Course 395: Machine Learning – Lectures

- Lecture 1-2: Concept Learning (*M. Pantic*)
- Lecture 3-4: Decision Trees & CBC Intro (*M. Pantic*)
- Lecture 5-6: Artificial Neural Networks (*S. Zafeiriou*)
- Lecture 7-8: Instance Based Learning (*M. Pantic*)
- Lecture 9-10: Genetic Algorithms (*M. Pantic*)
- ➢ Lecture 11-12: Evaluating Hypotheses (*M.F. Valstar*)
- Lecture 13-14: Bayesian Learning (*S. Zafeiriou*)
- Lecture 15-16: Bayesian Learning (*S. Zafeiriou*)
- Lecture 17-18: Inductive Logic Programming (*S. Muggleton*)

Evaluating Hypothesis – Lecture Overview

- Measures of classification accuracy
  - Classification Error Rate
  - Cross Validation
  - Recall, Precision, Confusion Matrix
  - Receiver Operator Curves, two-alternative forced choice
- Estimating hypothesis accuracy
  - Sample Error vs. True Error
  - Confidence Intervals
- Sampling Theory Basics
  - Binomial and Normal Distributions
  - Mean and Variance
  - One-sided vs. Two-sided Bounds
- Comparing Hypotheses
  - t-test
  - Analysis of Variance (ANOVA) test

Classification Measures – Error Rate

- Common performance measure for classification problems
  - *Success*: instance's class is predicted correctly (True Positives (TP) / Negatives (TN))
  - *Error*: instance's class is predicted incorrectly (False Positives (FP) /Negatives (FN))
  - *Classification error rate*: proportion of instances misclassified over the whole set of instances.

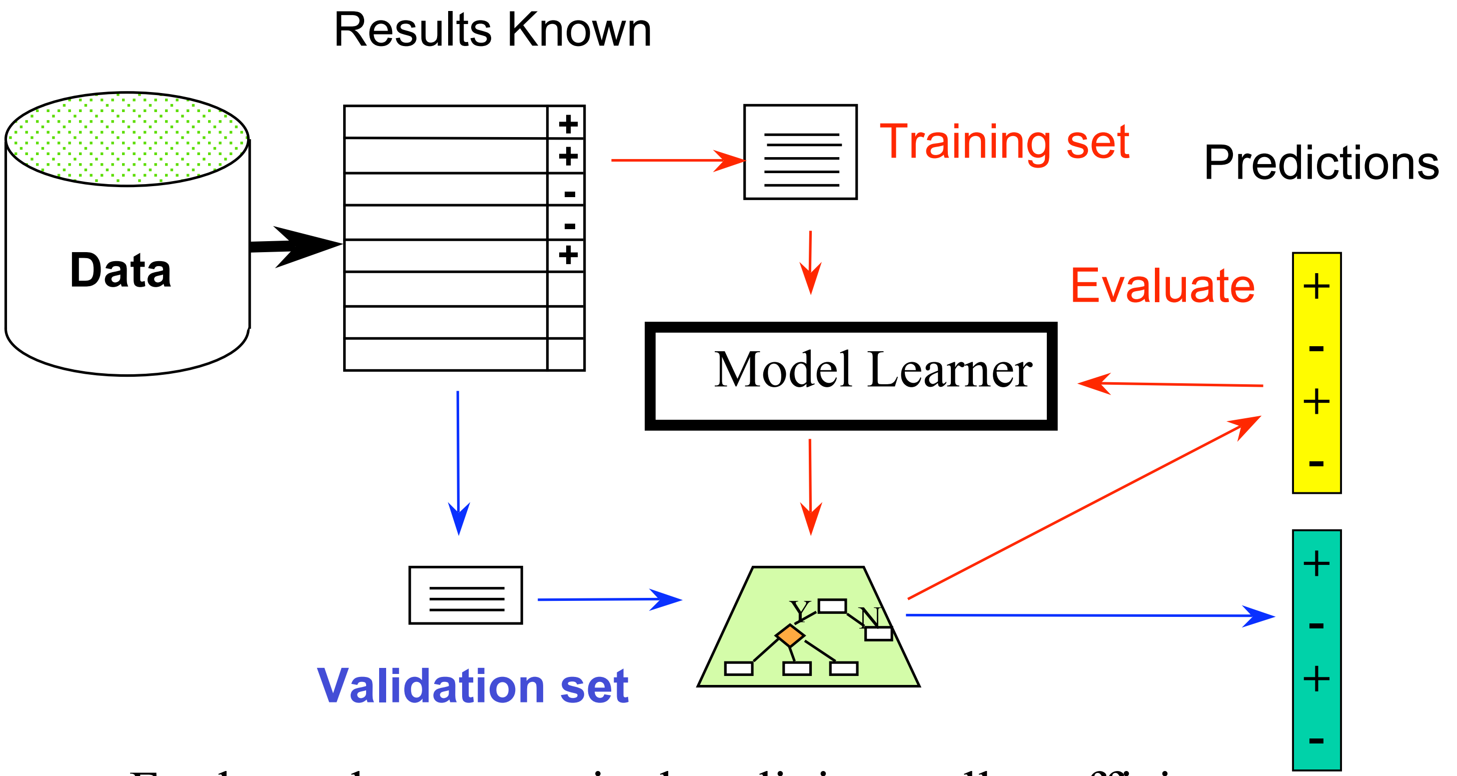
$$e = \frac{FP + FN}{TP + TN + FP + FN}$$

- Classification Error Rate on the **Training Set** can be too optimistic!
  - Unbalanced data sets

- Randomly split data into training and test sets (e.g. 2/3 for train, 1/3 for test)

The test data must not be used *in any way* to train the classifier!

