Performance Evaluation and Hypothesis Testing

add hints on how to evaluate with unbalanced data, see howtosvm.pdf

Motivation

- Evaluating the performance of learning systems is important because:
  - Learning systems are usually designed to predict the class of "future" unlabeled data points.
  - In some cases, evaluating alternative models (that we call hypotheses) is an integral part of the learning process.
  - For example, when pruning a decision tree, alternative pruned trees represent different "hypotheses" on how to interpret our data; in neural networks, different network architectures – with different numbers of hidden layers – also represent alternative hypotheses. **Which one is the best predictor of the reality**?

Issues

- **Which performance measure we should use?**
- How well can a classifier be expected to perform on "novel" data, not used for training?
- Since a performance measure is an ESTIMATE on a sample, how accurate is our estimate?
- How to compare performances of different hypotheses or those of different classifiers?

Performances of a given hypothesis

- Performances are usually reported in the form of a CONFUSION MATRIX (also called contingency table)
- The table has four cells (in case of binary classifiers):
  - TP: "true positive", i.e., number (or %) of positive instances classified as positive by the system
  - FP: "false positive", should be negative, the system classified as positive
  - TN: "true negative" negative instances classified as negative
  - FN: "false negative" positive instances classified as negative

$$ACCURACY = \frac{TP + TN}{|T|}$$

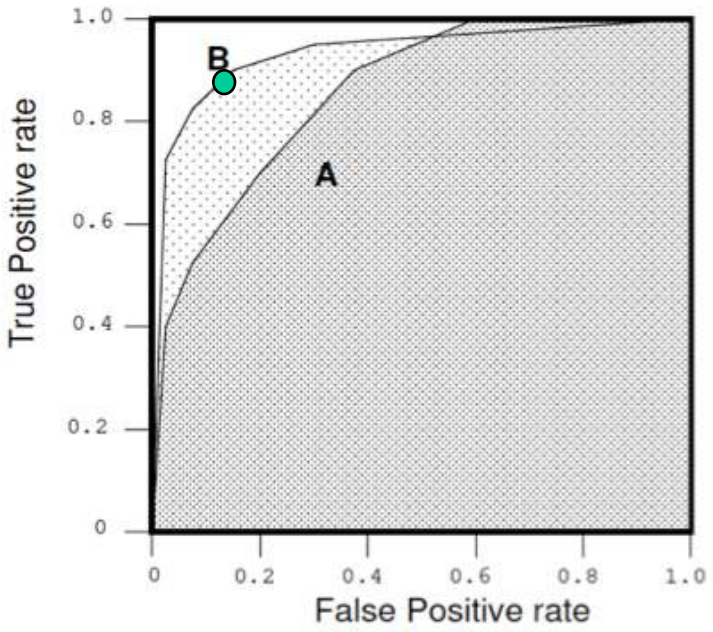
$$T = TP + TN + FP + FN$$

Performances of a given hypothesis

- Precision, Recall, F-measure

- $P=TP/(TP+FP)$
- $R=TP/(TP+FN)$
- $F=2(P \times R)/(P+R)$

ROC curves plot precision and recall

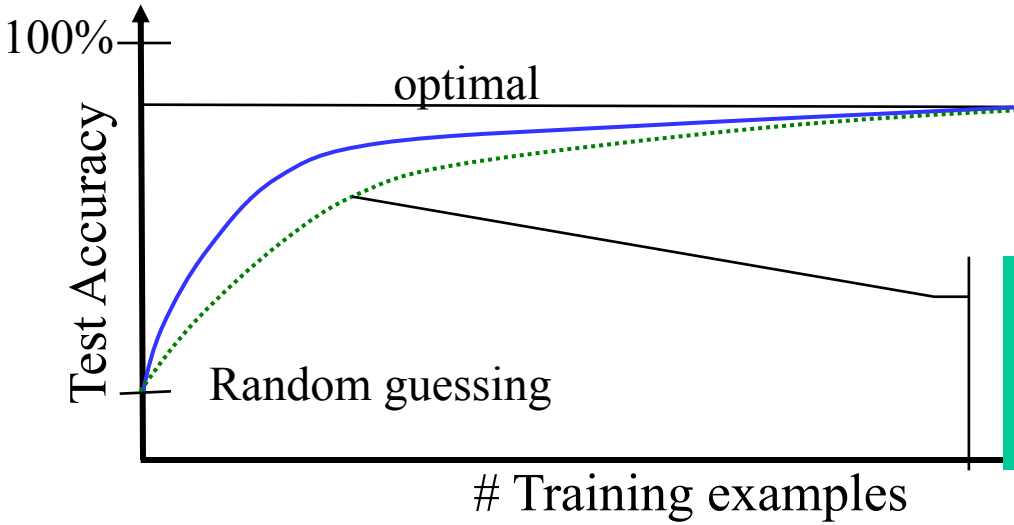


A and B here are two alternative runs of a ML algorithms (with different parameters and settings)

Receiver Operating Characteristic curve (or ROC curve.) is a graphical plot that illustrates the performance of a binary classifier system
The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various system settings (parameters, dimension of learning set, etc). One would obviouly aim at high TPR and low FPR.

Learning Curves

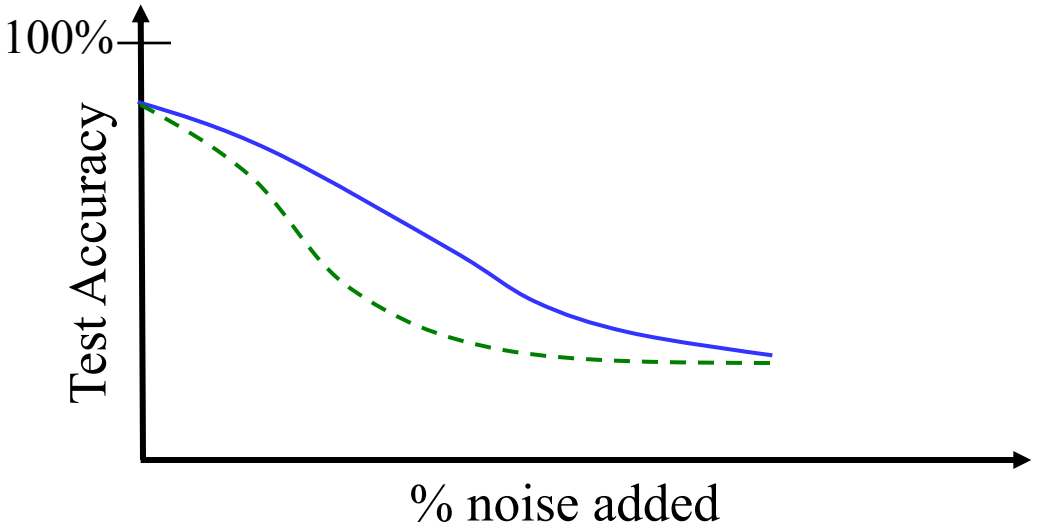
- Plots accuracy vs. **size of training set**.
- Has maximum accuracy nearly been reached or will more examples help?
- Is one system better when training data is limited?
- Most learners eventually converge to optimal given sufficient training examples.



Dotted line is some **baseline**, e.g random guess

Noise Curves (to test robustness)

- Plot accuracy versus noise level to determine **relative resistance to noisy training data**.
- Artificially add category or feature noise (i.e. instances with wrong classification or wrong/missing feature values) by randomly replacing some specified fraction of category or feature values with random values.



Issues

- Which performance measure should we use?
- **How well can a classifier be expected to perform on "novel" data, not used for training?**
- Since a performance measure is an ESTIMATE on a sample, how accurate is our estimate?
- How to compare performances of different hypotheses or those of different classifiers?

Evaluating an hypothesis

- ROC and Accuracy not enough
- How well will the learned classifier perform on novel data?
- Performance on the training data **is not a good indicator** of performance on future data

Example

Learning set

Test: is it a lion?

Testing on the training data is not appropriate.
The learner will try to fit at best on available data, and will not learn to **generalize**.
Possibly, it will misclassify new unseen istances.

Difficulties in Evaluating a learned model when only limited data are available

- ***Bias in the estimate:*** The observed accuracy of the learned "hypothesis" (model) over the training examples D is a **poor estimator** of its accuracy over future examples ==> we must test the hypothesis on a test set S chosen **independently** of the training set and the hypothesis.
- ***Variance in the estimate:*** Even with a separate test set, the measured accuracy can vary from the true accuracy, depending on the makeup of the particular set of test examples. The smaller the test set, the greater the expected variance.

Variance (intuitive)







High variance

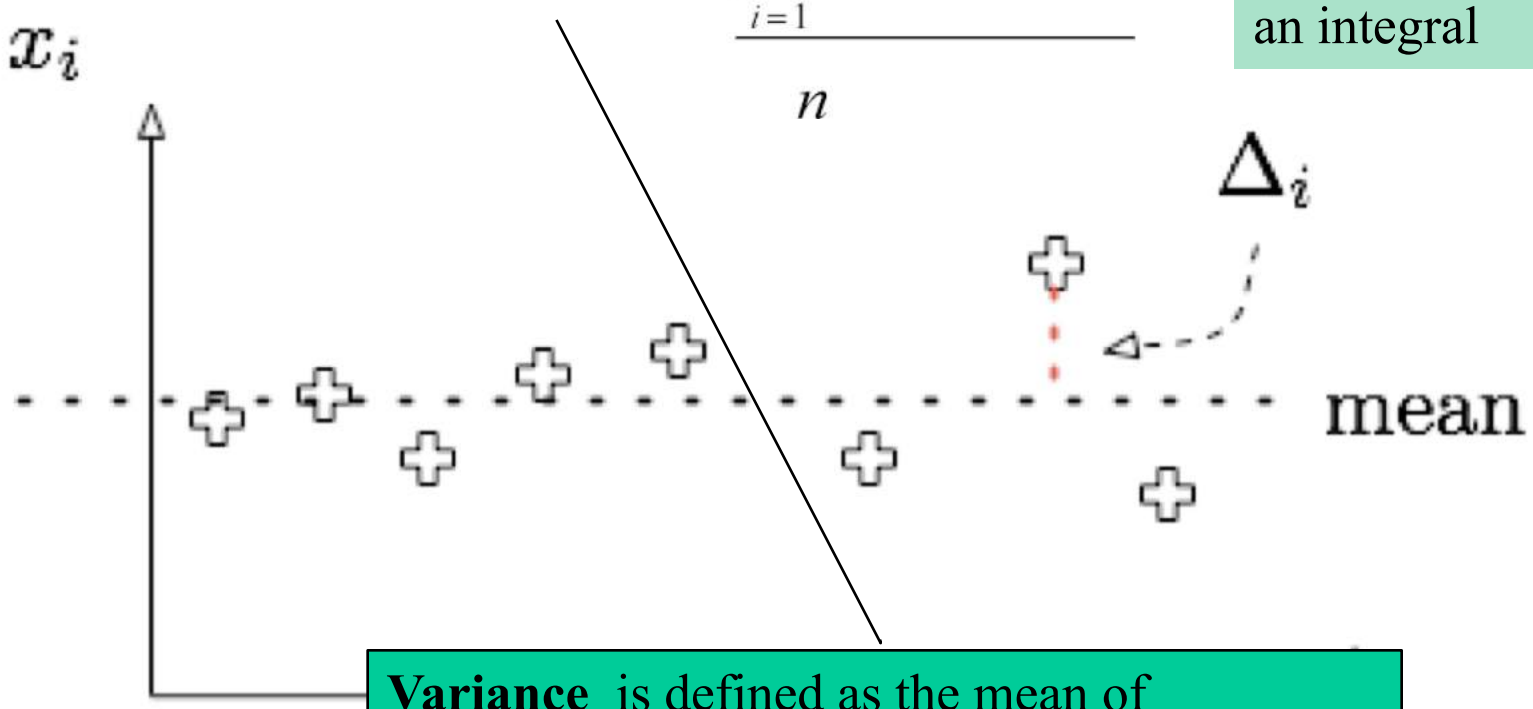



Low variance

Variance (for any discrete distribution)

$$\text{var}_x = \frac{\sum_{i=1}^{n} (x_i - x)^2}{n}$$

Note: for infinite values the sum becomes an integral

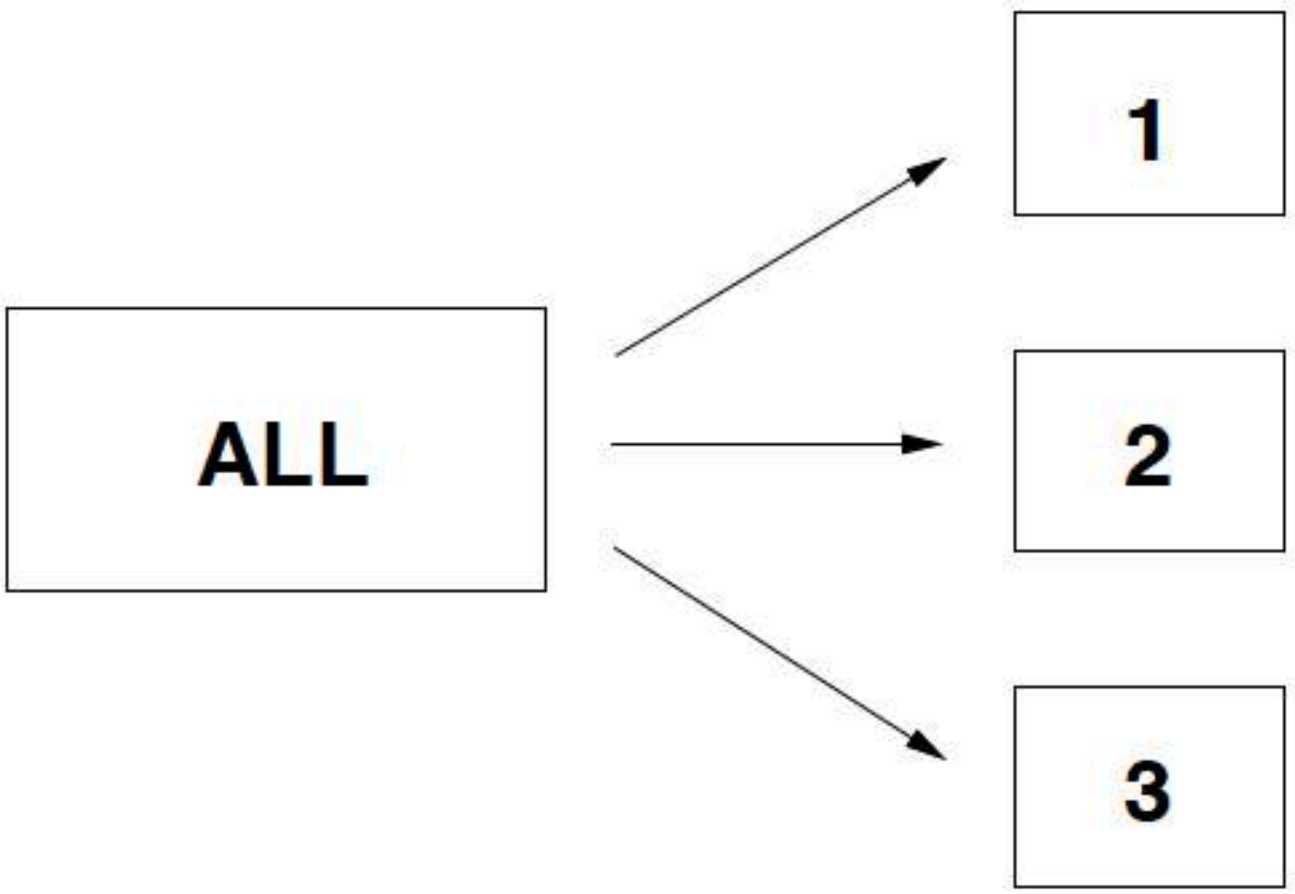


Variance is defined as the mean of **square differences** between values of *n* individual outcomes X_i and the mean (x), i.e. the *dispersion around the mean*.

