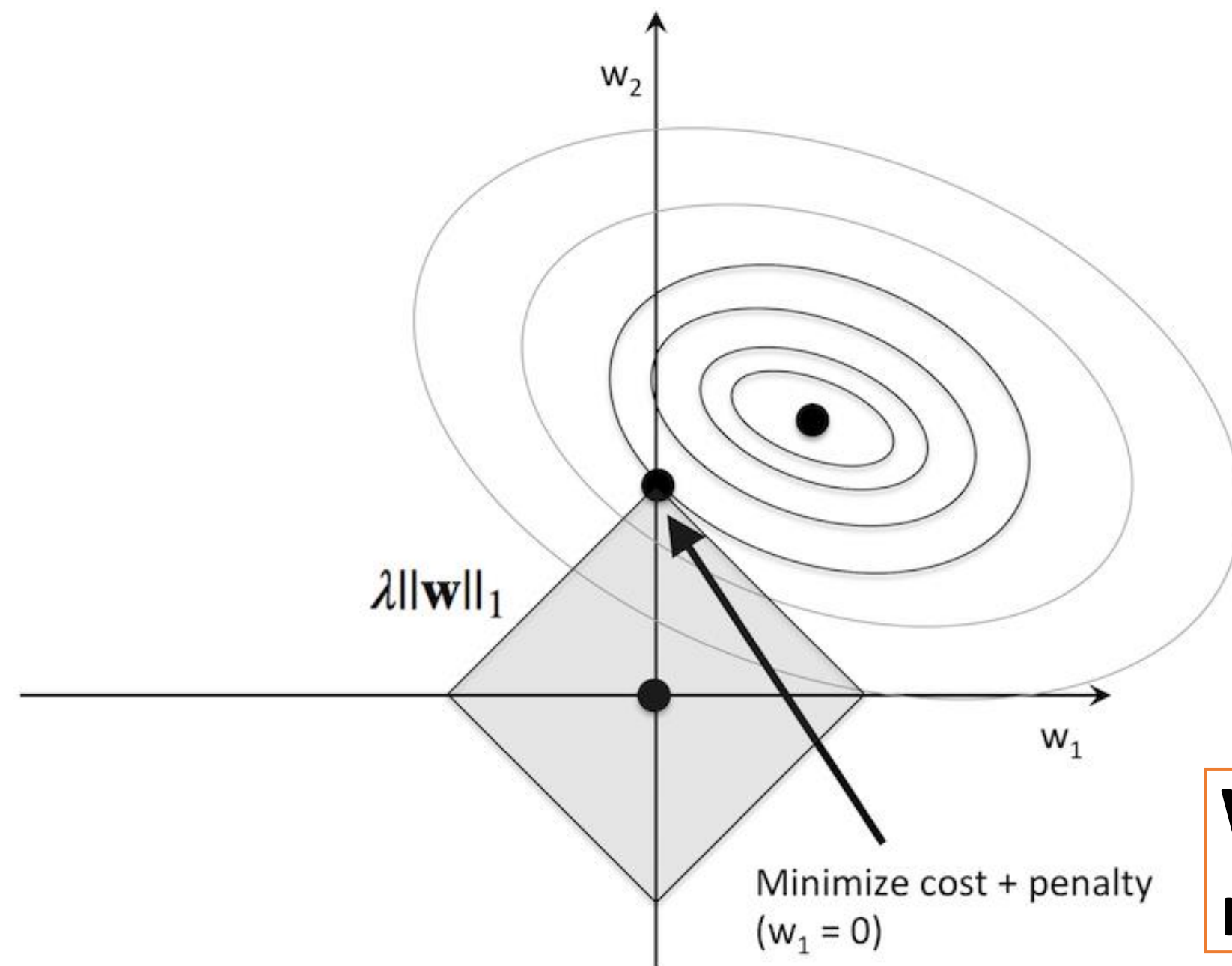
- Recursive Feature Elimination (RFE)
- Sequential Feature Selection (SFS)
- Permutation importance
- ...

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



Cost function adjusted for L1 Regularization



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

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

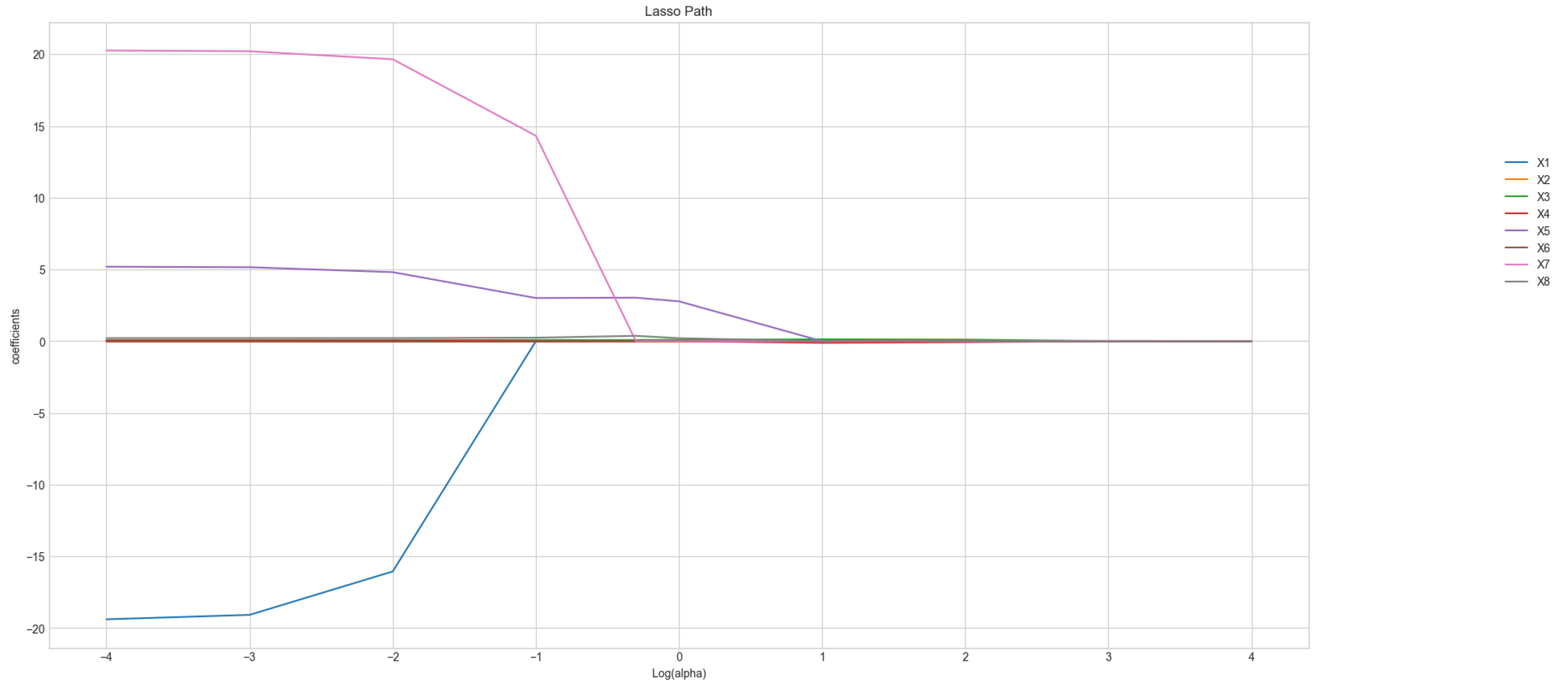
$$\mathbf{w} = \mathbf{w} - \eta \nabla_w \mathcal{J} \quad \mathbf{w} = \mathbf{w} - \eta \nabla_w \mathcal{J} - \eta \lambda$$

**Without
regularization**

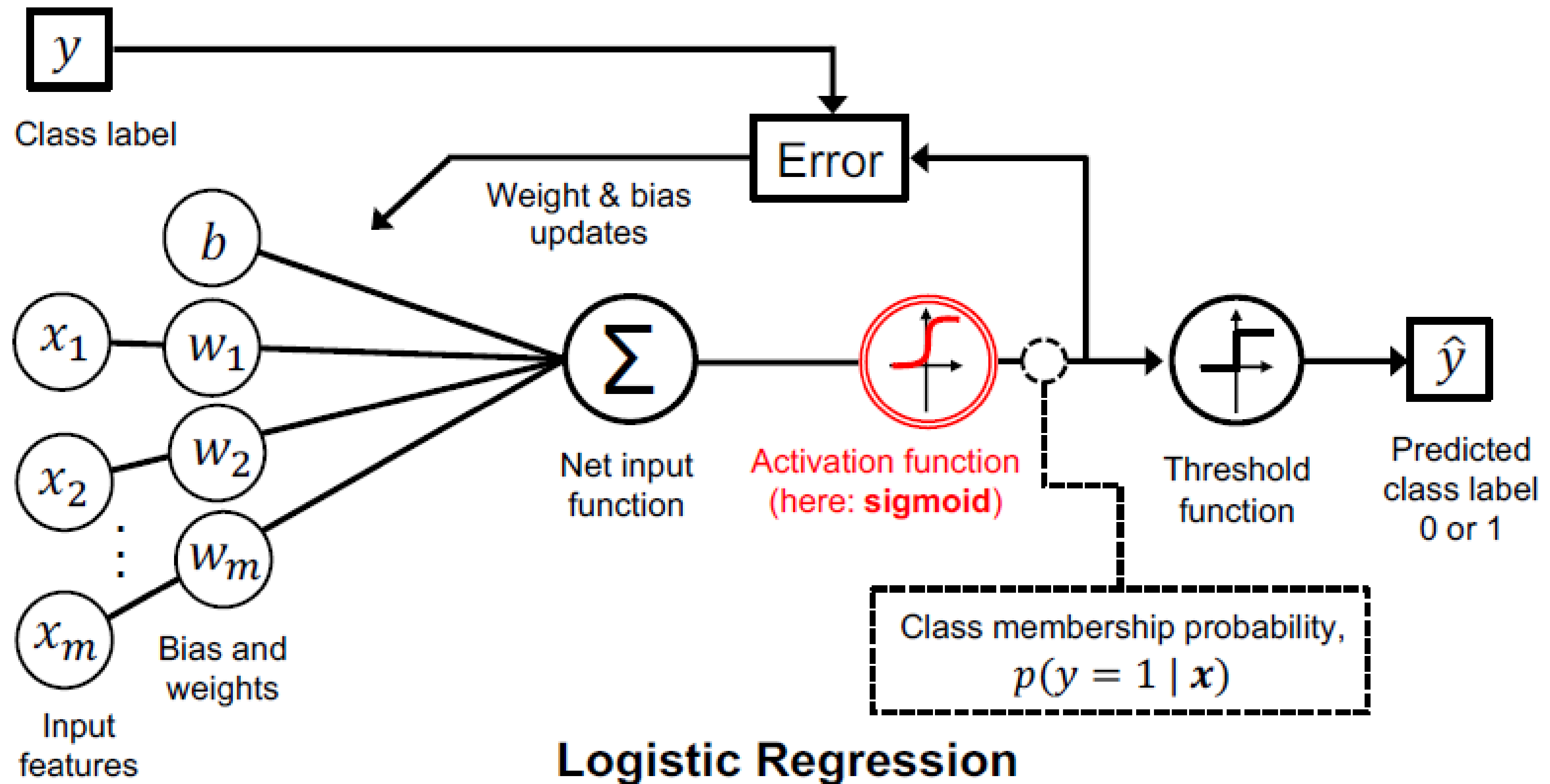
$$\mathbf{w} = (\mathbf{w} - \eta \lambda) - \eta \nabla_w \mathcal{J}$$

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

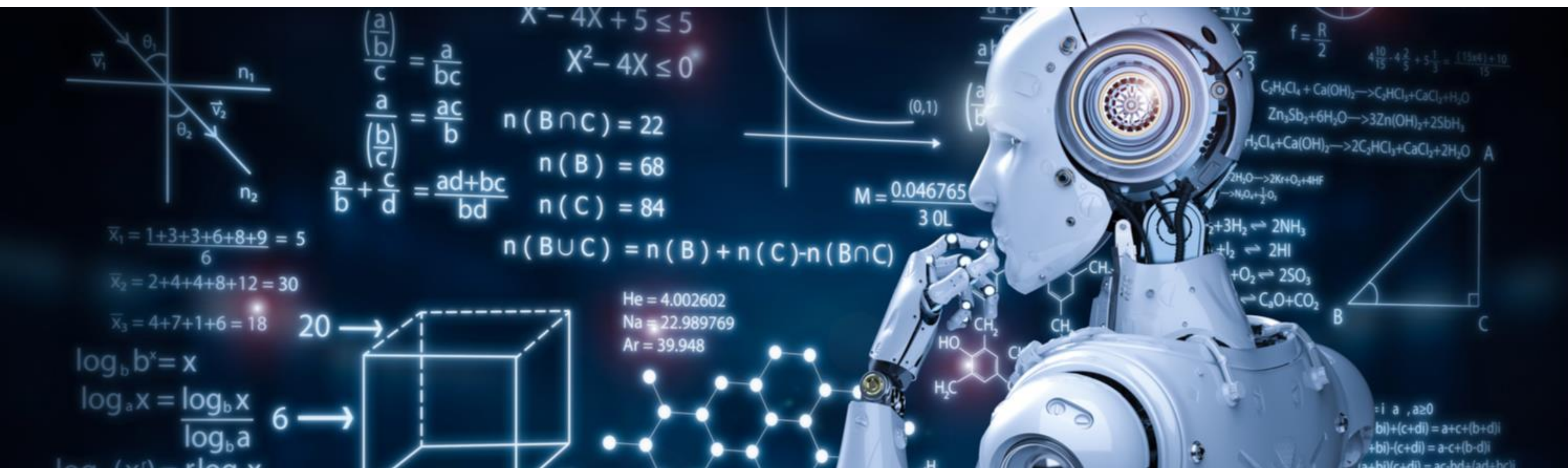
Lasso path



$$\arg \min_w \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|_1 \quad 8$$



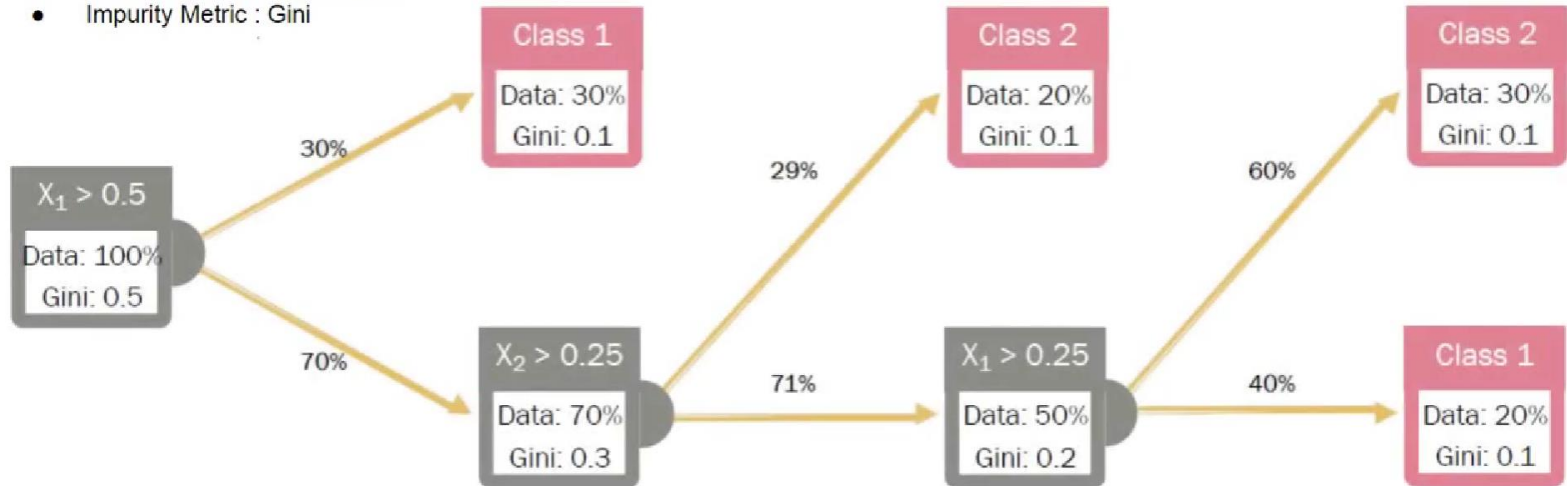
$$J = - \sum_{i=1} \left[y^{(i)} \log \left(\sigma \left(z^{(i)} \right) \right) + \left(1 - y^{(i)} \right) \log \left(1 - \sigma \left(z^{(i)} \right) \right) \right] + \lambda \|w\|_1$$



