Evaluating what has been learned

Lecture 03.01

Intuition → numeric evaluation

- How to measure the quality of the classifier
- How to quantify this measure
- How to compare the quality of two different classifiers

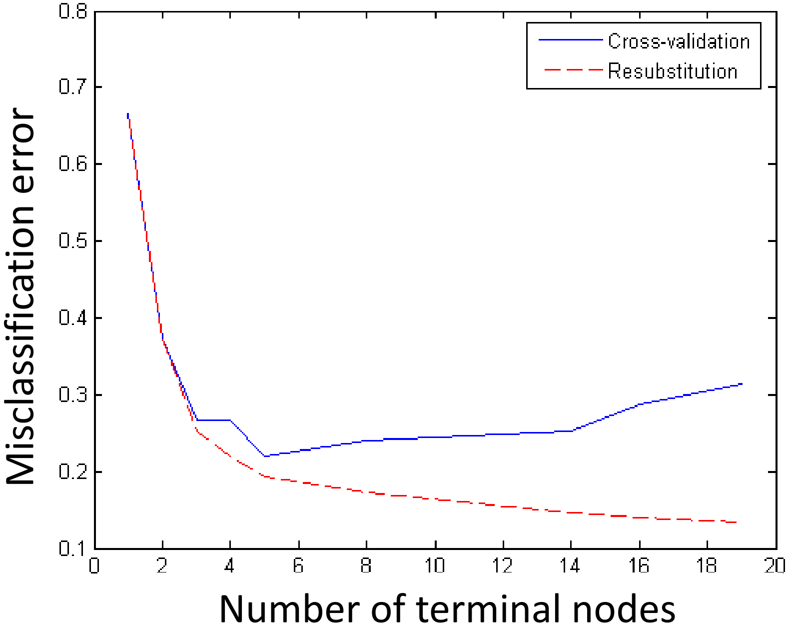
Natural performance measure: *error rate*

- *Success*: instance’s class is predicted correctly
- *Error*: instance’s class is predicted incorrectly
- *Error rate*: proportion of errors made over the whole set of test instances

Resubstitution (training) error

- Training error - error rate obtained from training data
- Example:

Error rate for different number of leaf nodes in a decision tree



Training error is (hopelessly) optimistic!

Error for a test set

- *Test set*: independent instances that played no part in formation of classifier
  - Assumption: both training data and test data are representative samples of the underlying problem
- Generally, the larger the training data, the better the classifier
- The larger the test data the more accurate the error estimate

Where to get the test set?

- Simple solution that can be used if lots of (labeled) data is available:
  - Split data into training and test set
- However: (labeled) data is usually limited
  - More sophisticated techniques need to be used
  - We need to make the most from the available data

Holdout

- *Holdout procedure*: method of splitting original data into training and test set
  - Dilemma: ideally both training set *and* test set should be large!

- The *holdout* method reserves a certain amount for testing and uses the remainder for training
  - Usually: one third for testing, the rest for training

- Problem: the samples might not be representative
  - Example: one class might be missing in the test data

- Advanced version uses *stratification*
  - Ensures that each class is represented with approximately equal proportions in both subsets

Repeated holdout

- Holdout estimate can be made more reliable by repeating the process with different subsamples
  - In each iteration, a certain proportion is randomly selected for training (possibly with stratification)
  - The error rates on the different iterations are averaged to yield an overall error rate

- This is called the *repeated holdout* method
- Still not optimum: different test sets overlap
  - Can we prevent overlapping?

Cross-validation

- *Cross-validation* avoids overlapping test sets
  - **First step:** split data into *k* subsets of equal size
  - **Second step:** use each subset in turn for testing, the remainder for training

k-fold cross-validation

- Often the subsets are stratified before the cross-validation is performed
- The error estimates are averaged to yield an overall error estimate

- Standard method: **stratified 10-fold cross-validation**

Leave-One-Out cross-validation

- Leave-One-Out: an extreme form of cross-validation
  - Set number of folds to **number of training instances:**

  for *n* training instances, build classifier *n* times using n-1 instances for training, and record the error rate of the left-out instance

✓ Makes best use of data

✓ Involves no random subsampling

❖ But, computationally expensive

Leave-One-Out-CV and stratification

- In the Leave-One-Out-CV: stratification is not possible

  It *guarantees* a non-stratified sample because there is only one instance in the test set!

- Extreme example: completely random dataset split equally into two classes
  - The classifier predicts majority class
  - 50% accuracy on fresh data
  - Leave-One-Out-CV estimate gives 100% error!

Bootstrap

- Cross-Validation uses sampling *without replacement*
  - The same instance, once selected, can not be selected again for a particular training/test set
- The *bootstrap* uses sampling *with replacement* to form the training set:
  - Randomly sample a dataset of *n* instances *n* times *with replacement* to form a new dataset of *n* instances
  - Use this data as the training set
  - Use the instances from the original dataset that don't occur in the new training set for testing
- Also called the *0.632 bootstrap (Why?)*

The 0.632 bootstrap

- A particular instance has a probability of 1–1/*n* of *not* being picked
- Thus, its probability of ending up in the test data is:

$$\left(1-\frac{1}{n}\right)^n \approx e^{-1} = 0.368$$

- This means the training data will contain approximately 63.2% of the instances

Estimating error with the bootstrap

- The error estimate on the test data will be very pessimistic: trained on just ~63% of the instances
- Therefore, combine it with the optimistic training error:

$$err = 0.632 \cdot e_{\text{test instances}} + 0.368 \cdot e_{\text{training instances}}$$

The training error gets less weight than the error on the test data

- Repeat process several times with different replacement samples; average the results
- This is the best way of estimating performance for very small datasets

Statistics rules!

ESTIMATING THE MEAN OF SUCCESS/ERROR RATE WITH CONFIDENCE

Predicting true performance

- Assume the estimated success rate is 75%. How close is this to the true success rate on an unknown future population?
  - Depends on the amount of test data

- Prediction is just like tossing a (biased!) coin
  - "Head" is a "success", "tail" is an "error"
  - And we want to approximate the real probability p("head") from a set of experiments

- In statistics, a succession of independent events like this is called a *Bernoulli process*
  - Statistical theory provides us with confidence intervals for the true underlying proportion of probabilities

Predicting performance *interval*

- We can say: p – *probability of success* of a classifier – lies within a certain specified interval with a certain specified confidence
- Example: S=750 successes in N=1000 trials
  - Estimated success rate: 75%
  - How close is this to the true success rate p?
    - Answer: with 80% confidence $p \in [73.2, 76.7]$
- Another example: S=75 and N=100
  - Estimated success rate: 75%
  - With 80% confidence $p \in [69.1, 80.1]$
    - I.e. the probability that $p \in [69.1, 80.1]$ is 0.8.
- The bigger the N – the more precise we are in our evaluation, i.e. the surrounding interval is smaller.
  - Above, for N=100 we were less confident than for N=1000.

Predicting performance *interval*

- How do we compute the predicted interval of classifier's success for a certain level of confidence?

- There is a large unknown number of samples to be classified in the future

- Out of this whole population we tested classifier only on N instances (N-the size of our test set)

Success as a random variable

- Let Y be the random variable with possible values

  1 for success and

  0 for error.
- Let probability of success be p.
- Then probability of error is $q=1-p$.

- What's the mean of the Y distribution?

