# Evaluating Hypotheses

# Motivation

Evaluating the performance of learning systems is important because:

- One reason is simply to understand whether to use the hypothesis. (Learning systems are usually designed to predict the class of "future" unlabeled data points.)
- In some cases, evaluating hypotheses is an integral part of the learning process (example, when pruning a decision tree)

# Difficulties in Evaluating Hypotheses when only limited data are available

- ***Bias in the estimate:*** The observed accuracy of the learned hypothesis over the training examples is a poor estimator of its accuracy over future examples ==> we test the hypothesis on a test set chosen independently of the training set and the hypothesis.

- ***Variance in the estimate:*** Even if the hypothesis accuracy is measured over an unbiased set of test examples independent of the training examples, the measured accuracy can still vary from the true accuracy, depending on the makeup of the particular set of test examples. The smaller the set of test examples, the greater the expected variance.

# Questions Considered

- Given the observed accuracy of a hypothesis over a limited sample of data, how well does this estimate its accuracy over additional examples?
- Given that one hypothesis outperforms another over some sample data, how probable is it that this hypothesis is more accurate, in general?
- When data is limited what is the best way to use this data to both learn a hypothesis and estimate its accuracy?

# Estimating Hypothesis Accuracy

**Two Questions of Interest:**

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

# Sample Error and True Error

- **<u>Definition 1:</u>** The ***sample error*** (denoted $error_s(h)$) of hypothesis *h* with respect to target function *f* and data sample *S* is:

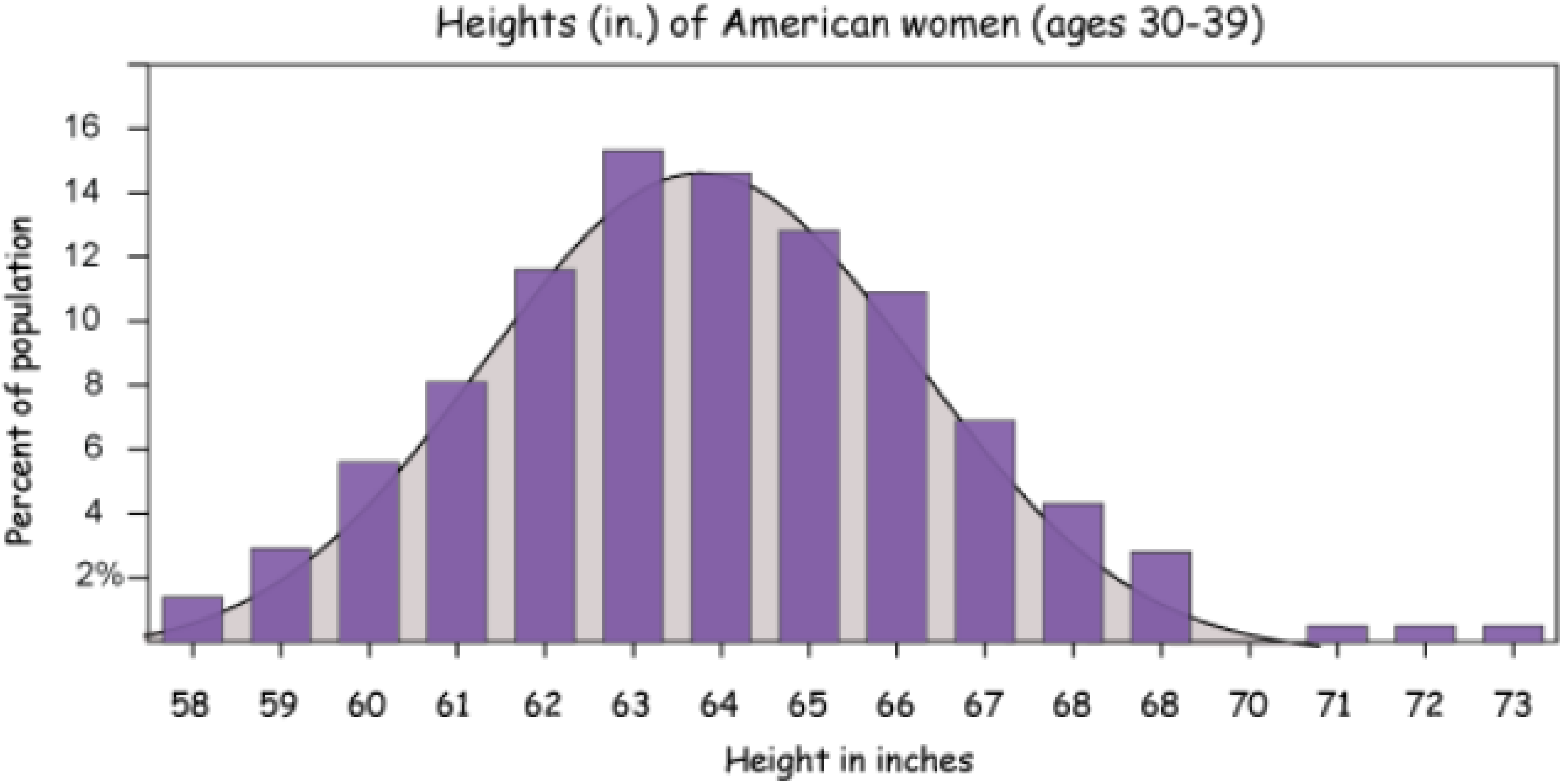
$$\boldsymbol{error_s(h) = 1/n \ \Sigma_{x \in S} \ \delta( f(x) , h(x) )}$$

where *n* is the number of examples in *S*, and the quantity $\delta(f(x),h(x))$ is 1 if $f(x) \neq h(x)$, and 0, otherwise.