Feature Selection with Decision Tree & Random Forest

Sample Tree:

- 2 Features : X_1 & X_2
- 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

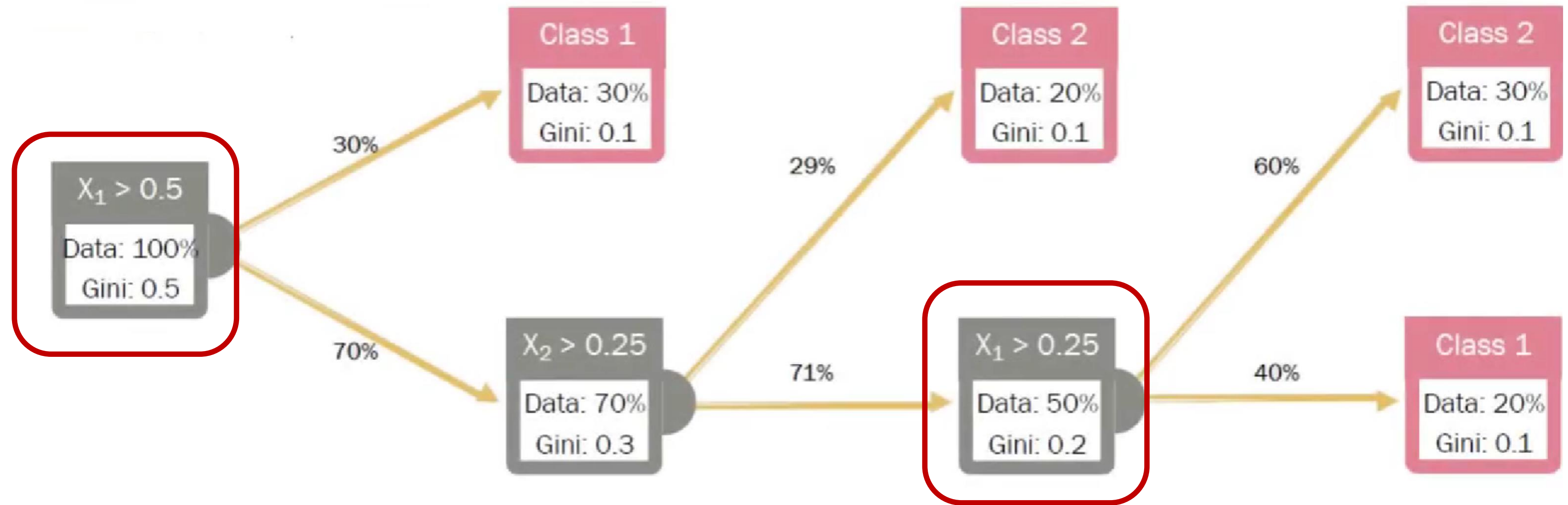


• Node Importance

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

• Feature Importance

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



- Decrease in impurity for X1 at top:
 - Impurity in parent node – impurity in child node
- Weighed by the ratio of data N_{in}/N
- How many times?
- 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)

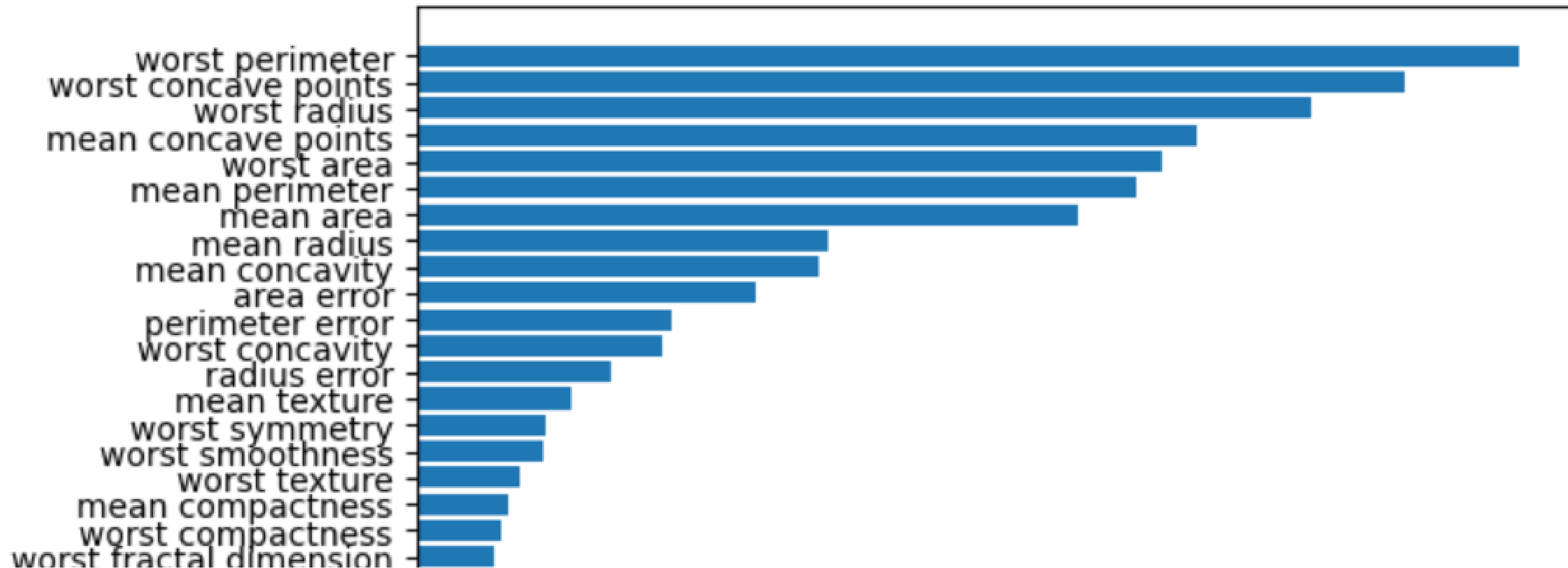
- for a given feature
 - for each tree
 - compute impurity decrease (Gini, Entropy)
 - weight by number of examples at that node
 - averaged over all trees
 - normalize importances so that sum of feature importances sum to 1

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

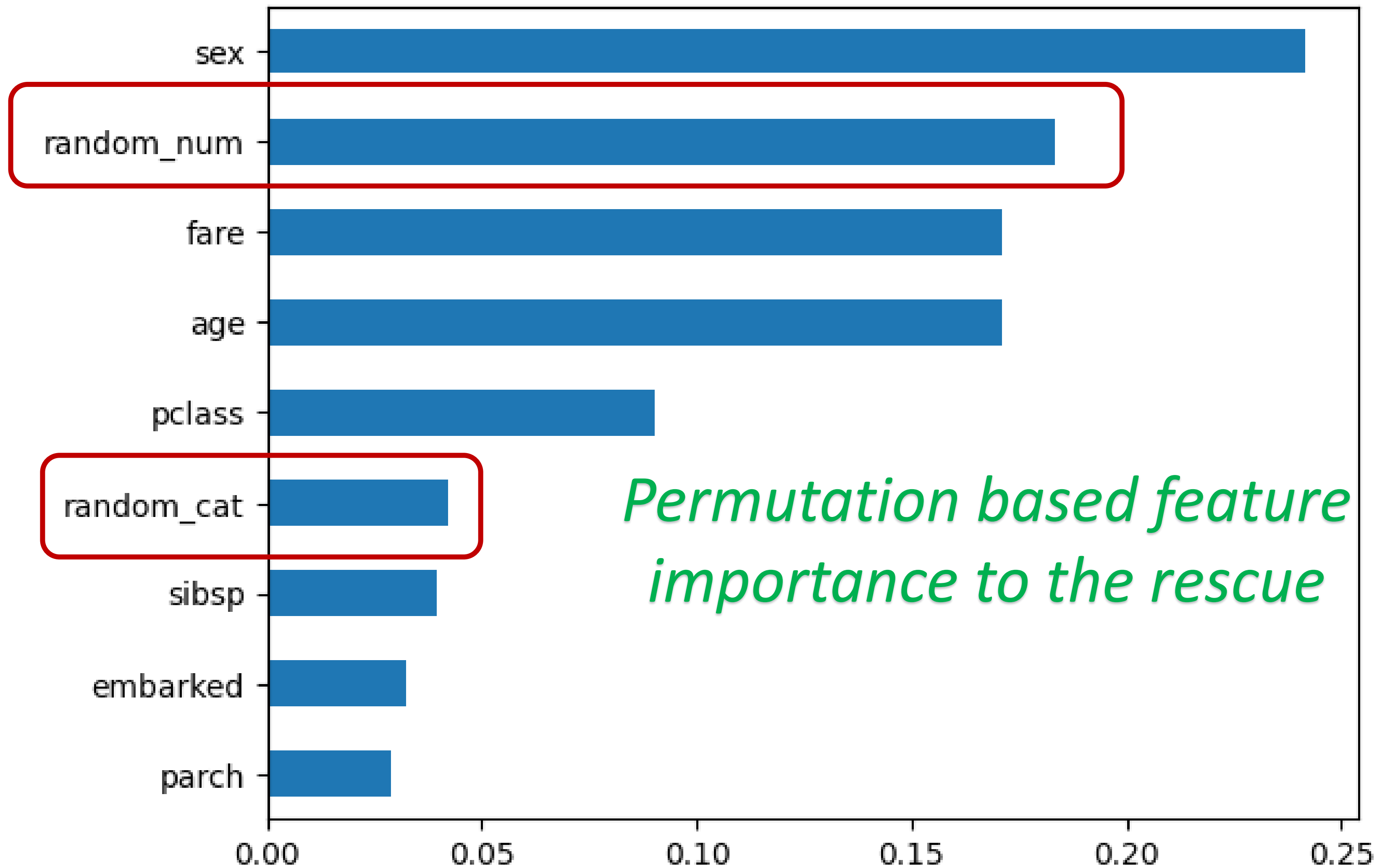
```
Text(0.5, 0, 'Random Forest Feature Importance')
```



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*



Dimensionality Reduction

Feature Selection

Filter Methods

- Information gain
- Correlation with target
- Pairwise correlation
- Variance threshold
- ...

Embedded Methods

- L1 (LASSO) regularization
- Decision tree
- ...

Wrapper Methods

- Recursive Feature Elimination (RFE)
- Sequential Feature Selection (SFS)
- Permutation importance

Wrapper methods

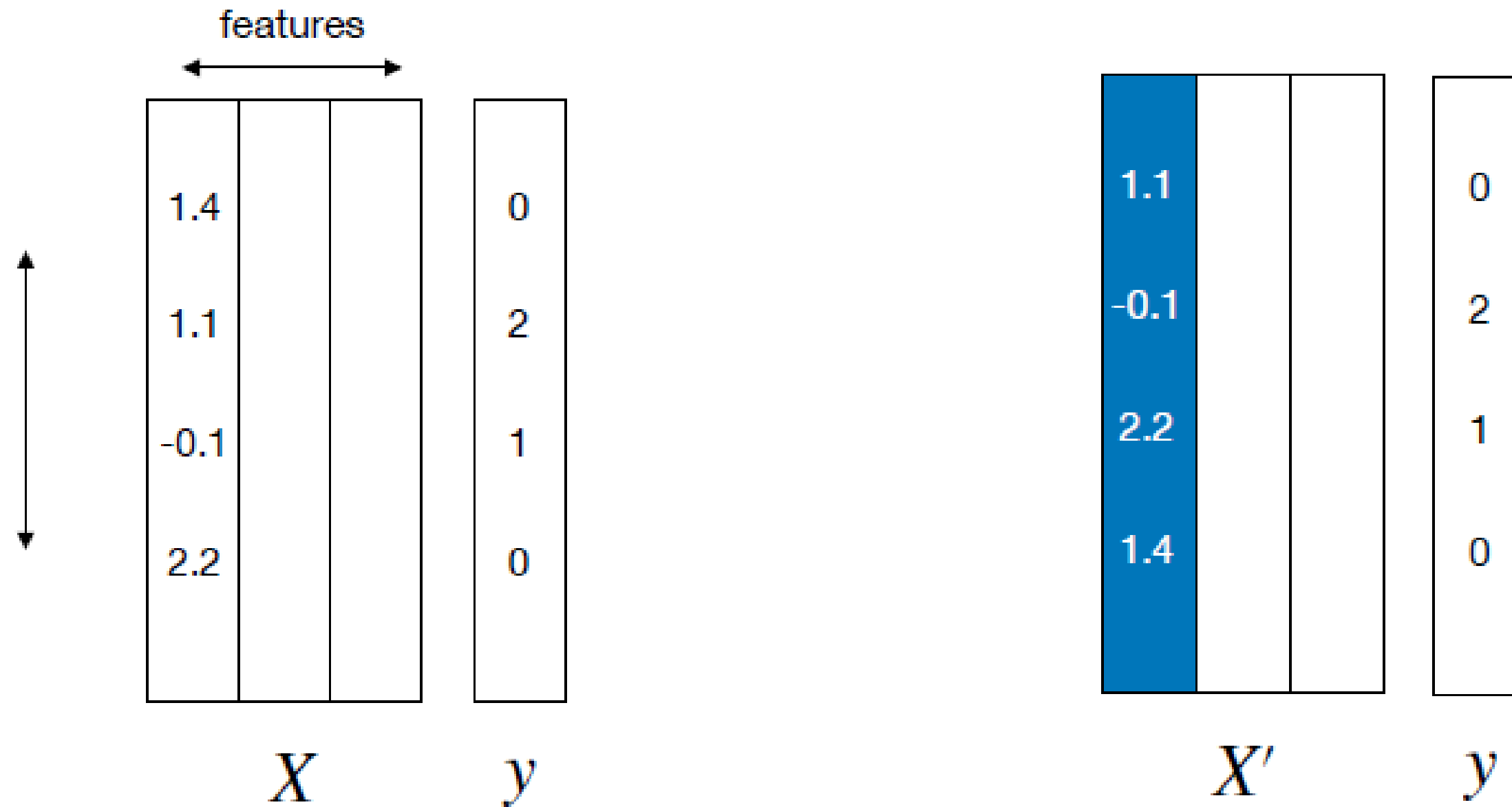
- General working
 - Wrap any algorithm to measure metrics
 - Make slight changes and run the algorithm again
 - The quantum of “slight changes” gives idea of feature importance
- E.g.:
 - RFE
 - Permutation Importance
 - Selective Feature Selection



