

## What you'll learn today

- The difference between **sample error** and **true error**
- Confidence intervals for sample error
- How to estimate confidence intervals
- Binomial distribution, Normal distribution, Central Limit Theorem
- Paired  $t$  tests and cross-validation
- Comparing learning methods

Slides largely pilfered from Tom

## A practical problem

Suppose you've trained a classifier  $h$  for your favorite problem (YFP), tested it on a sample  $S$ , and the error rate on the sample was 0.30.

- How good is that estimate?
- Should you throw away your old classifier for YFP, which has an error rate of 0.35 on sample  $S$ , and replace it with  $h$ ?
- Can you write a paper saying that you've reduced the best-known error rate for YFP from 0.35 to 0.30? Will it get accepted?

## Two Definitions of Error

The **true error** of hypothesis  $h$  with respect to target function  $f$  and distribution  $\mathcal{D}$  is the probability that  $h$  will misclassify an instance drawn at random according to  $\mathcal{D}$ .

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}} [f(x) \neq h(x)]$$

The **sample error** of  $h$  with respect to target function  $f$  and data sample  $S$  is the proportion of examples  $h$  misclassifies

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

Where  $\delta(f(x) \neq h(x))$  is 1 if  $f(x) \neq h(x)$ , and 0 otherwise.

Usually, you don't know  $error_{\mathcal{D}}(h)$ . The big question is: how well does  $error_S(h)$  estimate  $error_{\mathcal{D}}(h)$ ?

## Problems Estimating Error

1. *Bias*: If  $S$  is the training set,  $error_S(h)$  is (almost always) optimistically biased

$$bias \equiv E[error_S(h)] - error_{\mathcal{D}}(h)$$

This is also true if *any part* of the training procedure used *any part* of  $S$ , e.g. for feature engineering, feature selection, parameter tuning, ...

For an unbiased estimate,  $h$  and  $S$  must be chosen independently

2. *Variance*: Even with unbiased  $S$ ,  $error_S(h)$  may still *vary* from  $error_{\mathcal{D}}(h)$

Variance of  $X$  is  $Var(X) \equiv E[(X - E[X])^2]$

## Example

Hypothesis  $h$  misclassifies 12 of the 40 examples in  $S$

$$error_S(h) = \frac{12}{40} = .30$$

What is  $error_{\mathcal{D}}(h)$ ?

## Example

Hypothesis  $h$  misclassifies 12 of the 40 examples in  $S$

$$error_S(h) = \frac{12}{40} = .30$$

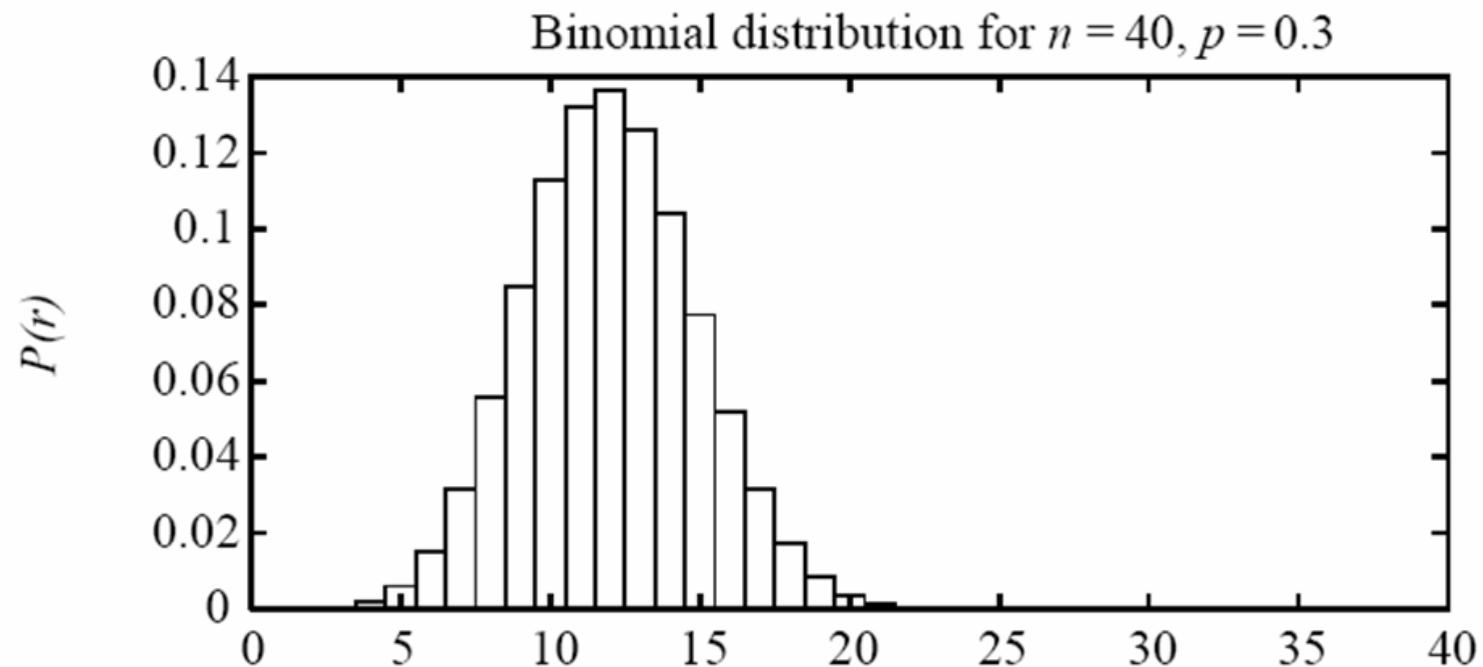
What is  $error_{\mathcal{D}}(h)$ ?