How to reduce the variance?

- A common method (if enough labeled data are available) is to perform several independent split on learning and test set, and then average the perfomances over these different split
- Known as k-fold cross evaluation

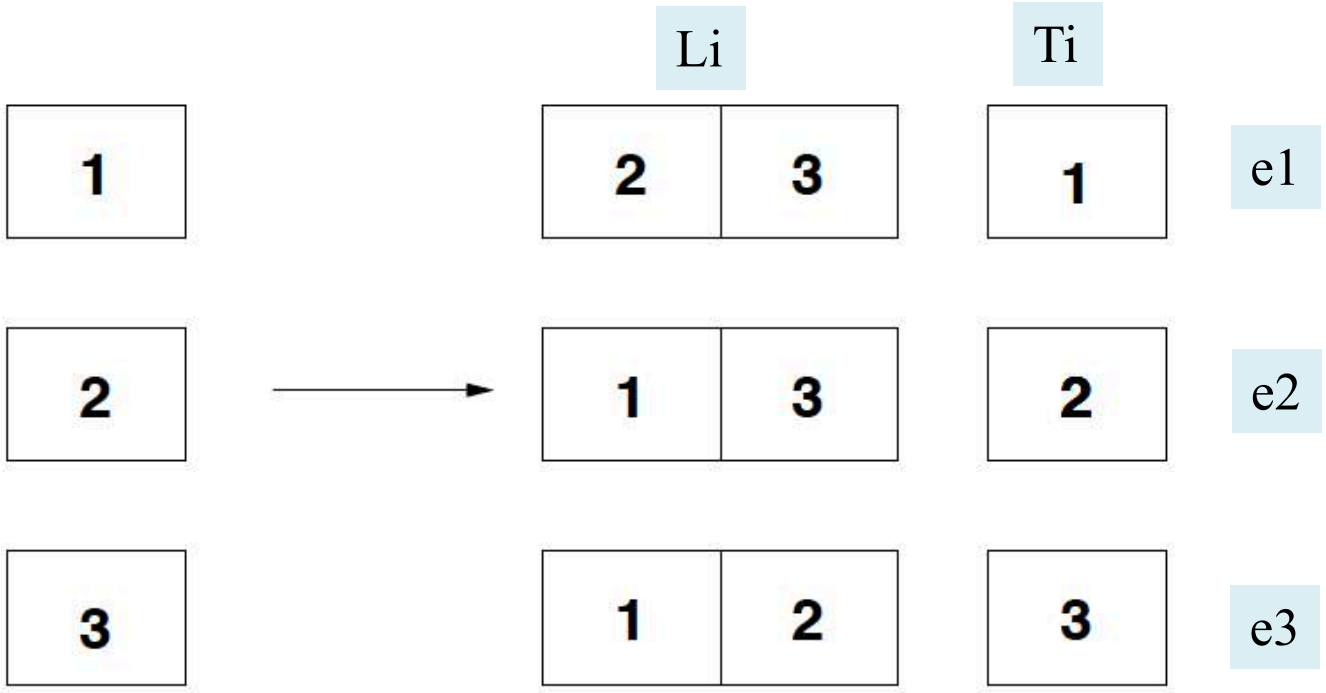
K-fold cross validation of an hypothesis



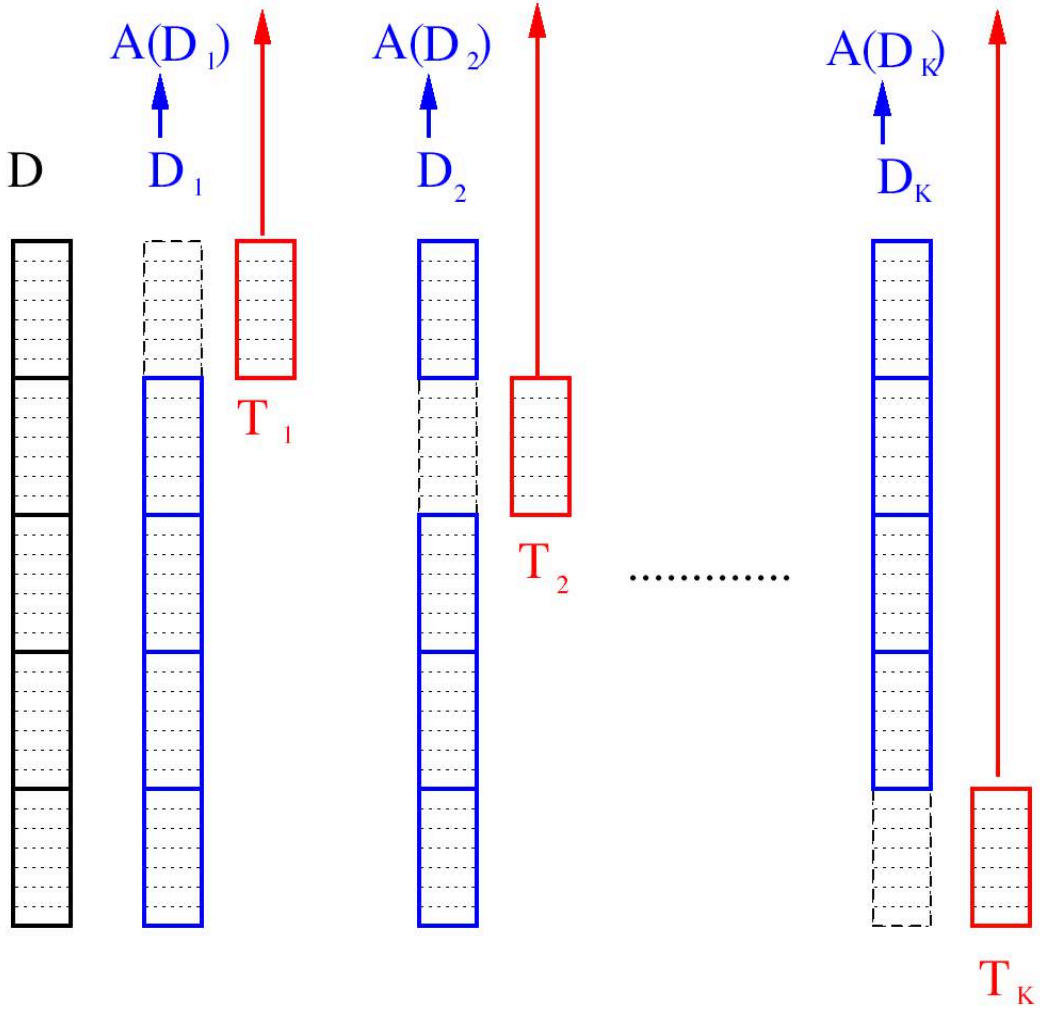
Partition all the available labeled data in **k equally sized** random samples

K-fold cross validation (2)

At each step, learn from Li and test on Ti, then compute the error on Ti



K-FOLD CROSS VALIDATION



Why k-fold reduces the variance?

- Intuitively, it reduces the probability of being lucky, or unlucky, in selecting the test set
- To understand the issue more in detail, we need to introduce the next topic: testing the accuracy of an error estimate

Issues

- Which performance measure we should use?
- How well can a classifier be expected to perform on "novel" data, not used for training?
- **Since a performance measure is an ESTIMATE on a sample, how accurate is our estimate?**
- How to compare performances of different hypotheses or those of different classifiers?

Questions to be considered in hypothesis testing

Let *h* be a model learned by a specific ML algorithm *L* with some specific setting of hyper-parameters. We denote *h* as an "hypothesis", and the objective is **to estimate its prediction accuracy**. The following are relevant questions:

Q1: Given the observed accuracy of *h* over a limited sample of test data S, **how well does this estimate** its accuracy over additional (unseen) instances?

Q2: Given that one hypothesis *h1* outperforms another (*h2*) over some sample data S, **how probable is** it that this hypothesis is more accurate in general (= over the full instance space X)?

Q3. When available classified data is limited, what is the **best way to use** this data to both learn a hypothesis and estimate its accuracy?

1. Estimating Hypothesis Accuracy

A better formulation of Q1:

- Given a hypothesis *h* and a data sample containing *n* examples (instances) drawn at random according to distribution $\mathcal{D}$, **what is the best estimate** of the accuracy of *h* over **future instances** drawn from the same distribution? => ***sample error*** vs. ***true error***
- What is the probable error in this accuracy estimate? => ***confidence intervals= error in estimating error***

In other terms, if we measure an error rate of, say, 20% on a sample test set, the true error rate of *h* on any sample is NOT guaranteed to be exactly 20%. Let's say that it is $20\% \pm \Delta$. Can we estimate the value of Δ?

Sample Error and True Error

- **Definition 1:** The ***sample error*** (denoted $error_S(h)$ or $e_S(h)$) of hypothesis h with respect to target (true) classification function f, on a data sample S is:

$$error_S(h) = 1/n \times \Sigma_{x \in S} \delta(c(x), h(x)) = r/n$$

where n is the number of instances in sample S, $c(x)$ is the ground-truth classification of instances in S, $h(x)$ is the classification produced by our current model h, and the quantity $\delta(c(x), h(x))$ is 1 if $c(x) \neq h(x)$, and 0, otherwise.

- **Definition 2:** The ***true error*** (denoted $error_{\mathcal{D}}(h)$, or p) of hypothesis h with respect to target (unknown) classification function f and distribution $\mathcal{D}$ *of instances*, is the **probability** that h will misclassify an instance x in X drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) = p = Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$

Sample error is an **estimate** on S of the true error on $\mathcal{D}$, which is a **probability**

Estimate, probability and random variables

- We are given a sample S of *n* instances, we classify S with *h(x)* and we measure *r* errors, we then estimate the error probability of *h(x)*: P(r errors in n instances)=*r/n*
- *Note: we call S "sample" since it can be **any** subset X' of X sampled according to $\mathcal{D}$.*
- However, *r (or r/n)* is a RANDOM variable, governed by **chance**. If we get another sample S' of *n* **different** instances, we may get a different number *r'* and a different estimate. In general $\mathbf{error_S(h) \neq error_{S'}(h)}$!!!
- **Simple experiment**: make *k* different sets of trials, in each toss a coin 10 times and measure the number of "head". Although, as the number of experiment, *k*, increases, the average number of "head" occurrences tend to ½, in each single trial you will likely obtain different numbers.
- In coin tossing, we know that the "real" head rate is 50%, but in hypothesis testing, we don't know what is the real error rate. So, how can we get an idea of $error_{\mathcal{D}}(h)$ on the entire population X, distributed according to $\mathcal{D}$?

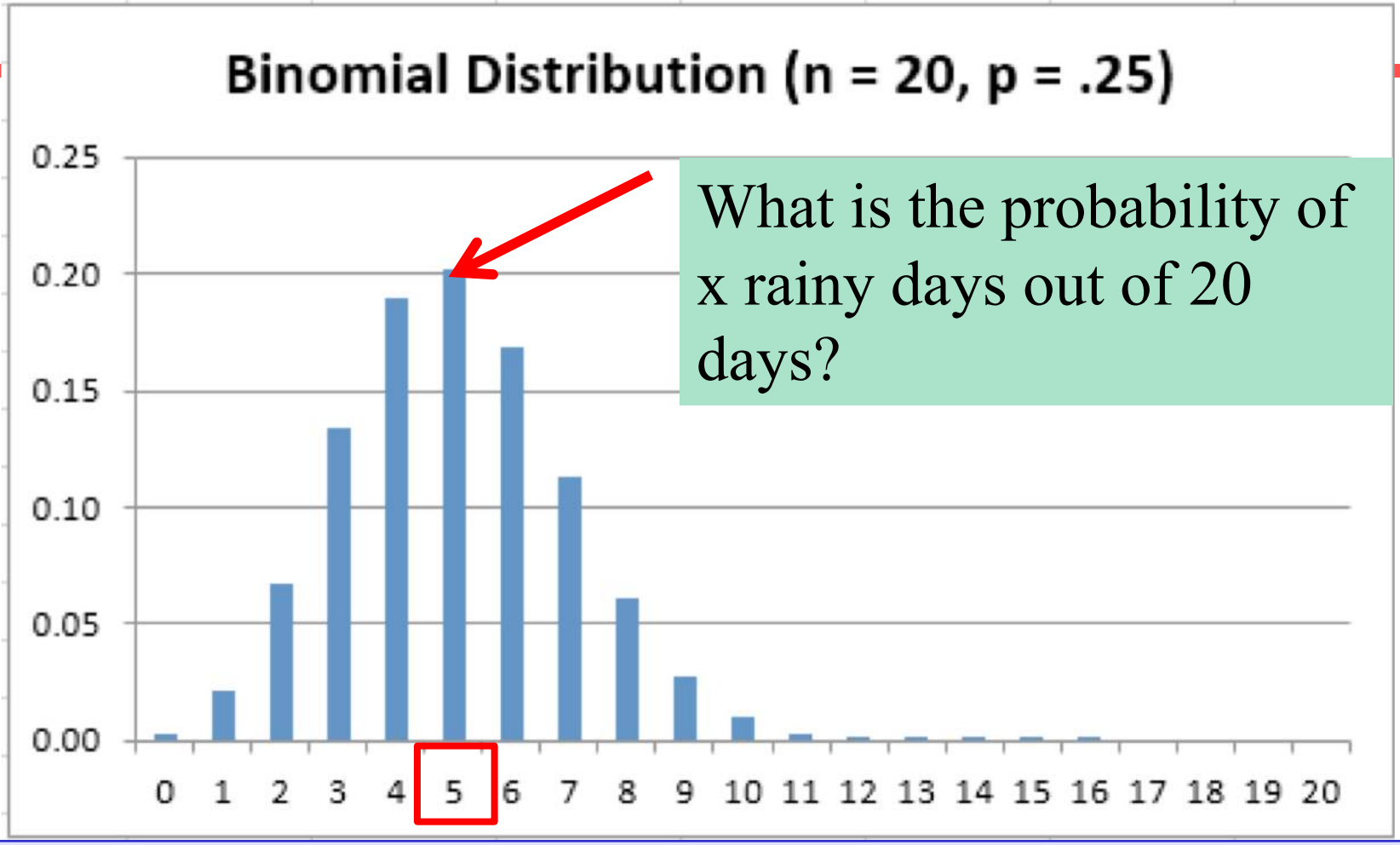
Sample error and true error

- We do not know the "true" error probability however we know that $error_{\mathcal{D}}(h)$ is a **random variable** that follows a **binomial distribution** with mean p (unknown)
- Why? And what is this "binomial"?
- Say p is the (unknown) "true" error probability of h(x) on X. If we have a sample of n instances, what is the probability that, given instances x in S, $c(x) \neq h(x)$ for r times??
- Even if we do not know the error rate p, each instance x in S has probability ***p*** of being misclassified by h(x) and ***(1-p)*** of being classified correctly.
- The probability of observing r misclassified exam instances is then:

$$P(error_D(h) = r/n) = \binom{r}{n} p^r (1-p)^{n-r} = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