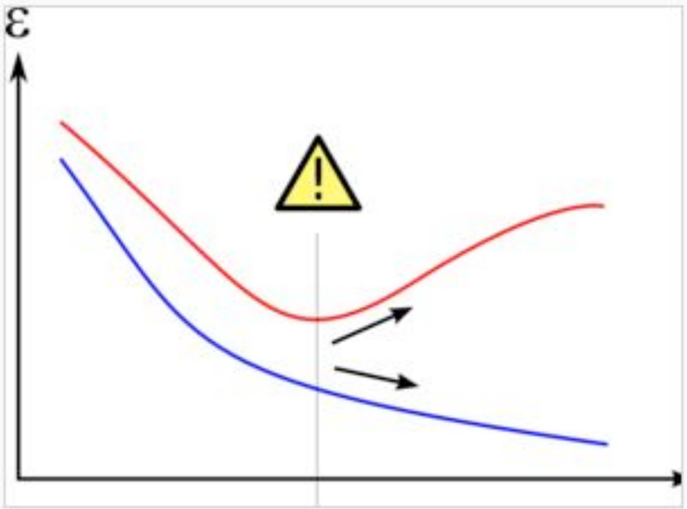
Classification Measures – Training/Test Sets



- For large datasets, a single split is usually sufficient.
- For smaller datasets, rely on cross validation

Cross Validation - Overfitting

- *Given a hypothesis space H, h∈H overfits the training data if ∃h'∈H such that h has smaller error over the training examples, but h' has smaller error than h over the entire distribution of instances.*

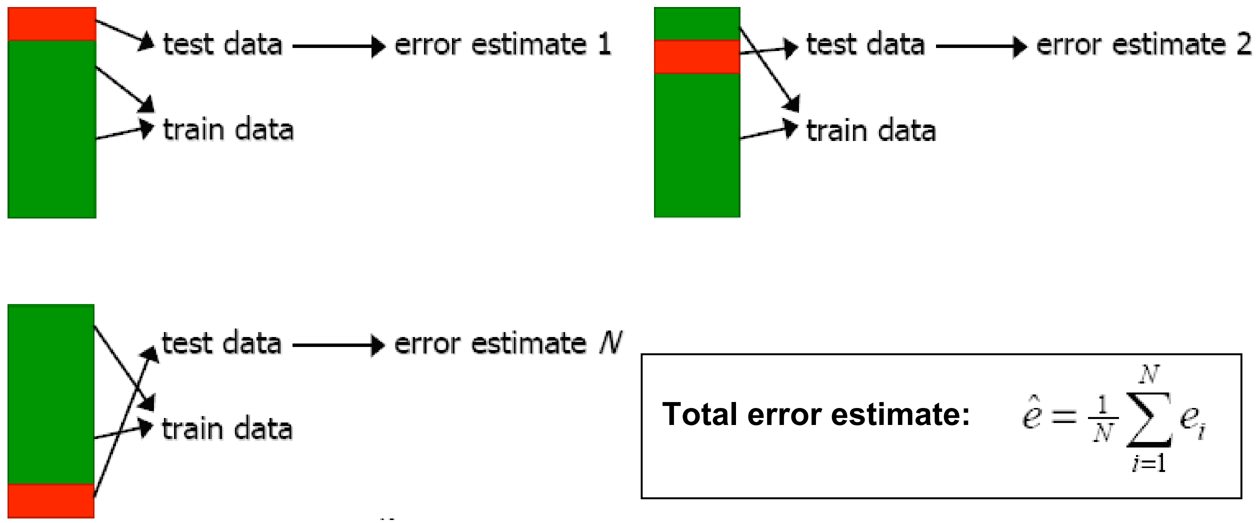


Cross Validation - Overfitting

- Overfitting can occur when:
  - Learning is performed for too long (e.g. in Neural Networks)
  - The examples in the training set are not representative of all possible situations (is usually the case!)
- The model is adjusted to uninformative features in the training set that have no causal relation to the true underlying target function!
- Cross Validation:
  - Leave one example out
  - Leave one attribute out
  - Leave n% out

Cross Validation

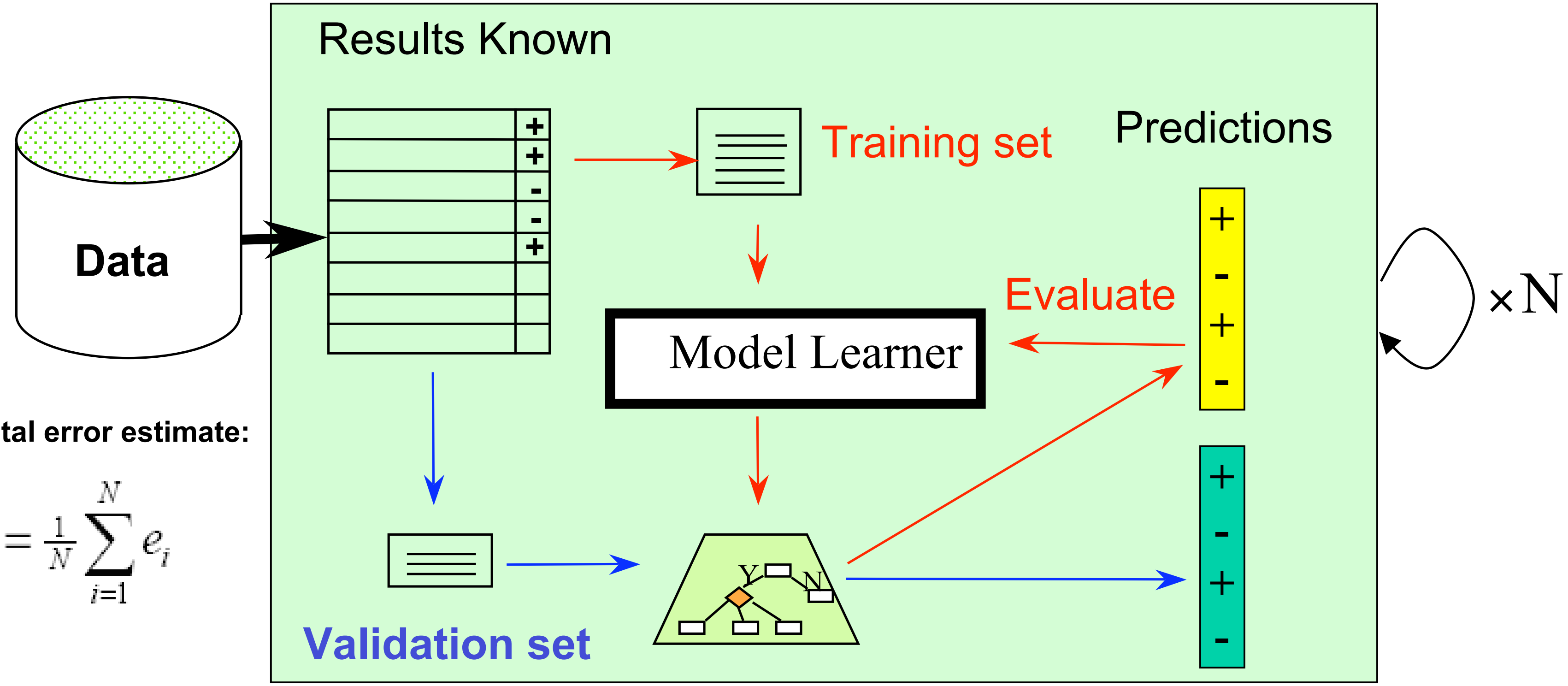
- Training Data segments between different folds should never overlap!
- Training and test data in the same fold should never overlap!