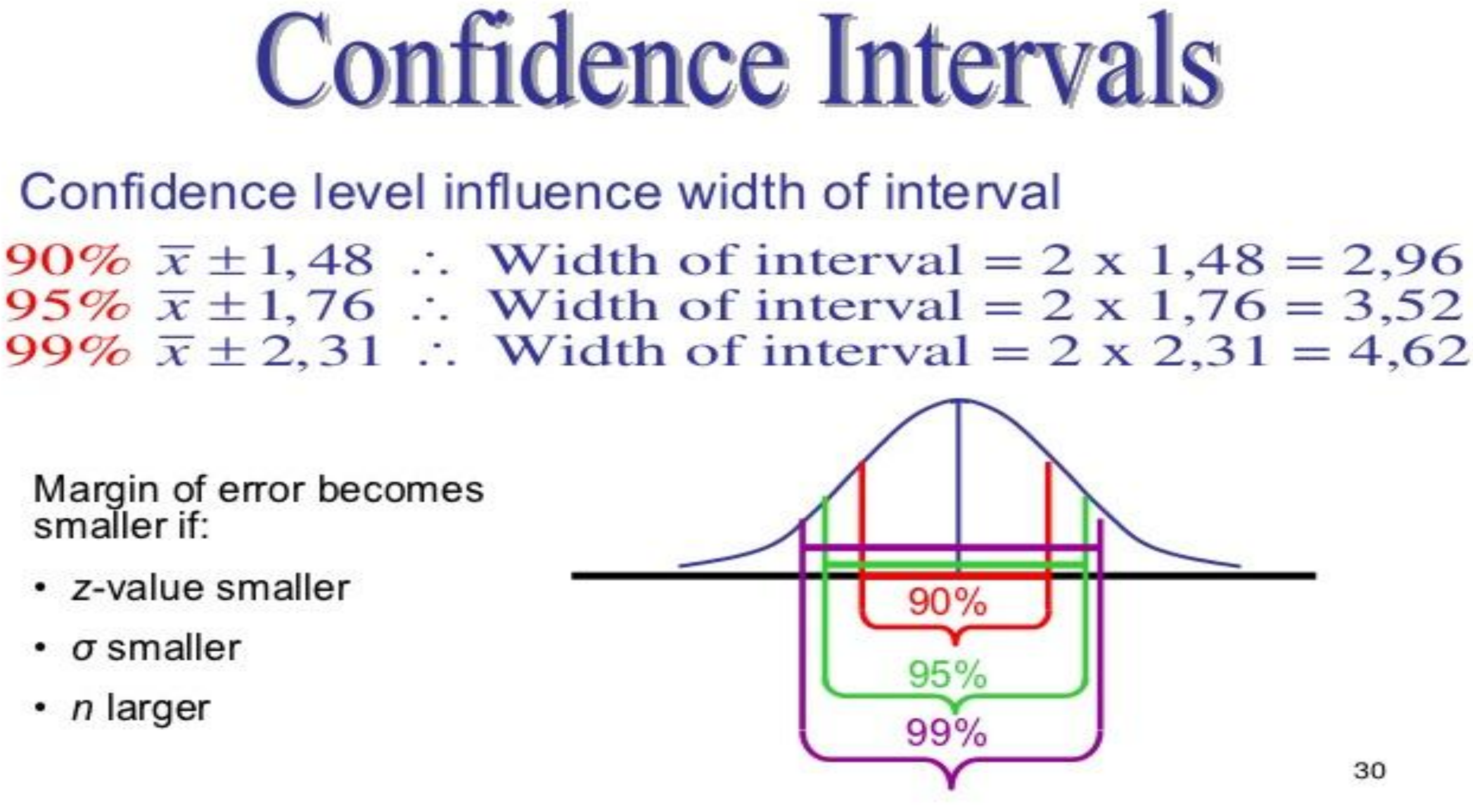
- **<u>Definition 2:</u>** The ***true error*** (denoted $error_D(h)$) of hypothesis *h* with respect to target function *f* and distribution *D*, is the probability that *h* will misclassify an instance drawn at random according to *D*.

$$\boldsymbol{error_D(h) = Pr_{x \in D} [ f(x) \neq h(x) ]}$$

a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.



A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times. Confidence intervals measure the degree of uncertainty or certainty in a sampling method.



# Confidence Intervals

Confidence level influence width of interval

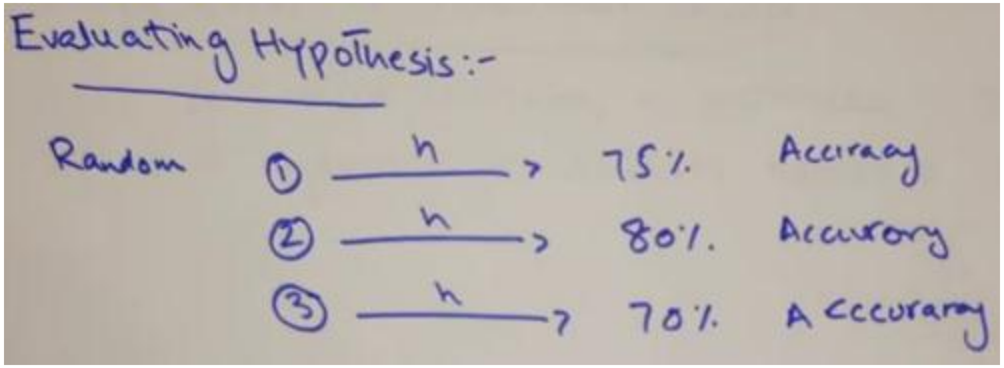
$90\%\ \overline{x} \pm 1,48 \ \therefore$ Width of interval $= 2 \times 1,48 = 2,96$

$95\%\ \overline{x} \pm 1,76 \ \therefore$ Width of interval $= 2 \times 1,76 = 3,52$

$99\%\ \overline{x} \pm 2,31 \ \therefore$ Width of interval $= 2 \times 2,31 = 4,62$

Margin of error becomes smaller if:

- *z*-value smaller
- *σ* smaller
- *n* larger

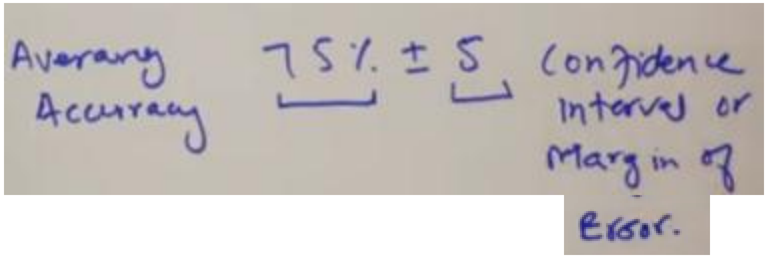


Evaluating Hypothesis :-

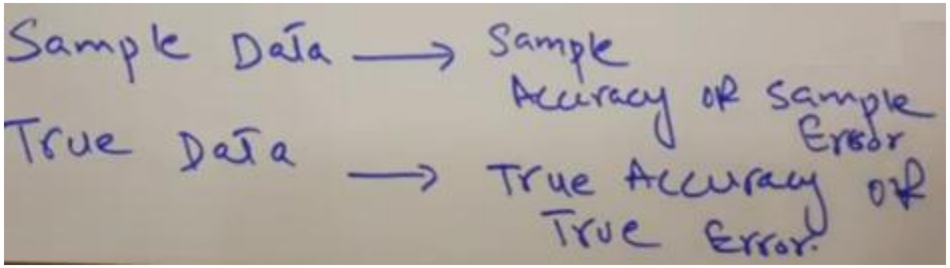
Random ① $\xrightarrow{h}$ 75% Accuracy

② $\xrightarrow{h}$ 80% Accurory

③ $\xrightarrow{h}$ 70% A cccuracy



Averany Accurracy $\underbrace{75\%}$ ± $\underbrace{5}$ confidence interval or Margin of Error.