  $\mu=1*p + 0*q = p$
- What's the standard deviation of Y distribution?

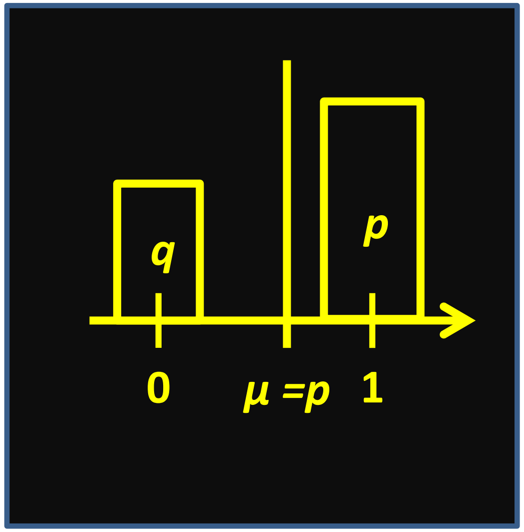
  $\sigma^2=(1-p)^2*p + (0-p)^2*q$

  $= q^2*p+p^2*q$

  $= pq(q+p)$

  $= pq(1- p + p)$

  $=pq$

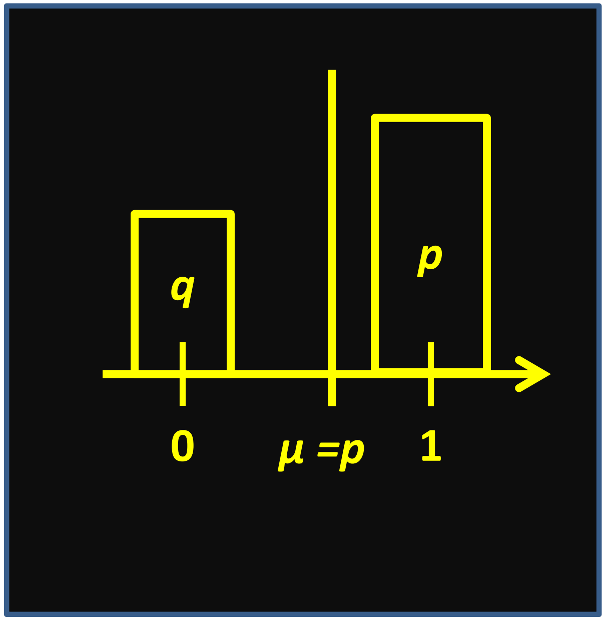


True distribution of classification success

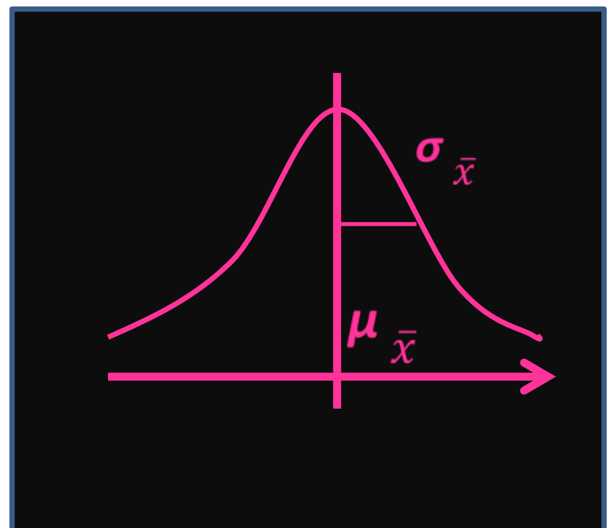
We do not know μ=p!

Distribution of sampling means

We can take a random sample of size N from the entire population of Y values. The average of this one sample, $\bar{x}$, might be close to the real mean μ, and might be not.
However, if we perform many random samplings, and plot the average of each sampling, the sampling averages would have normal distribution

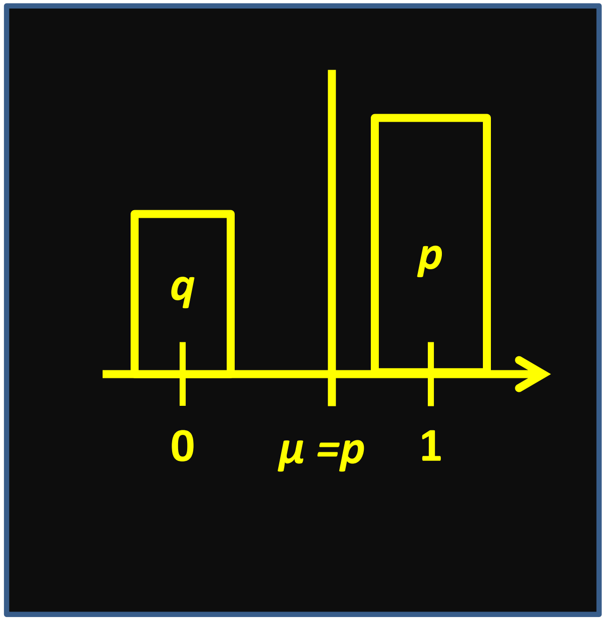


True probability distribution of *Y* in the entire population

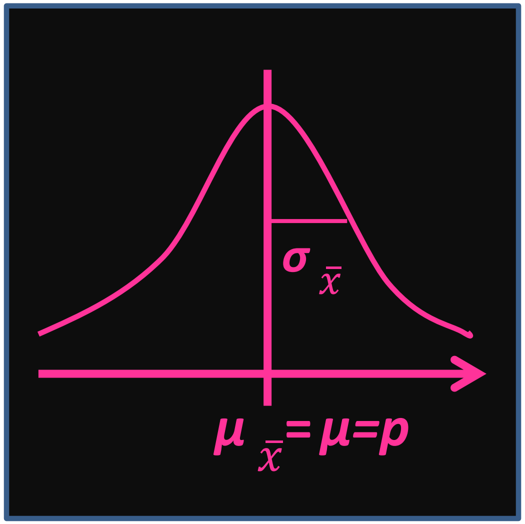


Distribution of sampling averages $\bar{x}$ for N=10

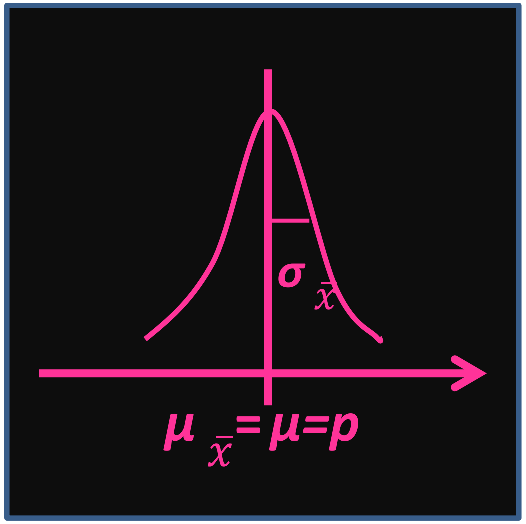
Distribution of sampling means



True distribution of classification success



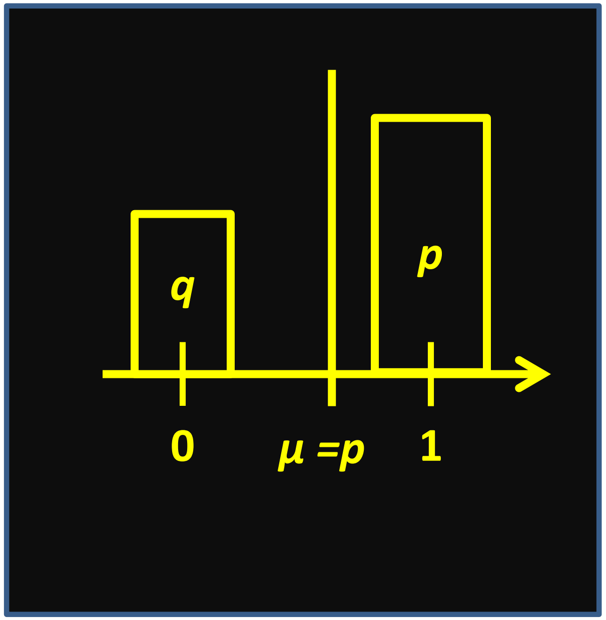
Distribution of sampling averages $\bar{x}$ for N=10