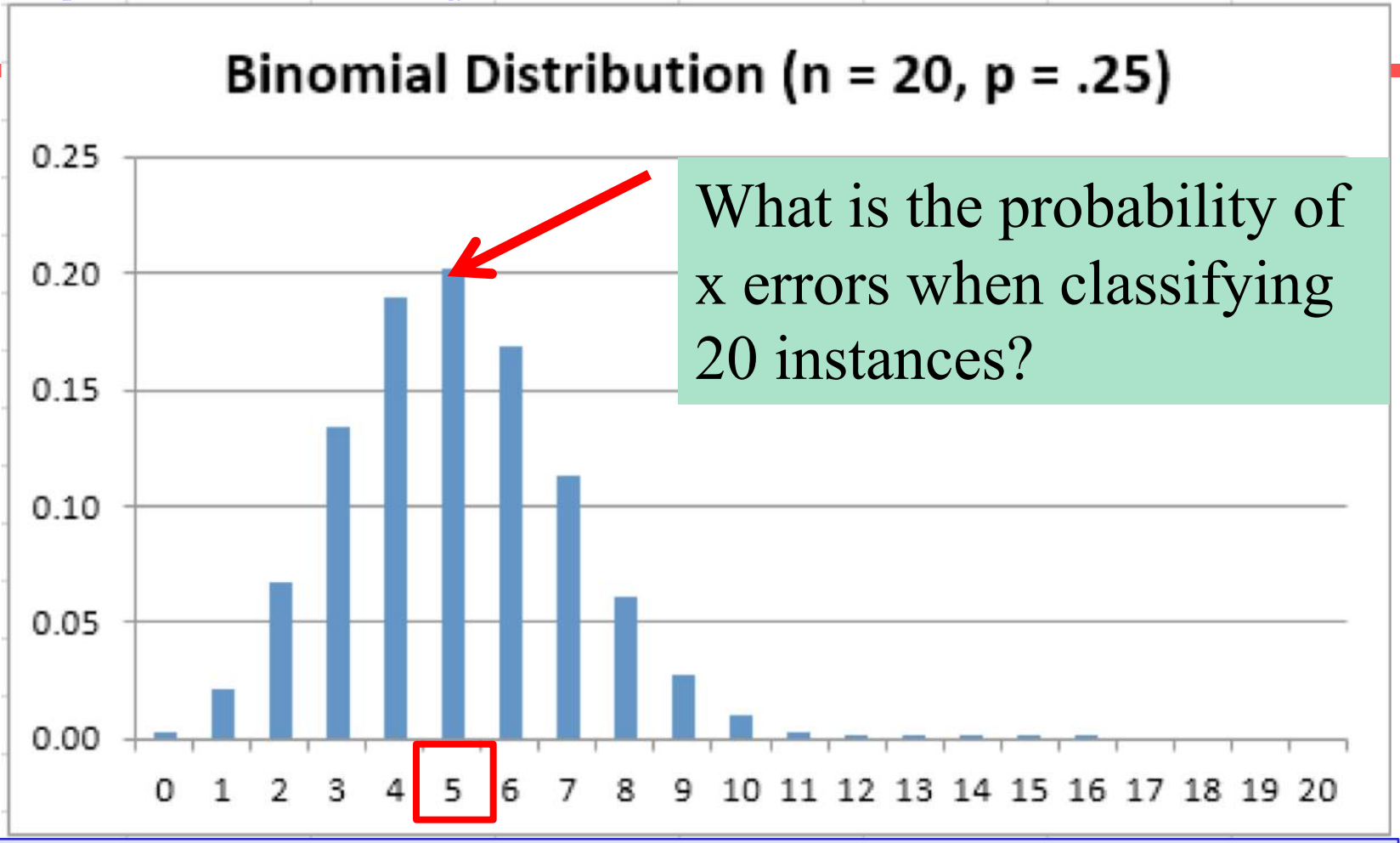
# of ways in which we can select r items from a population of n

Example- *p* is the probability of rain in january



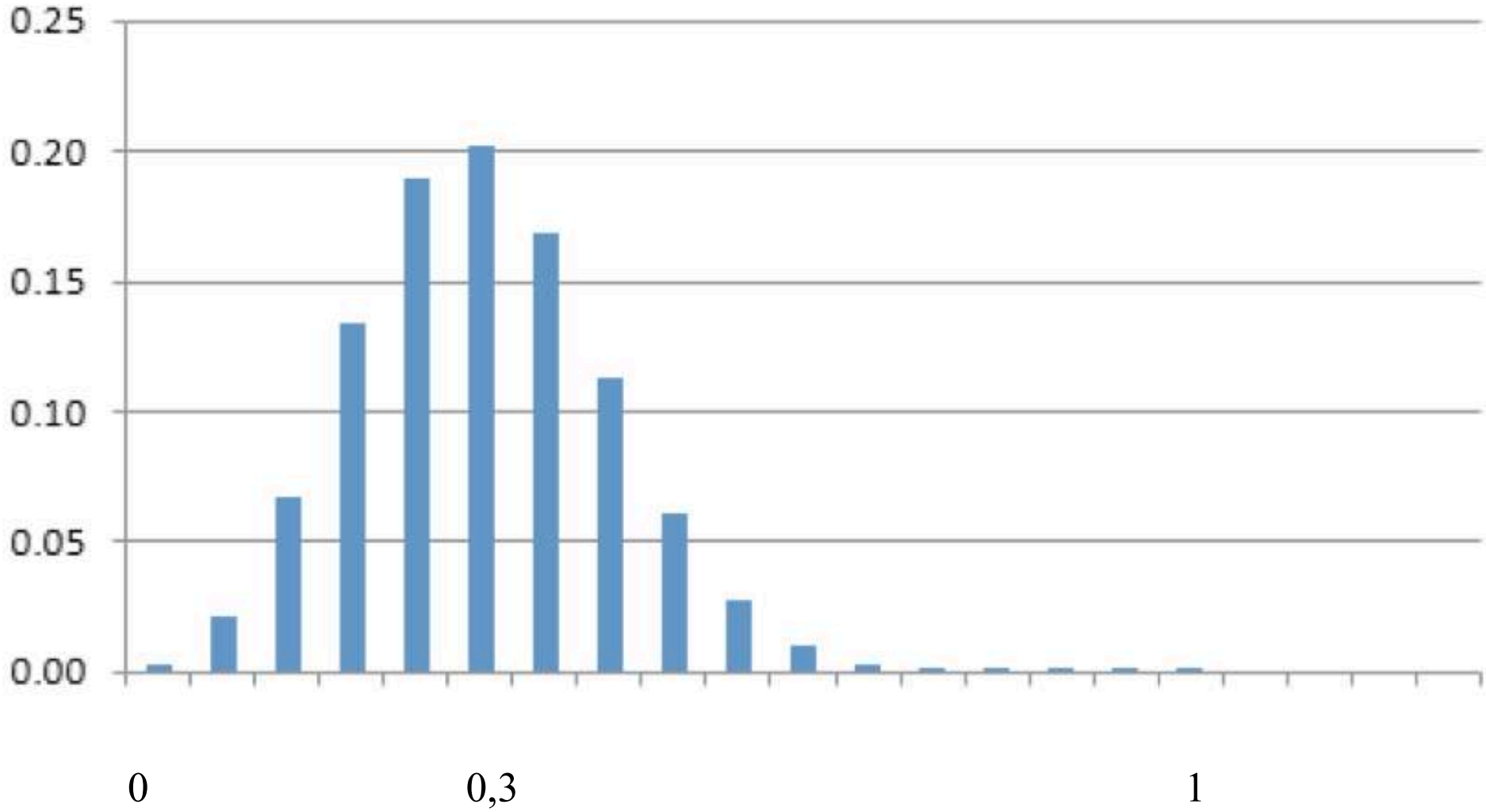
Abscissa is the value **r**, e.g. there is a 20% probability that there will be 5 rainy days out of 20, 6% probability of 8 out of 20, etc.

Example- *now p is the probability that our ML system misclassifies an instance x drawn at random*



Abscissa is the value **r**, e.g. there is a 20% probability that there will be 5 errors out of 20 classifications, 6% probability of 8 out of 20, etc.

We can normalize and plot r/n



Now x is the % of wrongly classified instances

How do we compute these probabilities?

- Say we know that *p(rain)*=25% (on january)
- However, if we watch the weather in 4 consecutive days, **we are not sure** we will get "rain" 1 time and "not rain" 3 times. The number of observed "rainy days" in each trial of 4 consecutive days **is governed by chance**.
- What is the probability of getting, instead, 2 rainy days in 4 days?

$$P(2\text{ "rain" } in\ 4\ observed\ days) = \binom{4}{2}(0.25)^2(1-0{,}25)^2 =$$

$$= \frac{4!}{2!\,(4-2)!}(0.25)^2(1-0{,}25)^2 = \frac{1\cdot 2\cdot 3\cdot 4}{1\cdot 2(1\cdot 2)}\,0.065(0.5625) = 0{,}21$$

Same formula to estimate the probability of 2 errors over 4 instances, given we know that the true error rate is 25%

Properties of Binomial distribution

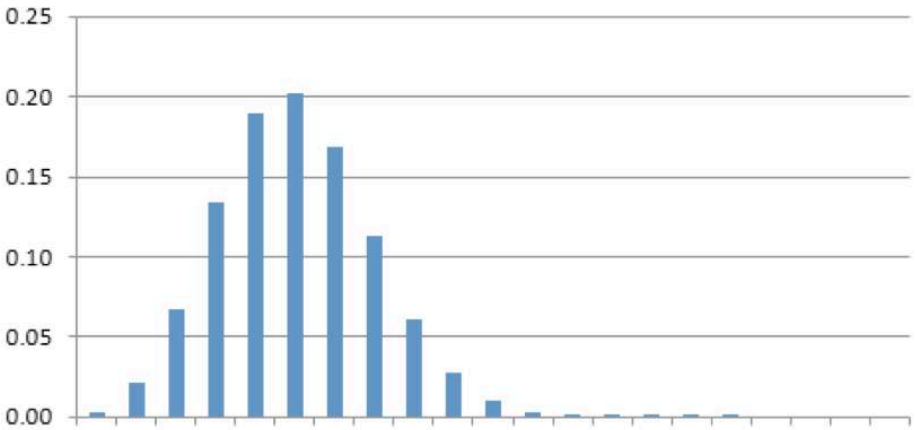
$$P\left(error_S = \frac{r}{n}\right) = \frac{n!}{r!\,(n-r)!} p^r (1-p)^{n-r}$$

$E_D(X) = \mu(X) = \boldsymbol{p}$ Expected value

where for short we use X to denote $error_S$ *,*
the random variable representing the observed errors in a trial

$Var_D(X) = p(1-p)$ variance

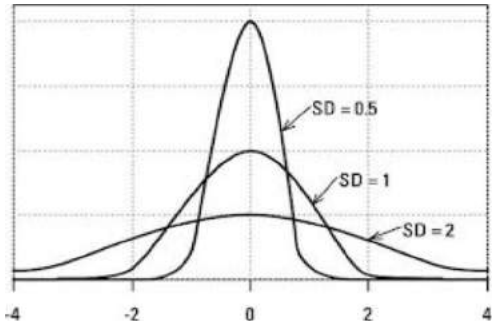
$\sigma_D(X) = \sqrt{p(1-p)}$ standard deviation



$E_{\mathcal{D}}(X)$= **expected value** of random variable X in distribution $\mathcal{D}$
(also denoted **μ(X)** or **X^**)
Var =variance
σ = standard deviation (also denoted SD)

Variance and Standard deviation of Binomial: why *p(1-p)*?

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - p)^2 = \frac{1}{n}(np(1-p)^2 + n(1-p)(0-p)^2) = p(1-p)$$



$$SD(X)=\sigma(X) = \sqrt{Var(X)} = \sqrt{p(1}$$

X_i is the outcome of a single trial. If the outcome is binary (e.g, X_i=1 means error and 0 non-error), and the probability of X_i=1 is p, then over n trials we expect that for *np* times X_i=1, for *n(1-p)* times X_i=0

Standard deviation of a random variable is the square root of its variance

Estimator of an error

So, we know that $error_{\mathcal{D}}(h)$ follows a **binomial distribution** with unknown mean **p**. If we sample the error on a set of instances S, we obtain a **value** r/n which is our current ESTIMATE $error_S(h)$ of $error_{\mathcal{D}}(h)$.

Note that the "estimator" $error_S(h)$ of p is also a random variable! *If we perform many experiments on different samples S we could get different values*

However, **for large enough dimension of the sample**, the mean (expected value) of $error_S(h)$ **is the same** es for $error_D(h)$!