Some things we know:

- If  $\theta = error_{\mathcal{D}}(h)$ , the sample error is a binomial with parameters  $\theta$  and  $|S|$  (i.e., it's like flipping a coin with bias  $\theta$  exactly  $|S|$  times.)
- Given  $r$  errors in  $n$  observations  $\hat{\theta} = \frac{r}{n}$  is the MLE for  $\theta = error_{\mathcal{D}}(h)$

## The Binomial Distribution

Probability  $P(R = r)$  of observing  $r$  misclassified examples



$$P(r) = \frac{n!}{r!(n-r)!} \text{error}_{\mathcal{D}}(h)^r (1 - \text{error}_{\mathcal{D}}(h))^{n-r}$$

Question: what's the random event here? what's the experiment?

## Aside: Credibility Intervals

From

$$P(R = r | \Theta = \theta) = \frac{n!}{r!(n-r)!} \theta^r (1-\theta)^{n-r}$$

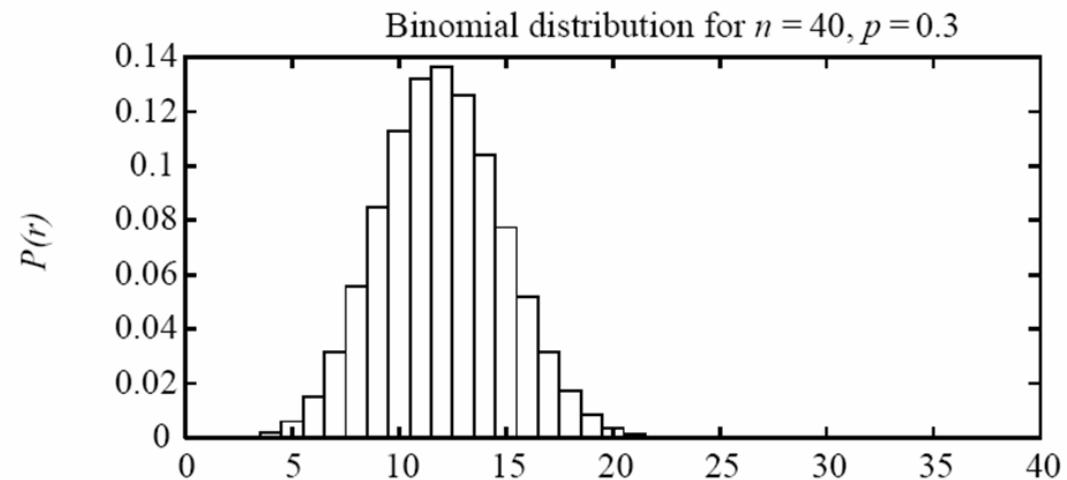
we could try and compute

$$P(\Theta = \theta | R = r) = \frac{1}{Z} P(R = r | \Theta = \theta) P(\Theta = \theta)$$

to get a MAP for  $\theta$ , or an interval  $[\theta_L, \theta_U]$  that probably contains  $\theta$  (probability taken over choices of  $\Theta$ )