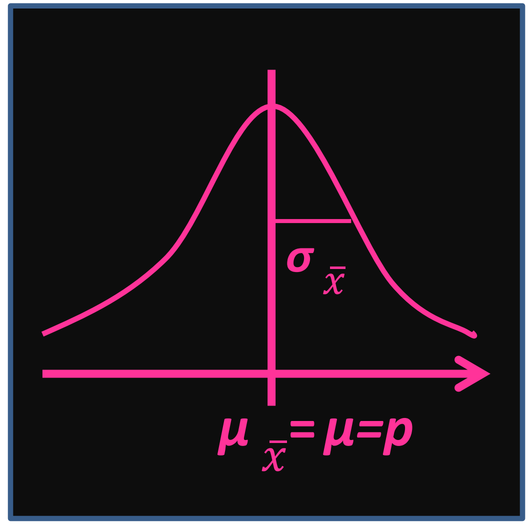
Distribution of sampling averages $\bar{x}$ for N=100

Given large enough number of samplings, the mean of sampling averages will approach the real mean of the entire population

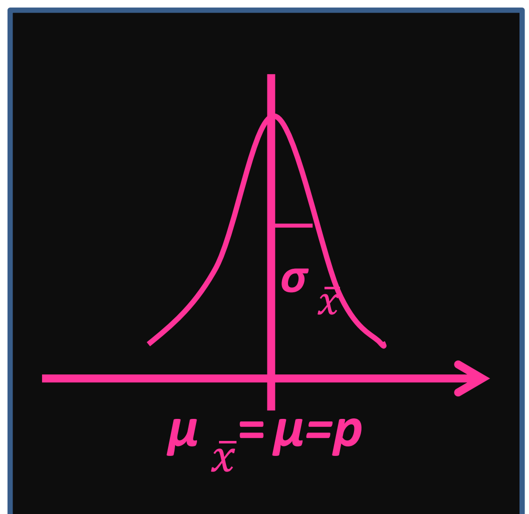
Standard deviation of sampling means



True distribution of
classification success



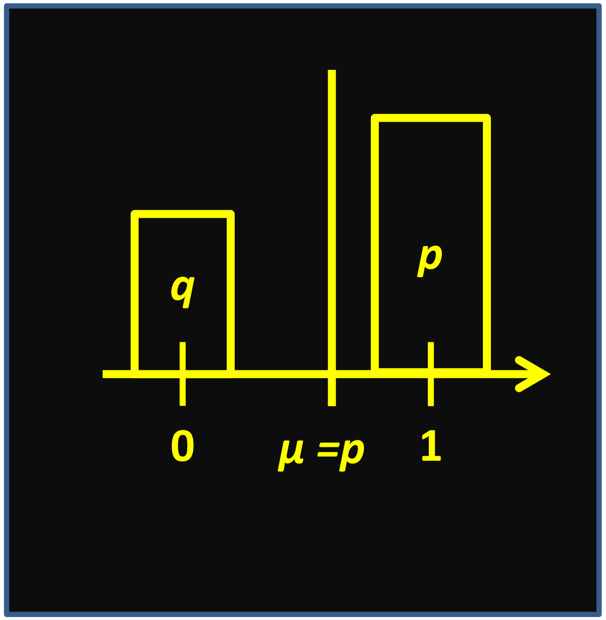
Distribution of
sampling averages $\bar{x}$
for N=10



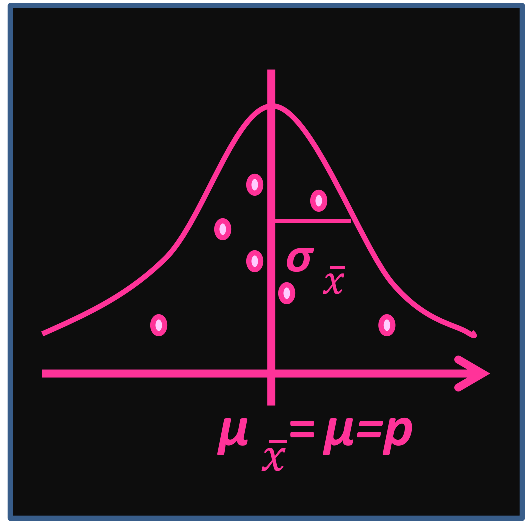
Distribution of
sampling averages $\bar{x}$
for N=100

The standard deviation will be smaller if the size of each sample is larger – the larger is each sample, the less is the error of estimating the real mean from this sample

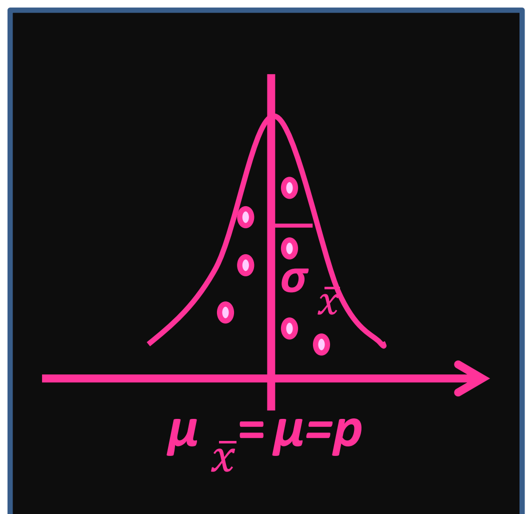
Standard deviation of sampling means



True distribution of classification success



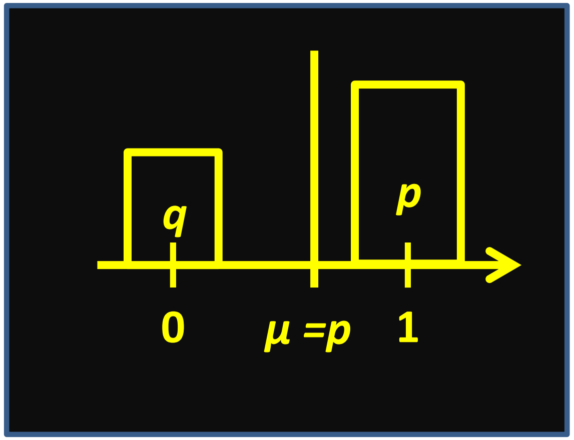
Distribution of sampling averages $\bar{x}$ for N=10



Distribution of sampling averages $\bar{x}$ for N=100

The dots, where each dot represents a mean of a particular sample, will fall closer to the real mean, if the size of each sample is large

Formula for standard deviation of the distribution of sampling means

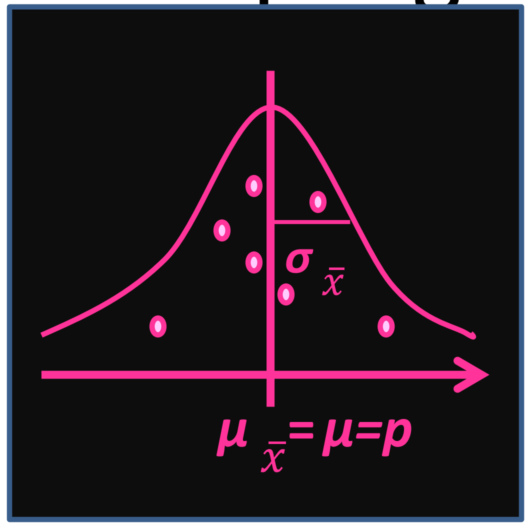


True distribution of classification success

If you take N=100 samples, you are much closer to the real mean than if you take N=2.