Permutation based feature importance

Permutation importance

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

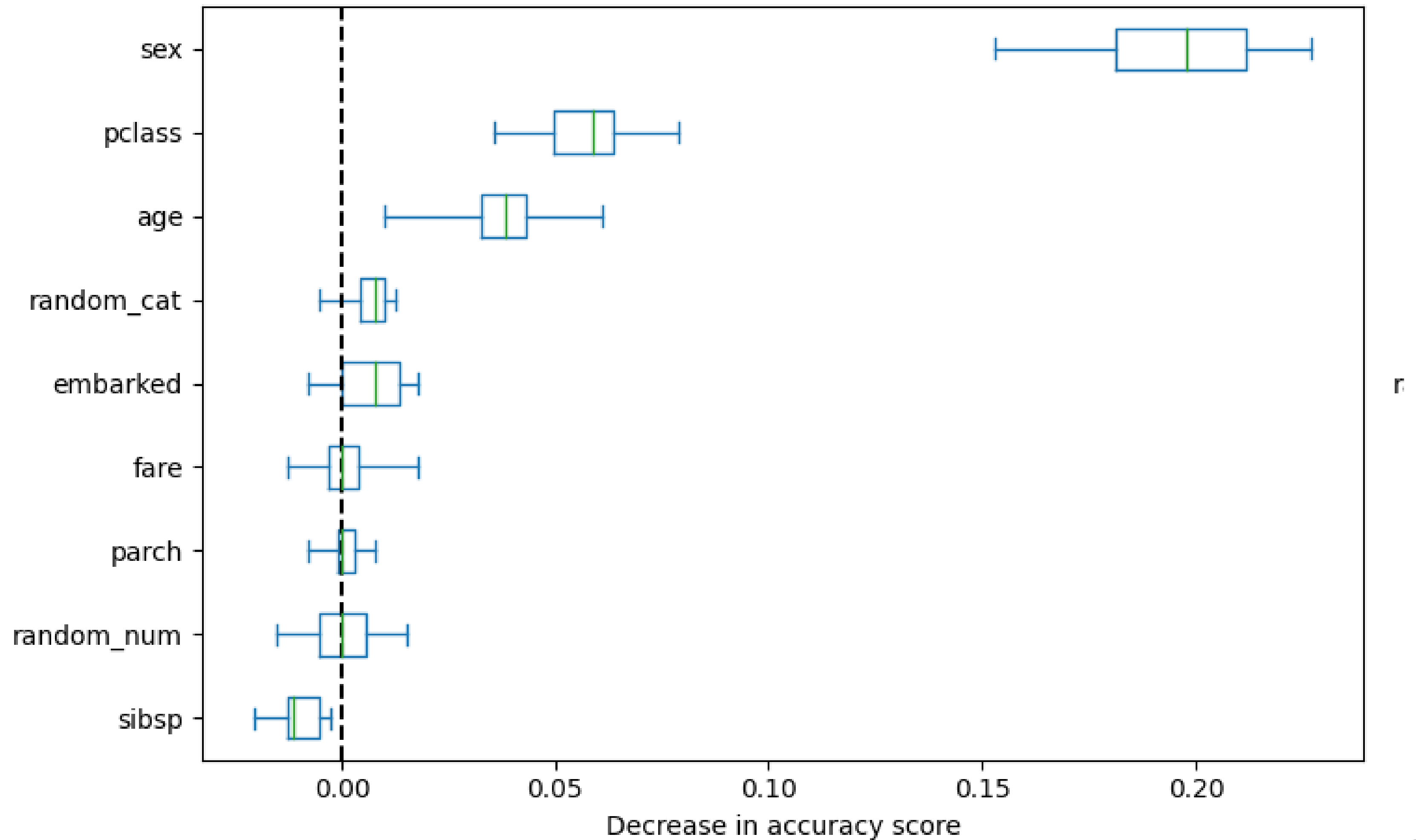


feature importance =
baseline performance – shuffled dataset performance

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)



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

- Information gain
- Correlation with target
- Pairwise correlation
- Variance threshold
- Chi-Squared ANOVA

Embedded Methods

- L1 (LASSO) regularization
- Decision tree
- ...

Wrapper Methods

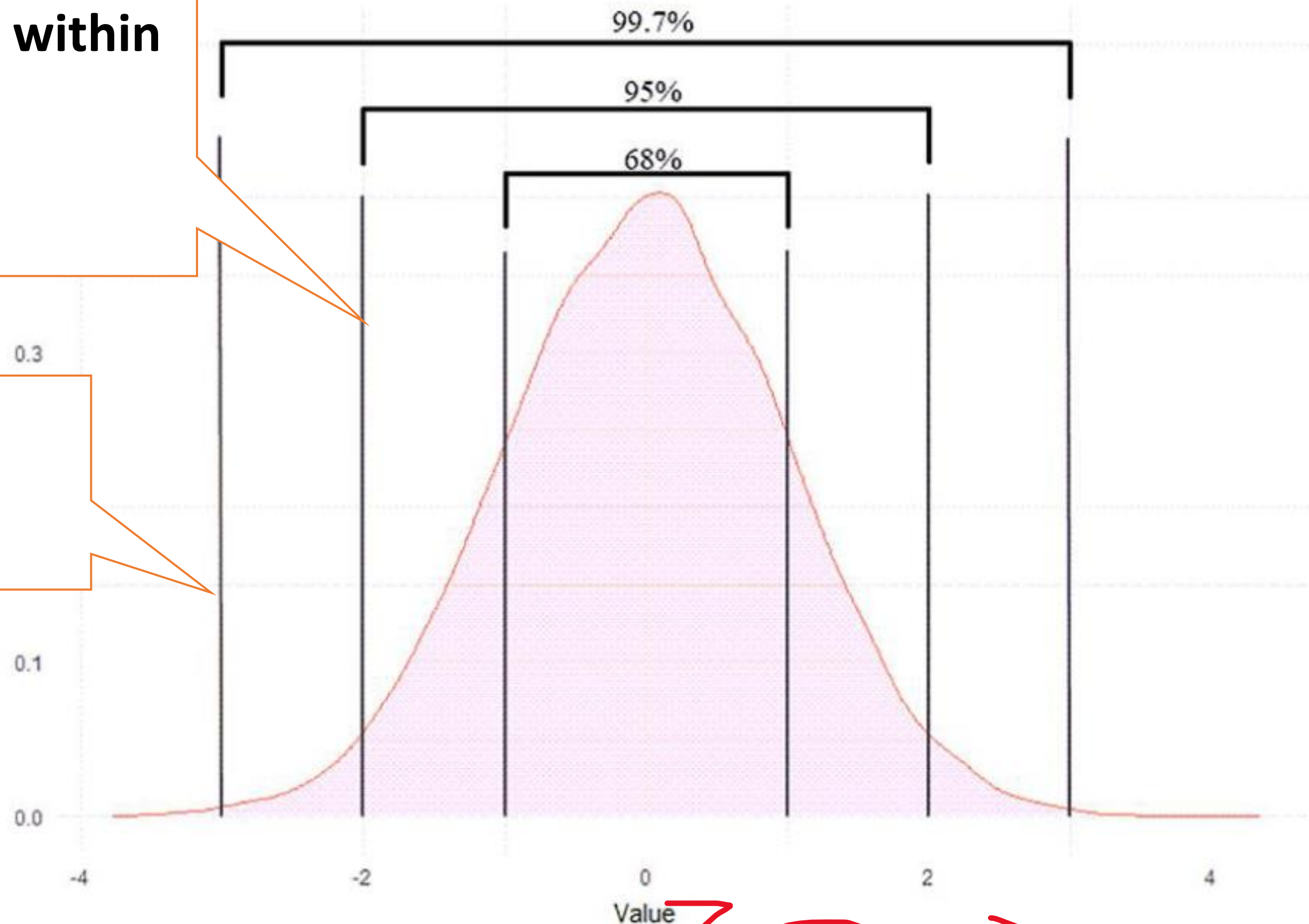
- Recursive Feature Elimination (RFE)
- Sequential Feature Selection (SFS)
- Permutation importance
- ...

Standard Normal Distribution

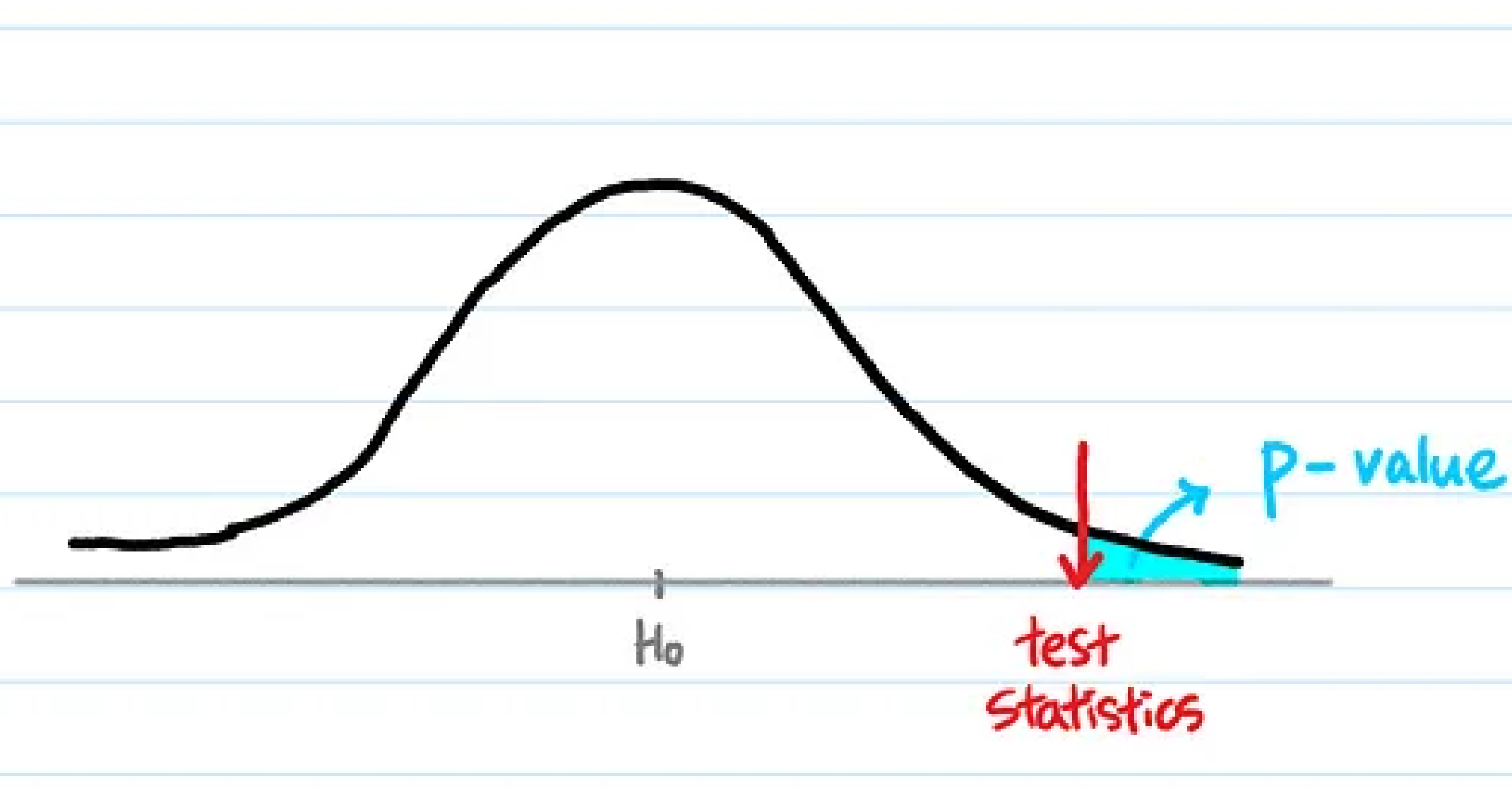
Probability of data within

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

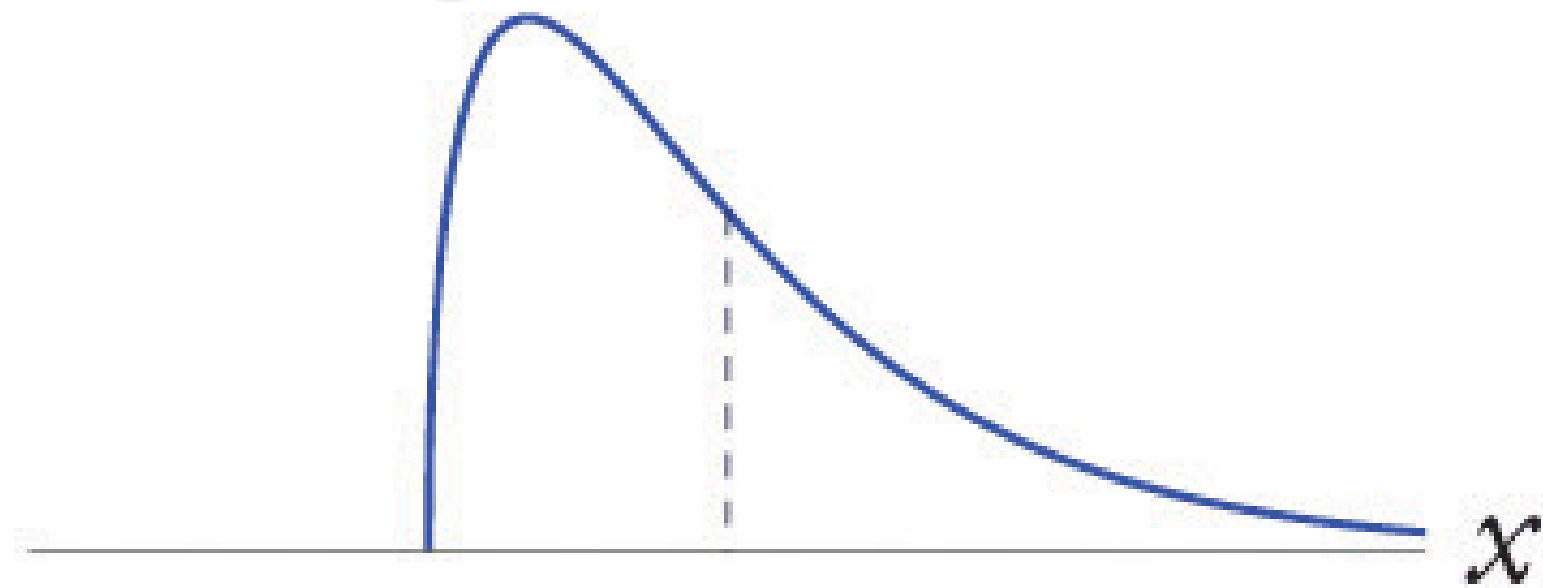


- $P(X > k)$

Why normal distribution?