Total error estimate: $\hat{e} = \frac{1}{N}\sum_{i=1}^{N} e_i$

Cross Validation

- Split the data into training and test sets in a repeated fashion.
- Estimate the total error as the average of each fold error.



Total error estimate:

$$\hat{e} = \frac{1}{N}\sum_{i=1}^{N} e_i$$

Classification Measures – Unbalanced Sets

- Even with cross validation, the classification rate can be misleading!
  - Balanced set: equal number of positive / negative examples

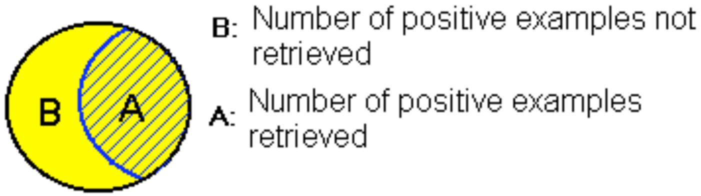
Classifier	TP	TN	FP	FN	Rec. Rate
A	25	25	25	25	50%
B	37	37	13	13	74%

  - Unbalanced set: unequal number of positive / negative examples

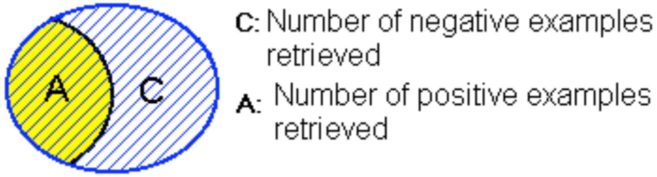
Classifier	TP	TN	FP	FN	Rec. Rate
A	25	75	75	25	50%
B	0	150	0	50	75%

Classifier B cannot classify any positive examples!

Classification Measures – Recall / Precision rates


B
A
B: Number of positive examples not retrieved
A: Number of positive examples retrieved

RECALL: $\frac{A}{A+B} \times 100\%$


A
C
C: Number of negative examples retrieved
A: Number of positive examples retrieved

PRECISION: $\frac{A}{A+C} \times 100\%$

- More insight over a classifier’s behaviour.
- For the positive class:
  - Classifier A: Recall = 50%, Precision = 25%
  - Classifier B: Recall = 0%, Precision = 0%

B classifier is useless!!

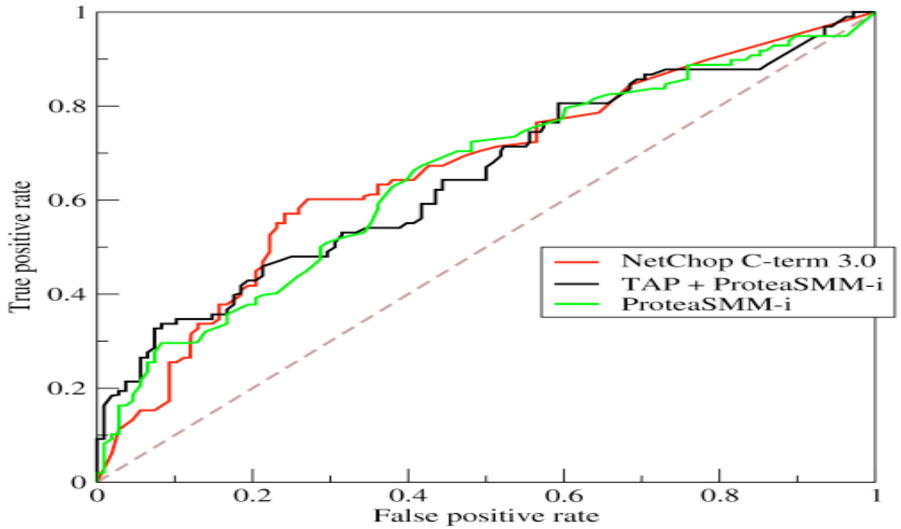
Classification Measures – F Measure

- Comparing different approaches is difficult when using two evaluation measures (e.g. Recall and Precision)
- F-measure combines recall and precision into a single measure:

$$f_\beta = \left(1+\beta^2\right)\frac{P\cdot R}{\left(\beta^2\cdot P\right)+R}$$

Classification Measures – ROC curves

Receiver Operator Characteristic (ROC) curves plot **true positive** rates against **false positive** rates



- Can be achieved by e.g. varying decision threshold of a classifier
- **Area under the curve** is an often used measure of goodness
- Two-forced alternative choice (**2AFC**) score is an easy to compute approximation of the area under the ROC curve

Classification Measures – Confusion Matrix

- A visualization tool used to present the results attained by a learner.
- Easy to see if the system is commonly mislabelling one class as another.

Predicted True	**A**	**B**	**C**
A	5	3	0
B	2	3	1
C	0	2	11

Recall	5/8	3/6	11/13
Precision	5/7	3/8	11/12

What are the recall and precision rates per class of this classifier?

Estimating accuracy of classification measures

- We want to know how well a machine learner, which learned the hypothesis *h* as the approximation of the target function *V*, performs in terms of classifying a novel, unseen example correctly.

- We want to assess the confidence that we can have in this classification measure.

Problem: we always have too little data!