## The Binomial Distribution

Probability  $P(R = r)$  of observing  $r$  misclassified examples

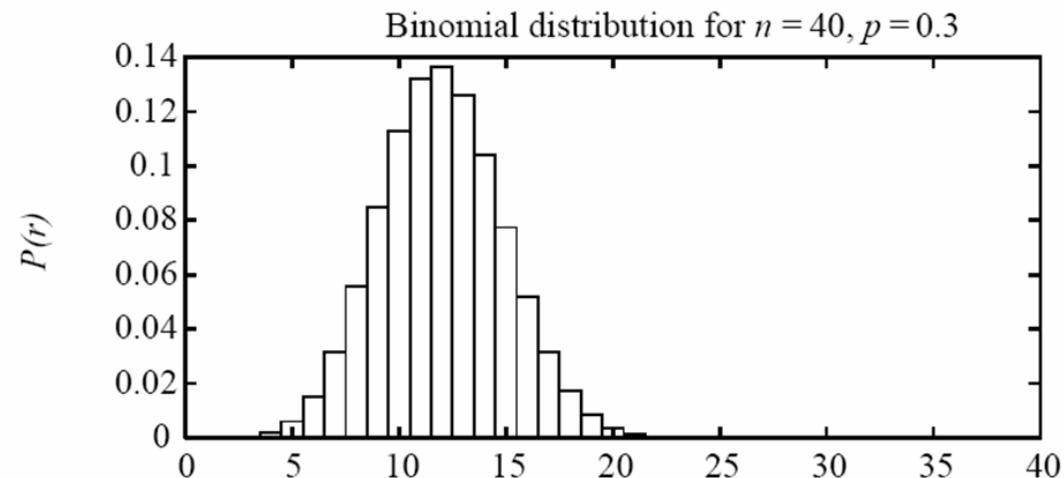


Usual interpretation:

- $h$  and  $error_{\mathcal{D}}(h)$  are fixed quantities (not random)
- $S$  is a random variable—i.e. the experiment is drawing the sample
- $R = error_S(h) \cdot |S|$  is a random variable depending on  $S$

## The Binomial Distribution

Probability  $P(R = r)$  of observing  $r$  misclassified examples



Suppose  $|S| = 40$  and  $error_S(h) = \frac{12}{40} = .30$ . How much would you bet that  $error_{\mathcal{D}}(h) < 0.35$  ?

Hint: the graph shows that  $P(R = 14) > 0.1$  and  $\frac{14}{40} = 0.35$ . So it would not be that surprising to see a sample error  $error_S(h) = .35$  given a true error of  $error_{\mathcal{D}}(h) < 0.30$ .

## Confidence Intervals for Estimators

Experiment:

1. choose sample  $S$  of size  $n$  according to distribution  $\mathcal{D}$
2. measure  $error_S(h)$

$error_S(h)$  is a random variable (i.e., result of an experiment)

$error_S(h)$  is an *unbiased estimator* for  $error_{\mathcal{D}}(h)$

Given observed  $error_S(h)$  what can we conclude about  $error_{\mathcal{D}}(h)$ ?

It's probably *not* true that  $error_{\mathcal{D}}(h) = error_S(h)$  but it probably is true that  $error_{\mathcal{D}}(h)$  is “close to”  $error_S(h)$ .

## Confidence Intervals: Recipe 1

If

- $S$  contains  $n$  examples, drawn independently of  $h$  and each other
- $n \geq 30$

Then

- With approximately 95% probability,  $error_{\mathcal{D}}(h)$  lies in interval

$$error_S(h) \pm 1.96 \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

Another rule-of-thumb: if the interval above is within  $[0, 1]$  then it's reasonable to use this approximation.