- Not everything follows normal distribution

Height distribution



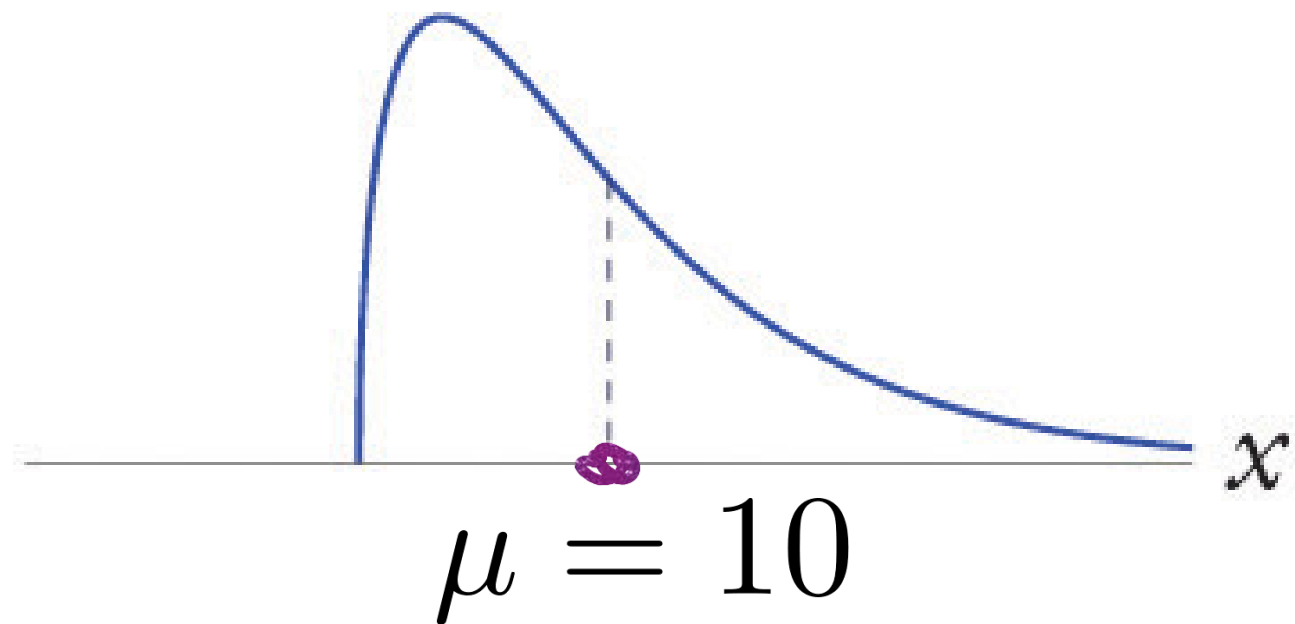
- Accidents on road follow Poisson distribution
- Volcanic eruption, asteroid strike: exponential distribution
- Why do we keep talking about normal distribution?
- The answer is in Central Limit theorem

Central Limit Theorem

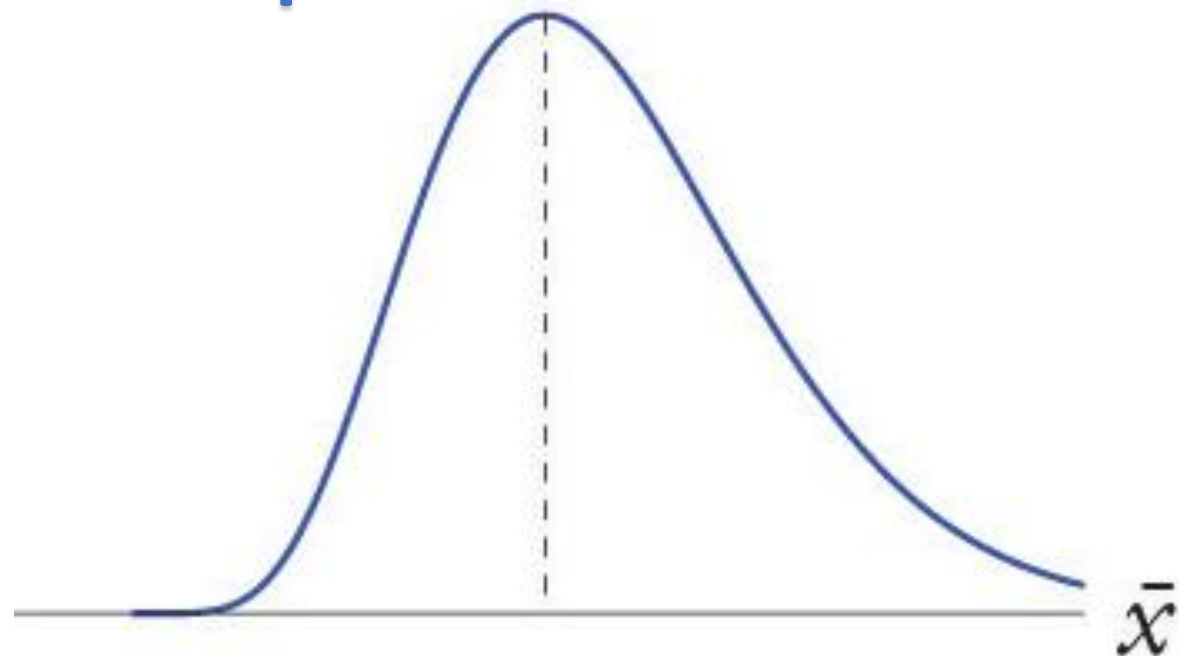
- Most fundamental to inferential statistics
- Provides convenient mechanism to apply normal distribution to everything

Central Limit Theorem – Sampling distribution of mean

Population distribution



Sampling distribution of sample mean with $n=5$

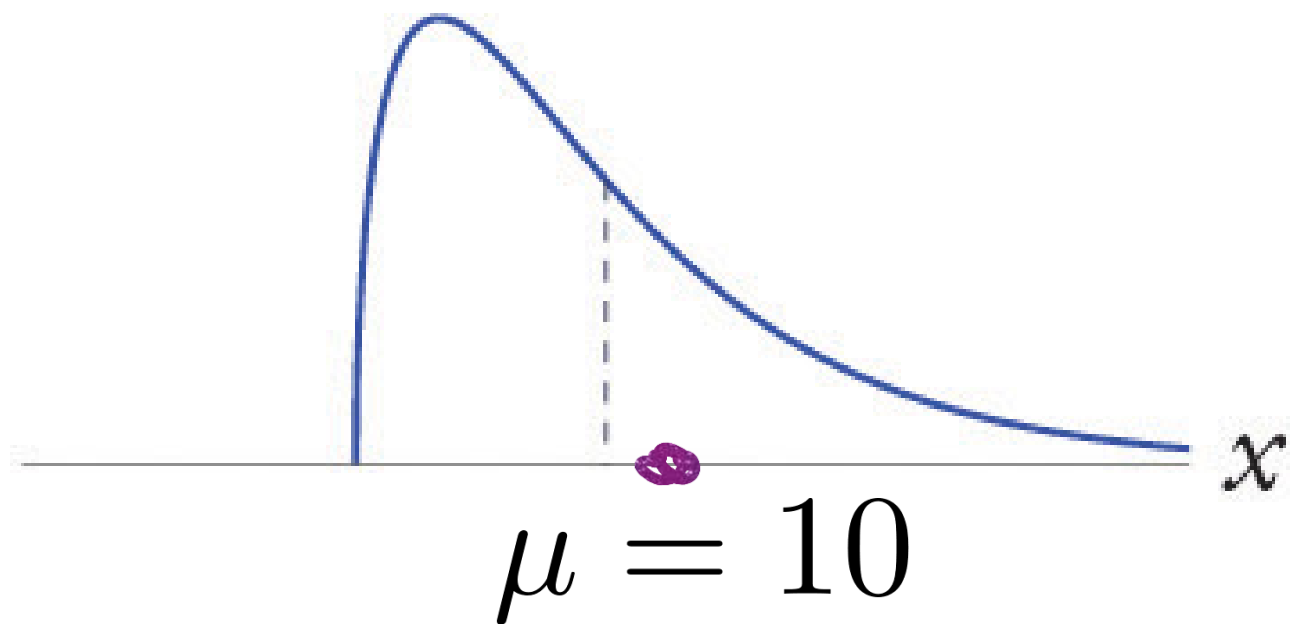


Sample size $n = 5$

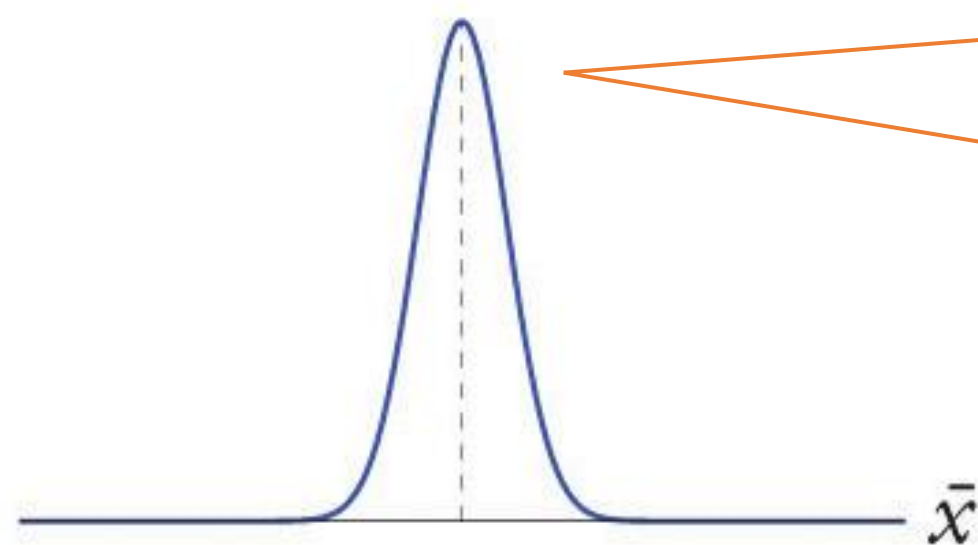
Sampling	Mean
1	7.5
2	9.5
3	11
4	9
5	11.5
6	10.5
7	9.75
8	9
9	9.25
10	9.8

Central Limit Theorem – Sampling distribution of mean

Population distribution



Sampling distribution of sample mean with $n=30$



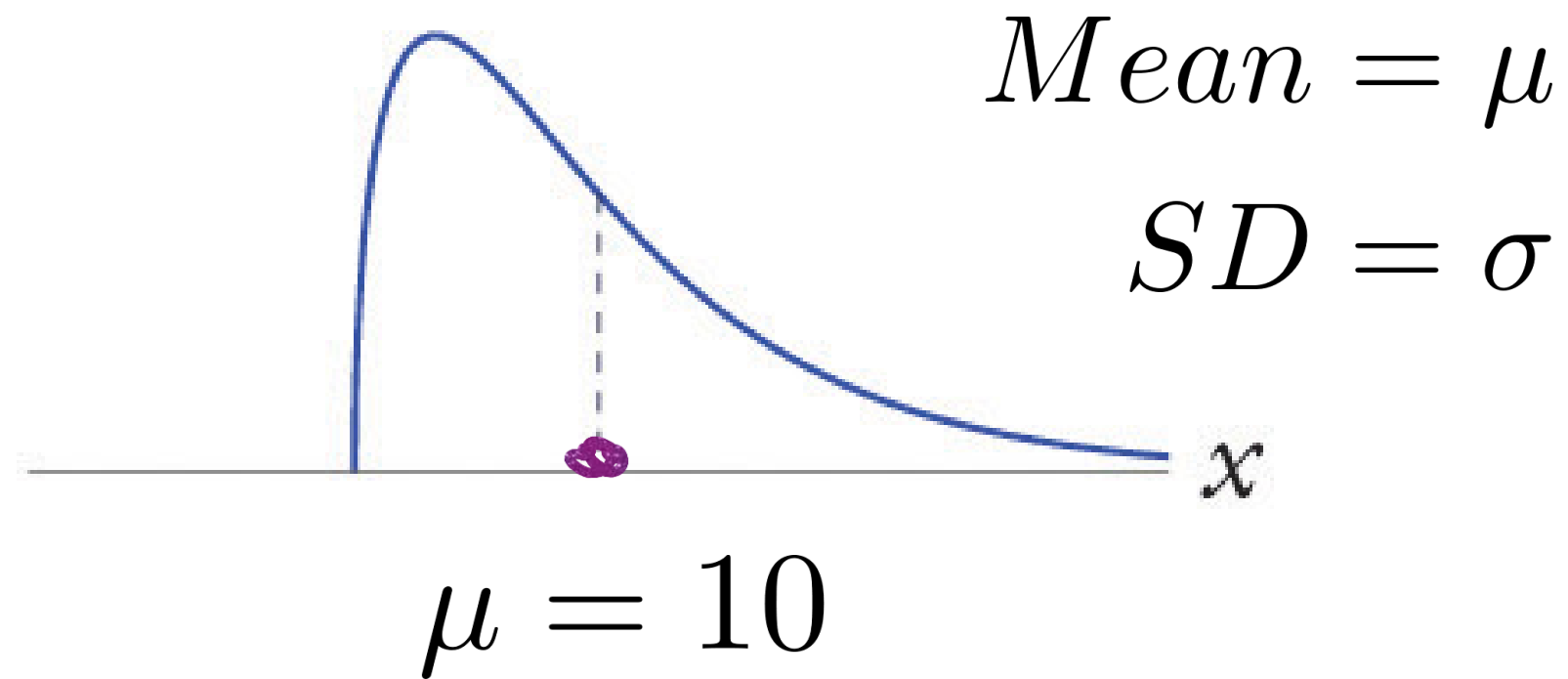
Sampling distribution approximates normal distribution when $n \geq 30$

Sample size $n = 30$

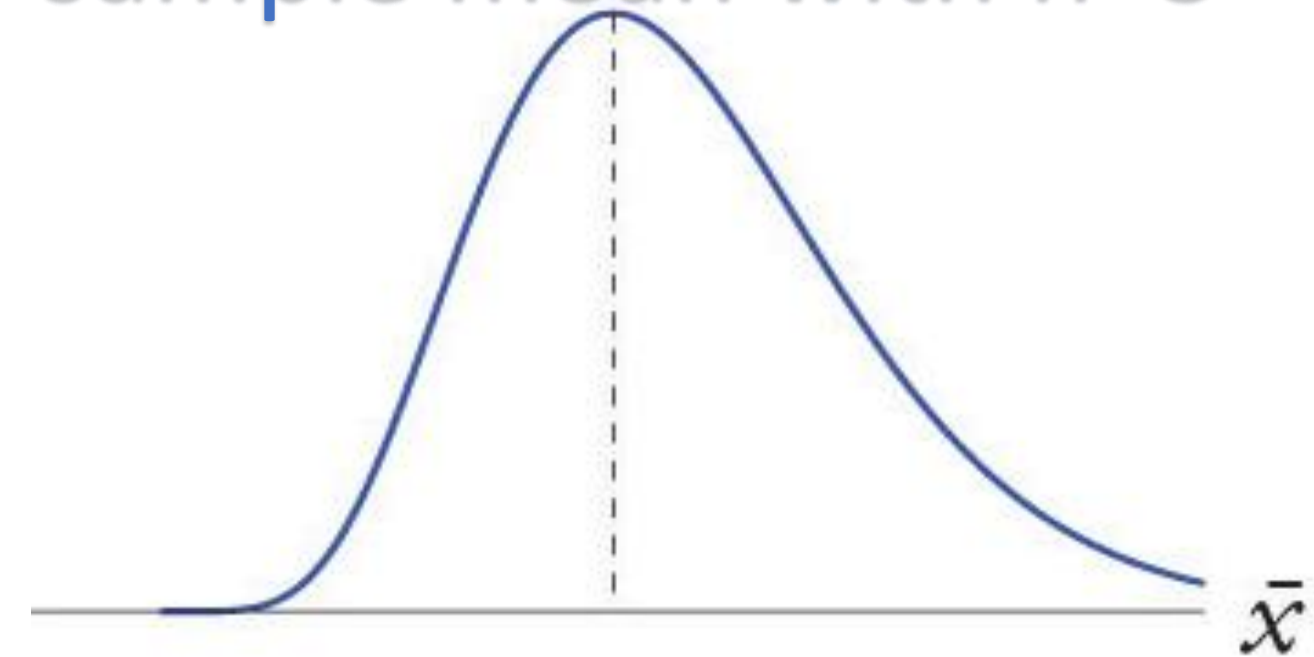
Sampling	Mean
1	7.5
2	9.5
3	11
4	12
5	11.5
6	10.5
7	9.75
8	10
9	10.25
10	9.8