Sample error & true error – True error

- The True error of hypothesis h is the probability that it will misclassify a randomly drawn example x from distribution D:

$$error_D(h) \equiv \Pr[V(x) \neq h(x)]$$

- However, we cannot measure the true error. We can only estimate $error_D$ by the Sample error $error_s$

Sample error & true error – Sample error

- Given a set S of n elements drawn i.i.d. from distribution D we empirically find the Sample error, a measure for the error of hypothesis h as:

$$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(V(x), h(x))$$

- The function $\delta(V(x), h(x))$ equals 1 if the hypothesis of an instance does not equal the target function of the same instance (i.e., makes an error) and is 0 otherwise

- Drawing n instances independently, identically distributed (i.i.d) means:

  - drawing an instance does not influence the probability that another instance will be drawn next

  - instances are drawn using the same underlying probability distribution D

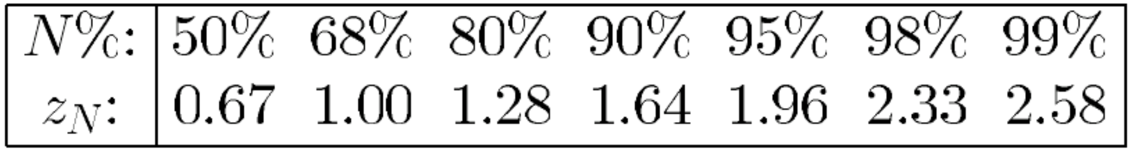
Confidence interval - Theory

Given a sample S of cardinality n >= 30 on which hypothesis h makes r errors, we can say that:

1. The most probable value of $error_D(h)$ is $error_s(h)$
2. With N % confidence, the true error lies in the interval:

$$error_s(h) \pm z_N \sqrt{\frac{error_s(h)\left(1 - error_s(h)\right)}{n}}$$

with:



$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Sampling theory - Basics

To evaluate machine learning techniques, we rely heavily on probability theory. In the next slides, basic knowledge of probability theory, including the terms mean, standard deviation, probability density function (pdf) and the concept of a Bernoulli trial are considered known.

Sampling theory – Mean, Std, Bernouilli

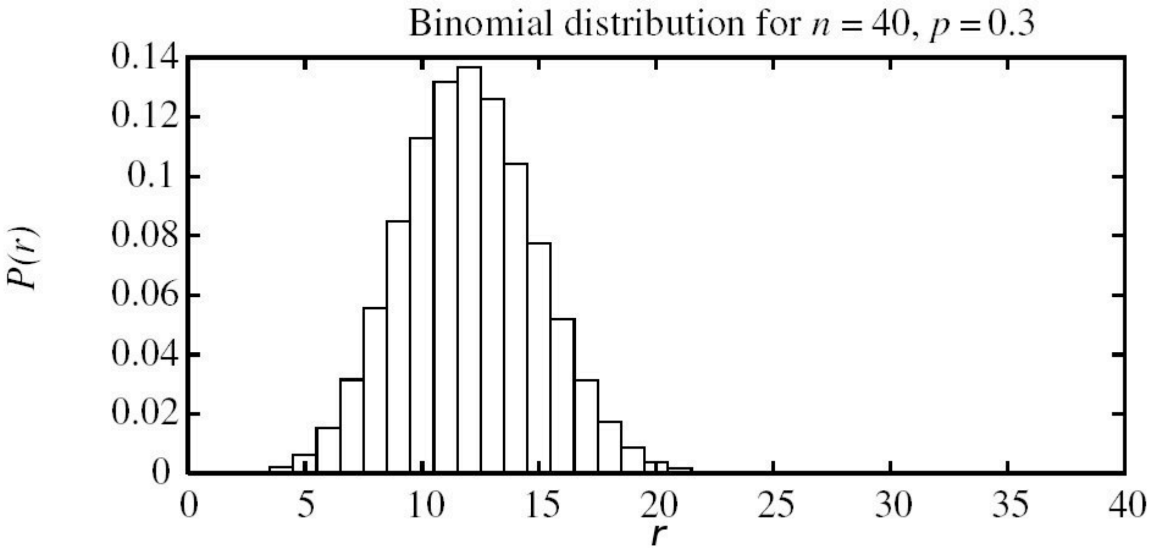
Given a random variable Y with a sequence of instances $y_1 \ldots y_n$,

- The expected or mean value of Y is: $E[Y] \equiv \sum_{i=1}^{n} y_i \Pr(Y = y_i)$
- The variance of Y is: $Var[Y] \equiv E\left[(Y - E[Y])^2\right]$
- The standard deviation of Y is: $\sigma = \sqrt{\mathrm{var}(Y)}$, and is the *expected error* in using a single observation of Y to estimate its mean.
- A Bernoulli trial is a trial with a binary outcome, for which the probability that the outcome is 1 equals p (think of a coin toss of an old warped coin with the probability of throwing heads being p).
- A Bernoulli experiment is a number of Bernoulli trials performed after each other. These trials are i.i.d. by definition.

Sampling theory - Binomial distribution

Let us run k Bernoulli experiments, each time counting the number of errors r made by h on a sample $S_{i,}$ $|S_i| = n$.

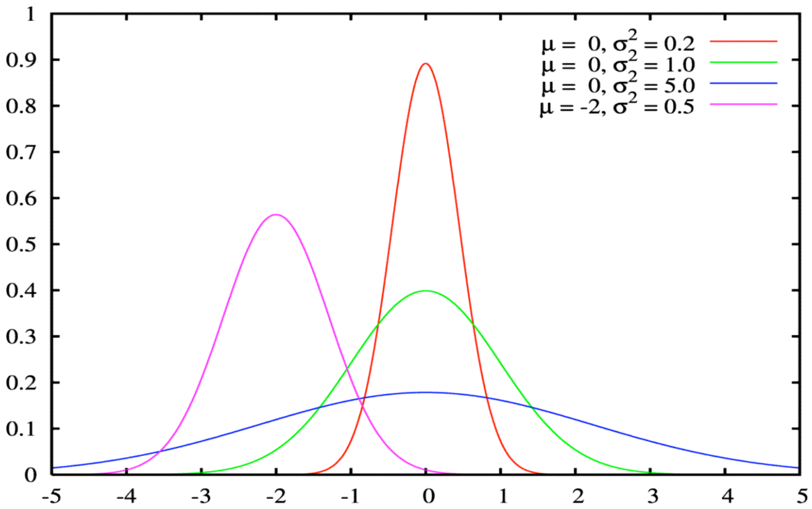
If k becomes large, the distribution of $error_{Si}(h)$ looks like:



This is called a Binomial distribution. The graph is an example of a pdf.