## Confidence Intervals: Recipe 2

If

- $S$  contains  $n$  examples, drawn independently of  $h$  and each other
- $n \geq 30$

Then

- With approximately  $N\%$  probability,  $error_{\mathcal{D}}(h)$  lies in interval

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

where

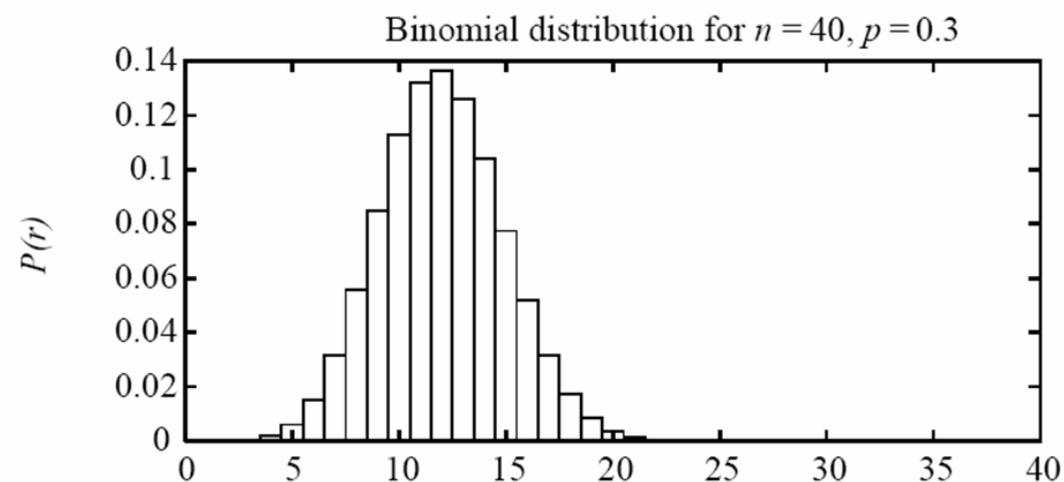
$N\%$ :	50%	68%	80%	90%	95%	98%	99%
$z_N$ :	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Why does this work?

## Facts about the Binomial Distribution

Probability  $P(r)$  of  $r$  heads in  $n$  coin flips, if  $p = \Pr(\text{heads})$

- Expected, or mean value of  $X$ ,  $E[X]$ , is  $E[X] \equiv \sum_{i=0}^n iP(i) = np$
- Variance of  $X$  is  $Var(X) \equiv E[(X - E[X])^2] = np(1 - p)$
- Standard deviation of  $X$ ,  $\sigma_X$ , is  $\sigma_X \equiv \sqrt{E[(X - E[X])^2]} = \sqrt{np(1 - p)}$



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

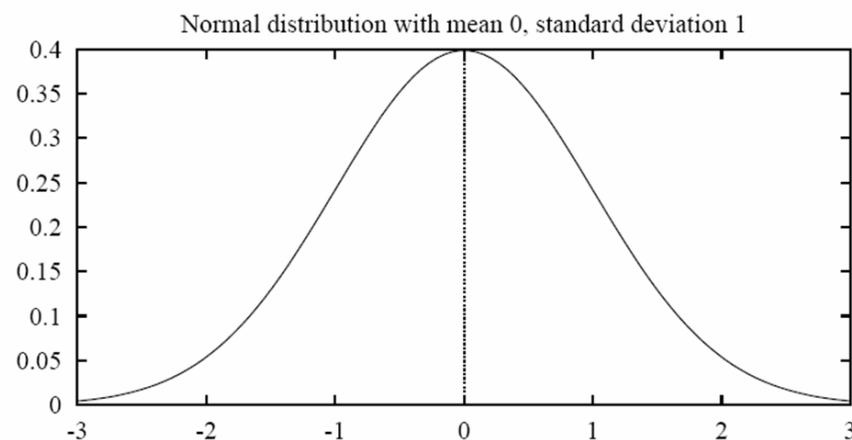
## Another Fact: the Normal Approximates the Binomial

$error_S(h)$  follows a *Binomial* distribution, with

- mean  $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
- standard deviation  $\sigma_{error_S(h)} = \sqrt{\frac{error_{\mathcal{D}}(h)(1-error_{\mathcal{D}}(h))}{n}}$

For large enough  $n$ , the binomial approximates a *Normal* distribution with

- mean  $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
- standard deviation  $\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$



## Central Limit Theorem

Consider a set of independent, identically distributed random variables  $Y_1 \dots Y_n$ , all governed by an arbitrary probability distribution with mean  $\mu$  and finite variance  $\sigma^2$ .