Central Limit Theorem – Sampling distribution of mean

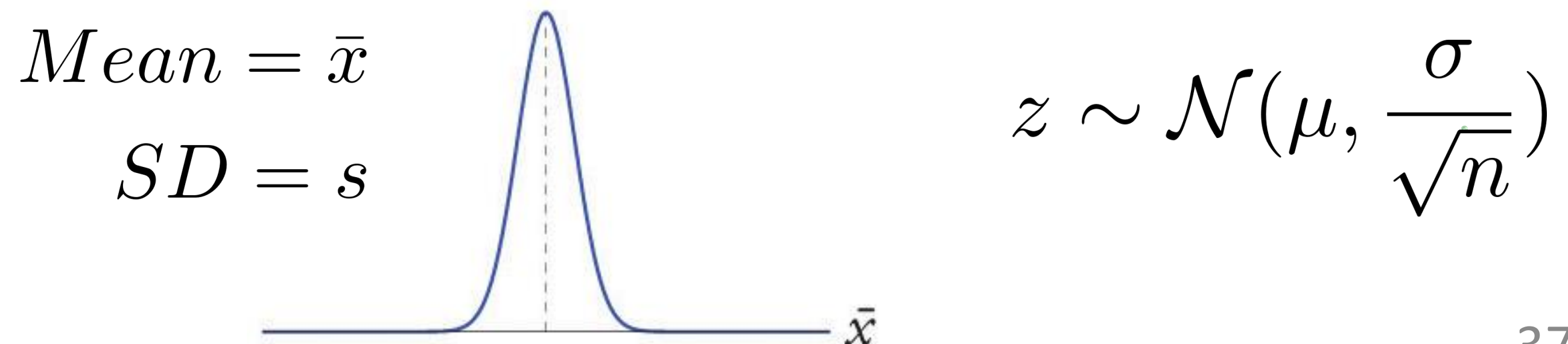
Population distribution



Sampling distribution of
sample mean with $n=5$



Sampling distribution of sample mean with $n=30$

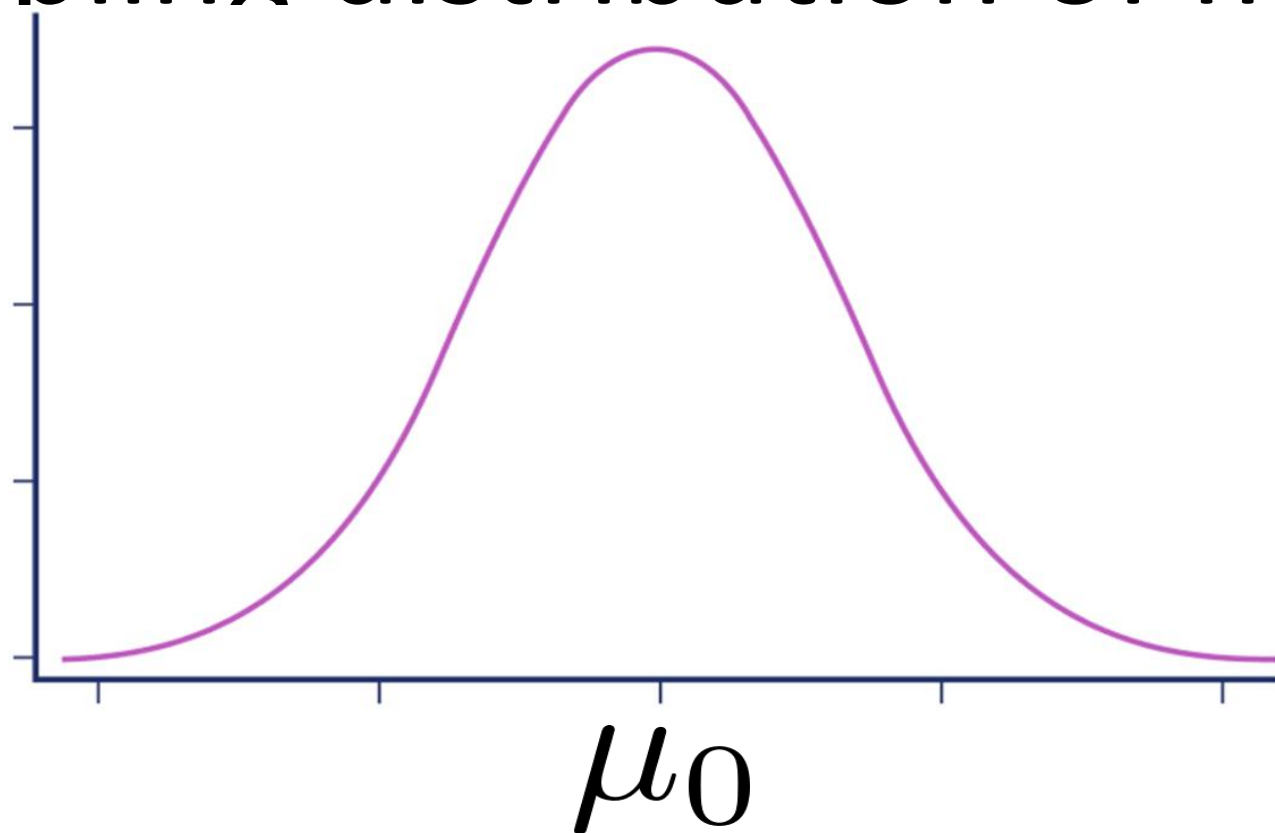


Made in India iPhone 15 battery life

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

$$X \sim \text{Unknown}(\mu_0, \sigma_0)$$

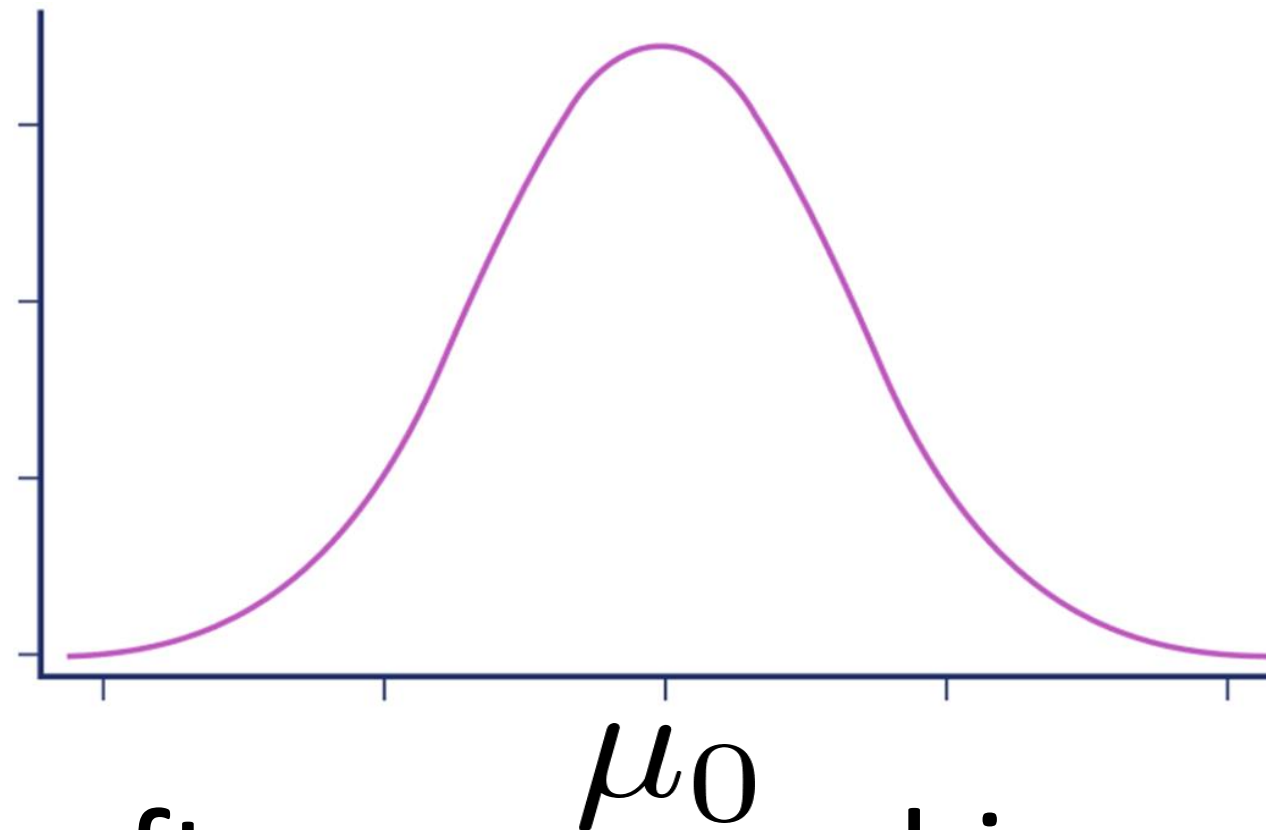
- How was population mean & variance found?
 - Draw samples (each of size ≥ 30) 10 times in population
 - Sampling distribution of mean battery life is gaussian



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

Made in India iPhone 15 battery life

- Historical sampling distribution



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

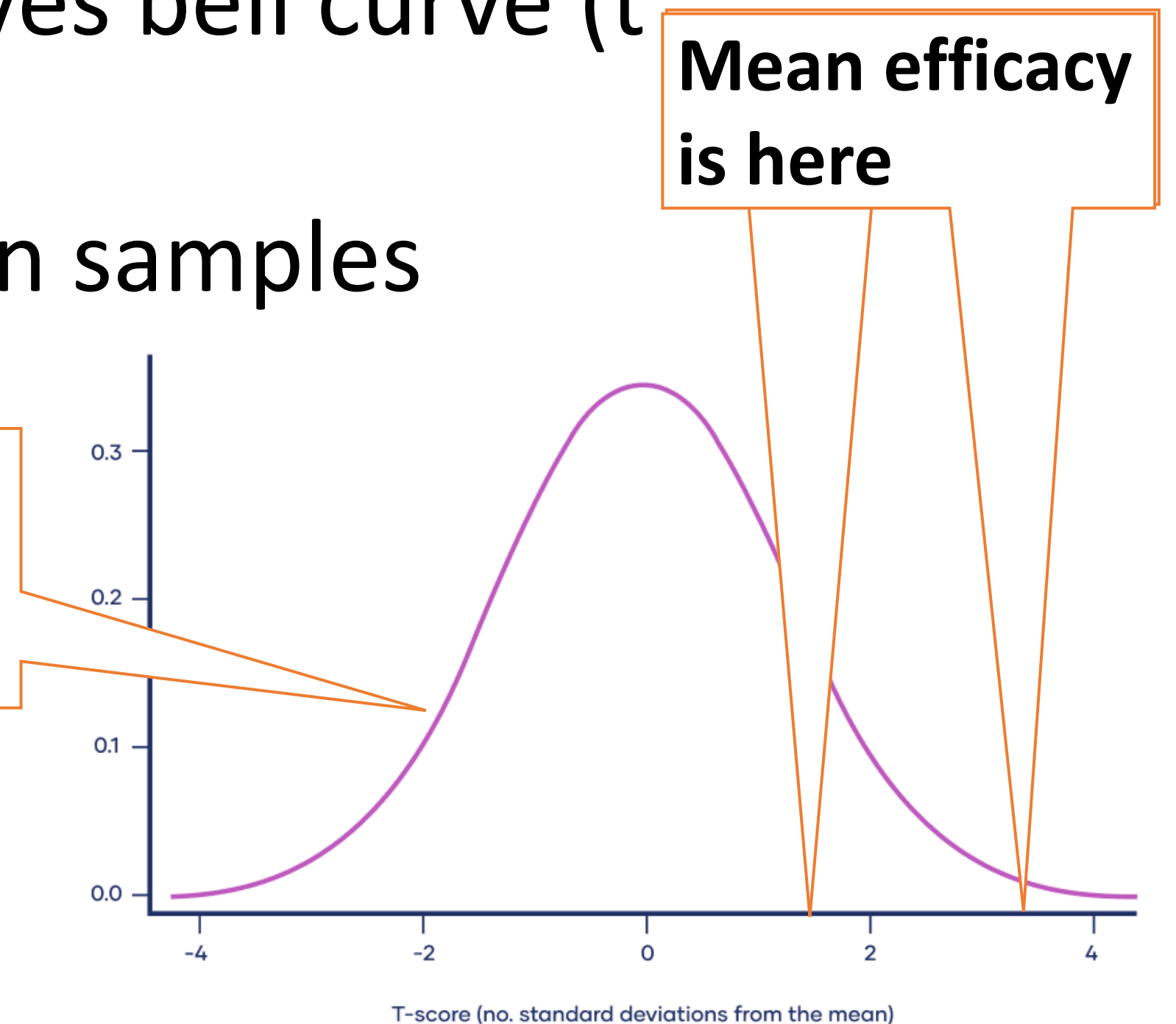
- v/s distribution after new machinery is installed
- Should new distribution exactly align with historical distribution? What is confidence interval?
- Where is the mean of new distribution?

Clinical Trial of a new cancer medicine

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

Efficacy of current medicine follows t distribution

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



T Distribution

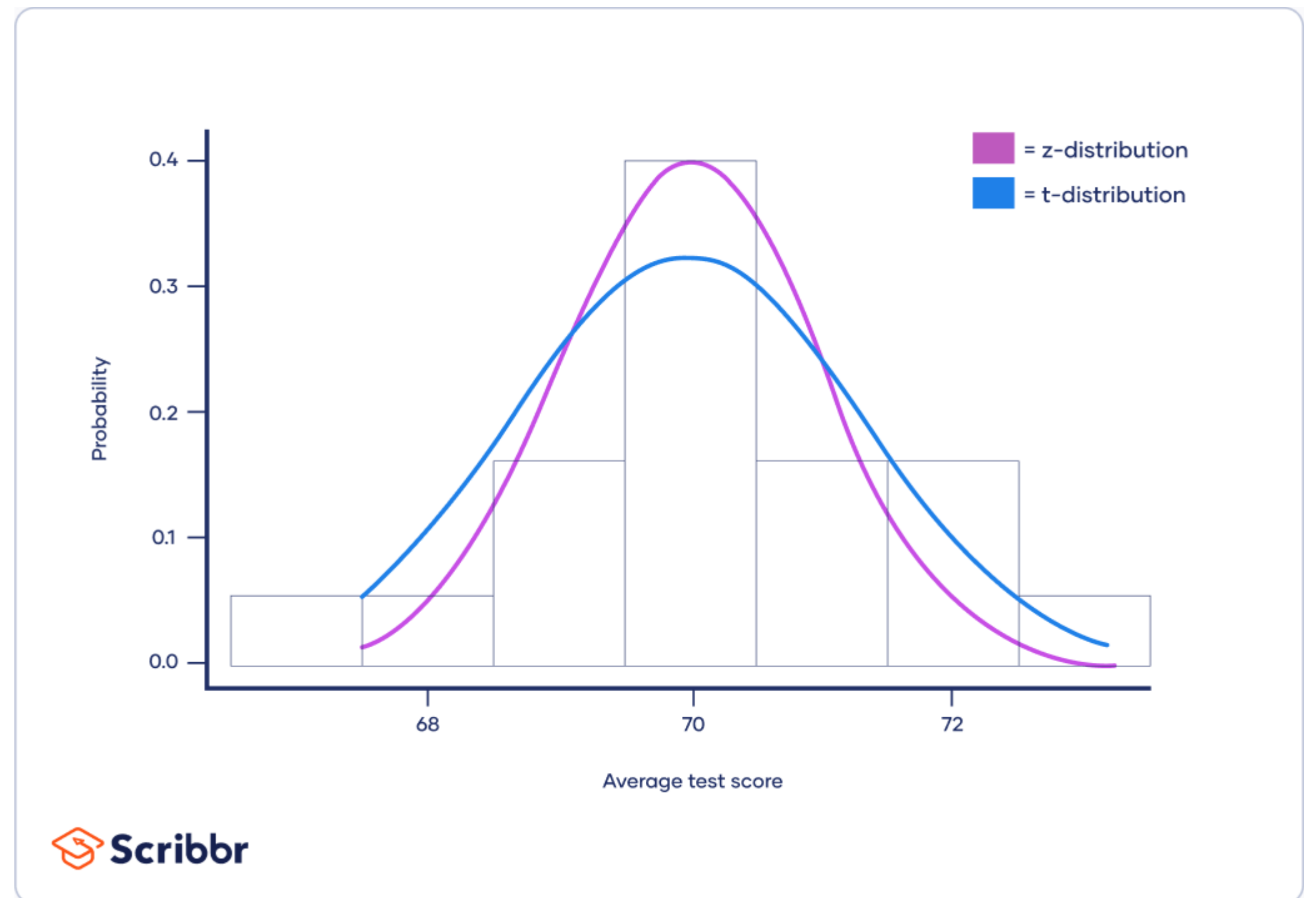
- Adjusted normal distribution for small sample sizes
- Number of samples = Degrees of freedom