Sample Data ⟶ Sample Accuracy or Sample Error

True Data ⟶ True Accuracy or True Error.

# Confidence Intervals for Discrete-Valued Hypotheses

- The general expression for approximate *N%* ***confidence intervals*** for $error_D(h)$ is:

$$\boldsymbol{error_S(h) \pm z_N \sqrt{error_S(h)\,(1 - error_S(h))\,/n}}$$

where $Z_N$ is given in [Mitchell, table 5.1]

- This approximation is quite good when

$$n\ error_S(h)(1 - error_S(h)) \geq 5$$

Suppose we wish to estimate the true error for some discrete valued hypothesis h, based on its observed sample error over a sample S, where

-- The sample S contains n examples drawn independent of one another, and independent of h, according to the probability distribution **D**

-- $n \geq 30$

-- Hypothesis h commits r errors over these n examples

(i.e., $error_s$ (h) = r/n).

Under these conditions, statistical theory allows to make the following assertions:

1. Given no other information, the most probable value of $error_D$ (h) is $error_s$(h)
2. With approximately 95% probability, the true error $error_D$ (h) lies in the interval

**Example:**

Suppose the data sample S contains n = 40 examples and that hypothesis h commits r = 12 errors over this data.

- The *sample error is* $error_s(h) = r/n = 12/40 = 0.30$

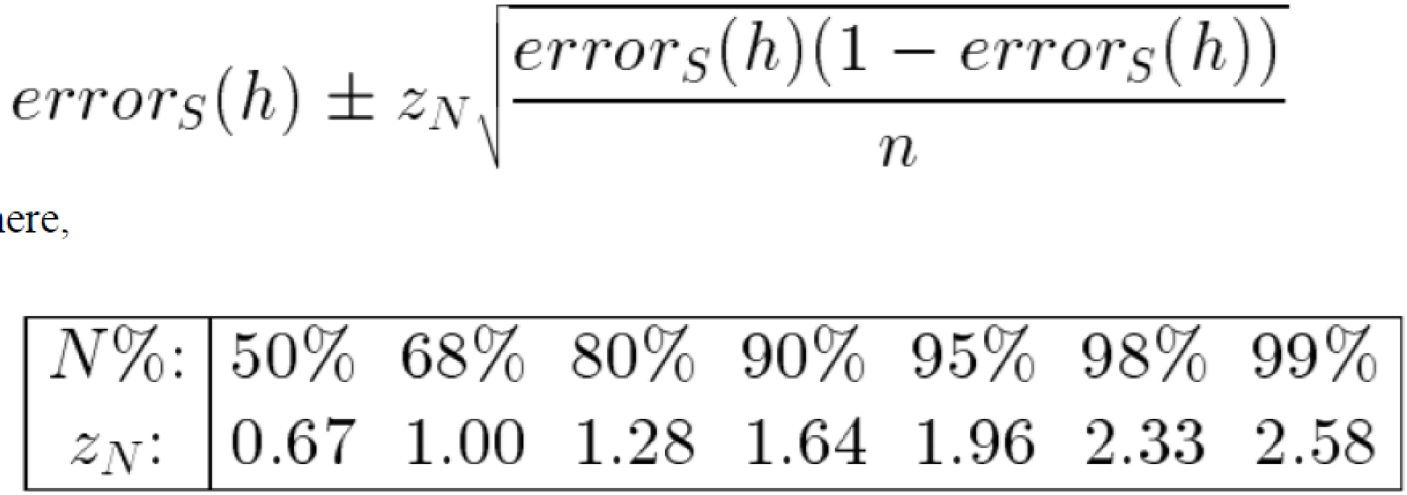
- Given no other information, *true error is* $error_D\ (h) = error_s(h)$*, i.e.,* $error_D\ (h) = 0.30$

- With the 95% confidence interval estimate for $\text{error}_\text{D}$ (h).

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

= 0.30 ± (1.96 * 0.07) = 0.30 ± 0.14

A different constant, $Z_N$, *is used to calculate the N% confidence interval. The general expression for approximate N% confidence intervals for* $error_D$ *(h) is*



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

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 |

The above equation describes how to calculate the confidence intervals, or error bars, for estimates of $error_D$ (h) that are based on $error_s$(h)

Example:

Suppose the data sample S contains n = 40 examples and that hypothesis h commits r = 12 errors over this data.

- The *sample error is* $error_s(h) = r/n = 12/40 = 0.30$
- With the 68% confidence interval estimate for $error_D$ (h).

$$error_S(h) \pm 1.00\sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

= 0.30 ± (1.00 * 0.07)

= 0.30 ± 0.07

- A *random variable* can be viewed as the name of an experiment with a probabilistic outcome. Its value is the outcome of the experiment.
- A *probability distribution* for a random variable $Y$ specifies the probability $\Pr(Y = y_i)$ that $Y$ will take on the value $y_i$, for each possible value $y_i$.
- The *expected value*, or *mean*, of a random variable $Y$ is $E[Y] = \sum_i y_i \Pr(Y = y_i)$. The symbol $\mu_Y$ is commonly used to represent E[Y].
- The *variance* of a random variable is $Var(Y) = E[(Y - \mu_Y)^2]$. The variance characterizes the width or dispersion of the distribution about its mean.
- The *standard deviation* of $Y$ is $\sqrt{Var(Y)}$. The symbol $\sigma_Y$ is often used used to represent the standard deviation of $Y$.
- The *Binomial distribution* gives the probability of observing $r$ heads in a series of $n$ independent coin tosses, if the probability of heads in a single toss is $p$.
- The *Normal distribution* is a bell-shaped probability distribution that covers many natural phenomena.
- The *Central Limit Theorem* is a theorem stating that the sum of a large number of independent, identically distributed random variables approximately follows a Normal distribution.
- An *estimator* is a random variable $Y$ used to estimate some parameter $p$ of an underlying population.
- The *estimation bias* of $Y$ as an estimator for $p$ is the quantity $(E[Y] - p)$. An unbiased estimator is one for which the bias is zero.
- A *N% confidence interval* estimate for parameter $p$ is an interval that includes $p$ with probability $N\%$.

Basic definitions and facts from statistics.