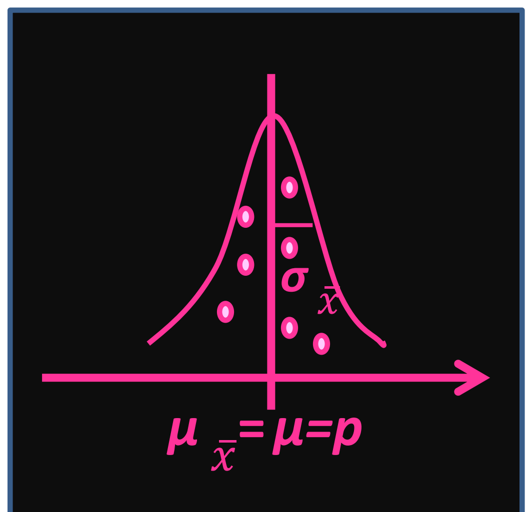
Turns out that: $\sigma^2_{\bar{x}} = \sigma^2/N$

Variance of the sampling mean distribution is inversely proportional to the size of the sample N



Distribution of sampling averages $\bar{x}$ for N=10

$\sigma_{\bar{x}} = \sigma/\sqrt{10}$



Distribution of sampling averages $\bar{x}$ for N=100

$\sigma_{\bar{x}} = \sigma/10$

Computing performance interval. Example

- How do we compute the predicted interval of classifier's success for a certain level of confidence?

- We sampled 100 instances: 75 correctly classified.
- Sample mean:

$\bar{x}$=(1*75+0*25)/100=0.75

- Sample variance:

S^2=[75*(1-0.75)^2+25*(0-0.75)^2]/ (N-1)=0.19

Adjustor – so we do not underestimate sample variance

Computing performance interval. Example

- How do we compute the predicted interval of classifier's success for a certain level of confidence?

- We sampled 100 instances: 75 correctly classified.
- Sample mean:

$\bar{x}$=(1*75+0*25)/100=0.75

- Sample variance:

s^2=[75*(1-0.75)^2+25*(0-0.75)^2]/ (N-1)=0.19

- Sample standard deviation:

s=sqrt(0.19)=0.435

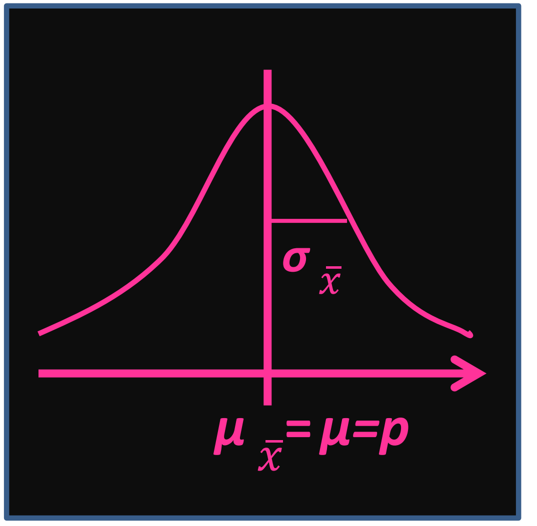
Computing performance interval. Example

- N=100 instances: 75 correctly classified.
- Sample standard deviation: ***s***=0.435

- We estimate the true standard deviation **σ** by sample standard deviation ***s***

- Now we can estimate one standard deviation of the distribution of sampling means:

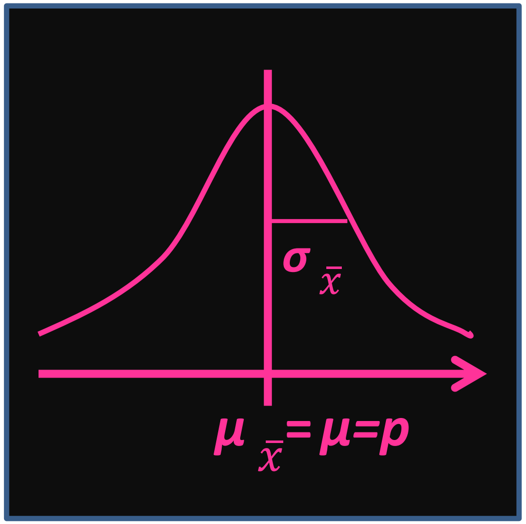
$\sigma_{\bar{x}}$= s/sqrt(N)=0.435/10=0.0435



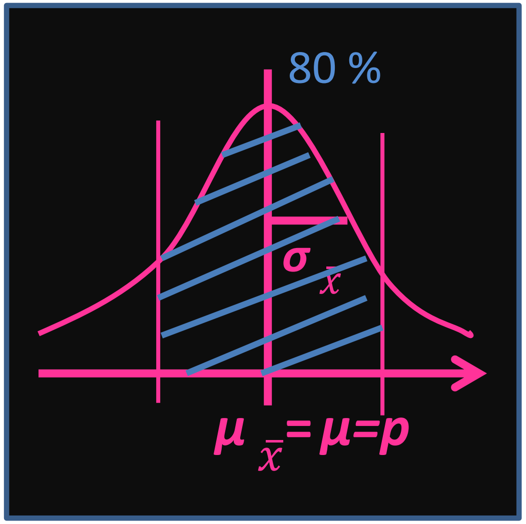
Computing performance interval. Example

$\sigma_{\bar{x}} = 0.0435$

How many such standard deviations away from the samplings mean we need to be to have 80% confidence that any random sample mean is within this interval?