Why? Because of the ***central limit theorem***

Central Limit Theorem

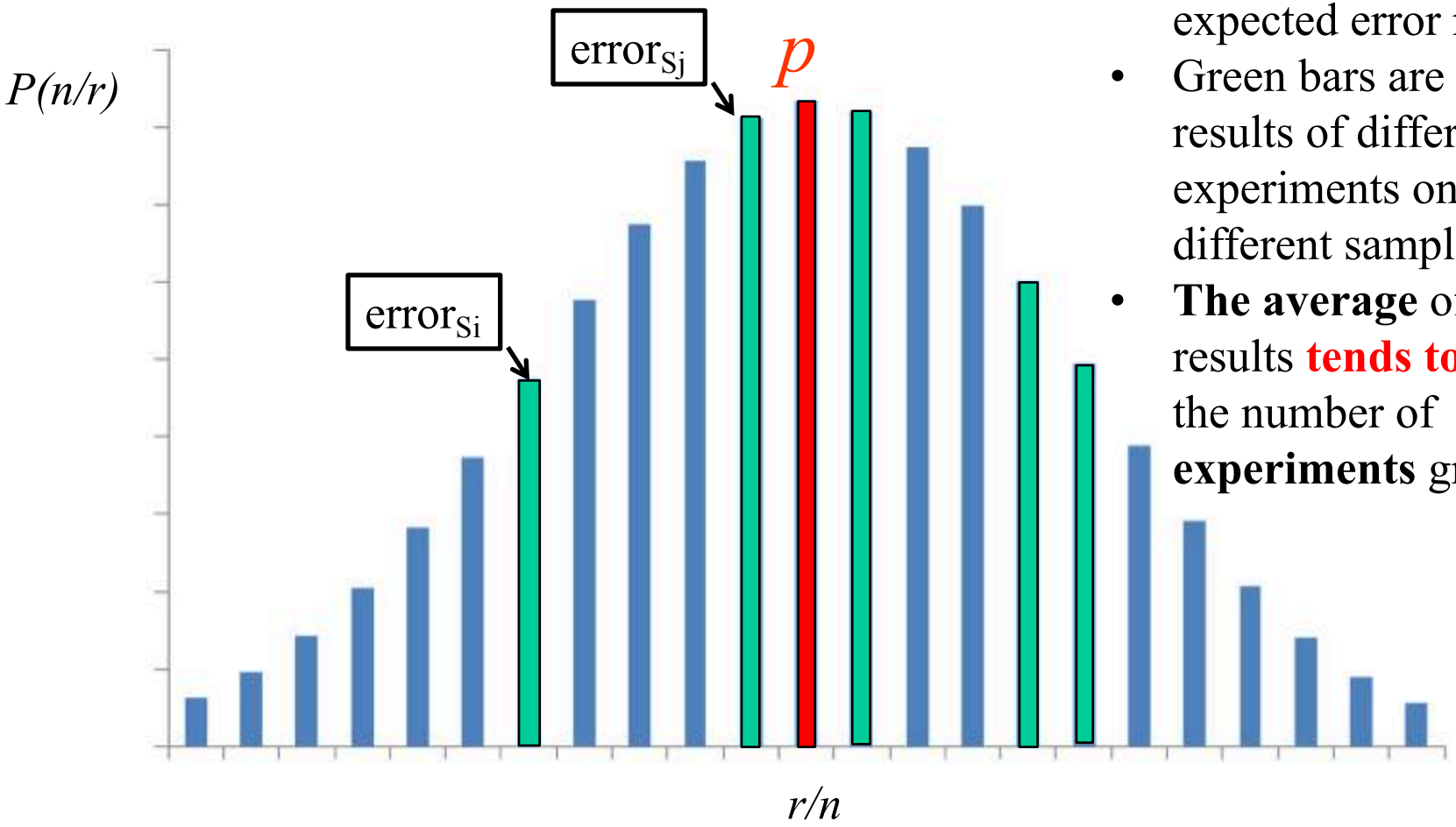
- (General formulation) The theorem states that the arithmetic mean of a sufficiently large number of experiments of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed
- This will hold true regardless of whether the source population is normal or skewed, provided the samples size is sufficiently large (usually $n \geq 30$).
- Futhermore, **the mean of all such experiments will (tend to) be the same as the "real" population mean**

Applied to our case (testing error rate)

In our case:

- **a) experiments are accuracy test on data samples S_i;**
- **b) the random variables are the error rates r_i/n_i observed on these samples S_i, they are independent from each other, and they follow a binomial;**
- **c) for sufficiently large number of experiments, the values r_i/n_i will be approximately normally distributed, and**
- **d) their mean value will tend to the "real" (unknown) error rate p over the entire set of instances X**

$avg(error_S(h)) \rightarrow p$

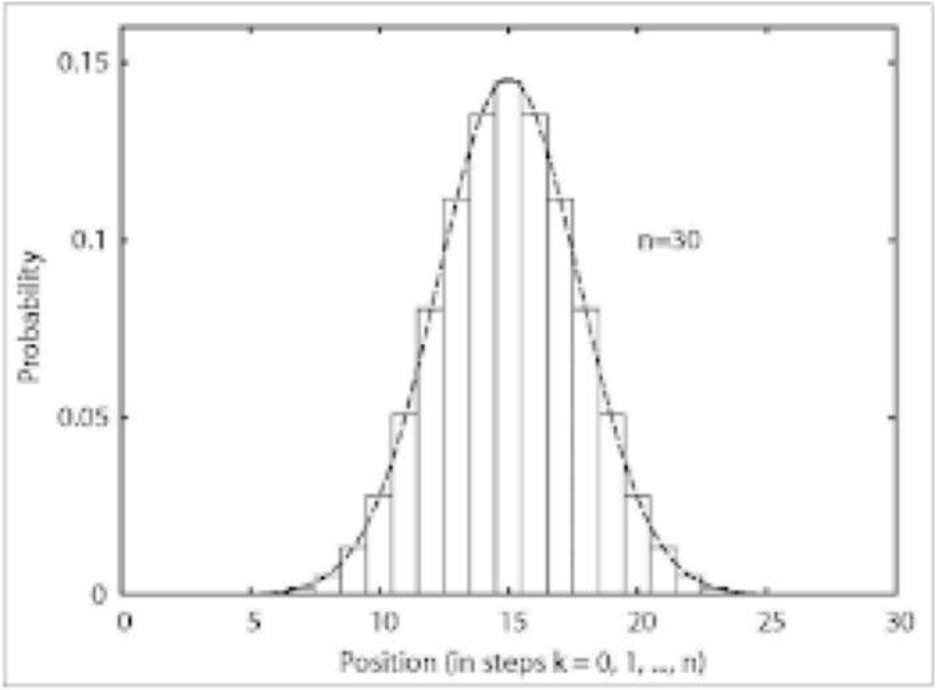


- *p* is the unknown expected error rate.
- Green bars are the results of different experiments on different samples S_i
- **The average** of these results **tends to p** as the number of **experiments** grows

In other words, if you perform "several" experiments, $E(error_S(h)) \rightarrow p$

Central Limit Theorem

- *Therefore, the theorem ensures that the* ***distribution of estimated error$_{Si}$(h) for different samples*** *Si of at least 30 instances follows a Gaussian (Normal) distribution whose mean is the true error mean error$_{D}$(h)=p*



What is a Gaussian Distribution?

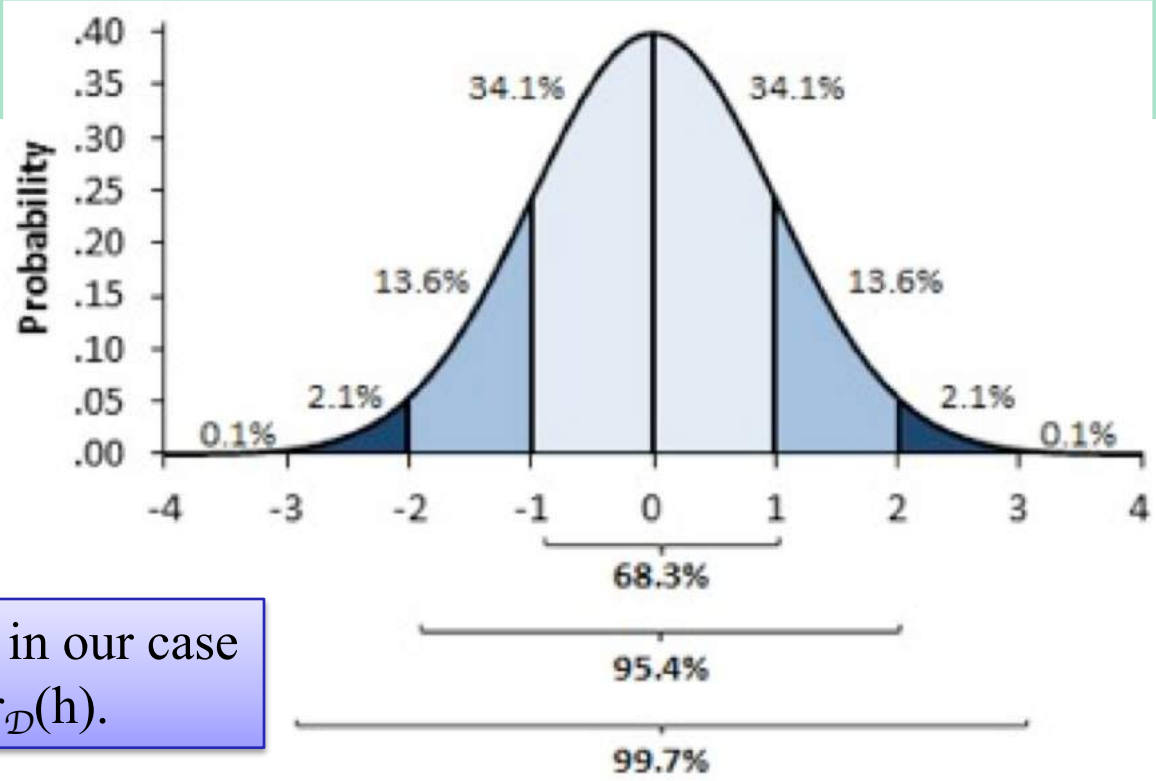
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The curve parameters are the mean **μ** (e.g., **p** in our specific case of error rate) and standard deviation **σ.** In a gaussian, for any μ and σ it holds that:

99.7% of the probability mass lies in the area below the mean value **μ ± 3** times the standard deviation **σ**

95.4% in the area below **μ ± 2 σ**

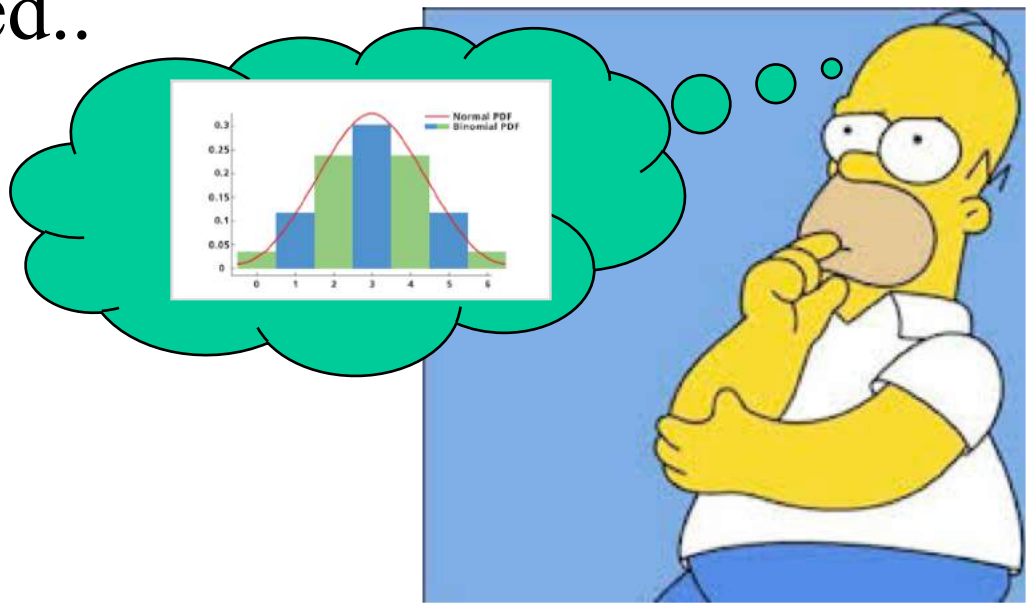
68.3% in the area below **μ ± σ**



The mean μ in our case is ***p***, or $\text{error}_{\mathcal{D}}(h)$.

..but..

- WHY is it important to know that the distribution of error rates is approximately normal (Gaussian)??
- Stay tuned..



Standard deviation of a sample

- Thanks to CL Theorem, we know that E($error_S(h)$)=E($error_D(h)$)=p (**mean of "true" error probability and mean of estimated error rate on different samples are the same**)
- *What about the standard deviation of $error_S(h)$??*
- Standard deviation **of a random variable** is the square root of its variance, as we said.
- The standard deviation **of a sample of n instances** is defined as:

$$\sigma_S = \frac{\sigma_D}{\sqrt{n-1}}$$

n-1 is called Bessel's correction
For large n **we can ignore the "-1"**

- Note that for n➔∞ (very large samples) **σ** tends to zero (since r/n➔p i.e., the observed error will converge to the real error rate)
- However, we don't know σ_D , since we don't know p!!But here comes the advantage of the central limit theorem..