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

z statistic follows from this

t statistic follows from this

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

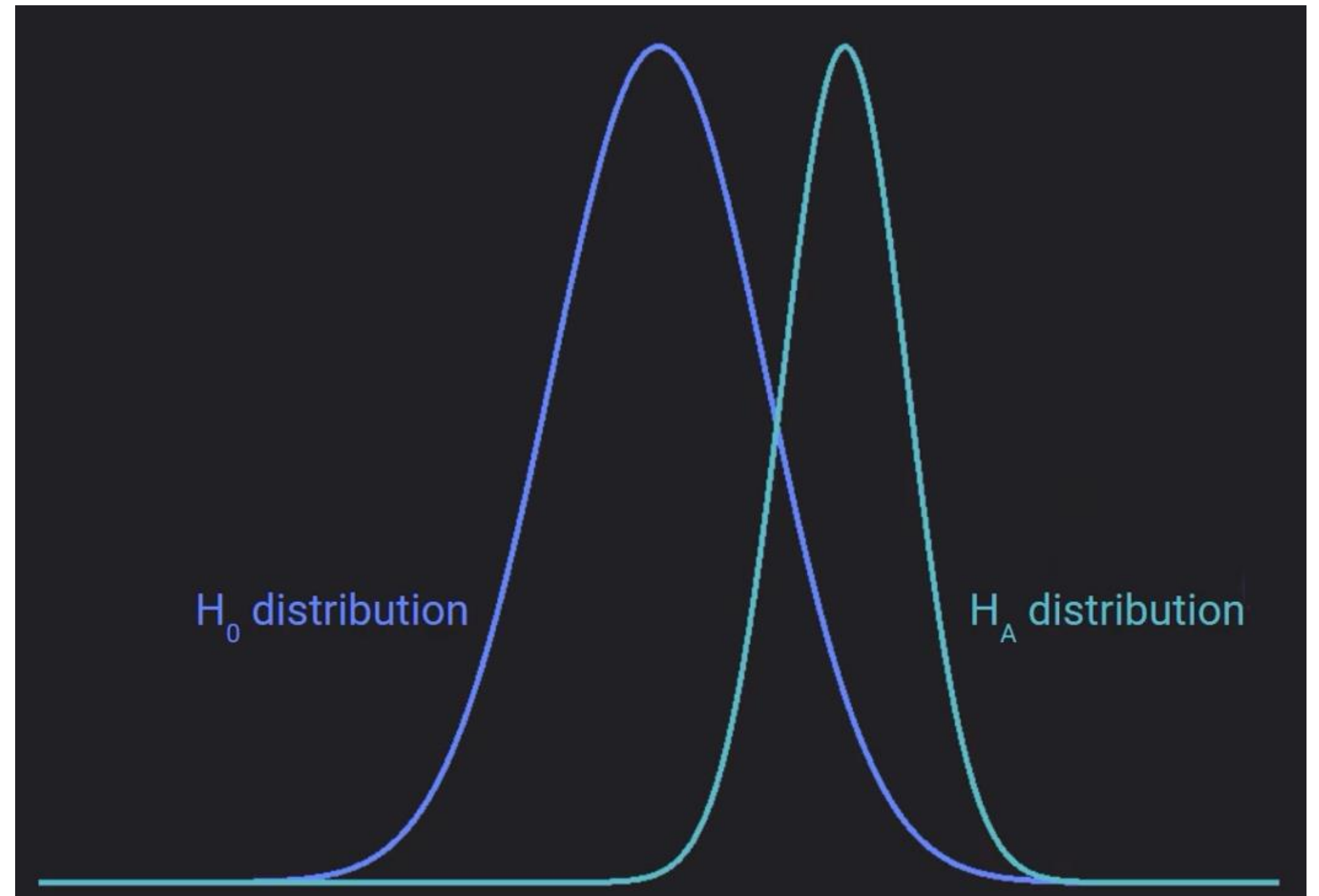


Intro to Hypothesis Test

- Null hypothesis – H_0 ,
- Alternate hypothesis – H_a

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

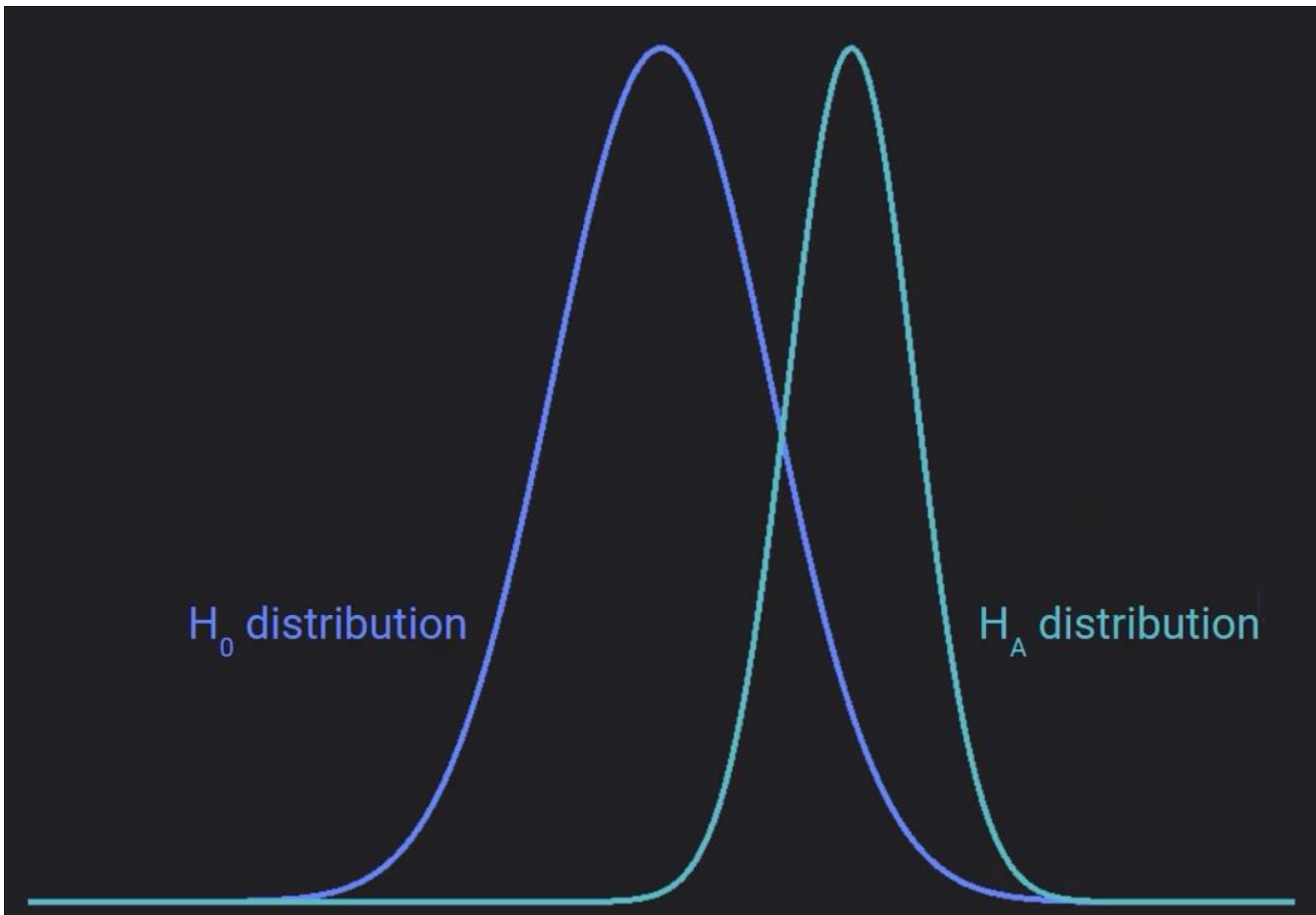
$$t_{n-1} \sim \mathcal{T}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



- Convert x to z/t

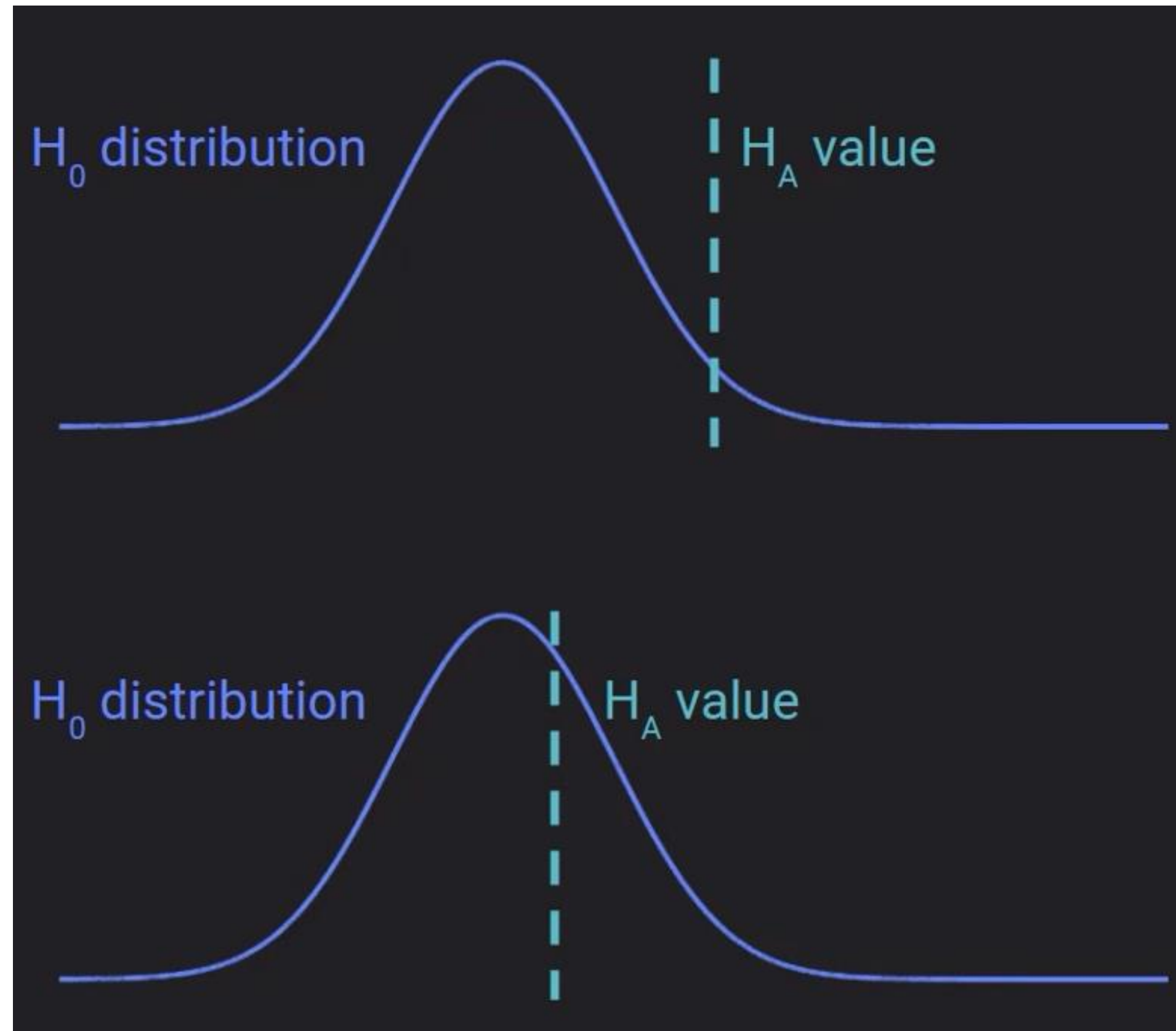
$$t_{n-1} = \frac{x - \mu}{\sigma / \sqrt{n}}$$

Intro to p-values in Hypothesis Test



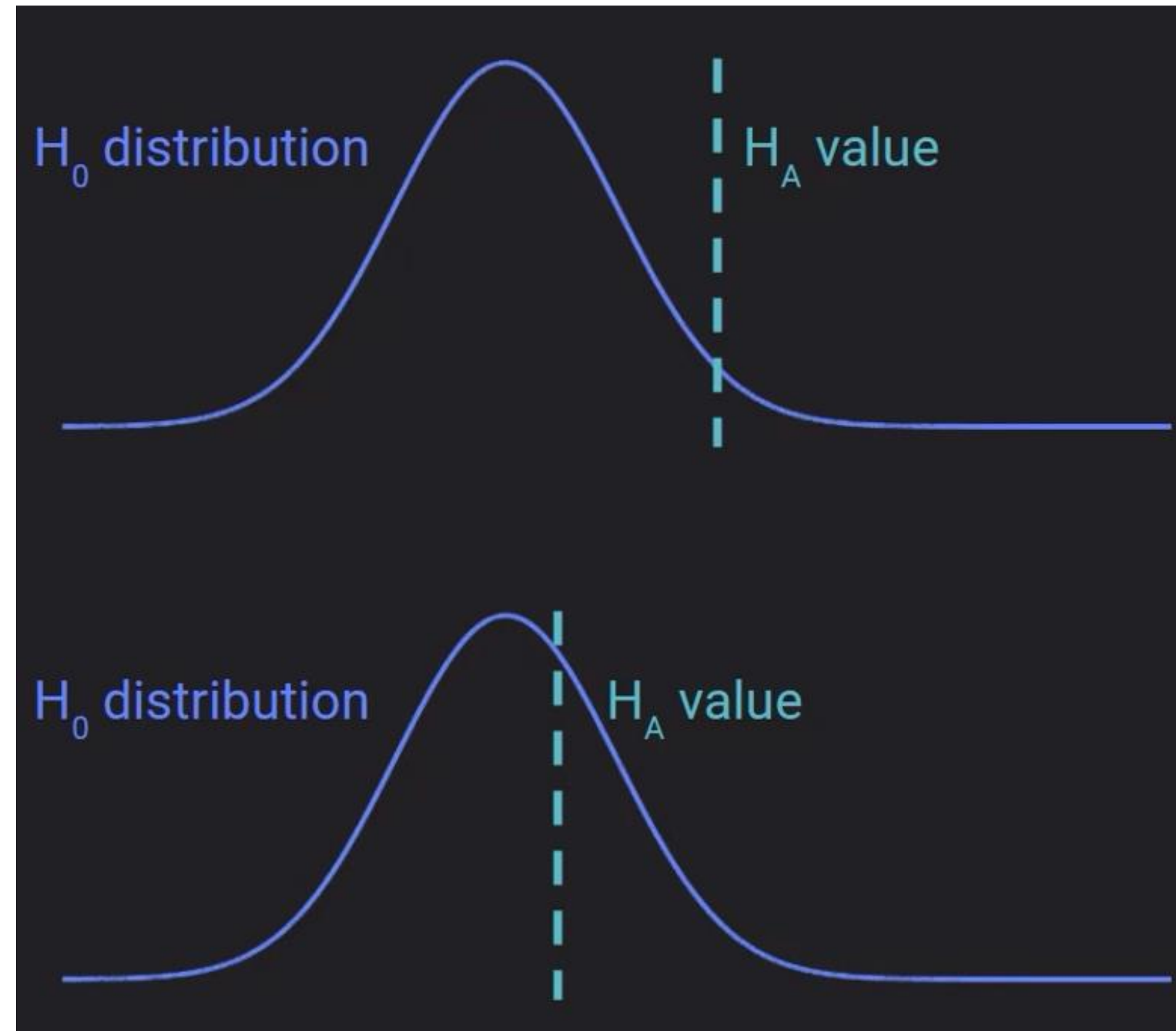
Indicates what is
the probability
value \geq t-statistic