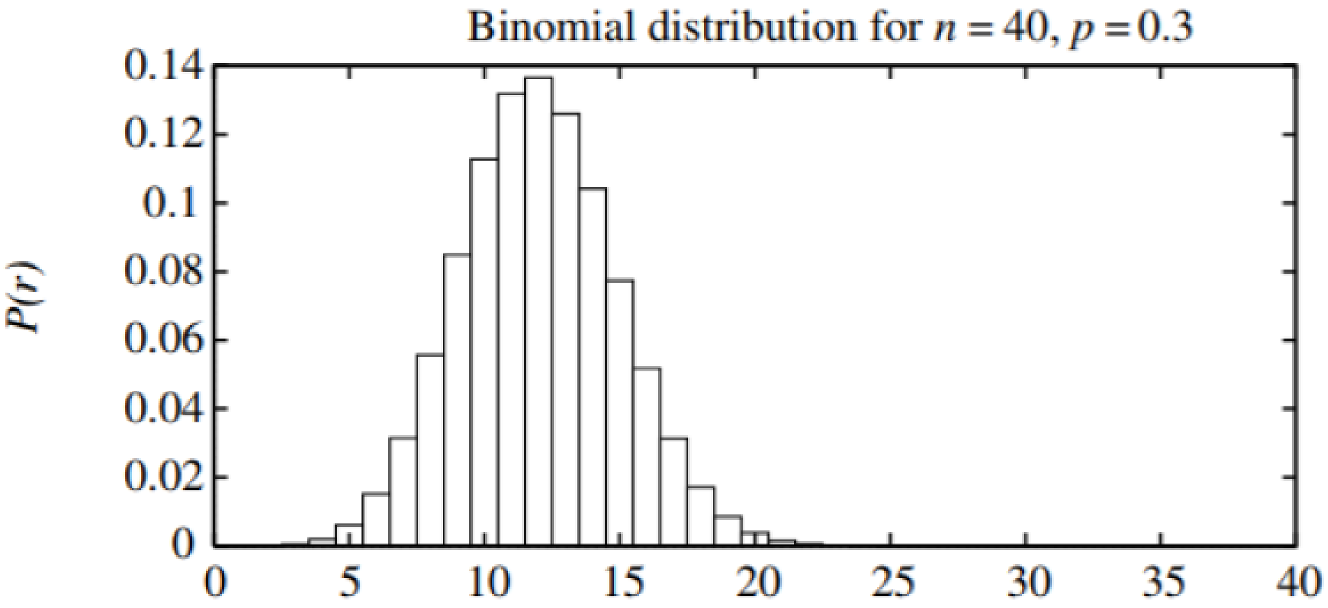
## BASICS OF SAMPLING THEORY

### Error Estimation and Estimating Binomial Proportions

- Collect a random sample S of n independently drawn instances from the distribution D, and then measure the sample error $error_s(h)$. Repeat this experiment many times, each time drawing a different random sample $S_i$ of size n, we would expect to observe different values for the various $error_{si}(h)$, depending on random differences in the makeup of the various Si. We say that $error_{si}(h)$, the outcome of the $i^{th}$ such experiment, is a random variable**.**

- Imagine that we were to run k random experiments, measuring the random variables $error_{s1}(h)$, $error_{s2}(h)$ . . . $errors_{sk}(h)$ and plotted a histogram displaying the frequency with which each possible error value is observed.

*As k grows, the histogram would approach a particular probability distribution called the Binomial distribution which is shown in below figure.*



A Binomial distribution is defined by the probability function

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

# Mean and Variance

- **<u>Definition 1:</u>** Consider a random variable *Y* that takes on possible values $y_1, \ldots, y_n$. The ***expected value*** (or ***mean value***) of *Y*, ***E[Y],*** is:

$$E[Y] = \Sigma_{i=1}^{n} y_i \, Pr(Y=y_i)$$

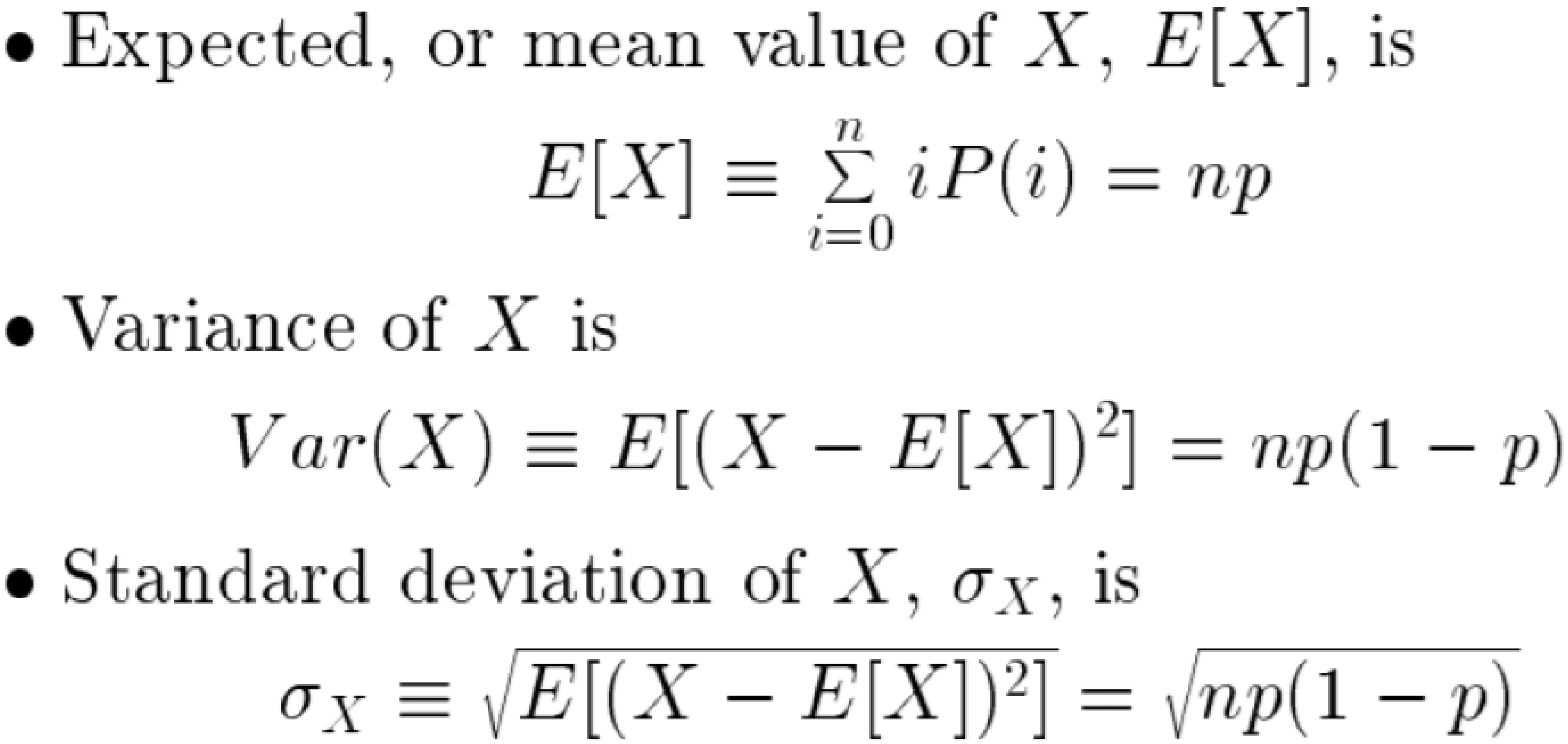
- **<u>Definition 2:</u>** The ***variance*** of a random variable *Y*, ***Var[Y],*** is: .

$$Var[Y] = E[(Y-E[Y])^2]$$

- **<u>Definition 3:</u>** The ***standard deviation*** of a random variable *Y* is the square root of the variance.