Estimating the standard deviation of $error_S(h)$ on the sample S

$$\sigma_S = \frac{\sigma_D}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}} \simeq \sqrt{\frac{\frac{r}{n}\left(1-\frac{r}{n}\right)}{n}} = \sqrt{\frac{error_S(h)\left(1-error_S(h)\right)}{n}}$$

We replace the (unknown) p with our computed mean value r/n
This is an ESTIMATE since we assume that r/n is a good approximation of the real error rate p, which holds approximately true for large enough n, according to CLT!

$$|\, error_D(h) - error_S(h) \,| = |\, p - error_S(h) \,| = |\, p - \frac{r}{n} \,| = \Delta$$

the **bias** of an estimator is the difference **Δ** between this estimator's expected value and the true value of the parameter being estimated

Example

- Why we can set r/n(1-r/n)≅p(1-p)??
- Say p=0.6 and r/n=0.7 (difference is 0.1, 10%)
- However, p(1-p)=0.24 r/n(1-r/n)=0.21 (difference is 0.03, only 3%))
- ➔Although approximating the real error with the estimated error can lead to a significant over or under-estimate, approximating the real SD with the estimated SD is much less critical
- In general if n is sufficiently large, the probability that our estimate is truly very far from real sigma is sufficiently low

Summary

- $error_{\mathcal{D}}(h)=|h(x)-c(x)|$ is a random boolean variable and $P(error_{\mathcal{D}}(h)$ follows a Binomial probability distribution with mean $\boldsymbol{p}$, over the full population $\mathcal{D}$.
- The sample error, $error_S(h)$, is also a random variable (depending on the selected sample S) following a Binomial distribution. In general, $error_S(h) \neq error_D(h)$ and

  $|error_D(h) \neq error_S(h)| = |p - r/n| = \Delta$ is called **bias**.
- If the number of examples n in any sample S is sufficiently large (>30), then *according to Central Limit Theorem (CLT)* the underlying binomial distribution for $error_S(h)$ **approximates a Gaussian (Normal) distribution** and has the same mean $\boldsymbol{p}$ as for $error_{\mathcal{D}}(h)$ in $\mathcal{D}$
- THIS DOES NOT MEAN that for n>30 $error_D(h)=error_S(h)$!! It means that, if we would repeat the experiment on different independent samples S_i of the same dimension n, we would obtain different values $\mathbf{r_i/n}$ for $error_{Si}(h)$ and the DISTRIBUTION of these values approximates a Gaussian with mean **p.** Furthermore, for a single experiment with sufficiently large n we have:

$$\sigma_S = \sqrt{\frac{p(1-p)}{n-1}} \approxeq \sqrt{\frac{r/n(1-\frac{r}{n})}{n}}$$

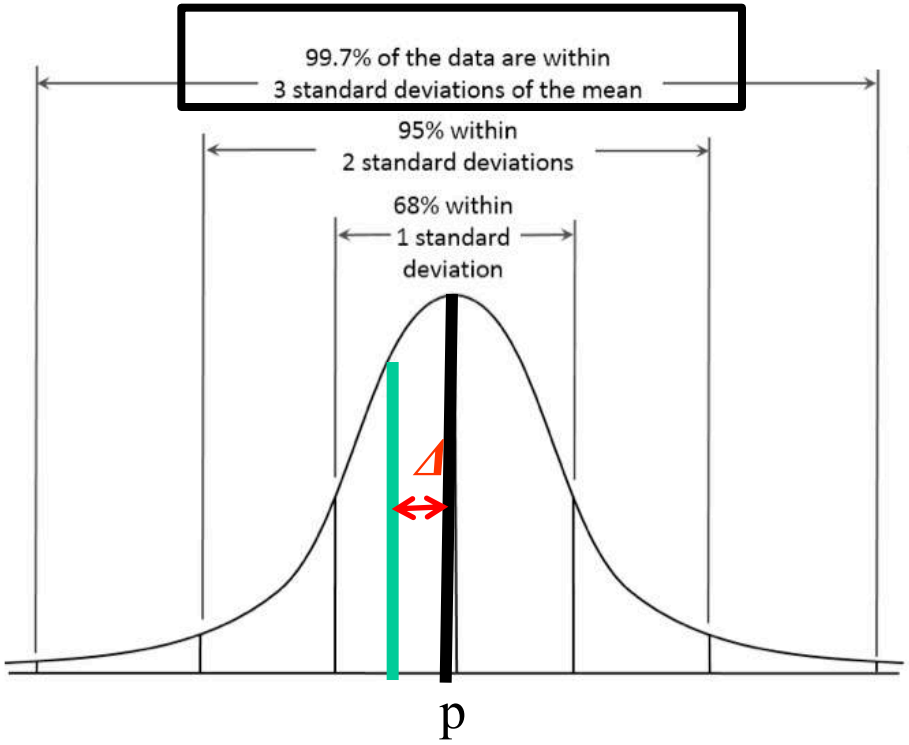
Confidence interval for an error estimate

- **Confidence interval**

$$LB \le \left| error_{\mathsf{D}}(h(x)) - error_{S}(h(x)) \right| = \left| p - r/n \right| \le UB$$

- LB and UB provide an estimate of the minimum and maximum expected **discrepancy** between the **measured and real** error rate
- In general terms, a CI provides **bounds for the bias Δ of an estimator**
- *Def*: an **N% confidence interval** for an estimate Δ is an interval [LB,UB] that includes Δ with probability N% (*with probability N% we have* $LB \le \Delta \le UB$)
- In our case, $\Delta = \left| error_{\mathsf{D}}(h(x)) - error_{S}(h(x)) \right|$ is a random variable (again!) representing the **difference** between true and estimated error If $error_S$ *and* $error_{\mathcal{D}}$ obey a **gaussian distribution**, then also Δ does, and the confidence interval can be easily estimated!!

Confidence intervals



$$LB \leq |error_S(h) - error_D(h) = |\frac{r}{n} - p| \leq UB$$

$$\frac{r}{n} - LB \leq p \leq \frac{r}{n} + UB$$

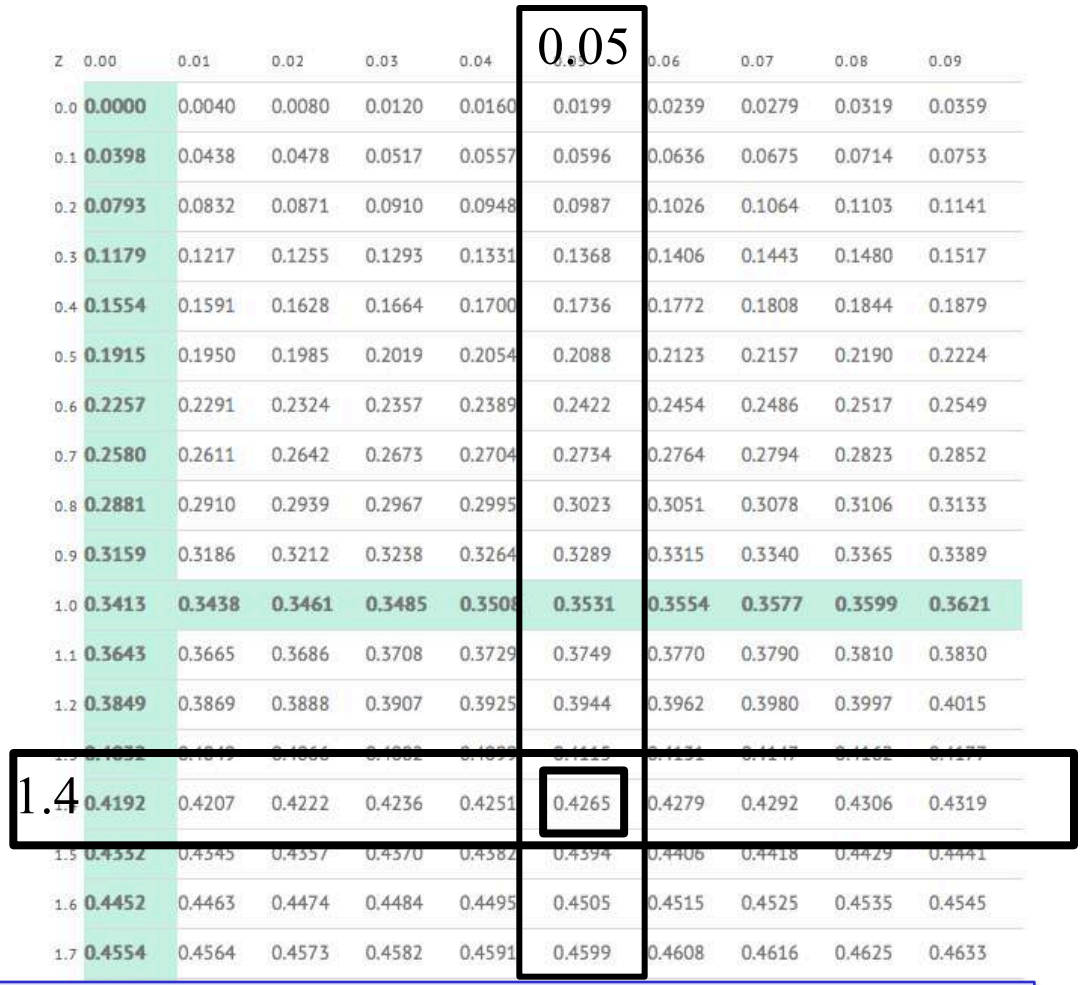
$$P(error_S(h) = x) \; follows \; a \; Gaussian \; and \; \sigma_S \cong \sqrt{\frac{r/n(1-\frac{r}{n})}{n}}$$

$$P(error_D(h) = y) \;\; also \; follows \; a \; Gaussian \; with \; SAME \; mean \; p$$

For any gaussian, we can say e.g "with a probability of 68% (95%, 99.7%) any value **x** we measure for $error_S$ will lie in the interval $\pm 1\sigma$ ($\pm 2\sigma$, $\pm 3\sigma$) around the mean **p**" .
More in general, **with an N% probability it will lie in the $\pm z\sigma$ interval**

The z table for Gaussian distributions

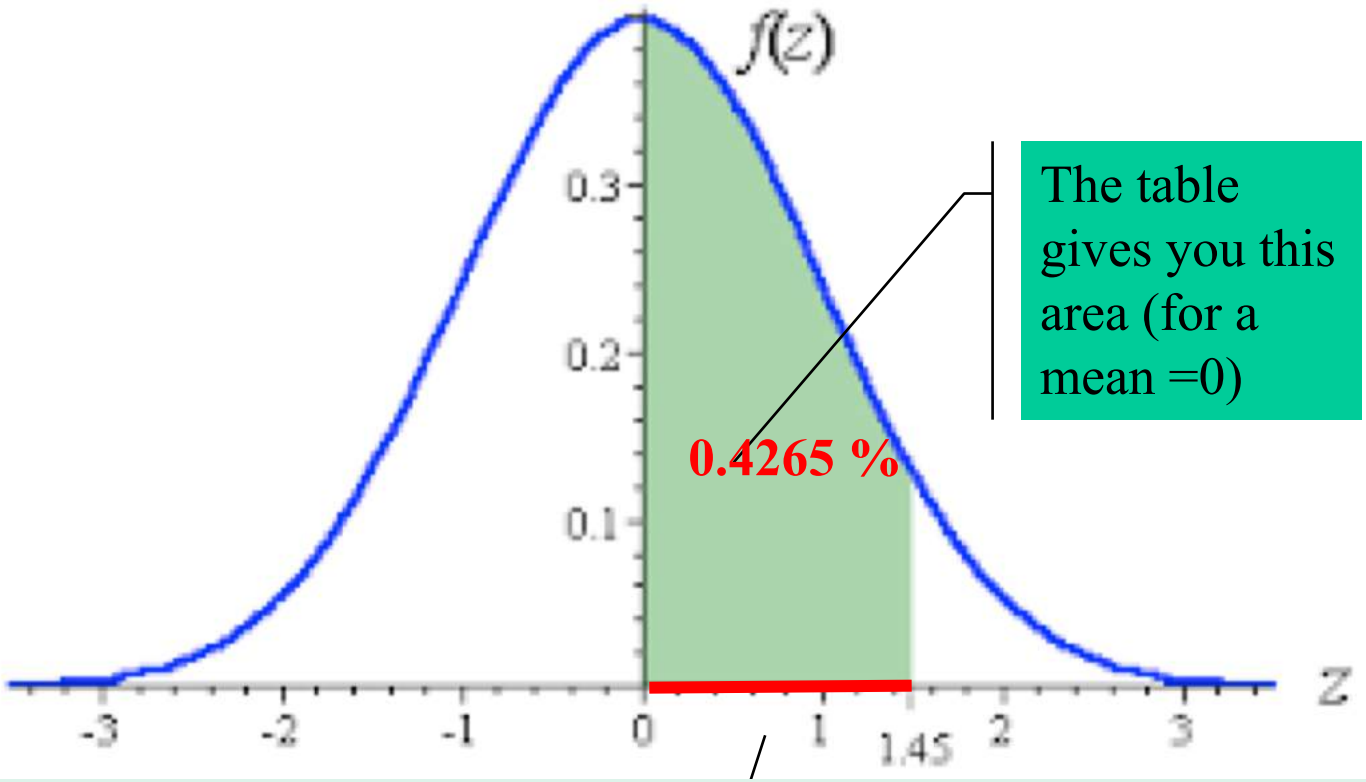
- z is **half of the lenght of the interval around the mean p that includes N% of the total probability mass.** A z-score tells you "*how many standard deviations*" (**z × σ**) from the mean p your result r/n is.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

Highlighted cell says that N=0.4265x2=0.913 % of the probability mass lies z=(1.4+0.05)=1.45 times the standard deviations around (±) the mean

Z- table

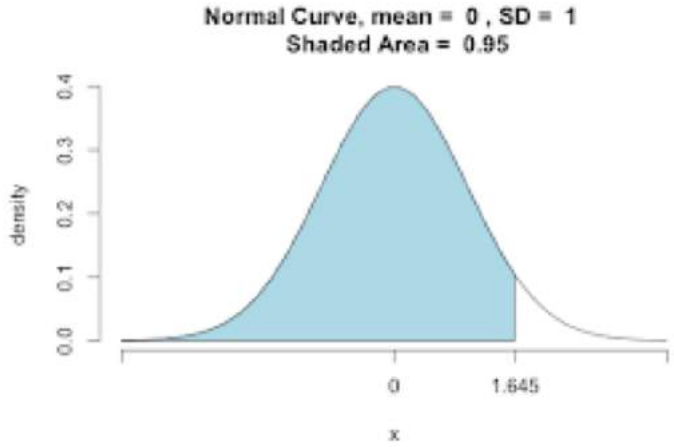


Or viceversa: you input the desired probability mass (say, N=0.913), you divide by 2 (0.4265) and obtain the z value from the table, to calculate the interval

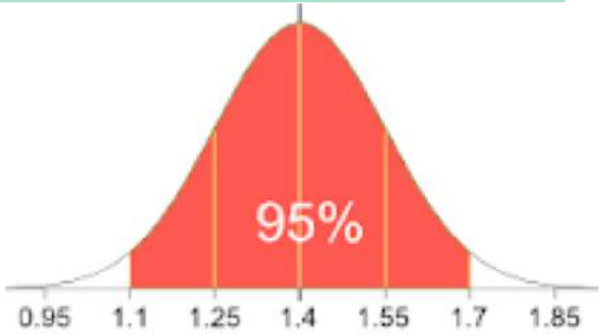
z score (e.g. 1.45)

There are different z-tables! Be aware

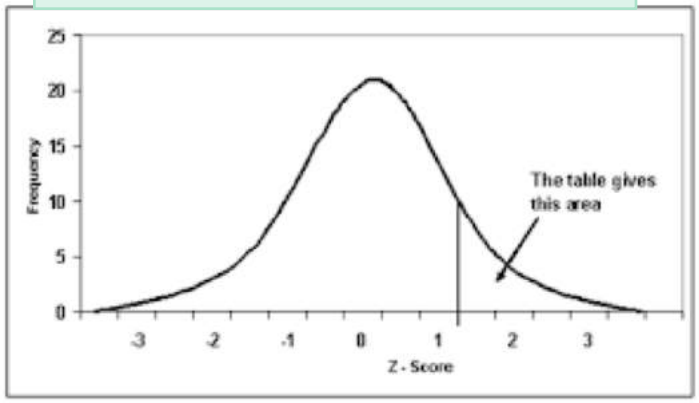
The area to the left of **zσ**



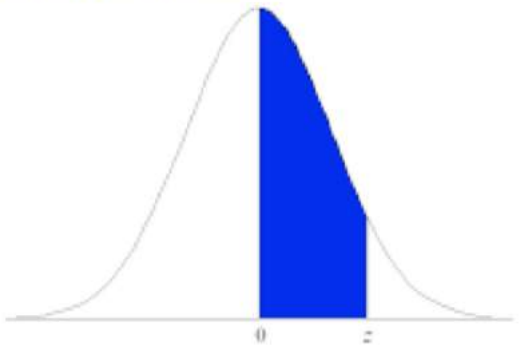
The area in between **±zσ**



The area to the right of **zσ**

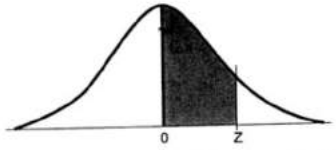


The area between **the mean** and **zσ**



The table contains the area under the standard normal curve between 0 and a specific value of z.