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



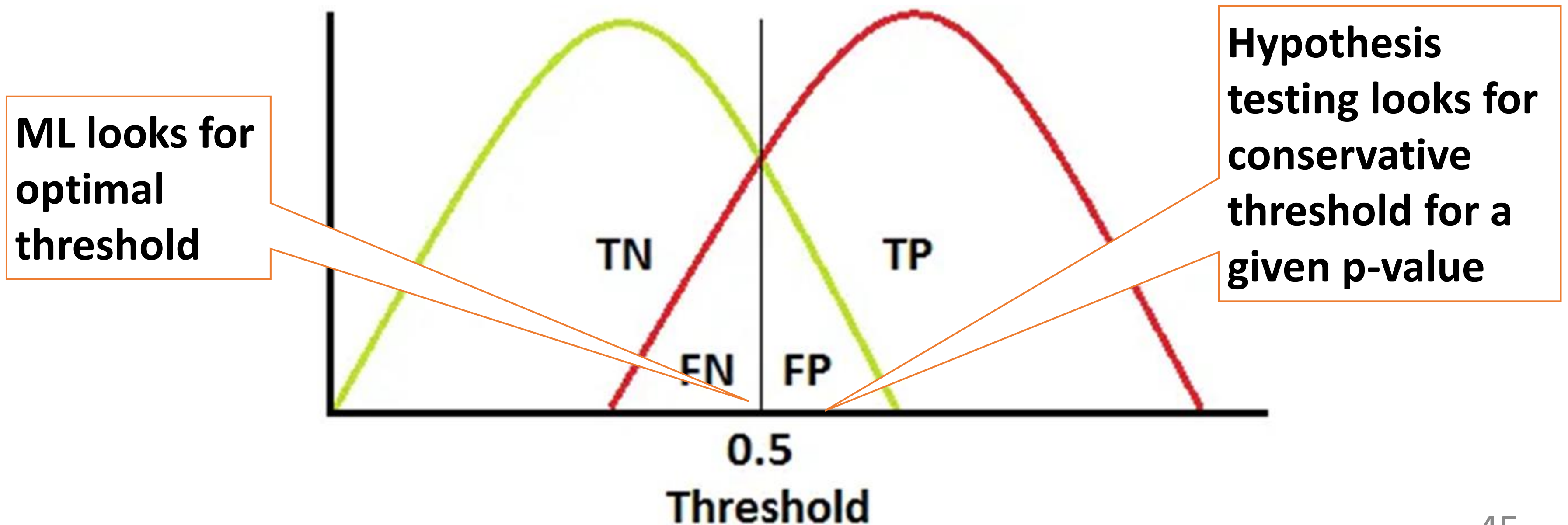
p-values in Hypothesis Test

- How likely for the H_A value to occur if H_0 is true?
- What is the probability of observing a value of H_A larger than current value if there was no true effect?
- $P(H_A \mid H_0)$
- Definite Integral
 - Lookup table or coding



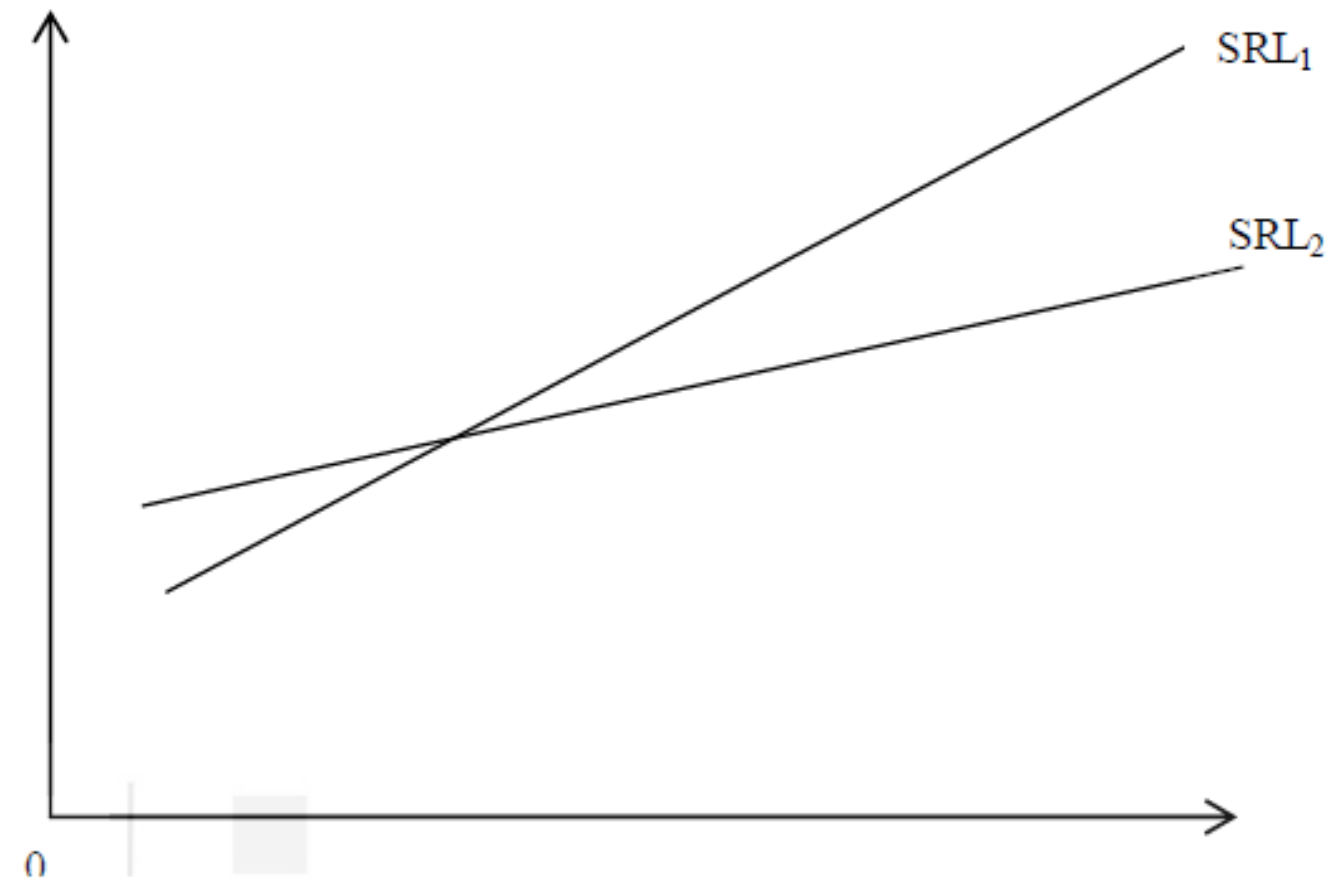
Hypothesis Test in Machine Learning

- p values & statistical significance
- Machine Learning
 - Prediction on each X_{test} record is hypothesis testing



Regression - Population versus Sample

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



**Hypothesis
function(s)**

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

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

Null Hypothesis is coefficient = 0

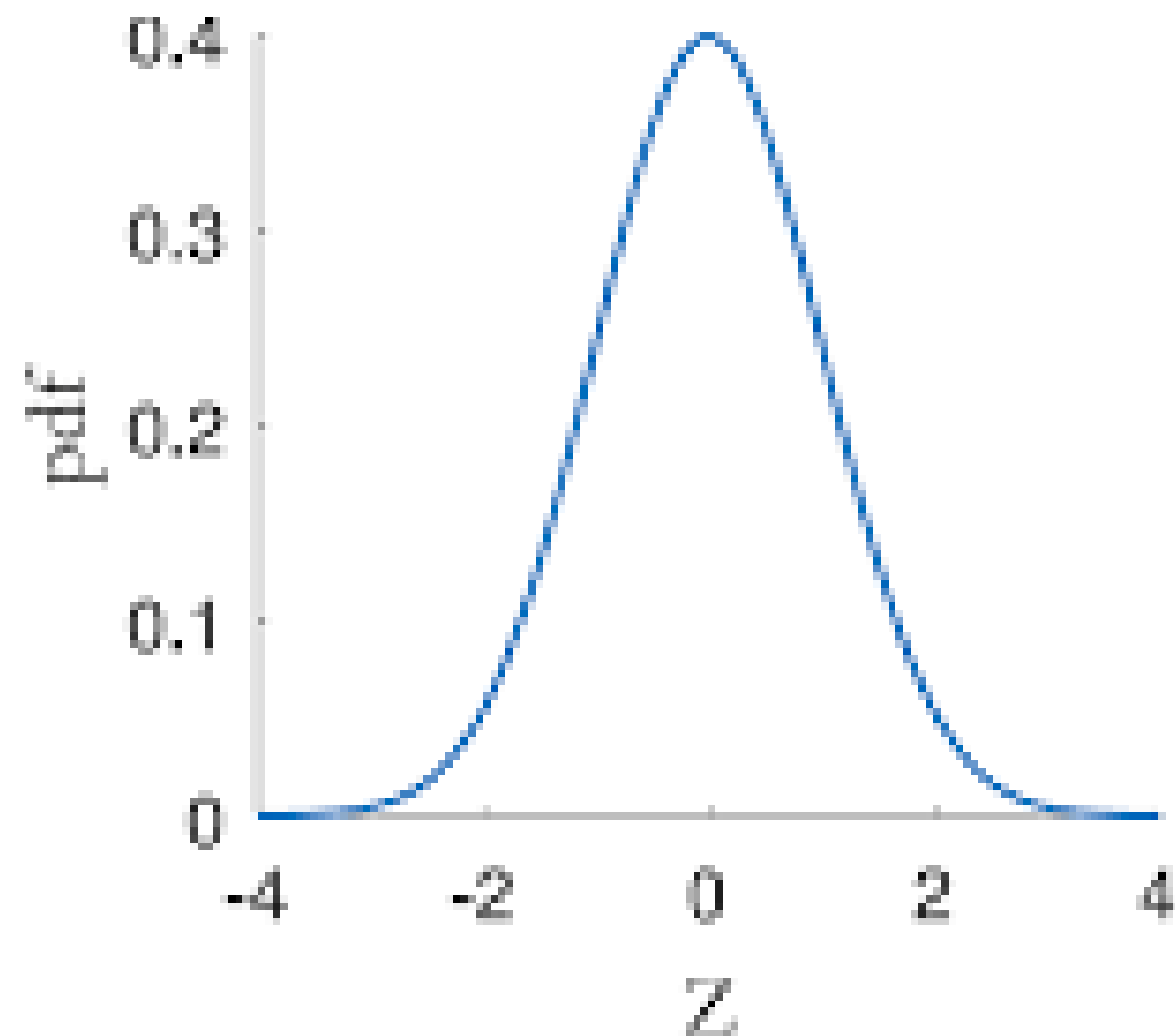
Alternate Hypothesis is coefficient not 0

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	0.0550	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			