If the random variable ***X*** follows a Binomial distribution, then:

- The probability ***Pr(X = r)*** that ***X*** will take on the value ***r*** is given by ***P(r)***



- 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

$$\sigma_X \equiv \sqrt{E[(X - E[X])^2]} = \sqrt{np(1-p)}$$

## The Binomial Distribution

Consider the following problem for better understanding of Binomial Distribution

- Given a worn and bent coin and estimate the probability that the coin will turn up heads when tossed.
- Unknown probability of heads ***p***. Toss the coin ***n*** times and record the number of times ***r*** that it turns up heads.

$$\text{Estimate of } p = r / n$$

- If the experiment were *rerun*, generating a new set of ***n*** coin tosses, we might expect the number of heads ***r*** to vary somewhat from the value measured in the first experiment, yielding a somewhat different estimate for ***p***.
- The Binomial distribution describes for each possible value of ***r*** (i.e., from 0 to n), the probability of observing exactly ***r*** heads given a sample of ***n*** independent tosses of a coin whose true probability of heads is ***p***.

The general setting to which the Binomial distribution applies is:

1. There is a base experiment (e.g., toss of the coin) whose outcome can be described by a random variable ‘Y’. The random variable Y can take on two possible values (e.g., Y = 1 if heads, Y = 0 if tails).

2. The probability that Y = 1 on any single trial of the base experiment is given by some constant p, independent of the outcome of any other experiment. The probability that Y = 0 is therefore (1 - p). Typically, p is not known in advance, and the problem is to estimate it.

3. A series of n independent trials of the underlying experiment is performed (e.g., n independent coin tosses), producing the sequence of independent, identically distributed random variables $Y_1, Y_2, \ldots, Y_n$. Let R denote the number of trials for which $Y_i = 1$ in this series of n experiments

$$R \equiv \sum_{i=1}^{n} Y_i$$

4. The probability that the random variable R will take on a specific value r (e.g., the probability of observing exactly r heads) is given by the Binomial distribution

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

**Mean, Variance and Standard Deviation**

The Mean (expected value) is the average of the values taken on by repeatedly sampling the random variable

***Definition:*** Consider a random variable Y that takes on the possible values $y_1, \ldots y_n$. The expected value (Mean) of Y, E[Y], is

$$E[Y] \equiv \sum_{i=1}^{n} y_i \Pr(Y = y_i)$$

The Variance captures how far the random variable is expected to vary from its mean value.

***Definition:*** The variance of a random variable Y, Var[Y], is

$$Var[Y] \equiv E[(Y - E[Y])^2]$$

The variance describes the expected squared error in using a single observation of Y to estimate its mean E[Y].

The square root of the variance is called the standard deviation of Y, denoted $\sigma_y$

***Definition:*** The standard deviation of a random variable Y, $\sigma_y$, is

$$\sigma_Y \equiv \sqrt{E[(Y - E[Y])^2]}$$

In case the ***random variable Y is governed by a Binomial distribution***, then the Mean, Variance and standard deviation are given by

$$E[Y] = np$$

$$Var[Y] = np(1 - p)$$

$$\sigma_Y = \sqrt{np(1 - p)}$$

# Estimators, Bias and Variance

Since $error_S(h)$ (an **estimator** for the true error) obeys a Binomial distribution (See, [Mitchell, Section 5.3]),

we have:

- $error_S(h) = r/n$

- $error_D(h) = p$

Where

$n$ is the number of instances in the sample $S$,

$r$ is the number of instances from $S$ misclassified by $h$

$p$ is the probability of misclassifying a single instance drawn from $D$.

Definition:

Estimator:

$\text{error}_s$(h) an estimator for the true error $\text{error}_{D(}$h): An estimator is any random variable used to estimate some parameter of the underlying population from which the sample is drawn

Estimation bias: is the difference between the expected value of the estimator and the true value of the parameter.

The ***estimation bias*** (≠ from the inductive bias) of an estimator *Y* for an arbitrary parameter *p* is ***E[Y] – p***

The ***standard deviation*** for $\text{error}_S(h)$ is given by

$$\sqrt{p(1-p)/n} \approx \sqrt{\text{error}_S(h)(1-\text{error}_S(h))/n}$$

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

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

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

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

If the estimation bias is zero, we say that Y is an *unbiased estimator for p. Notice* this will be the case if the average of many random values of Y generated by repeated random experiments (i.e., E[Y]) converges toward p.

Is *error$_s$(h)  an unbiased estimator for error$_D$(h)? Yes, because for a Binomial* distribution the expected value of *r is equal to np (Equation 5.4). It* follows, given that *n is a constant, that the expected value of r/n is p.*

Two quick remarks are in order regarding the estimation bias. First, when we mentioned at the beginning of this chapter that testing the hypothesis on the training examples provides an optimistically biased estimate of hypothesis error, it is exactly this notion of estimation bias to which we were referring. In order for *errors(h) to give an unbiased estimate of* $error_D(h)$*, the hypothesis h and sample S must be chosen independently.*

Second, this notion of *estimation bias should* not be confused with the *inductive bias of a learner introduced in Chapter 2. The* estimation bias is a numerical quantity, whereas the inductive bias is a set of assertions.

A second important property of any estimator is its variance. Given a choice among alternative unbiased estimators, it makes sense to choose the one with least variance. By our definition of variance, this choice will yield the smallest expected squared error between the estimate and the true value of the parameter.