How do I know?



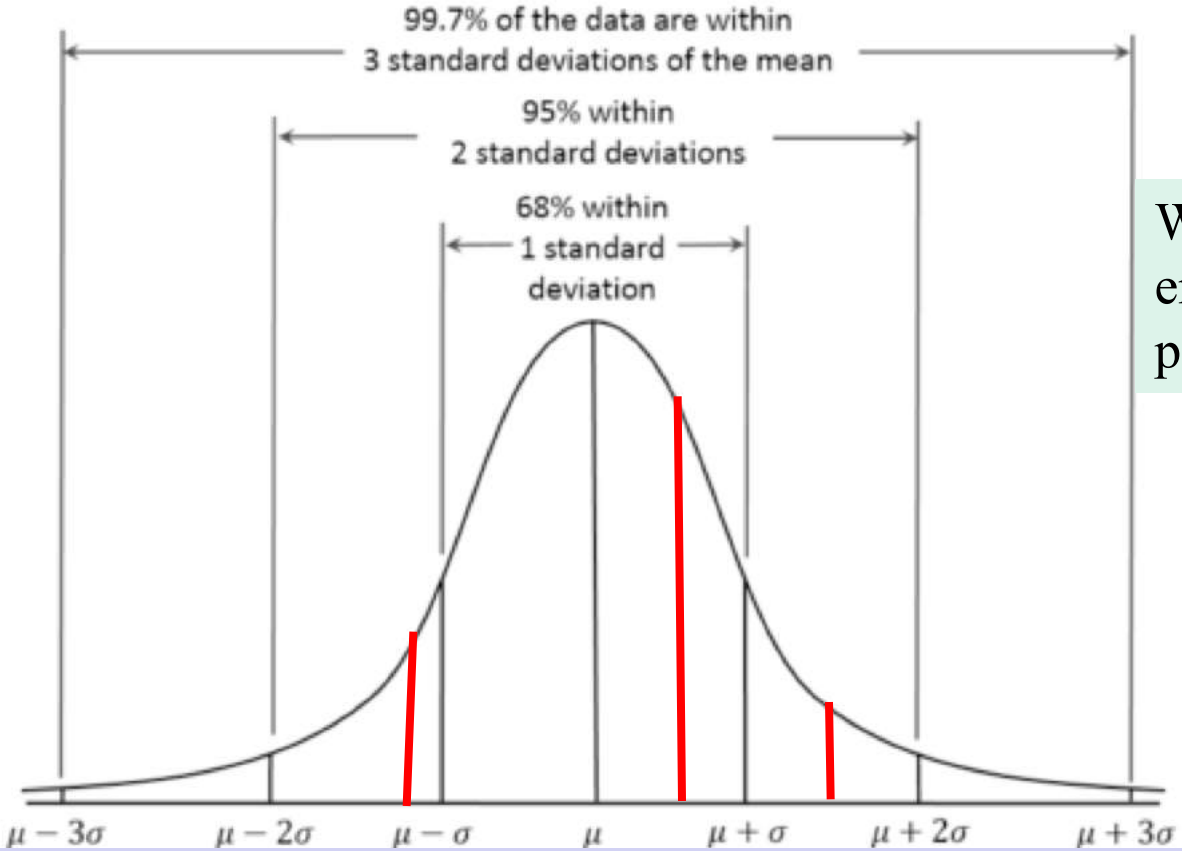
This table presents the area between the mean and the Z score . When Z=1.96, the shaded area is 0.4750.

Areas Under the Standard Normal Curve

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000									

Source: Adapted by permission from *Statistical Methods* by George W. Snedecor and William G. Cochran, sixth edition © 1967 by The Iowa State University Press, Ames, Iowa, p. 548.

Finding the N% confidence interval



Where is our error estimate placed?

We don't know **where** is our error estimate r/n (the red line) is placed, but we know that, e.g. with 68.26% probability (34.13+34,13) it will be at a distance $\pm 1\sigma$ from the mean error, with probability 95,44 (68,26+2x13.59) it will be at a distance $\pm 2\sigma$
and in general, **with probability N% at a distance $\pm Z_N\sigma$ (this is because it follows a gaussian)**

Finding the N% Confidence Interval

- To determine the confidence interval **we need to fix either z or N**
- Say N% is the known PARAMETER (we want to compute, e.g. the 80% confidence interval).
- We know that N% of the probability mass lies between $p \pm z\sigma$
- The Gaussian table gives us the z value for any N%,
- e.g., if N=80% then z=1,28: **we know that with 80% probability our estimated error is $\pm 1{,}28\sigma$ far from the true error**.

N%	**50**	**68**	**80**	**90**	**95**	**98**	**99**
$\mathbf{z_N}$	**0,67**	**1.00**	**1.28**	**1.64**	**1.96**	**2.33**	**2.58**

And: $$\sigma_S = \frac{\sigma_D}{\sqrt{n}} = \frac{1}{n}\sqrt{np(1-p)} \cong \sqrt{\frac{\frac{r}{n}(1-\frac{r}{n})}{n}} = \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

Case 1: finding the interval with a given confidence N

- We know the formula to compute the interval, given the estimated error rate:

$$[LB, UB] = \left[\frac{r}{n} - z\sqrt{\frac{r/n(1-\frac{r}{n})}{n}}, \frac{r}{n} + z\sqrt{\frac{r/n(1-\frac{r}{n})}{n}}\right]$$

- In this formula, **z** is unknown. But we know N, so we look in the table and we obtain z for the desired N, and compute the interval.

Example

- We have a classifier which produced an hypothesis model h(x), and a test set S of 100 instances
- We apply h(x) on the sample test set S and compute 13% (0.13) error rate (r/n)
- Since n>30 we assume that the error distribution follows a gaussian with mean 0,13 and standard deviation σ_S: $\sqrt{0.13(1-0.13)/100}$
- If we wish, e.g., to compute the N=90% confidence interval, on the table we find Z=1.64

N%	50	68	80	90	95	98	99
z_N	0,67	1.00	1.28	1.64	1.96	2.33	2.58

Calculating the N% Confidence Interval: Example (2)

- We then have:
- Z=1.64

$$\sigma_S \simeq \sqrt{0.13(1-0.13)/100}$$

- The 90% confidence interval is estimated using the previous formula:

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

Example 2 (finding 95% CI on a face recognition task)

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 $error_s(h)$?

N%	**50**	**68**	**80**	**90**	**95**	**98**	**99**
z_N	**0,67**	**1.00**	**1.28**	**1.64**	**1.96**	**2.33**	**2.58**

Solution

Precision is 0.78 hence **error rate r/n** is 0.22; the test set has 50 instances, hence **n**=50

Choose e.g. to compute the N% confidence interval with N=0.95

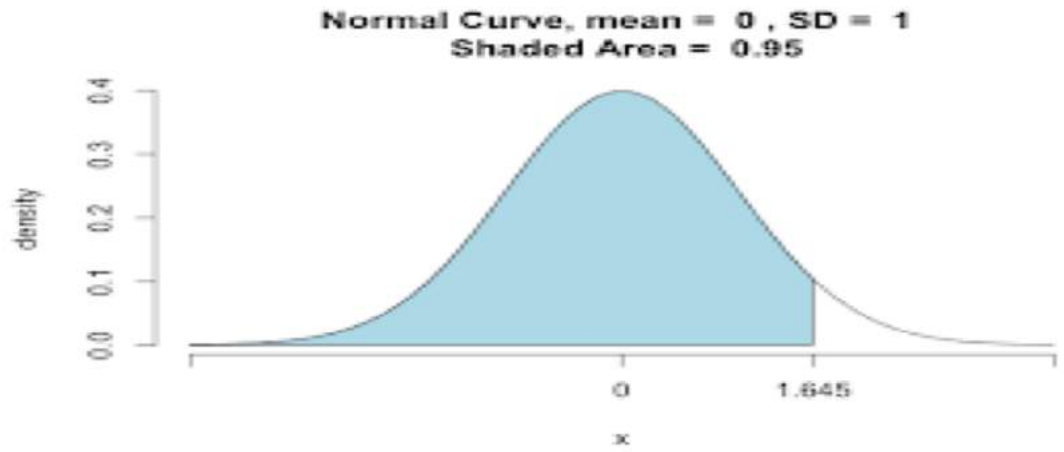
Given that $error_s(h)$=0.22 and n= 50, and z_N=1.96 for $N = 95$, we can now say that with 95% probability $error_D(h)$ will lie in the interval:

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

$$[0.11, 0.34]$$

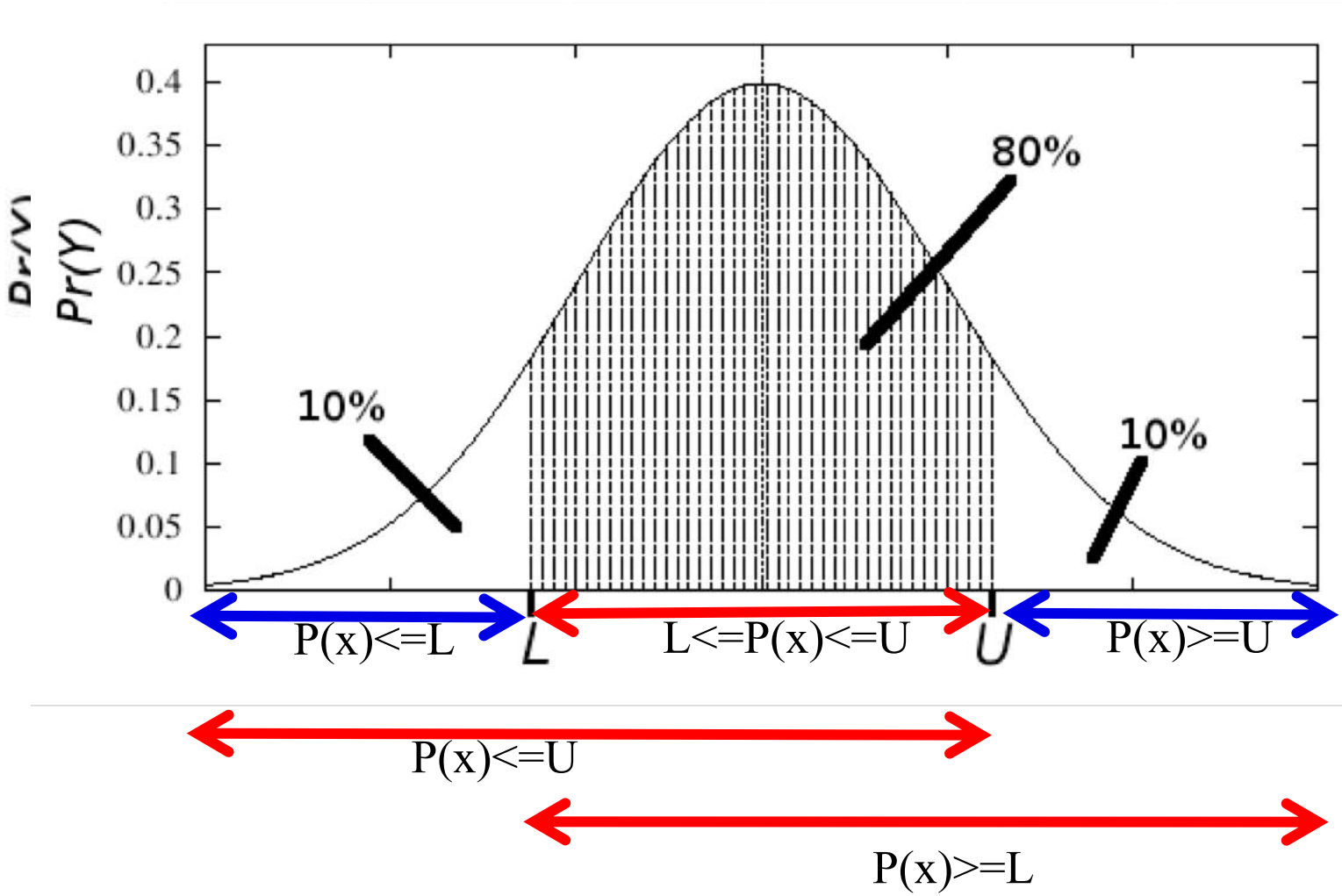
One side bound

- We might be interested in computing the probability that the error of our ML system is "at most" a given value, rather than within a given range



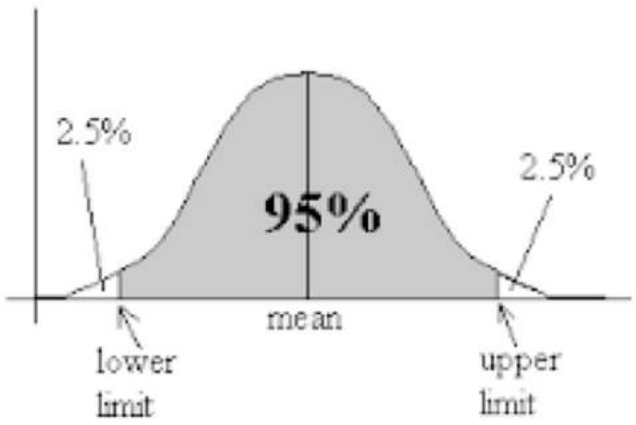
- Which amounts to computing the blue area
- Now N% is the area for which $error_S \leq \mathbf{z\sigma}$

One sided / two sided bounds. Gaussian is symmetric!