Because the mean of the distribution of the sampling means is equal to the real mean μ, answering the previous question will answer: how big an interval should we allocate around μ, such that any random sampling of size N will have its mean within this interval



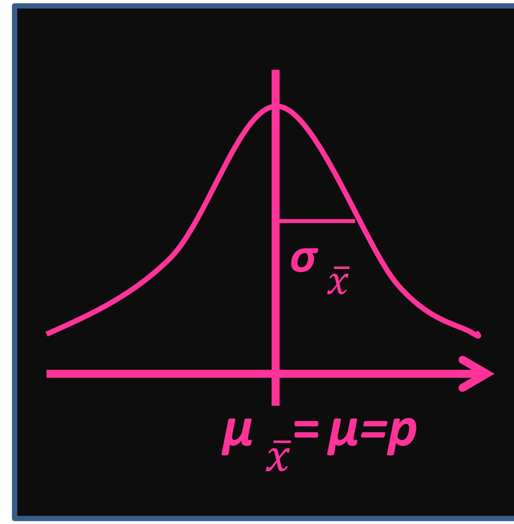
Computing performance interval. Example

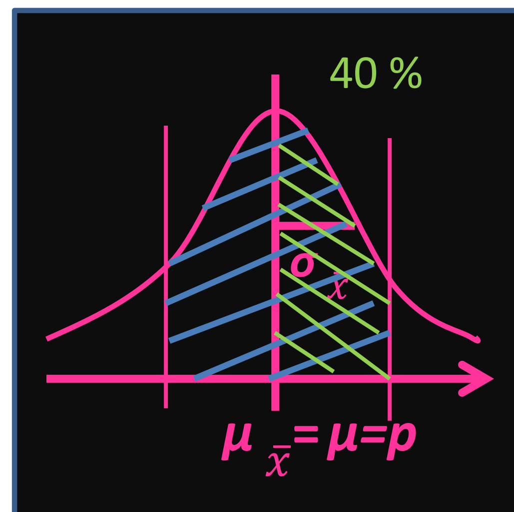
$\sigma_{\bar{x}} = 0.0435$

How many such standard deviations away from the samplings mean we need to be to have 80% confidence that any random sample mean is within this interval?

Because the mean of the distribution of the sampling means is equal to the real mean μ, answering the previous question will answer: how big an interval should we allocate around μ, such that any random sampling of size N will have its mean within this interval

We want the upper part (above mean) to be 40%, since normal distribution is symmetric.





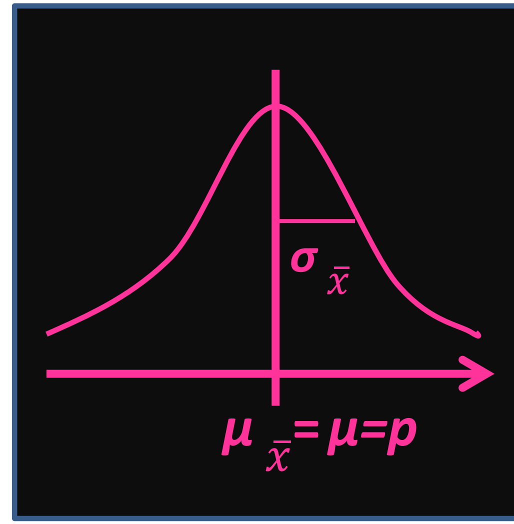
Computing performance interval. Example

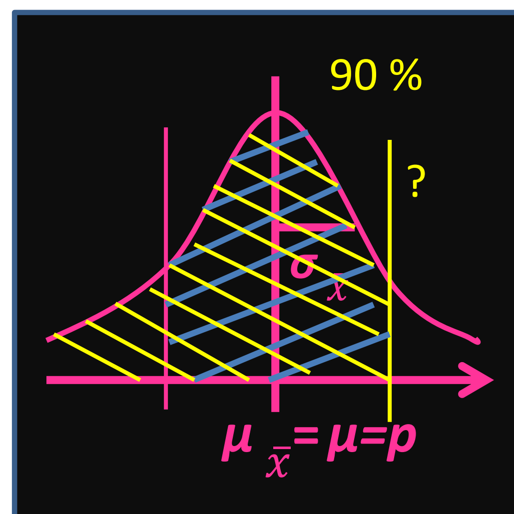
$\sigma_{\bar{x}} = 0.0435$

How many such standard deviations away from the samplings mean we need to be to have 80% confidence that any random sample mean is within this interval?

Because the mean of the distribution of the sampling means is equal to the real mean μ, answering the previous question will answer: how big an interval should we allocate around μ, such that any random sampling of size N will have its mean within this interval

The probability of the variable to be less than the upper mark is 40+50=90%





Computing performance interval. Example

$\sigma_{\bar{x}} = 0.0435$

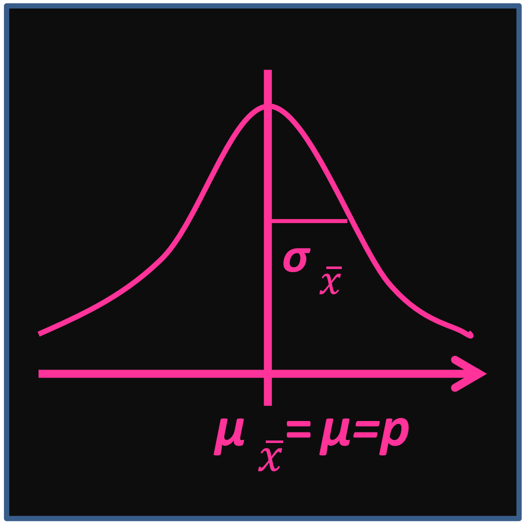
The probability of the variable to be less than the upper mark is 40+50=90%

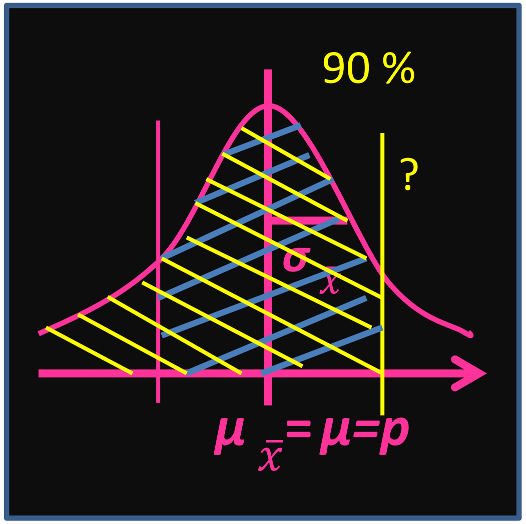
Cumulative probability up to this point

How many standard deviations above the mean

Z-table

z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.500	.504	.508	.512	.516	.520	.524	.528	.532	.536
0.1	.540	.544	.548	.552	.556	.560	.564	.568	.571	.575
0.2	.580	.583	.587	.591	.595	.599	.603	.606	.610	.614
0.3	.618	.622	.626	.630	.633	.637	.641	.644	.648	.652
0.4	.655	.659	.663	.666	.670	.674	.677	.681	.684	.688
0.5	.692	.695	.699	.702	.705	.709	.712	.716	.719	.722
0.6	.726	.729	.732	.736	.740	.742	.745	.749	.752	.755
0.7	.758	.761	.764	.767	.770	.773	.776	.779	.782	.785
0.8	.788	.791	.794	.797	.800	.802	.805	.808	.811	.813
0.9	.816	.819	.821	.824	.826	.829	.832	.834	.837	.839
1.0	.841	.844	.846	.849	.851	.853	.855	.858	.850	.862
1.1	$\boxed{.864}$	.867	.869	.871	.873	.875	.877	.879	.881	.883
1.2	.885	.887	.889	.891	.893	.894	.896	.898	.900	.902