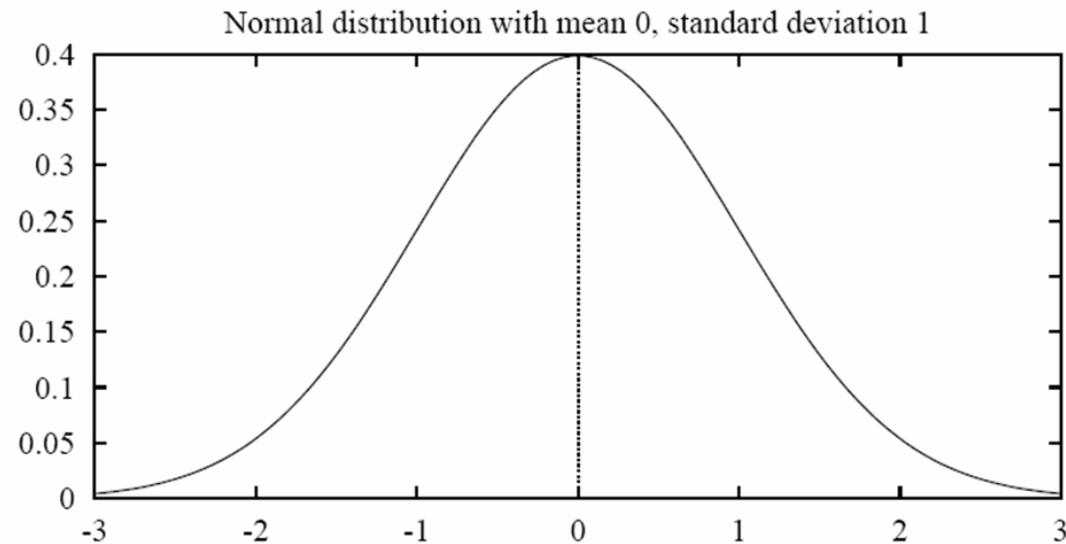
Define the sample mean,

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

**Central Limit Theorem.** As  $n \rightarrow \infty$ , the distribution governing  $\bar{Y}$  approaches a Normal distribution, with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ .

Notice that the standard deviation for  $Y$  is  $\sigma$  but the standard deviation for  $\bar{Y}$  is  $\frac{\sigma}{\sqrt{n}}$  (aka the *standard error of the mean*)

## Fact about the Normal Distribution

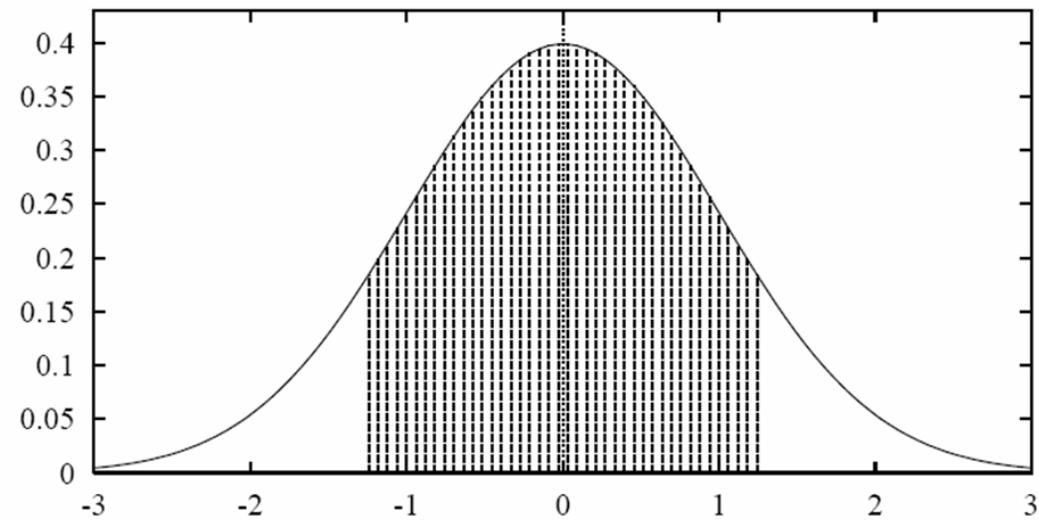


$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that  $X$  will fall into the interval  $(a, b)$  is given by  $\int_a^b p(x)dx$

- Expected, or mean value of  $X$ ,  $E[X]$ , is  $E[X] = \mu$
- Variance of  $X$  is  $Var(X) = \sigma^2$
- Standard deviation of  $X$ ,  $\sigma_X$ , is  $\sigma_X = \sigma$

## Facts about the Normal Probability Distribution



80% of area (probability) lies in  $\mu \pm 1.28\sigma$

N% of area (probability) lies in  $\mu \pm z_N\sigma$

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

## Confidence Intervals, More Correctly

If

- $S$  contains  $n$  examples, drawn independently of  $h$  and each other
- $n \geq 30$