Example

- In the previous emotion recognition example, we said that with 95% probability (confidence) true error lies in the [0.11,0.34] interval
- There is a 5% area outside this interval, of which, 2.5% to the left and 2.5% to the right
- Therefore, we can also say that there is a 2.5% probability that the true error is higher than 0.34 (the upper bound)
- There is a (95+2.5=97.5)% probability that it is below 0.34
- There is a 2.5% probability that the true error is lower than 0.11 (the lower bound)
- There is a (95+2.5=97.5)% that it is higher than 0.11

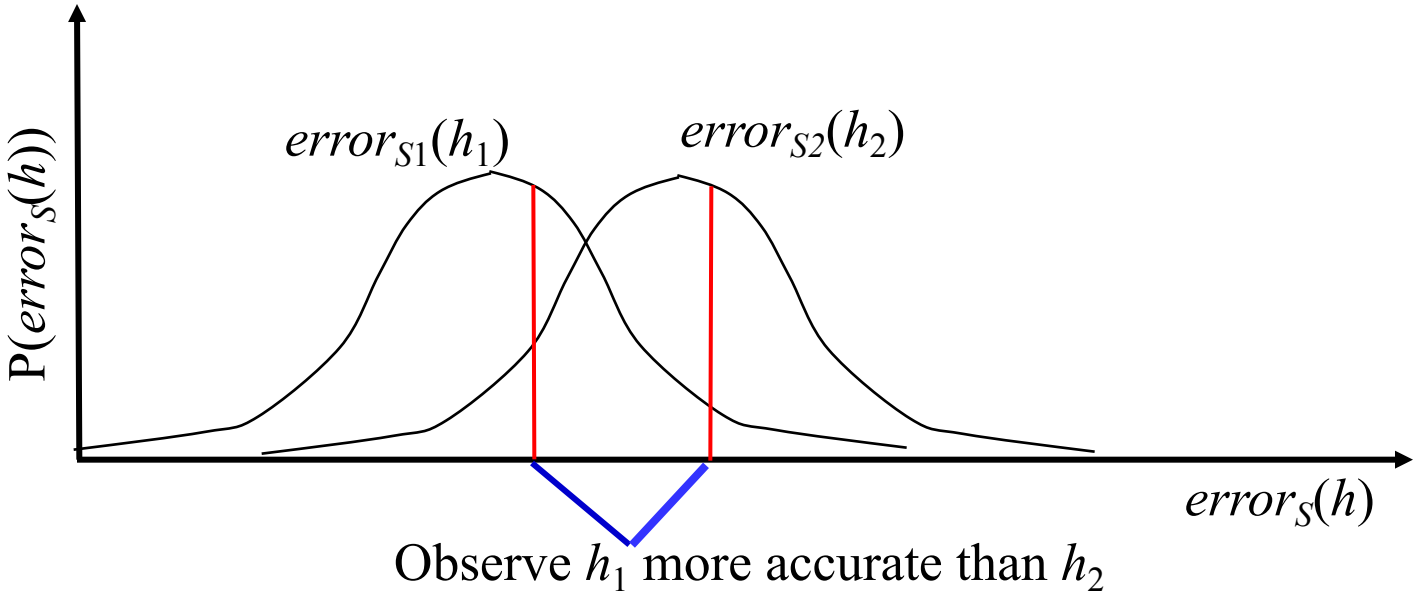


Issues

- Which performance measure we should use?
- How well can a classifier be expected to perform on "novel" data, not used for training?
- Since a performance measure is an ESTIMATE on a sample, how accurate is our estimate?
- **How to compare performances of different hypotheses or those of different classifiers?**

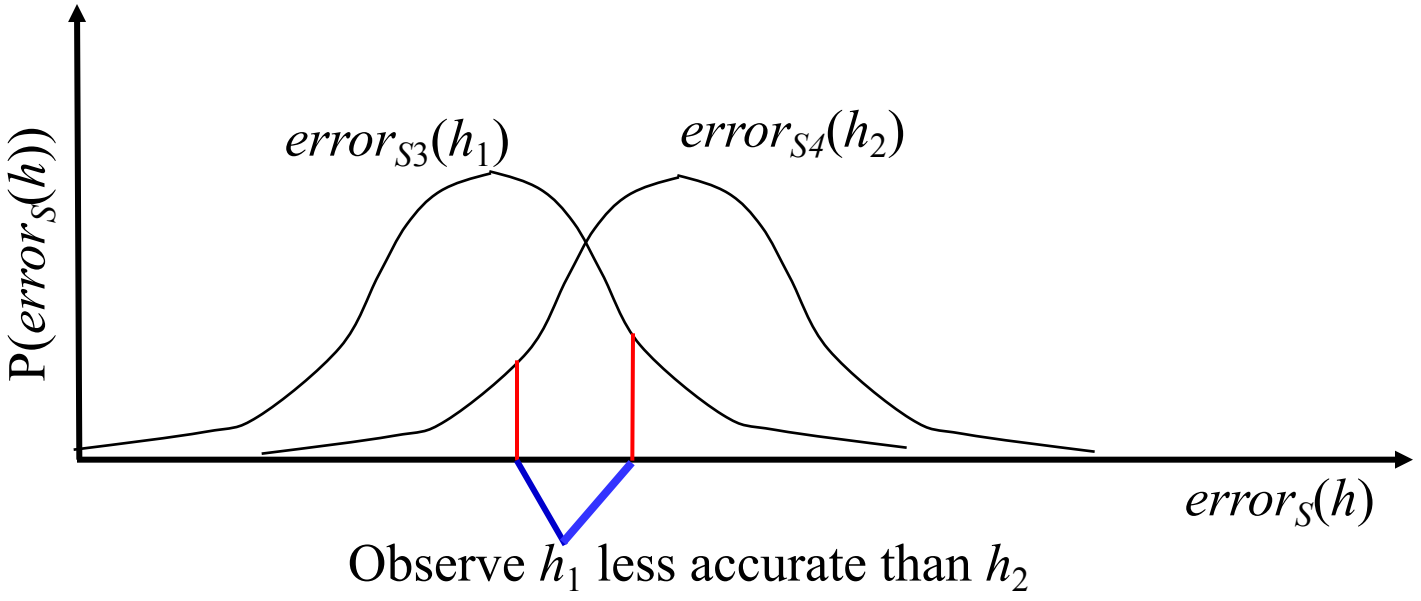
a) Comparing Two Learned Hypotheses

- When evaluating two hypotheses (e.g. using different hyper-parameters on the same ML algorithm), their observed ordering with respect to accuracy **may or may not** reflect the ordering of their **true** accuracies.
  - Assume h_1 is tested on test set S_1 of size n_1
  - Assume h_2 is tested on test set S_2 of size n_2



Comparing Two Learned Hypotheses

- When evaluating two hypotheses, their observed ordering with respect to accuracy may or may not reflect the ordering of their true accuracies.
  - Assume h_1 is tested on test set S_3 of size n_1
  - Assume h_2 is tested on test set S_4 of size n_2



Alternative Hypotheses testing

- When we wish to understand **how much we can rely on a statistical finding** (for example, that a ML model *h2* is more precise than *h1* on a sample dataset), we need **to list alternatives** (e.g. that h2 in NOT more precise than h1 on the entire population).
- One of these alternatives is called the NULL HYPOTHESIS H0
- Usually, the null hypothesis is one that **disconfirms** our findings

Alternative Hypotheis testing (2)

- Suppose we measure the error rate of h1 and the error rate of h2, and find a non zero difference **d**=$error_{S1}$(h1)-$error_{S2}$(h2)
- **Two-tail test** (we obtain a value |d|≠0):
  - **H0: although we measure a value** |d|≠0, this value **does not support** that there is a difference between h1 and h2, hence $error_{\mathcal{D}}$(h1)-$error_{\mathcal{D}}$ (h2) could actually be 0
  - **H1:** there is indeed a (**statistically significant**) difference between h1 and h2 (either positive or negative): with high confidence our finding is true.
- **One-tail right-test** (we find that d>0)
  - H0: data do not support that h2>h1
  - H1: h2>h1 (error of h1 is significantly lower)
- **One-tail left-test** (we find that d<0)
  - H0: data do not support that h2<h1
  - H1: h1>h2 (error of h1 is significantly higher)

Z-Score Test for Comparing alternative classifiers (= Hypotheses)

- Assumes h_1 is tested on test set S_1 of size n_1 and h_2 is tested on test set S_2 of size n_2. Assume both n>30 for the Central Limit Theorem to hold.
- Compute the difference between the accuracy of h_1 and h_2: $\hat{d} = \left| error_{S_1}(h_1) - error_{S_2}(h_2) \right|$
- The difference is a random varia
  the difference between two varia
  gaussian distribution, it also follows a gaussian, with standard deviation:

Note: the SD of the sum or difference of random variables, is the sum of SDs.

$$\sigma_d = \sqrt{\frac{\sigma_{h1}^D}{n_1} + \frac{\sigma_{h2}^D}{n_2}} \cong \sqrt{\frac{\sigma_{h1}^{S1}}{n_1} + \frac{\sigma_{h2}^{S2}}{n_2}} = \sqrt{\frac{error_{S_1}(h_1)\cdot(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)\cdot(1-error_{S_2}(h_2))}{n_2}}$$

Testing for the null hypothesis

- If H0 (null hypothesis) holds true, then we must have:

  $error_{\mathcal{D}}(h1) = error_{\mathcal{D}}(h2)$ and therefore $d_{\mathcal{D}} = 0$

  i.e., although if in our experiments we observe that e.g., $error_S(h1) < error_S(h2)$, the "true" mean of error differences of h1 and h2 on $\mathcal{D}$ is zero.

- To test the likelyhood of H0, we t

$$\hat{d} = |error_{s1}(h1) - error_{s2}(h2)|$$

Error estimates on the samples

$$d_D = |error_D(h1) - error_D(h2)| = 0$$

d_D must be zero if H0 holds true

$$|\hat{d} - d_D| = |d| \leq z \times \sigma_d$$

Error bounds in estimating d^

$$z = \frac{\hat{d}}{\sigma_d}$$

We know both **d** and **σ** so we compute **z** and look on a z-table, to see"how many times" our result d^ is far from the expected mean difference (which is zero according to H0)

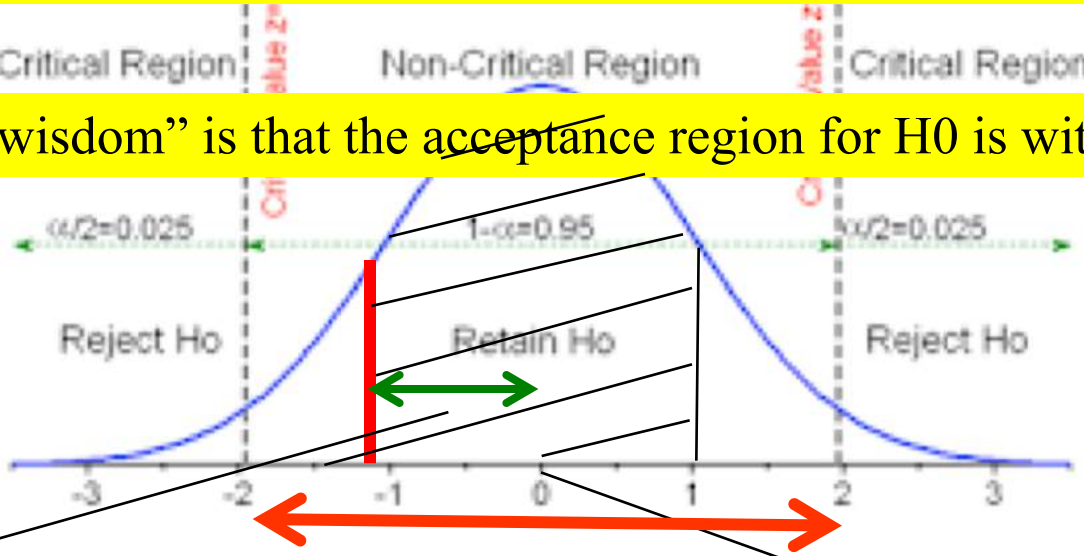
Two-tail test

$|d_S-0|=z\times\sigma$ ➔ $z=|d_S|/\sigma$

Given z, using the table we can compute N (the confidence area).

If the area lies **within** the non-critical region (N≤95%), the null Hypothesis is ACCEPTED (= there is no significant difference between the two hypotheses)

The "common wisdom" is that the acceptance region for H0 is within -2σ and +2σ



We estimate d_S on the sample

If H0 holds, $d_D=0$

Two tail test

- In other terms: the far-est our measured distance d_S is from the "expected" distance (d_D=0) in case the null hypothesis H0 holds true, the less confident we are in H0.
- For any measured value of d_S , the y axis give us **the probability of observing that value**
- If d_S is more far than $\pm 2\sigma$ from d_D, then we may conclude that the probability of having observed the value d_S in case d_D=0 is too small. And hence we reject H0 as being very unlikely .