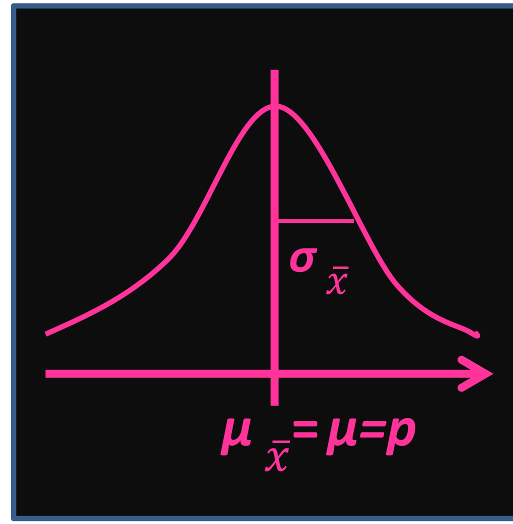




Computing performance interval. Example

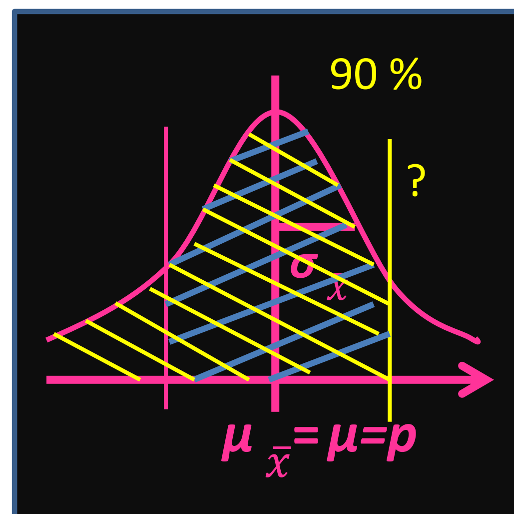
$\sigma_{\bar{x}} = 0.0435$

Our sample mean is less than real mean plus 1.28 standard deviations with probability 90%



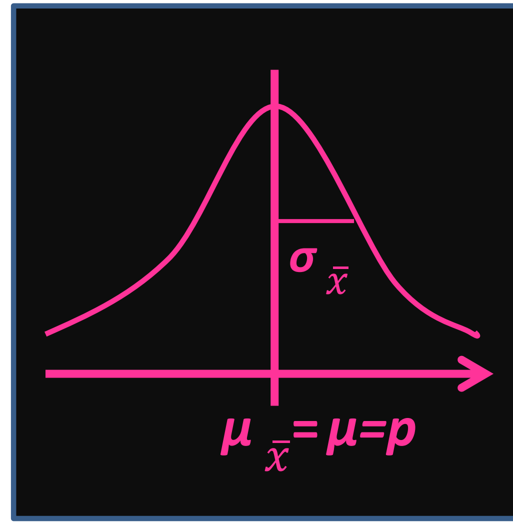
Z-table

z	**0.0**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	.500	.504	.508	.512	.516	.520	.524	.528	.532	.536
0.1	.540	.544	.548	.552	.556	.560	.564	.568	.571	.575
0.2	.580	.583	.587	.591	.595	.599	.603	.606	.610	.614
0.3	.618	.622	.626	.630	.633	.637	.641	.644	.648	.652
0.4	.655	.659	.663	.666	.670	.674	.677	.681	.684	.688
0.5	.692	.695	.699	.702	.705	.709	.712	.716	.719	.722
0.6	.726	.729	.732	.736	.740	.742	.745	.749	.752	.755
0.7	.758	.761	.764	.767	.770	.773	.776	.779	.782	.785
0.8	.788	.791	.794	.797	.800	.802	.805	.808	.811	.813
0.9	.816	.819	.821	.824	.826	.829	.832	.834	.837	.839
1.0	.841	.844	.846	.849	.851	.853	.855	.858	.850	.862
1.1	.864	.867	.869	.871	.873	.875	.877	.879	.881	.883
1.2	.885	.887	.889	.891	.893	.894	.896	.898	.900	.902



Computing performance interval. Example

$\sigma_{\bar{x}} = 0.0435$

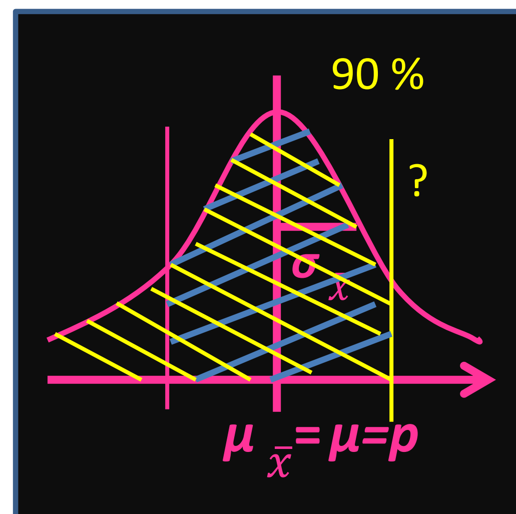


Our sample mean is less than real mean plus 1.28 standard deviations with probability 90%

Our sample mean $\bar{x}$=0.75 falls within 1.28 $\sigma_{\bar{x}}$ from the real mean $\mu=p$

or

the real mean $\mu=p$ is within 1.28 $\sigma_{\bar{x}}$ from the sample mean $\bar{x}$=0.75.



The real mean $\mu=p$ is between:

[$\bar{x}$ - 1.28 $\sigma_{\bar{x}}$, $\bar{x}$ - 1.28 $\sigma_{\bar{x}}$]

[0.75-1.28*0.0435, 0.75+1.28*0.0435]

[0.69, 0.805]

Computing performance interval. Result

The real mean $\mu=p$ is between:

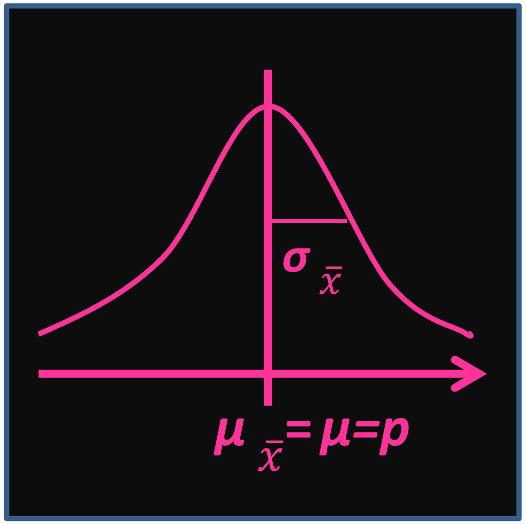
[0.69, 0.805] with the probability 80%

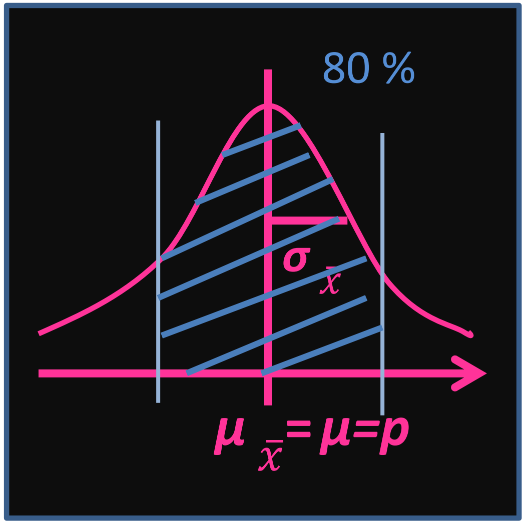
We can say that with confidence 80% the correctness of our classifier on real datasets is between 69% and 80.5%

Confidence – is a level of reliability of estimating the population parameter (in this case, the mean of the real population, $\mu=p$) from the sample data.