Sampling theory - Normal distribution

- The Normal or Gaussian distribution is a very well known and much used distribution. Its probability density function looks like a bell.



- The Normal distribution has many useful properties. It is fully described by it's mean and variance and is easy to use in calculations.
- The good thing: given enough experiments, a Binomial distribution converges to a Normal distribution.

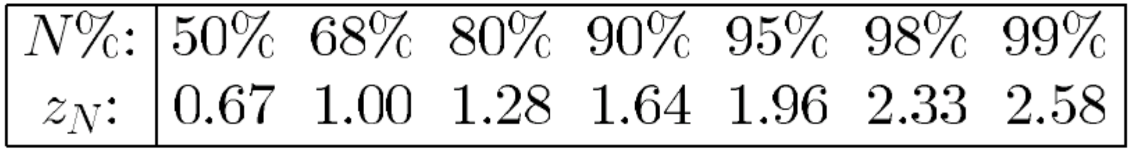
Confidence interval - Theory

Given a sample S of cardinality n >= 30 on which hypothesis h makes r errors, we can say that:

1. The most probable value of $error_D(h)$ is $error_s(h)$
2. With N % confidence, the true error lies in the interval:

$$error_s(h) \pm z_N \sqrt{\frac{error_s(h)\left(1 - error_s(h)\right)}{n}}$$

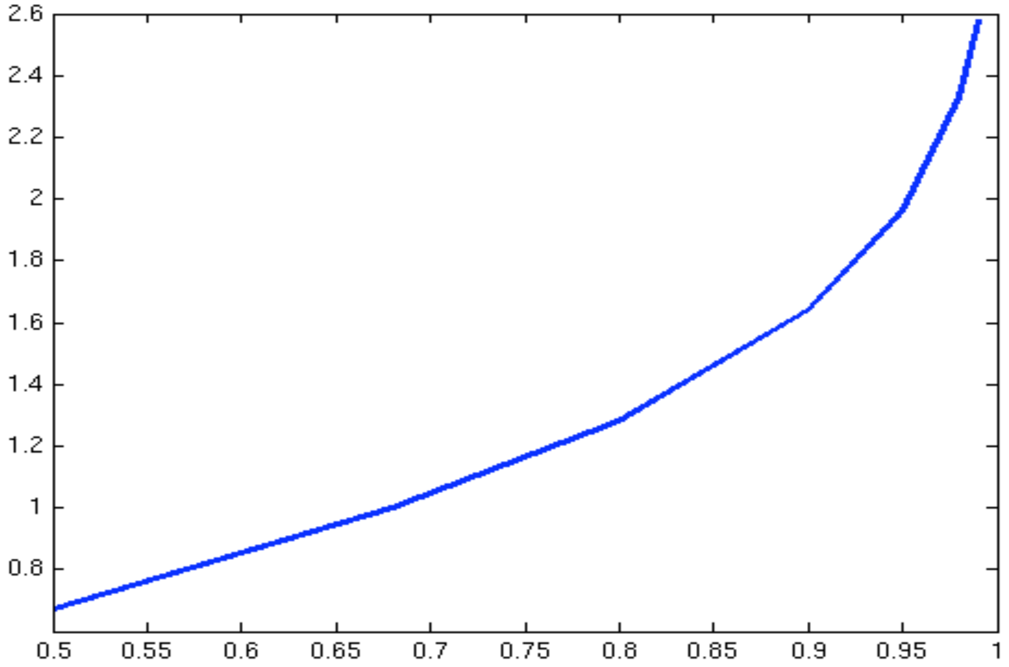
with:



$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Confidence interval – z_N

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58



Confidence interval – example (1)

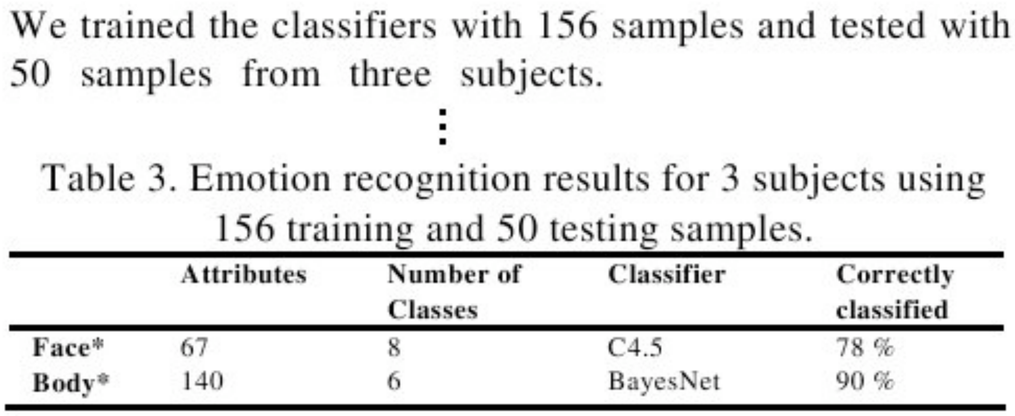
Consider the following example:

- A classifier has a 13% chance of making an error
- A sample *S* containing 100 instances is drawn
- We can now compute, that with 90% confidence we can say that the true error lies in the interval,

$$\left[0.13 - 1.64\sqrt{\frac{0.13(1-0.13)}{100}}, 0.13 + 1.64\sqrt{\frac{0.13(1-0.13)}{100}}\right] =$$
$$[0.075, 0.19]$$

Confidence interval – example (2)

Given the following extract from a scientific paper on multimodal emotion recognition:



We trained the classifiers with 156 samples and tested with 50 samples from three subjects.