Then

- With approximately 95% probability,  $error_S(h)$  lies in interval

$$error_{\mathcal{D}}(h) \pm 1.96 \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

equivalently,  $error_{\mathcal{D}}(h)$  lies in interval

$$error_S(h) \pm 1.96 \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

which is approximately

$$error_S(h) \pm 1.96 \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

## Calculating Confidence Intervals: Recipe 2

1. Pick parameter  $p$  to estimate
  - $error_{\mathcal{D}}(h)$
2. Choose an unbiased estimator
  - $error_{\mathcal{S}}(h)$
3. Determine probability distribution that governs estimator
  - $error_{\mathcal{S}}(h)$  governed by Binomial distribution, approximated by Normal when  $n \geq 30$
4. Find interval  $(L, U)$  such that  $N\%$  of probability mass falls in the interval
  - Use table of  $z_N$  values

## Estimating the Difference Between Hypotheses: Recipe 3

Test  $h_1$  on sample  $S_1$ , test  $h_2$  on  $S_2$

1. Pick parameter to estimate

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

2. Choose an estimator

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

3. Determine probability distribution that governs estimator

$$\sigma_{\hat{d}} \approx \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

4. Find interval  $(L, U)$  such that  $N\%$  of probability mass falls in the interval

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

A Tastier Version of Recipe 3: Paired  $z$ -test to compare  $h_A, h_B$

1. Partition data into  $k$  disjoint test sets  $T_1, T_2, \dots, T_k$  of equal size, where this size is at least 30.

2. For  $i$  from 1 to  $k$ , do

$$Y_i \leftarrow \text{error}_{T_i}(h_A) - \text{error}_{T_i}(h_B)$$

3. Return the value  $\bar{Y}$ , where  $\bar{Y} \equiv \frac{1}{k} \sum_{i=1}^k Y_i$

By the Central Limit Theorem,  $\bar{Y}$  is approximately Normal with variance

$$s_{\bar{Y}} \equiv \frac{1}{k} \left( \frac{1}{k} \sum_{i=1}^k (Y_i - \bar{Y})^2 \right)$$

Yet another Version of Recipe 3: Paired  $t$ -test to compare  $h_A, h_B$

1. Partition data into  $k$  disjoint test sets  $T_1, T_2, \dots, T_k$  of equal size,  
where ~~this size is at least 30~~