To illustrate these concepts, suppose we test a hypothesis and find that it commits $r = 12$ errors on a sample of $n = 40$ randomly drawn test examples. Then an unbiased estimate for $error_{\mathcal{D}}(h)$ is given by $error_S(h) = r/n = 0.3$. The variance in this estimate arises completely from the variance in $r$, because $n$ is a constant. Because $r$ is Binomially distributed, its variance is given by Equation (5.7) as $np(1-p)$. Unfortunately $p$ is unknown, but we can substitute our estimate $r/n$ for $p$. This yields an estimated variance in $r$ of $40 \cdot 0.3(1 - 0.3) = 8.4$, or a corresponding standard deviation of $\sqrt{8.4} \approx 2.9$. This implies that the standard deviation in $error_S(h) = r/n$ is approximately $2.9/40 = .07$. To summarize, $error_S(h)$ in this case is observed to be 0.30, with a standard deviation of approximately 0.07. (See Exercise 5.1.)

In general, given $r$ errors in a sample of $n$ independently drawn test examples, the standard deviation for $error_S(h)$ is given by

$$\sigma_{error_S(h)} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}} \tag{5.8}$$

which can be approximated by substituting $r/n = error_S(h)$ for $p$

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}} \tag{5.9}$$

To illustrate these concepts, suppose we test a hypothesis and find that it commits $r = 12$ errors on a sample of $n = 40$ randomly drawn test examples. Then an unbiased estimate for $error_{\mathcal{D}}(h)$ is given by $error_S(h) = r/n = 0.3$. The variance in this estimate arises completely from the variance in $r$, because $n$ is a constant. Because $r$ is Binomially distributed, its variance is given by Equation (5.7) as $np(1-p)$. Unfortunately $p$ is unknown, but we can substitute our estimate $r/n$ for $p$. This yields an estimated variance in $r$ of $40 \cdot 0.3(1 - 0.3) = 8.4$, or a corresponding standard deviation of $\sqrt{8.4} \approx 2.9$. This implies that the standard deviation in $error_S(h) = r/n$ is approximately $2.9/40 = .07$. To summarize, $error_S(h)$ in this case is observed to be 0.30, with a standard deviation of approximately 0.07. (See Exercise 5.1.)

In general, given $r$ errors in a sample of $n$ independently drawn test examples, the standard deviation for $error_S(h)$ is given by

$$\sigma_{error_S(h)} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}} \tag{5.8}$$

which can be approximated by substituting $r/n = error_S(h)$ for $p$

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}} \tag{5.9}$$

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

## Confidence Intervals

describe the uncertainty associated with an estimate is to
give an interval within which the true value is expected to fall, along with the probability with which it is expected to fall into this interval. Such estimates are called *conjdence interval estimates.*

***Definition: An N% confidence interval for some parameter p is an interval that is*** expected with probability ***N% to contain p.***

# Examples

Suppose you test a hypothesis $h$ and find that it commits $r = 12$ errors on a sample $S$ of $n = 40$ randomly drawn test examples. An unbiased etimate for $error_{\mathcal{D}}(h)$ is given by $error_S(h) = r/n = 0.3$.
The variance in this estimate arises from $r$ alone ($n$ is a constant).

From the Binomial distribution, this variance is $np(1 - p)$.

We can substitute $r/n$ as an estimate for $p$. Then the variance for $r$ is estimated to be $40 \times 0.3(1 - 0.3) = 8.4$ and the standard deviation is $\sqrt{8.4} \approx 2.9$.

Therefore the standard deviation in $error_S(h) = r/n$ is approximately $2.9/40 = 0.07$.

$error_S(h)$ is observed to be $0.30$ with standard deviation of approximately $0.07$.

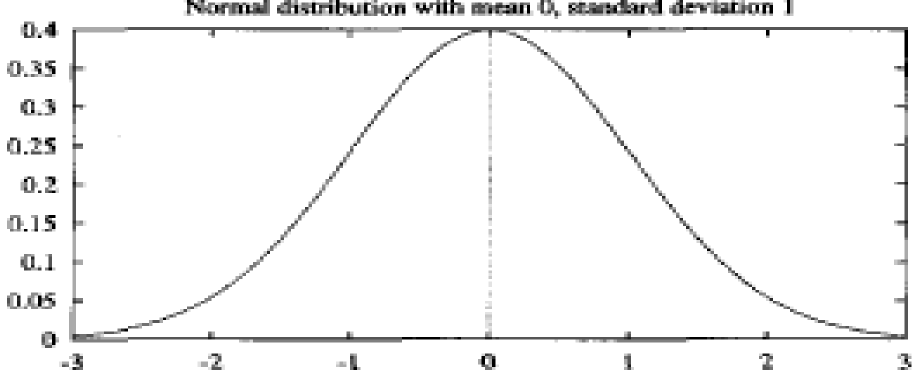
# Examples

Suppose you test a hypothesis $h$ and find that it commits $r = 300$ errors on a sample $S$ of $n = 1000$ randomly drawn test examples. What is the standard deviation in $error_S(h)$ ?

The standard deviation for $r$ is estimated to be $\sqrt{1000 \times 0.3(1 - 0.3)} \approx 14.5$.

Therefore the standard deviation in $error_S(h) = r/n$ is approximately $14.5/1000 = .0145$.

$error_S(h)$ is observed to be $0.30$ with standard deviation of approximately $.0145$.



A Normal distribution (also called a Gaussian distribution) is a bell-shaped distribution defined by the probability density function

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

A Normal distribution is fully determined by two parameters in the above formula: $\mu$ and $\sigma$.

If the random variable $X$ follows a normal distribution, then:

- The probability that $X$ will fall into the interval $(a, b)$ is given by

$$\int_a^b p(x)dx$$

- The expected, or mean value of $X$, $E[X]$, is

$$E[X] = \mu$$

- The variance of $X$, $Var(X)$, is

$$Var(X) = \sigma^2$$

- The standard deviation of $X$, $\sigma_X$, is

$$\sigma_X = \sigma$$

The Central Limit Theorem (Section 5.4.1) states that the sum of a large number of independent, identically distributed random variables follows a distribution that is approximately Normal.

The Normal or Gaussian distribution.