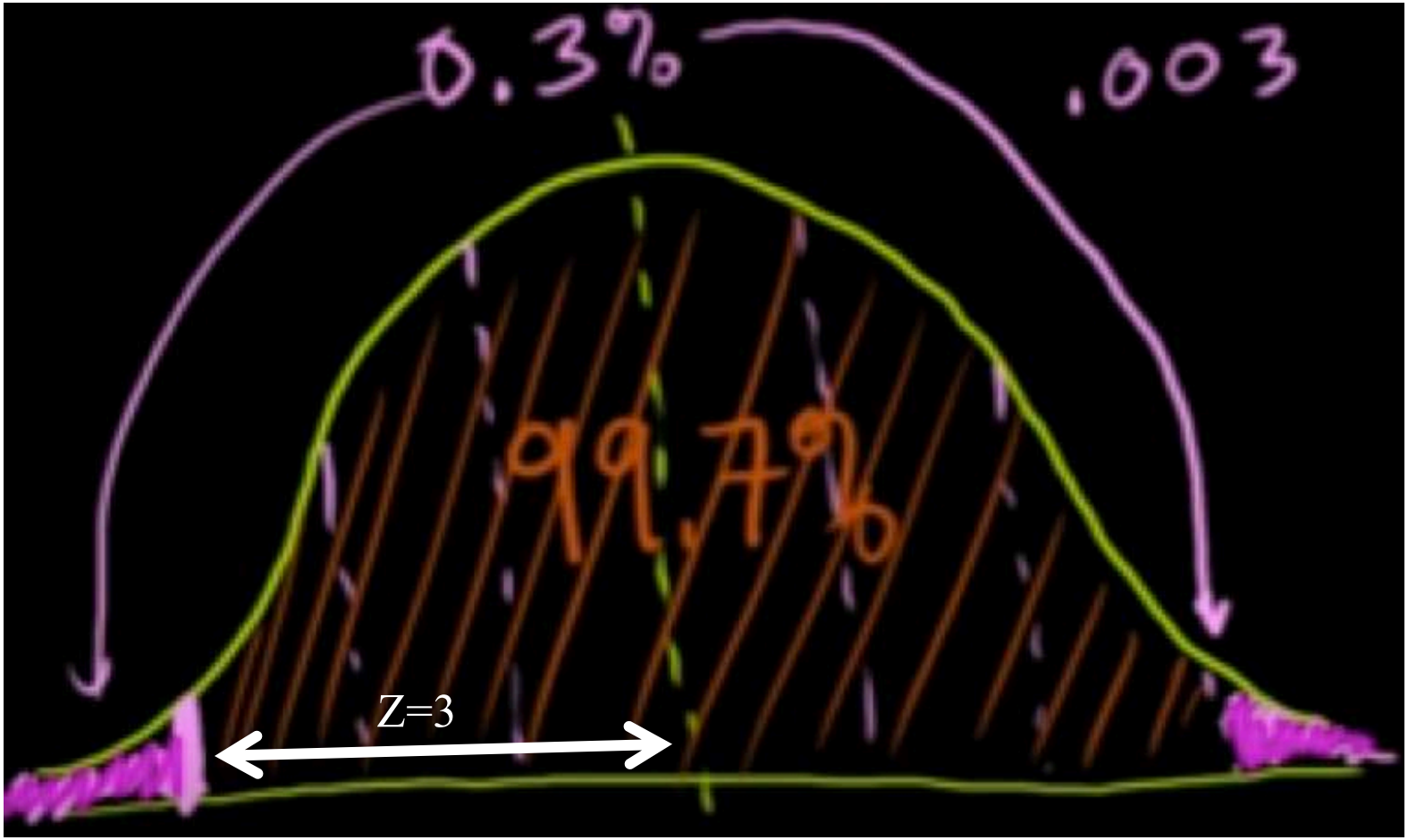
Example: Testing for the null hypothesis

- Assume that $d_S=0.15$ and $\sigma_S=0.05$ then $z=d_S/\sigma=3$
- Then,

N=99,87%

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	[illegible]	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
[illegible]	[illegible]	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
[illegible]	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

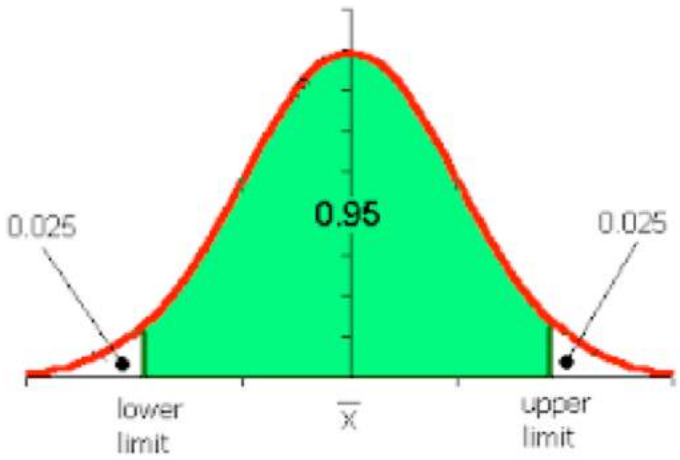
We should reject H0!!



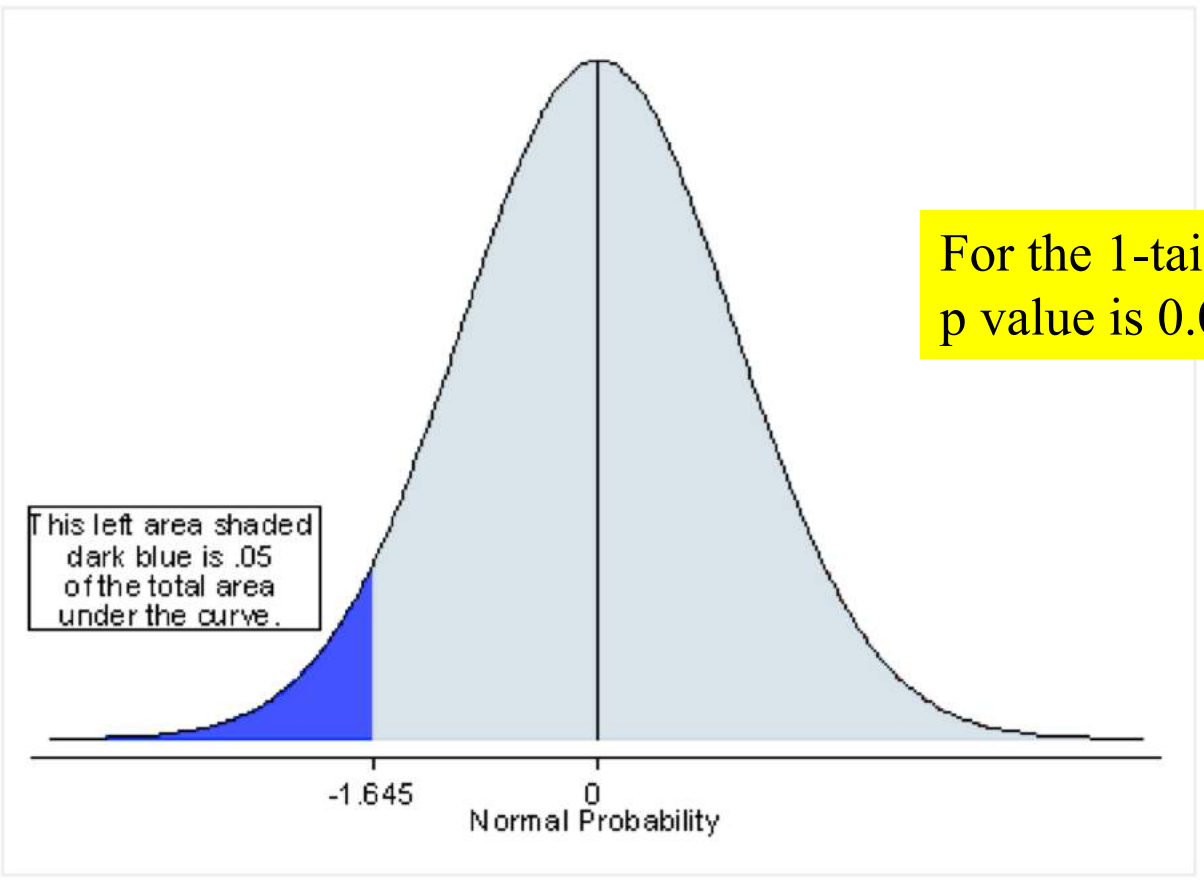
Our z-test says that, if the mean difference is zero, the **probability** to obtain the value |d|=0.15 or more, is **less than 0.03** (100-99,87)!! **So H0 is VERY UNLIKELY**

p-value

- The p-value is the "probability value" of observing our estimate, **given that H0 holds true**
- Common wisdom is to reject the null hypotesis if $p<0.05$ (5%) (same as saying that the estimated value lies outside the $\pm 2\sigma$ interval, or outside the 95% probability mass around the mean)
- In previous example we obtained $p<0.03$



One-tail test



For the 1-tail test the p value is 0.05

For example, if we test h1>h2 we state the null hypothesis as follows:
H0: data do not support that h1>h2 (hence h1≤h2)
H1: h1>h2 (in this case we should get an estimate of d)

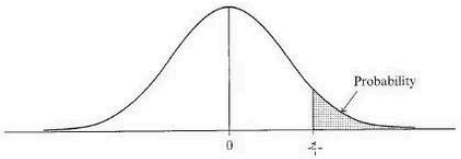
Example (one-tail left test: is truly h2>h1?)

- $error_{s1}(h1)$= x1=17.5%, $error_{s2}(h2)$= x2=12.4%, d=5.1% (0.51)
- n_1=50, n_2=50

$$\sigma_d \cong \sqrt{\frac{error_{S_1}(h_1)\cdot(1-error_{S_1}(h_1))}{n_1}+\frac{error_{S_2}(h_2)\cdot(1-error_{S_2}(h_2))}{n_2}}$$

$$=\sqrt{\frac{0.175\cdot(1-0.175)}{50}+\frac{0.124\cdot(1-0.124)}{50}}=\sqrt{0.005}$$

$$z=\frac{(x1-x2)+(error_D(h_1)-error_D(h_2))}{0.07}=(0.051-0)/0.07=0.73$$

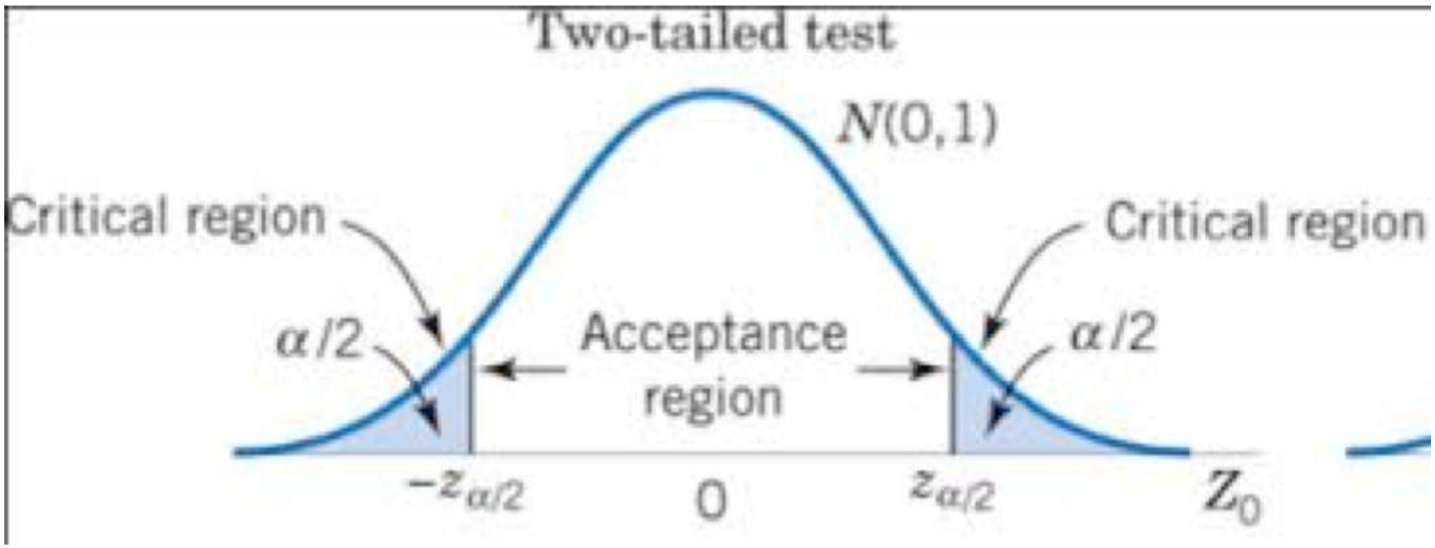


z	Second Decimal Place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	[illegible]	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170

N=0.2327 ➔ p>0.05

The null hypothesis is **accepted**: difference is not large enough to support h1<h2 (p **is not lower** than 0.05)

Summary: two-side test



b) Comparing 2 Learning Algorithms

- Comparing the average accuracy of hypotheses produced **by two different ML algorithms** is more difficult. Ideally, we want to measure:

$$E_{S \subset D}(error_D(L_A(S)) - error_D(L_B(S)))$$

where $L_X(S)$ represents the hypothesis learned by learning algorithm L_X from training data S.

- To accurately estimate this, we need to average over **multiple, independent training and test sets**.
- However, since labeled data is limited, generally must average over multiple splits of the overall data set into training and test sets (**K-fold cross validation**).

K-Fold Cross Validation: summary

- Every example in D used as a test example once and as a training example k–1 times.
- All test sets are independent; however, training sets overlap significantly (see two previous slides).
- In total we test on $[(k–1)/k]\cdot|D|$ training examples.
- Standard method is 10-fold.
- If k is low, not sufficient number of train/test trials; if k is high, test set may be too small and test variance is high and run time is increased.
- If k=|D|, method is called ***leave-one-out*** cross validation (at each step, you leave out one example). Used for specific cases (e.g. learning recommendations)

How to use K-Fold Cross Validation to evaluate different learning algorithms

Randomly partition dataset D into k disjoint equal-sized (N) subsets $P_1 \ldots P_k$

For i from 1 to k do:

Use P_i for the test set and remaining data for training

$S_i = (D - P_i)$

$h_A = L_A(S_i)$

$h_B = L_B(S_i)$

$\delta_i = \text{error}_{Pi}(h_A) - \text{error}_{Pi}(h_B)$

Return the average difference in error:

$$\bar{\delta} = \frac{1}{k}\sum_{i=1}^{k}\delta_i$$

Error bound is computed as:

$$\bar{\delta} \pm Z \cdot \sigma_{\bar{\delta}}$$

$$\sigma_{\bar{\delta}} = \sqrt{\frac{1}{k(k-1)}\sum_{i=1}^{k}\left(\delta_i - \bar{\delta}\right)^2}$$

Is L_A better than L_B?

- K-fold cross validation improves confidence in our estimate of δ since we are performing many experiments and computing δ as the AVERAGE of δ_i.
- As K grows this average tends to the true mean difference (however we cannot make K too big since individual samples should be large enough for the CLT to apply)
- We can in any case apply hypothesis testing as before

Sample Experimental Results

Which experiment provides better evidence that SystemA is better than SystemB?

Experiment 1

	SystemA	SystemB	δ
Trial 1	87%	82%	+5%
Trail 2	90%	85%	+5%
Trial 3	88%	83%	+5%
Trial 4	82%	77%	+5%
Trial 5	85%	80%	+5%
Average	85%	80%	**+5%**

Experiment 2

	SystemA	SystemB	δ
Trial 1	90%	82%	+8%
Trail 2	85%	90%	–5%
Trial 3	80%	85%	–5%
Trial 4	85%	75%	+10%
Trial 5	77%	82%	– 5%
Average	85%	80%	**+5%**

Experiment 1 mean δ has σ=0, therefore we have a perfect confidence in the estimate of δ

Experimental Evaluation Conclusions

- Good experimental methodology is important to evaluating learning methods.
- Important to test on a variety of domains to demonstrate a general bias that is useful for a variety of problems. Testing on 20+ data sets is common.
- Variety of freely available data sources
  - UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - KDD Cup (large data sets for data mining) http://www.kdnuggets.com/datasets/kddcup.html
  - CoNLL Shared Task (natural language problems) http://www.ifarm.nl/signll/conll/
- Data for real problems is preferable to artificial problems to demonstrate a useful bias for real-world problems.
- Many available datasets have been subjected to significant feature engineering to make them learnable.