⋮

Table 3. Emotion recognition results for 3 subjects using 156 training and 50 testing samples.

	Attributes	Number of Classes	Classifier	Correctly classified
Face*	67	8	C4.5	78 %
Body*	140	6	BayesNet	90 %

For the Face modality, what is n? What is $error_s(h)$?

Exercise: compute the 95% confidence interval for this error.

Confidence interval – example (3)

Given that *error*$_s$(*h*)=0.22 and n= 50, and z_N=1.96 for $N = 95$, we can now say that with 95% probability *error*$_D$(*h*) will lie in the interval:

$$\left[0.22-1.96\sqrt{\frac{0.22(1-0.22)}{50}},0.22+1.96\sqrt{\frac{0.22(1-0.22)}{50}}\right] =$$

$$[0.11,0.34]$$

What will happen when $n \to \infty$?

However, we are not only uncertain about the quality of *error*$_S$*(h)*, but also about how well *S* represents *D*!!!

Sampling theory - Two sided/one sided bounds

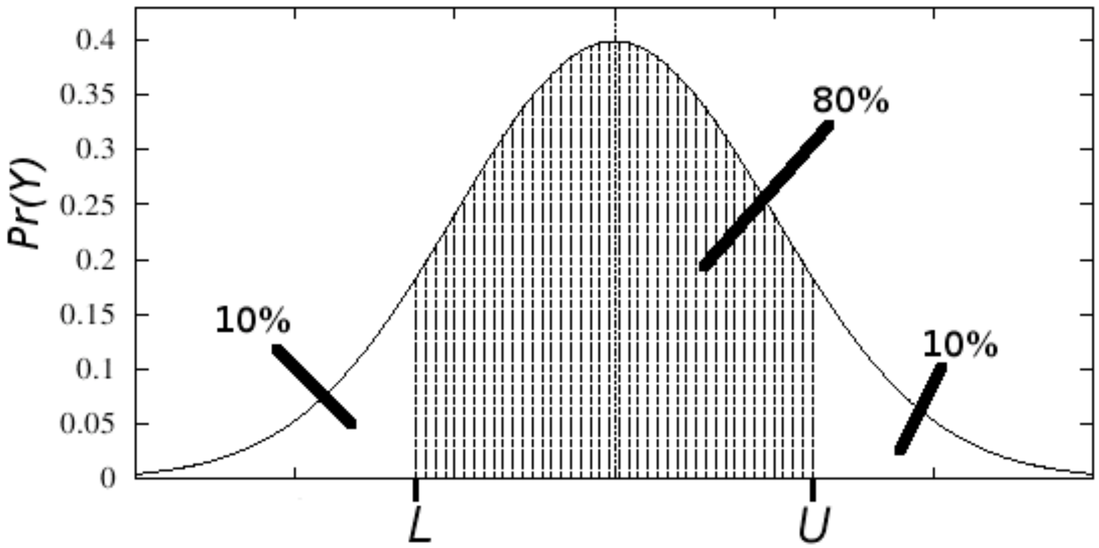
- We might be interested not in a confidence interval with both an upper and a lower bound, but instead in the upper or lower limit only. For instance, what is the probability that $error_D(h)$ is at most U.

- In the confidence interval example, we found that with (100-a)=95% confidence

$$L = 0.11 \le error_D\left(h\right) \le 0.34 = U$$

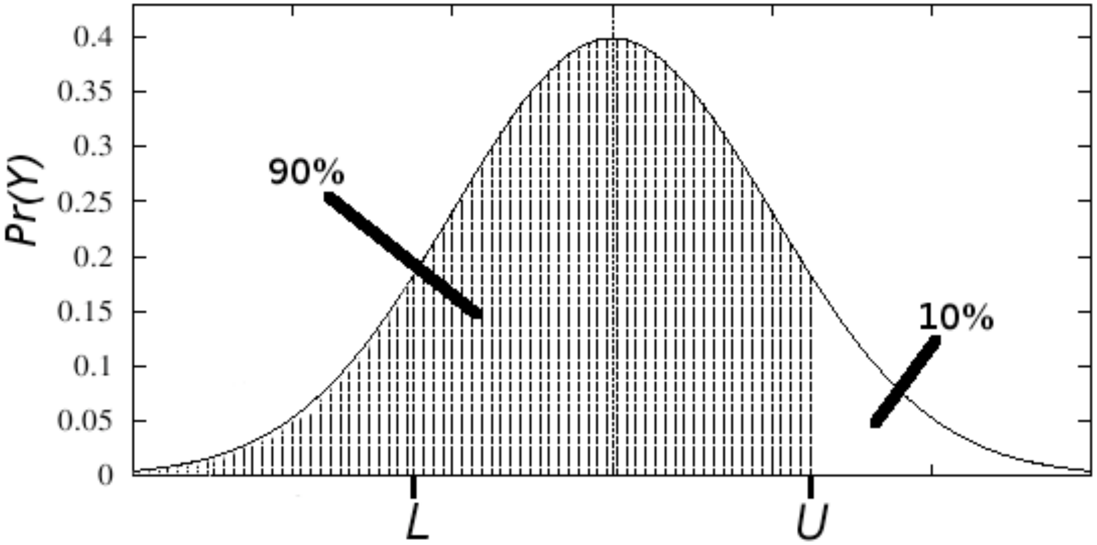
- Using the symmetry property of a normal distribution, we now find that $error_D(h) <= U=0.34$ with confidence (100-a/2)=97.5%.