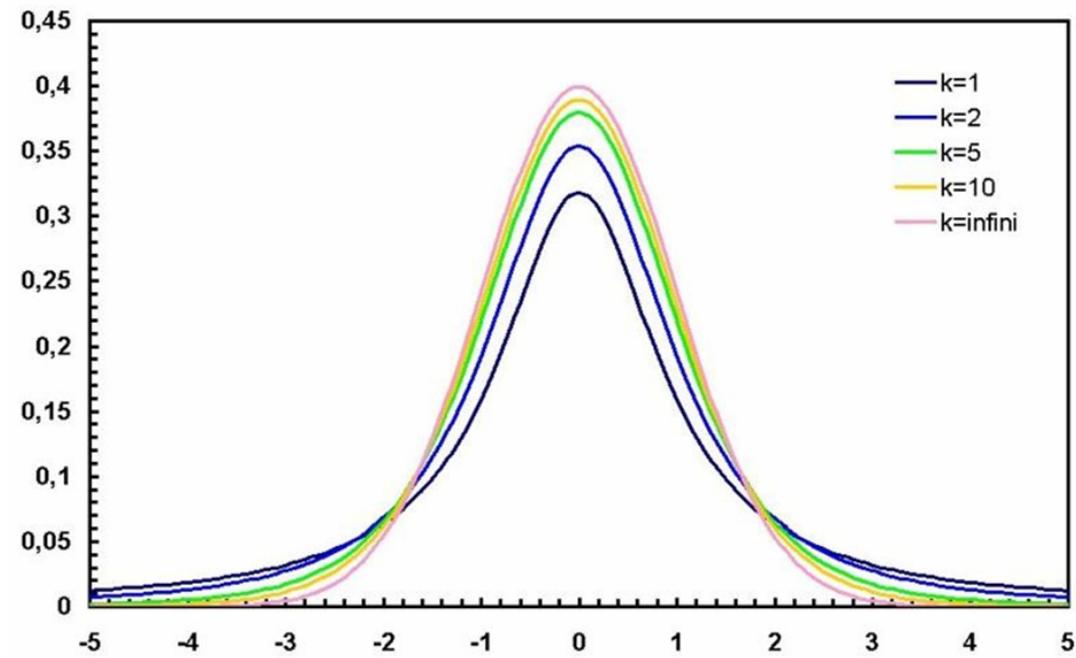
2. For  $i$  from 1 to  $k$ , do

$$y_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$$

3. Return the value  $\bar{y}$ , where  $\bar{y} \equiv \frac{1}{k} \sum_{i=1}^k y_i$

$\bar{Y}$  is approximately distributed as a  $t$  distribution with  $k - 1$  degrees of freedom.

## The $t$ -distribution



$\nu$	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
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

### Yet Another Version of Recipe 3

1. Formulate the *null hypothesis*: the expected value of the difference is zero: i.e., for  $Y = error_S(h_A) - error_S(h_B)$

$$E[Y] = 0$$

2. Use samples  $S_1, \dots, S_k$  to generate samples  $y_1, \dots, y_k$  of  $Y$ , and then  $\bar{y}$  a sample of  $\bar{Y} \tilde{N}(\mu, \sigma)$  where
  - $\sigma$  is estimated with the sample
  - $\mu = 0$  by the hypotheses
3. Assume  $\bar{y} > 0$ . You might compute
  - the probability  $p_1$  of seeing  $\bar{Y} \geq \bar{y}$  under the null hypothesis (one-tail test)
  - the probability  $p_2$  of seeing  $\bar{Y} \geq \bar{y}$  or  $\bar{Y} \leq -\bar{y}$  under the null hypothesis (two-tail test)
4. If  $p_1$  is low enough, then you *reject the null hypothesis*

## Recipe 4: Comparing learning algorithms $L_A$ and $L_B$

What we'd like to estimate:

$$E_{S \subset \mathcal{D}}[\text{error}_{\mathcal{D}}(L_A(S)) - \text{error}_{\mathcal{D}}(L_B(S))]$$

where  $L(S)$  is the hypothesis output by learner  $L$  using training set  $S$

i.e., the expected difference in true error between hypotheses output by learners  $L_A$  and  $L_B$ , when trained using randomly selected training sets  $S$  drawn according to distribution  $\mathcal{D}$ .

But, given limited data  $D_0$ , what is a good estimator?

- could partition  $D_0$  into training set  $S$  and training set  $T_0$ , and measure

$$\text{error}_{T_0}(L_A(S_0)) - \text{error}_{T_0}(L_B(S_0))$$

- even better, repeat this many times and average the results (next slide)

## Comparing learning algorithms $L_A$ and $L_B$

1. Partition data  $D_0$  into  $k$  disjoint test sets  $T_1, T_2, \dots, T_k$  of equal size.
2. For  $i$  from 1 to  $k$ , do  
*use  $T_i$  for the test set, and the remaining data for training set  $S_i$* 
  - $S_i \leftarrow \{D_0 - T_i\}$
  - $h_A \leftarrow L_A(S_i)$
  - $h_B \leftarrow L_B(S_i)$
  - $y_i \leftarrow \text{error}_{T_i}(h_A) - \text{error}_{T_i}(h_B)$
3. Return the value  $\bar{y}$ , where  $\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^k y_i$
4.  $\frac{1}{k} \sum_{i=1}^k \text{error}_{T_i}(L(S_i))$  is the cross-validated error rate of  $A$ , and the procedure is called  $k$ -fold cross-validation.

A special case: if  $k = |D_0|$  and  $|T_i| = 1$  this is leave-one-out cross-validation.

## Comparing learning algorithms $L_A$ and $L_B$

Notice we'd like to use the paired  $t$  test on  $\bar{y}$  to obtain a confidence interval (or reject the null, etc)

In practice this is a good approximation, but it's not really correct: because the training sets in this algorithm are not independent (they overlap!), the error rates are not independent

It's more correct to view algorithm as producing an estimate of

$$E_{S \subset D_0} [error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

instead of

$$E_{S \tilde{\mathcal{D}}} [error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

but even this approximation is better than no comparison

## Things to worry about

In real life:

- Do you understand the assumptions behind your recipes?
- Is your sample representative?
- Are your test cases independent?