Chi-squared Distribution

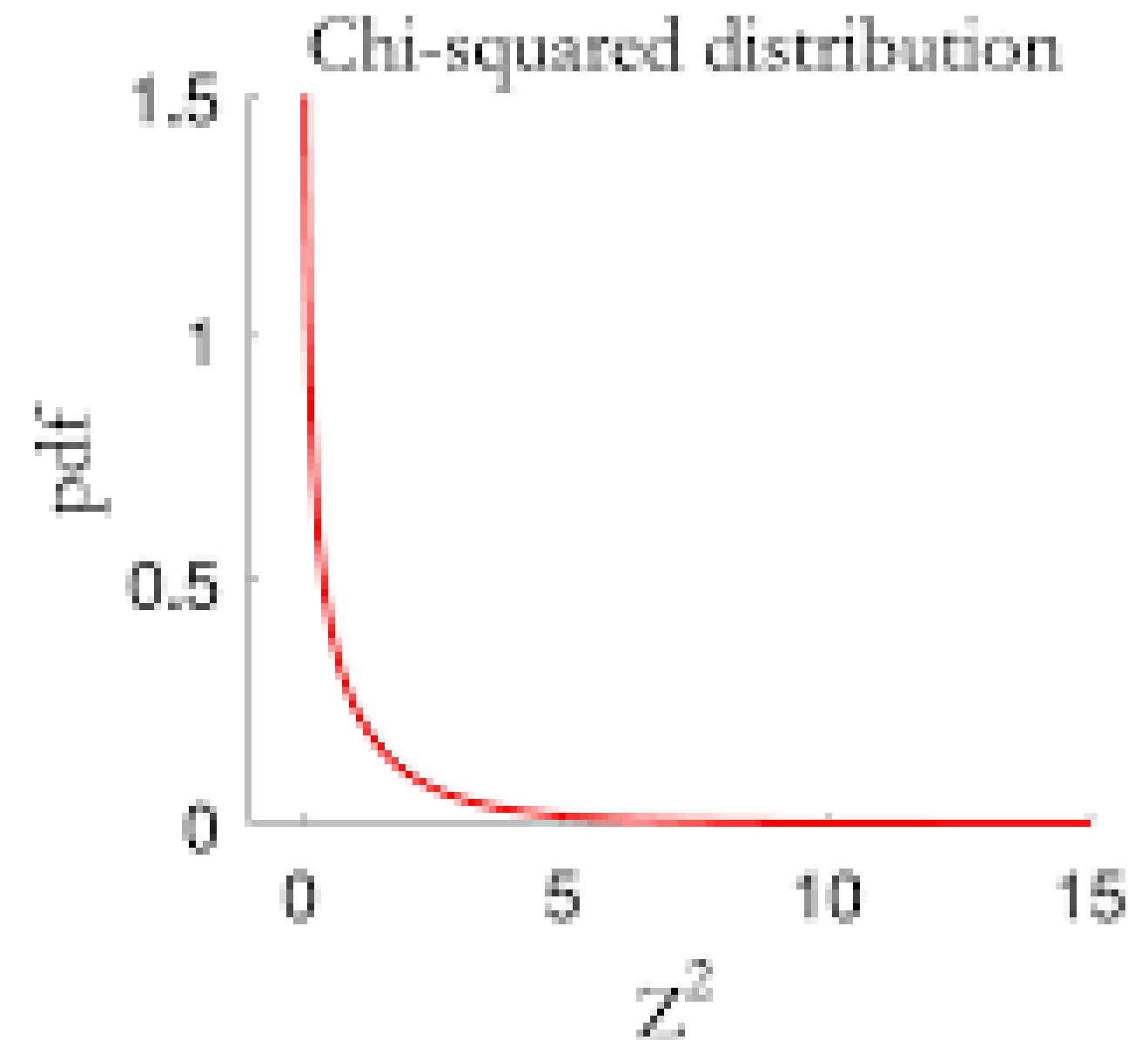
$$Z \sim N(0, 1)$$

Standard Normal Distribution



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

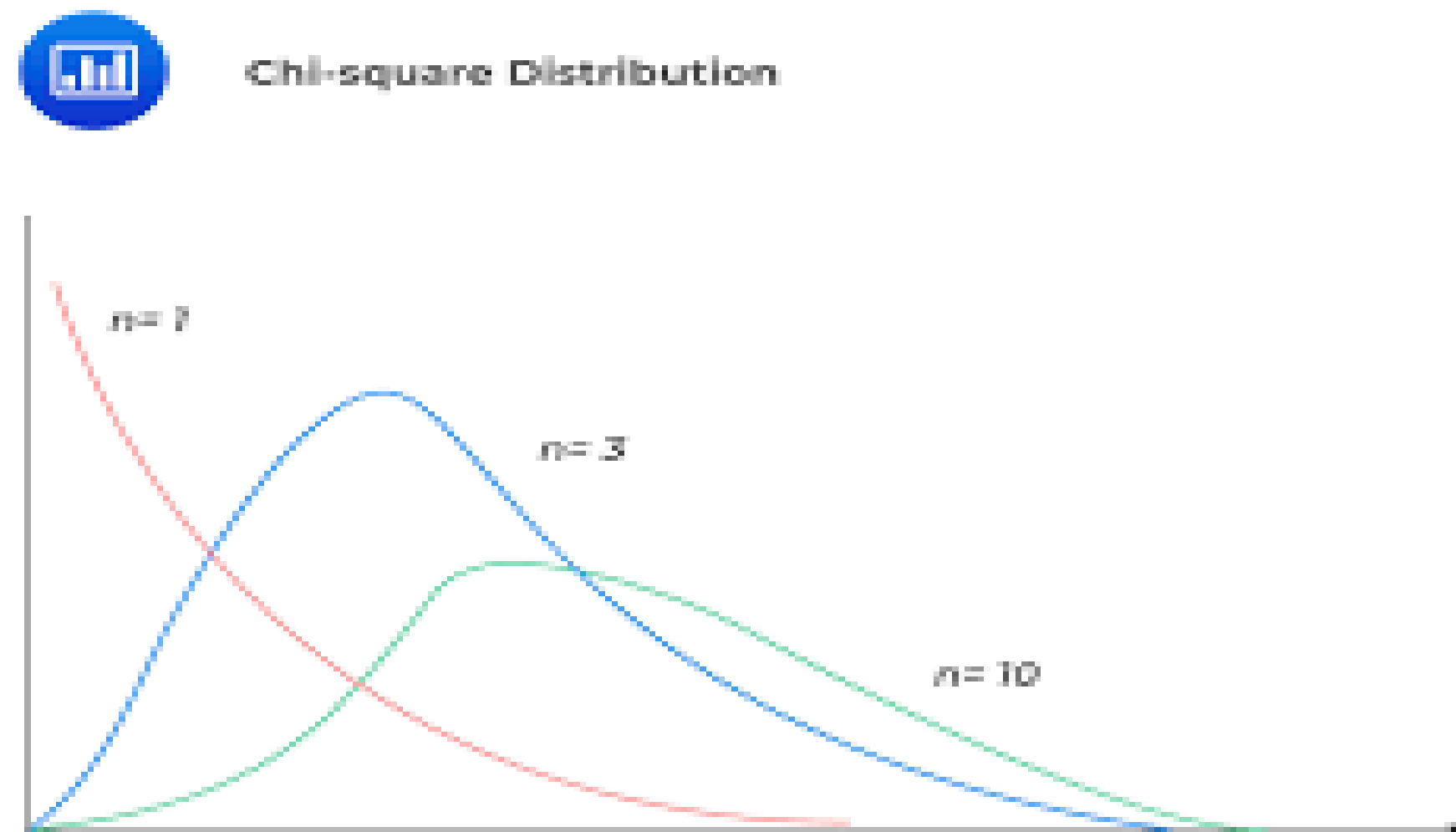
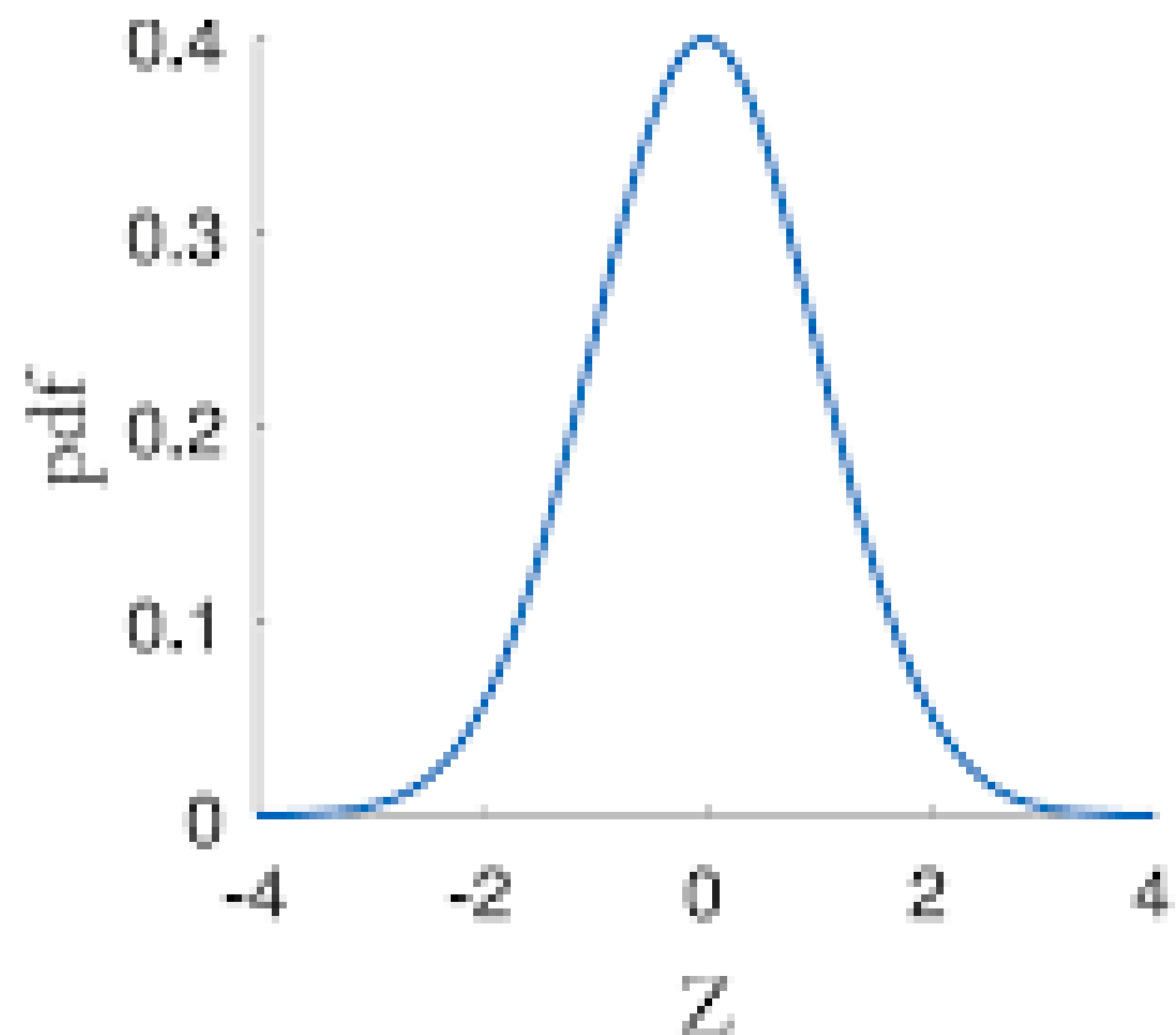
Chi-Squared Distribution



Chi-squared Distribution with two degrees of freedom

- Two random var Z_1, Z_2 with std normal distribution

$$Z_1 \sim N(0, 1) \quad Z_2 \sim N(0, 1) \quad Q = Z_1^2 + Z_2^2 \sim \chi_2^2$$

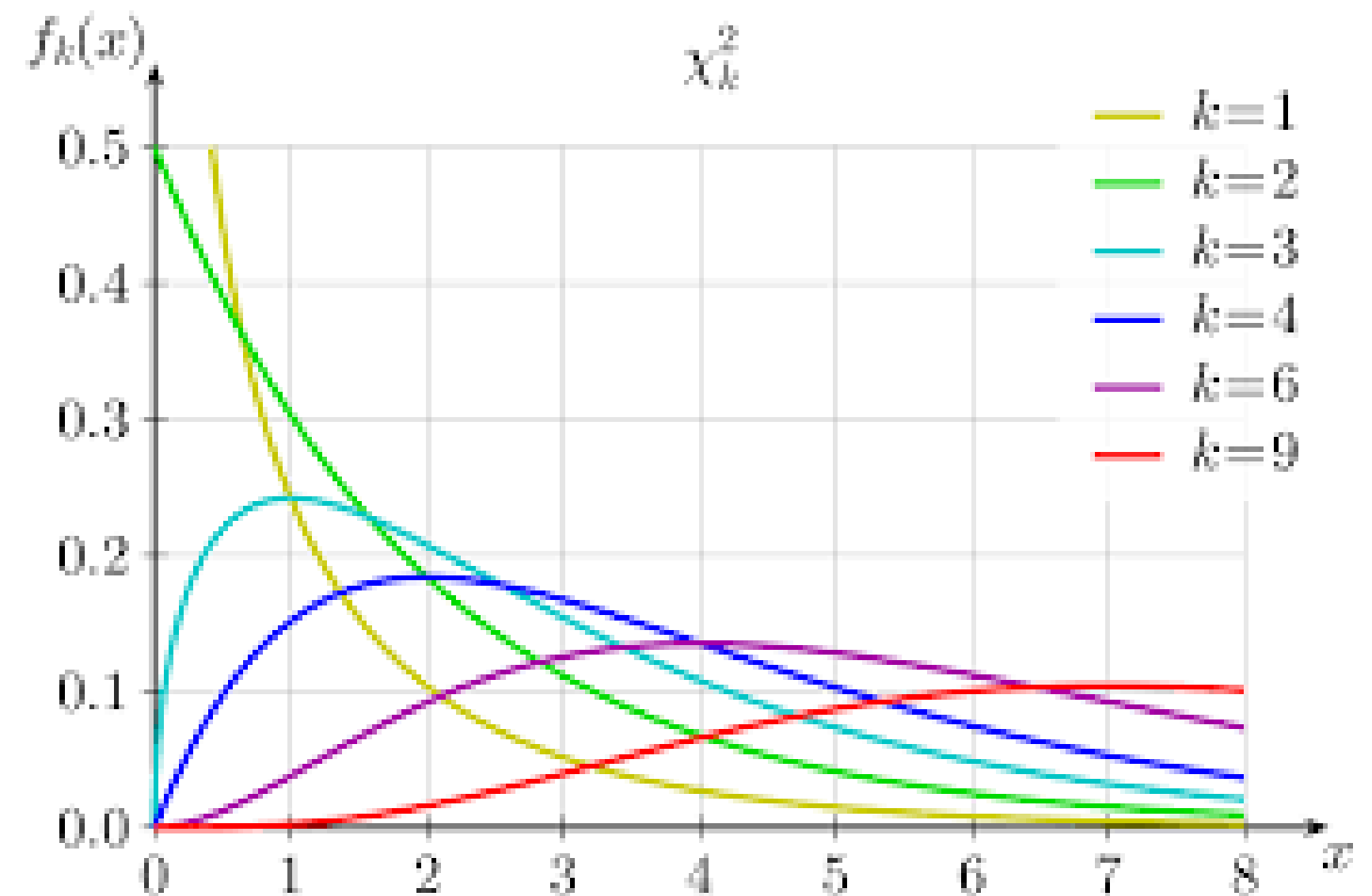
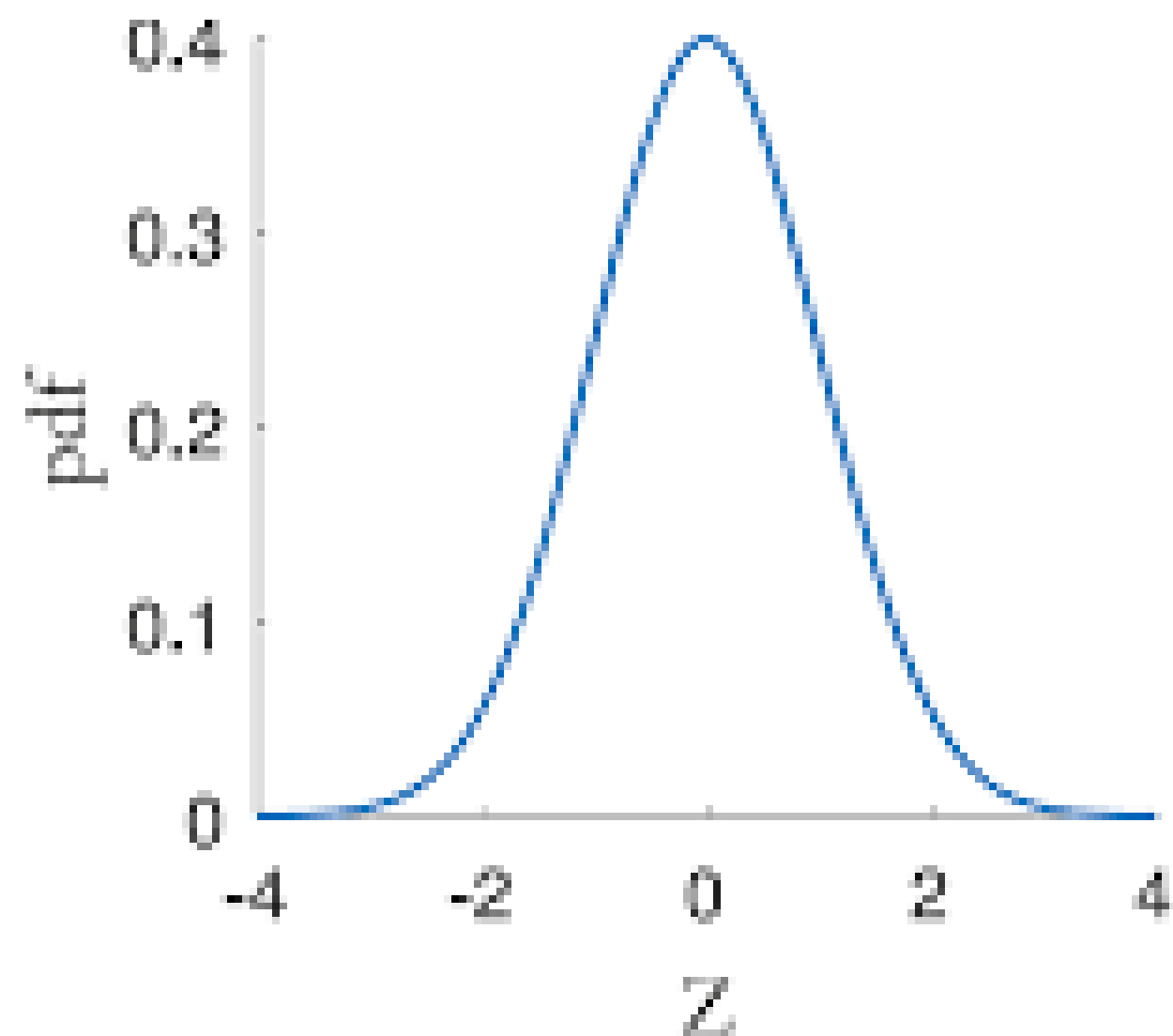


Chi-squared Distribution with n degrees of freedom

- n random var Z_1, \dots, Z_n with std normal distribution

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

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





QUESTIONS