Sampling theory - Two sided/one sided bounds



- The confidence of $L<=Y<=U$ can be found as: $\int_{L}^{U} \Pr(Y)$
- In this case the confidence for Y lying in this interval is 80%

Sampling theory - Two sided/one sided bounds



- The confidence of $Y<=U$ can be found as: $\int_{-\infty}^{U} \Pr(Y)$

- In this case the confidence for Y being smaller than U is 90%

Comparing hypotheses - Ideal case

We want to estimate the difference in errors *d* made by hypotheses h_1 and h_2, tested on samples S_1 and S_2

$$d \equiv error_D(h_1) - error_D(h_2)$$

The estimator we choose for this problem is:

$$\overline{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

Comparing hypotheses – ideal case

- As both *error*$_{S1}(h_1)$ and *error*$_{S2}(h_2)$ follow approximately a Normal distribution, so will d. Also, the variance of d is the sum of the variances of *error*$_{S1}(h_1)$ and *error*$_{S2}(h_2)$:

$$\sigma^2_{\bar{d}} \approx \frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S2}(h_2))}{n_2}$$

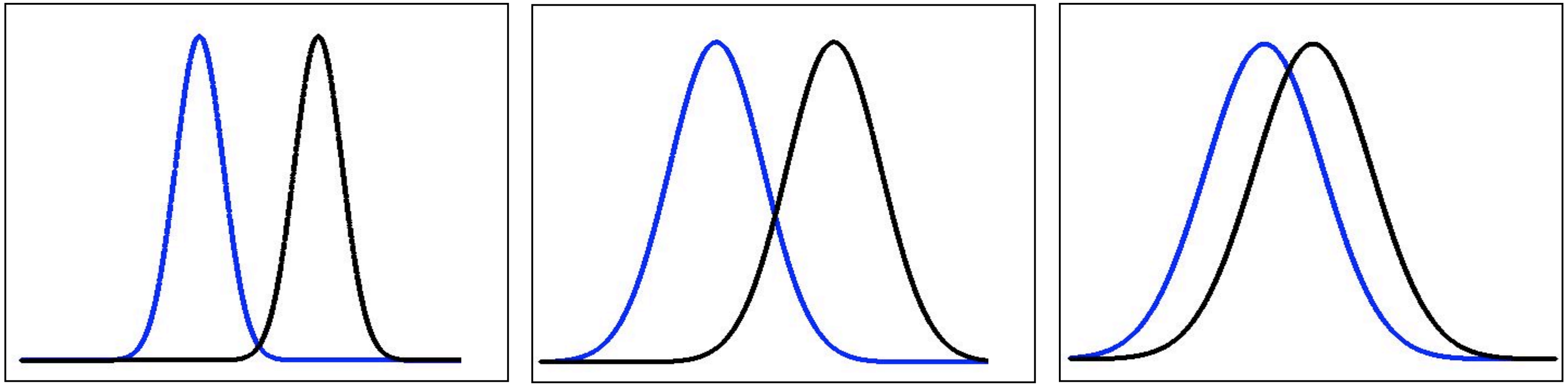
- Now we can find the *N*% confidence interval for the error difference d:

$$\bar{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S2}(h_2))}{n_2}}$$

$$= \pm z_N \sigma_{\bar{d}}$$

T-test

- Assess whether the means of two distributions are *statistically* different from each other.



Consider the distributions as the classification errors of two different classifiers, derived by cross-validation. Are the means of the distributions enough to say that one of the classifiers is better?

T-test

- The t test is a test on the null hypothesis

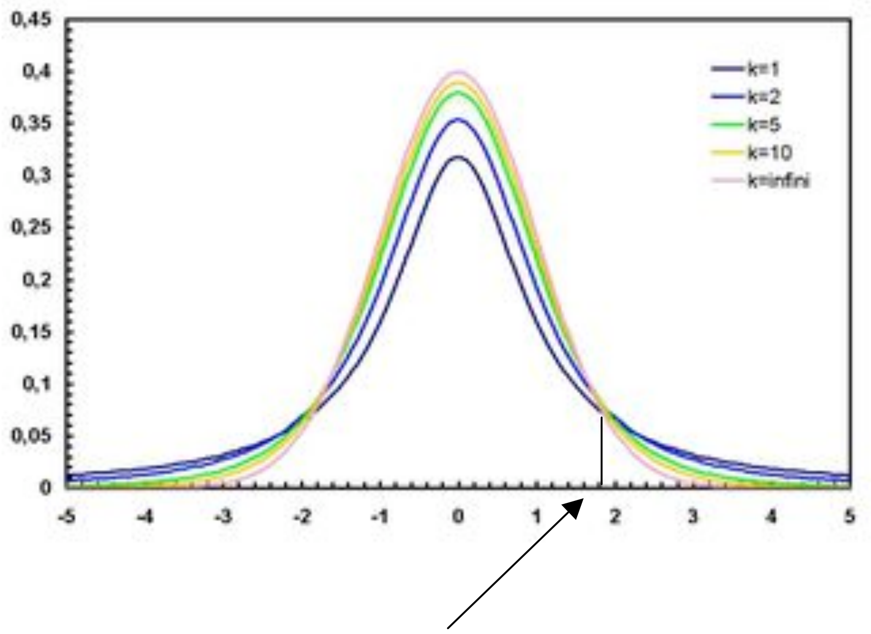
  H_0 : the means of the distributions are the same, against the alternative hypothesis

  H_α : at least two means of distributions are unequal.

T-test: $$t = \frac{\overline{x}_T - \overline{x}_C}{SE(\overline{x}_T - \overline{x}_C)}$$

$$SE(\overline{x}_T - \overline{x}_C) = \sqrt{\frac{\mathrm{var}_T}{n_T} + \frac{\mathrm{var}_C}{n_C}}$$

T distribution

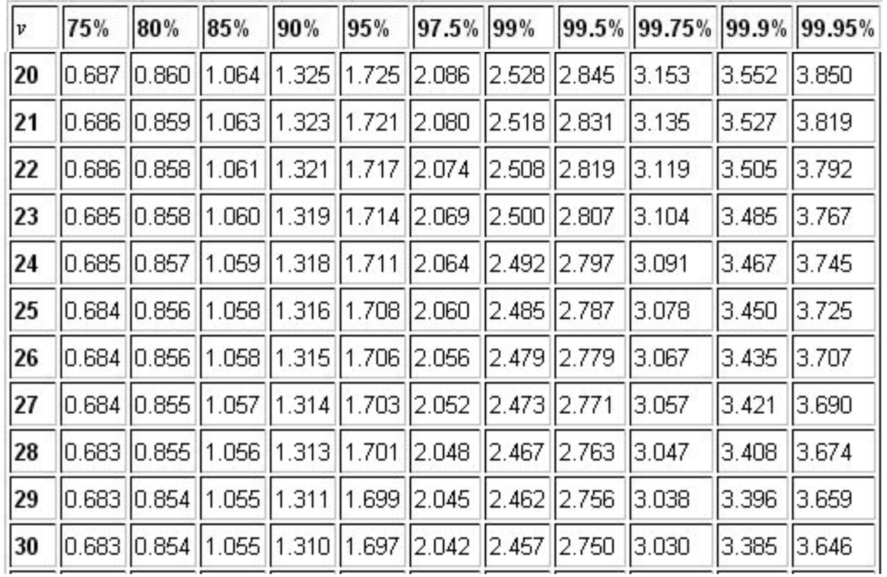


threshold

T-test

- Significance level α%: α times out of 100 you would find a statistically significant difference between the distributions even if there was none. It essentially defines our tolerance level.