# Normal Distribution Approximates Binomial

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

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

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

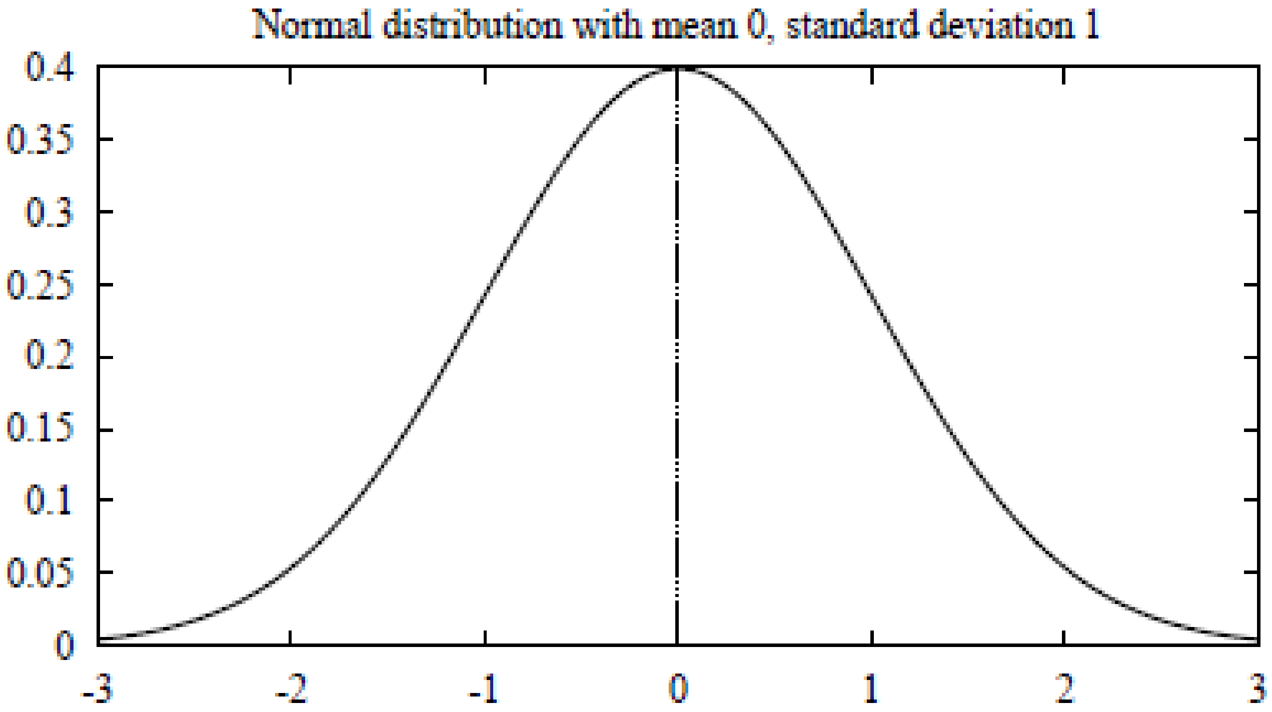
# Normal Distribution Approximates Binomial

Approximate this by a *Normal* distribution with

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

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

# Normal Probability Distribution



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

## 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 |

Note: with 80% confidence the value of the random variable will lie in the two-sided interval $[-1.28, 1.28]$.

With 10% confidence it will lie to the right of this interval (resp. left).

With 90% confidence it will lie in the one-sided interval $[-\infty, 1.28]$

Let $\alpha$ be the probability that the value lies *outside* the interval.

Then a $100(1-\alpha)\%$ two-sided confidence interval with lower-bound $L$ and upper-bound $U$ can be converted into a $100(1-(\alpha/2))\%$ one-sided confidence interval with lower bound $L$ and no upper bound (resp. upper bound $U$ and no lower bound).

# 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

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

# Central Limit Theorem

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

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

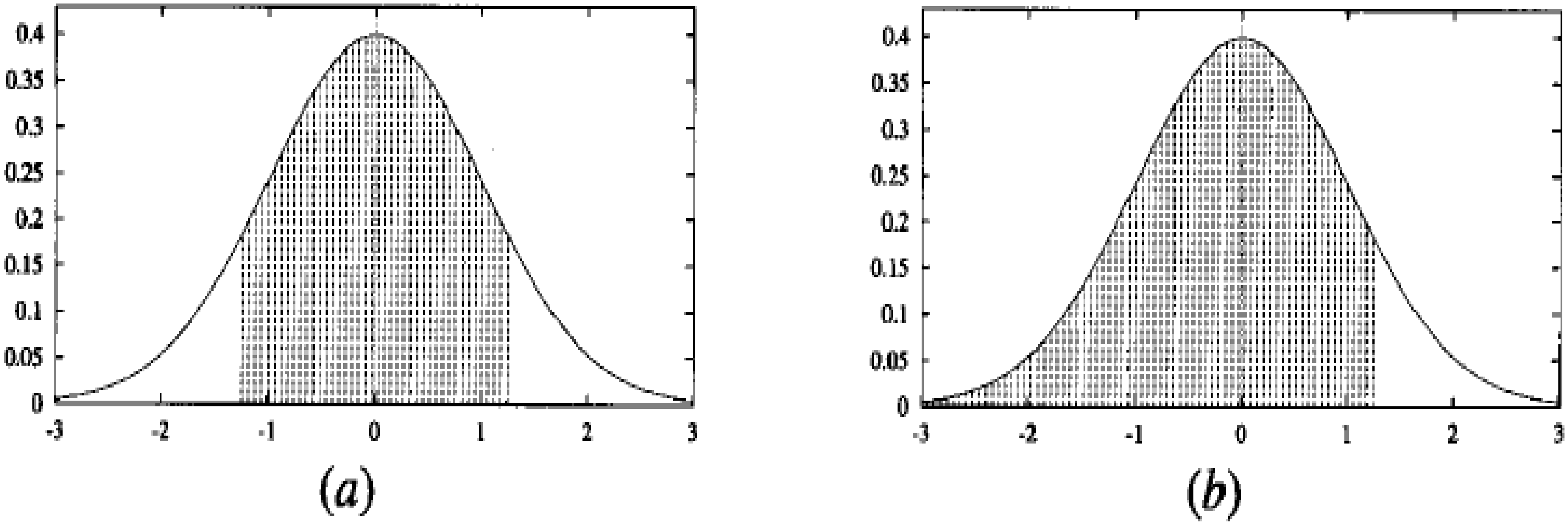
**Central Limit Theorem.** As $n \to \infty$, the distribution governing $\bar{Y}$ approaches a Normal distribution, with mean $\mu$ and variance $\frac{\sigma^2}{n}$.

*the sum of a large number of independent, identically distributed (i.i.d) random variables follows a distribution that is approximately Normal.*

# Calculating Confidence Intervals

1. Pick parameter $p$ to estimate
   - $error_{\mathcal{D}}(h)$
2. Choose an estimator
   - $error_S(h)$
3. Determine probability distribution that governs estimator
   - $error_S(h)$ 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

# Two-sided and One-sided Bounds



A Normal distribution with mean 0, standard deviation 1. (*a*) With 80% confidence, the value of the random variable will lie in the two-sided interval $[-1.28, 1.28]$. Note $z_{.80} = 1.28$. With 10% confidence it will lie to the right of this interval, and with 10% confidence it will lie to the left. (*b*) With 90% confidence, it will lie in the one-sided interval $[-\infty, 1.28]$.

confidence interval is a two-sided bound; that is, it bounds the estimated quantity from above and from below. In some cases, we will be interested only in a one-sided bound. For example, we might be interested in the question "What is the probability that $\text{error}_D$,(h) is at most U?' This kind of onesided question is natural when we are only interested in bounding the maximum error of h and do not mind if the true error is much smaller than estimated. procedure for finding such onesided error bounds. It follows from the fact that the Normal distribution is symmetric about its mean. Because of this fact, any two-sided confidence interval based on a Normal distribution can be converted to a corresponding one-sided interval with twice the confidence (see Figure 5.l(b)). That is, a 100(1- a)% confidence interval with lower bound L and upper bound U implies a 100(1- a/2)% confidence interval with lower bound L and no upper bound. It also implies a 100(1 -a/2)% confidence interval with upper bound U and no lower bound. Here a corresponds to the probability that the correct value lies outside the stated interval. In other words, a is the probability that the value will fall into the unshaded region in Figure 5.l(a), and a/2 is the probability that it will fall into the unshaded region in Figure 5.l(b).