We may also say that the result [0.69, 0.805] is statistically significant with significance level 10%:

significance=100%-confidence





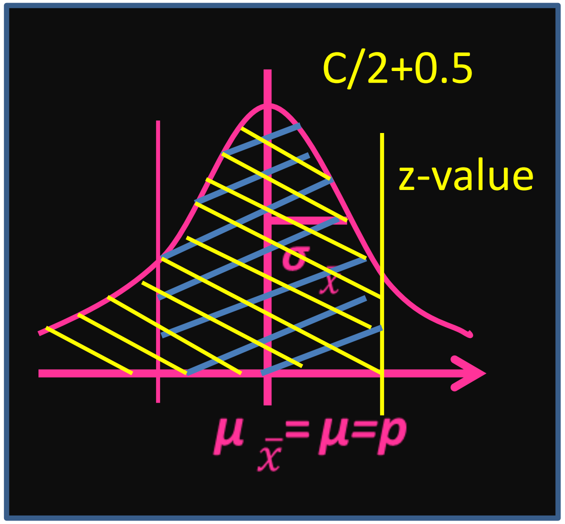
Computing confidence interval of classifier's success rate <u>in practice</u>

- Estimate real standard deviation by computing sample standard deviation:

$$\sigma^2 \approx \Sigma_i^N (mean_X - x_i)^2/(N-1)$$

- For confidence interval C, find z-value for C/2+0.5 (from the z-table)
- Real μ=p is within:

$$p = \bar{x} \pm z \frac{\sigma}{\sqrt{N}}$$

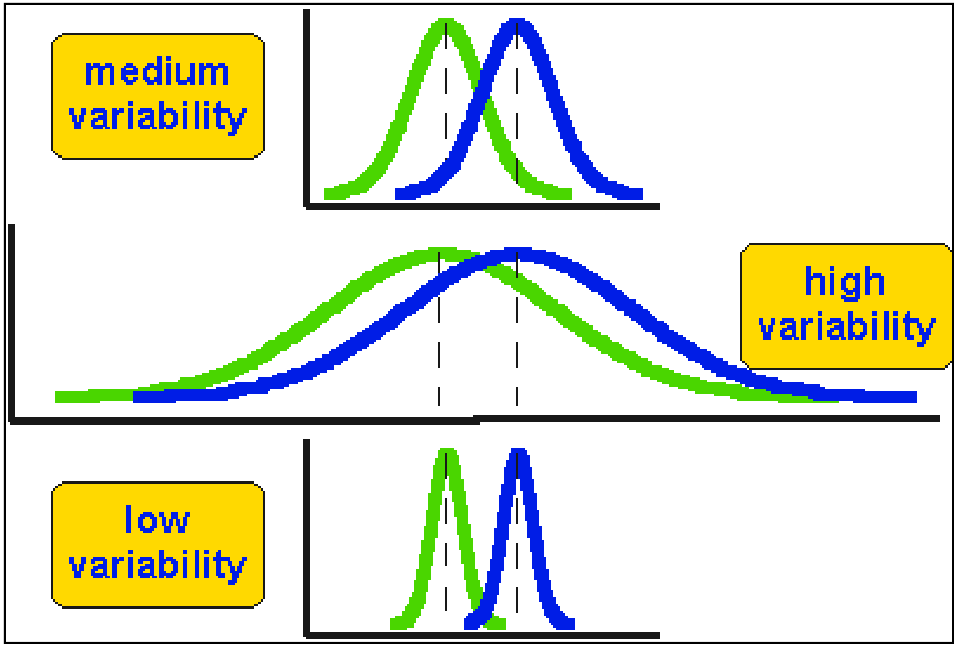


More statistics!

COMPARING PERFORMANCE OF TWO CLASSIFIERS

Comparing performance of learning schemes

- Which of two learning schemes perform better?
- Note: this is domain-dependent!
- Obvious way: compare error (success) rate on different test sets (for example, for different folds of cross-validation)
- Problem: variance in estimate



Statistical test for significant difference

- Question: are the means of two samples *significantly* different?
- In our case the samples are cross-validation accuracy for different folds from the same dataset
- The same Cross-Validation is applied twice: once for classifier A and once for classifier B

Probability distribution of sampling means

- Let m_X denote the mean of the probability of success of classifier A, and m_Y – the mean of the probability of success of classifier B

- We already know that the means of multiple samplings for each classifier are normally distributed around the real means μ_A and μ_B of classifier's success rate for the entire population

Probability distribution of sample mean differences

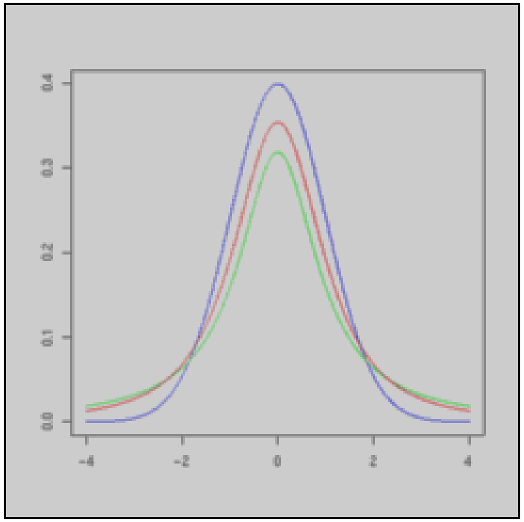
- We know how to estimate the intervals for the real means μ_A and μ_B for a certain confidence level
- *Suppose,* $\mu_A = 70 \pm 10$ *and* $\mu_B = 60 \pm 10$

- Which one is better?

Real means are somewhere inside these intervals.
Maybe they are just the same?

Probability distribution of sample mean *differences*

- If we take *k* samplings, and for each sample compute the difference of the means d_m, then for multiple samplings the distribution of the mean differences approaches the *Student's* distribution *T* with *k*-2 degrees of freedom



Student's distribution (red) for 2 degrees of freedom compared to normal distribution (blue)

Standard deviation of Student's distribution

- Student's distribution is very similar to the normal distribution
- Not surprisingly:
  - The experimentally estimated mean represents a mean μ_d of a real difference between X and Y for the entire population
  - The real standard deviation σ_d is inversely proportional to the sample size *N*:

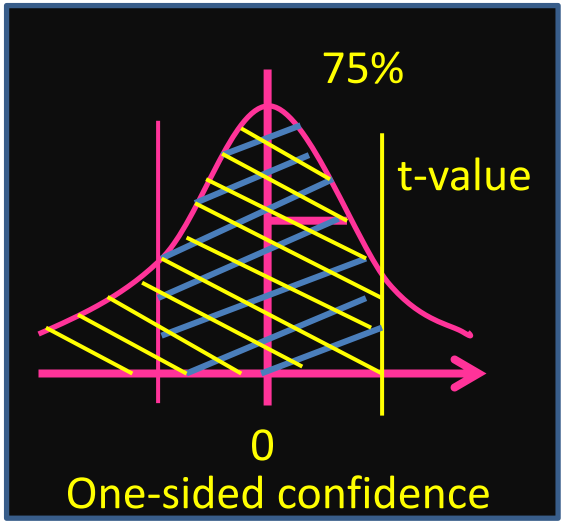
    $\sigma_d^2 = s_d^2/N$

Null-hypothesis

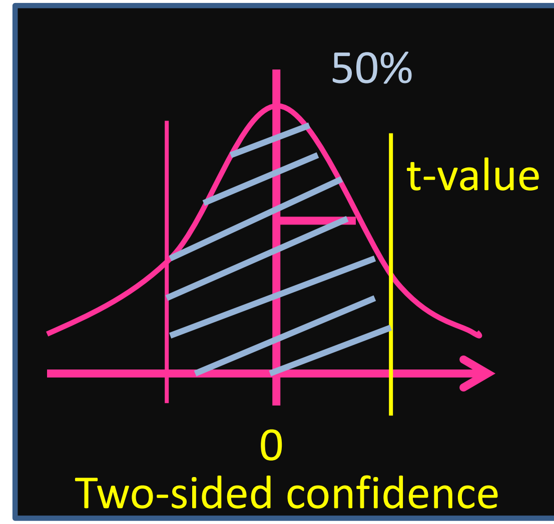
- We formulate our statistical hypothesis about the true value of μ_d:

$$\mu_d=0$$

Next, we select the level of significance (or confidence), and we find within how many standard deviations from the mean $\mu_d=0$ should be sample mean difference m_d of any random sampling in order to be still considered 0-difference (no statistically significant difference)



T-table

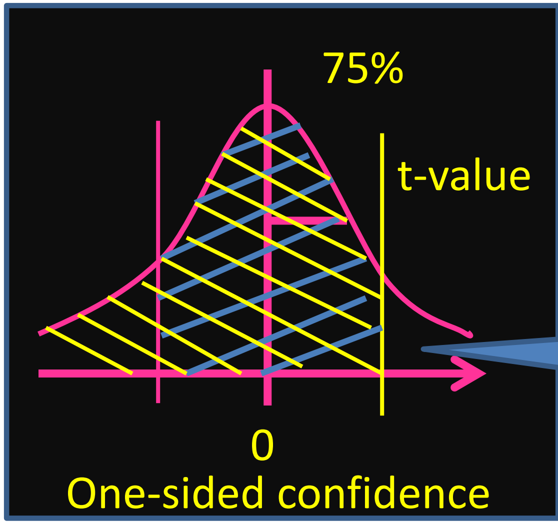


How many standard deviations from the mean – *t*-value

Degrees of freedom

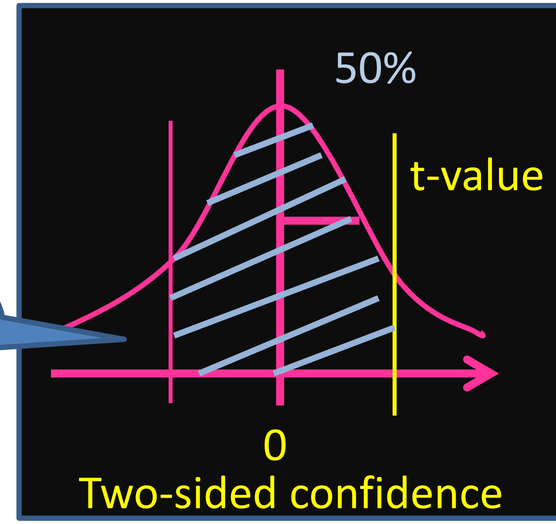
One Sided	**75%**	**80%**	**85%**	**90%**	**95%**	**97.5%**	**99%**	**99.5%**	**99.75%**	**99.9%**	**99.95%**
Two Sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437

T-table



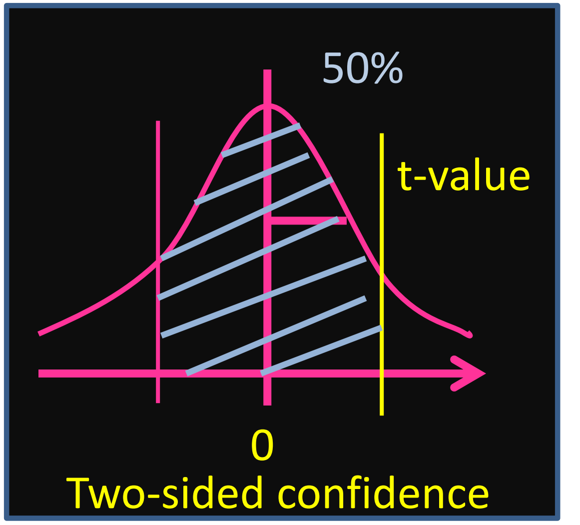
α=25%

α=50%



- One-sided test is used if we only interested if our difference is significantly greater than zero, or significantly smaller than zero, but not both
- Two-sided – if we are interested if our difference is significantly different from zero – both greater and smaller

T-test



- If the mean of differences of two samples is within the interval, then our Null-hypothesis is correct – there is no significant difference between two classifiers (for a given significance level)

- If the mean of differences is outside the interval, then the difference is significant (not by random chance), and we select the classifier with higher on average success rate

Comparing performance of two classifiers in practice

- Perform k classifications on each of k datasets using classifier A and classifier B in turn
- Compute difference of classification means for each dataset
- Find mean (average) and variance s of differences
- Fix a significance level α. Compute confidence for two-sided T-distribution: C=1.00 – α. Find t-value from the T-table for confidence C and k-2 degrees of freedom
- Find interval for the hypothesis μ_d=0: $\mu_d = 0 \pm t \frac{\sigma}{\sqrt{N}}$

- If the mean of differences is greater than $+t \frac{\sigma}{\sqrt{N}}$, then the first classifier is significantly better,
- if the mean of differences is less than $-t \frac{\sigma}{\sqrt{N}}$, then the second classifier is significantly better

Example. Input

- We have compared two classifiers through cross-validation on 10 different datasets (folds).
- The success rates are:

Dataset	Classifier A	Classifier B	Difference
1	89.4	89.8	-.4
2	90.2	90.6	-.4
3	87.7	88.2	-.5
4	90.3	90.9	-.6
5	91.2	91.7	-.5
6	89.4	89.8	-.4
7	90.2	90.6	-.4
8	87.7	88.3	-.5
9	90.3	90.9	-.6
10	91.2	91.7	-.5

Example. Mean and variance of differences

- m_d = -0.48
- s_d = 0.0789

$$\sigma_d = \frac{s_d}{\sqrt{k}} = \frac{0.0789}{\sqrt{10}} = 0.0249$$

Example. T-interval

σ_d=0.0249

The critical value of t for a two-tailed statistical test, α = 10% (c=90%) and k-2=8 degrees of freedom is: 1.86

The average difference should be outside the interval [-1.86*0.0249, 1.86*0.0249] in order to be significant

One Sided	**75%**	**80%**	**85%**	**90%**	**95%**	**97.5%**	**99%**	**99.5%**	**99.75%**	**99.9%**	**99.95%**
Two Sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781

Example. Solution

Significance $\alpha = 10\%$:

The average difference should be outside interval [-0.046, 0.046] in order to be significant

Our average difference is -0.48. The second classifier is significantly better than the first

The Inadequacy of success rates

- As the class distribution becomes more skewed, evaluation based on success rate breaks down.
  - Consider a dataset where the classes appear in a 999:1 ratio.
  - A simple rule, which classifies every instance as the majority class, gives a 99.9% accuracy – no further improvement is needed!
- Evaluation by classification success rate also assumes equal error costs---that a false positive error is equivalent to a false negative error.
  - In the real world this is rarely the case, because classifications lead to actions which have consequences, sometimes grave.