- Degrees of freedom: Sum of samples in the two groups - 2



ν	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646

If the calculated *t* value is above the threshold chosen for statistical significance then the null hypothesis that the two groups do not differ is rejected in favour of the alternative hypothesis, which typically states that the groups do differ.

T-test – MATLAB

```
H = TTEST2(X,Y,ALPHA)
```

- performs a T-test of the hypothesis that two independent samples, in the vectors X and Y, come from distributions with equal means, and returns the result of the test in H.
- H==0 indicates that the null hypothesis ("means are equal") cannot be rejected at the α% significance level.
- H==1 indicates that the null hypothesis can be rejected at the α% level. The data are assumed to come from normal distributions with unknown, but equal, variances. X and Y can have different lengths.

Analysis of Variance (ANOVA) test

- Similar to t-test, but compares several distribution simultaneously.

Notation:

- g is the number of groups we want to compare.
- $\mu_1, \mu_2, \ldots, \mu_g$ are the means of the distributions we want to compare.
- $n_1, n_2, \ldots, n_g$ are the sample sizes
- $\overline{Y}_{1,} \overline{Y}_{2,} \ldots, \overline{Y}_g$ are the sample means
- $\sigma_1, \sigma_2, \ldots, \sigma_g$ are the sample standard deviations

- The ANOVA test is a test on the null hypothesis H_0 : the means of the distributions are the same, against the alternative hypothesis H_α :at least two means are unequal.

Analysis of Variance (ANOVA) test

- **Basic principle** : compute two different estimates of the population variance:
  - The within groups estimate pools together the sums of squares of the observations about their means:

$$WSS = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$$

– The between groups estimate, calculated with reference to the grand mean, that is, the mean of all the observations pooled together:

$$BSS = \sum_{i=1}^{g} n_i (\overline{Y}_i - \overline{Y})$$

- ➢ only a good estimate of the sample variance if the null hypothesis is true
- ➢ only then will the grand mean be a good estimate of mean of each group.

Analysis of Variance (ANOVA) test

- The ANOVA *F* test statistic is the ratio of the between estimate and the within estimate:

$$F = \frac{\text{Between estimate}}{\text{Within estimate}} = \frac{BSS/(g-1)}{WSS/(N-g)}$$

- When the null hypothesis is false, the between estimate tends to overestimate the population variance, so it tends to be larger than the within estimate. Then, the *F* test statistic tends to be considerably larger than 1.

- The ANOVA test has an F probability distribution function as its sampling distribution. It has two degrees of freedom that determine its exact shape:

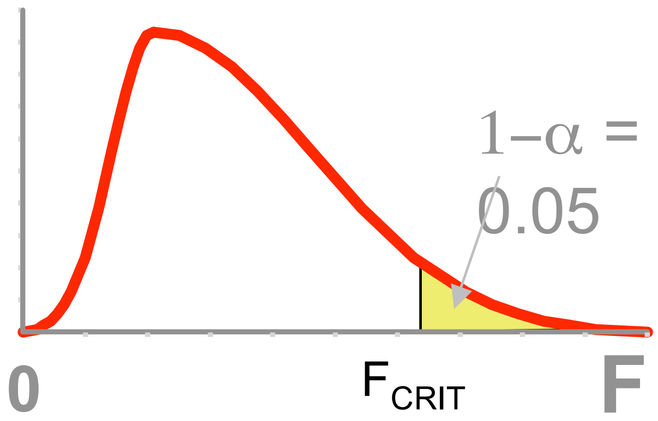
$$df_1 = g-1$$
$$df_2 = N-g.$$

Analysis of Variance (ANOVA) test

- $\boldsymbol{H_0}$: All Equal
- $\boldsymbol{H_1}$: Not All Equal

If $F > F_{CRIT}$:

There is evidence that at least one μ_i distribution differs from the rest.



The shaded area of the graph indicates the rejection region at the $\boldsymbol{\alpha}$ significance level

Analysis of Variance (ANOVA) test - MATLAB

```
P = ANOVA1(X)
```

- performs a one-way ANOVA for comparing the means of two or more groups of data. It returns the p-value for the null hypothesis that the means of the groups are equal.

- If X is a matrix, ANOVA1 treats each column as a separate group, and determines whether the population means of the columns are equal.

Sampling theory - Central limit theorem

The Central Limit theorem states that,

Given a set of i.i.d. random variables $Y_1 \dots Y_n$ governed by an arbitrary pdf with mean μ and finite variance σ^2. Define the sample mean

$$\overline{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Then, as $n \rightarrow \infty$, the distribution governing

$$\frac{\overline{Y}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

approaches a standard Normal distribution.

Course 395: Machine Learning – Lectures

- Lecture 1-2: Concept Learning (*M. Pantic*)
- Lecture 3-4: Decision Trees & CBC Intro (*M. Pantic*)
- Lecture 5-6: Artificial Neural Networks (*S. Zafeiriou*)
- Lecture 7-8: Instance Based Learning (*M. Pantic*)
- Lecture 9-10: Genetic Algorithms (*M. Pantic*)
- Lecture 11-12: Evaluating Hypotheses (*M.F. Valstar*)
- ➢ Lecture 13-14: Bayesian Learning (*S. Zafeiriou*)
- Lecture 15-16: Bayesian Learning (*S. Zafeiriou*)
- Lecture 17-18: Inductive Logic Programming (*S. Muggleton*)