To illustrate, consider again the example in which $h$ commits $r = 12$ errors over a sample of $n = 40$ independently drawn examples. As discussed above, this leads to a (two-sided) 95% confidence interval of $0.30 \pm 0.14$. In this case, $100(1 - \alpha) = 95\%$, so $\alpha = 0.05$. Thus, we can apply the above rule to say with $100(1 - \alpha/2) = 97.5\%$ confidence that $error_{\mathcal{D}}(h)$ is at most $0.30 + 0.14 = 0.44$, making no assertion about the lower bound on $error_{\mathcal{D}}(h)$. Thus, we have a one-sided error bound on $error_{\mathcal{D}}(h)$ with double the confidence that we had in the corresponding two-sided bound

# Difference in Error of two Hypotheses

- Let $h_1$ and $h_2$ be two hypotheses for some discrete-valued target function. $H_1$ has been tested on a sample $S_1$ containing $n_1$ randomly drawn examples, and $h_2$ has been tested on an independent sample $S_2$ containing $n_2$ examples drawn from the same distribution.

- Let's estimate the difference between the true errors of these two hypotheses, ***d***, by computing the difference between the sample errors: $\boldsymbol{\hat{d} = error_{S1}(h1)-error_{S2}(h2)}$

- The approximate ***N%*** confidence interval for ***d*** is:

$$\boldsymbol{d^\wedge \pm Z_N \sqrt{error_{S1}(h_1)(1-error_{S1}(h_1))/n_1 + error_{S2}(h_2)(1-error_{S2}(h_2))/n_2}}$$

# Difference Between Hypotheses

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

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

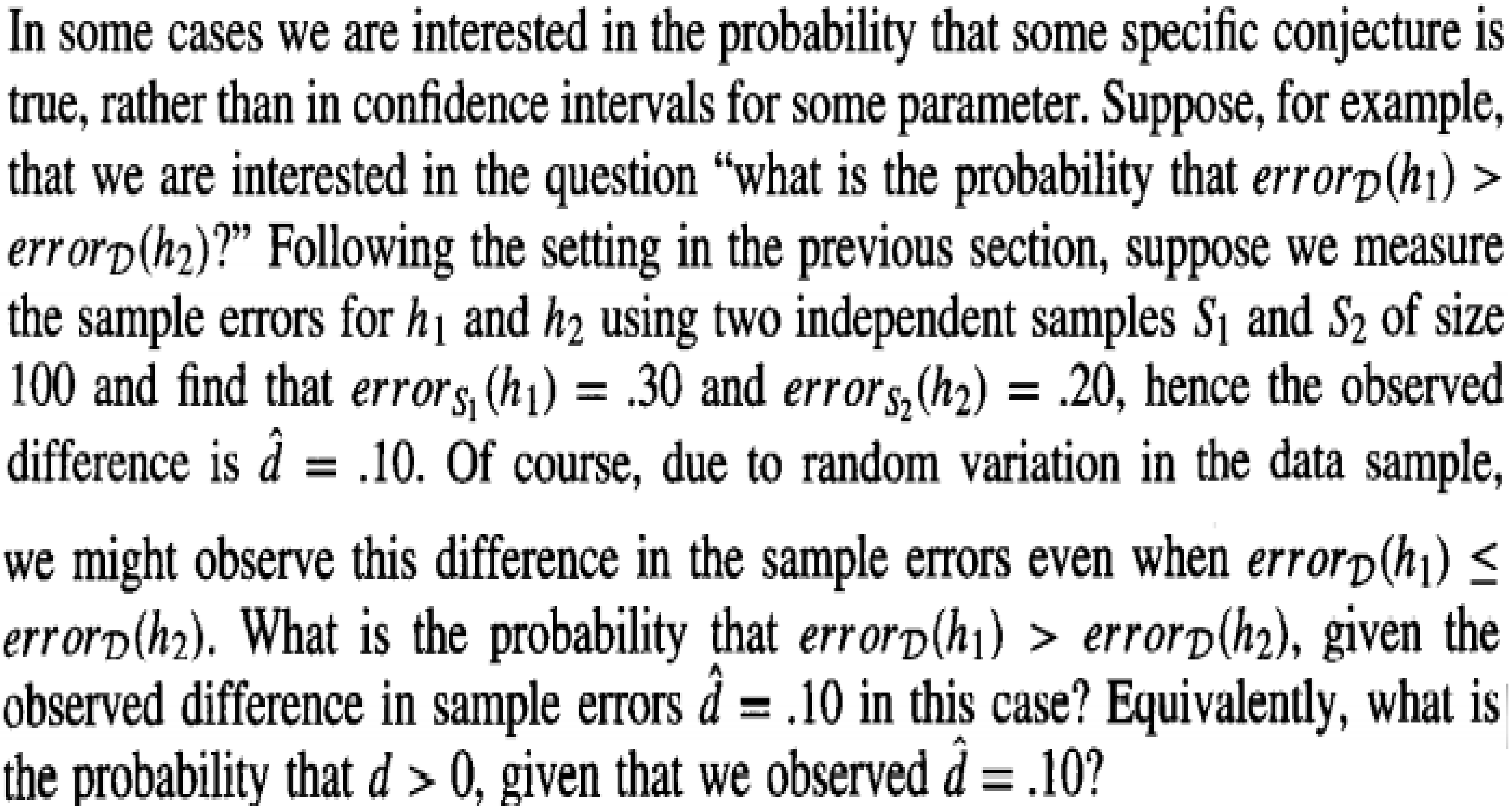
3. Determine probability distribution that governs estimator

$$\sigma_{\mathrm{d}} \approx \sqrt{\frac{\mathrm{errors_1(h_1)(1 - errors_1(h_1))}}{\mathrm{n_1}} + \frac{\mathrm{errors_2(h_2)(1 - errors_2(h_2))}}{\mathrm{n_2}}}$$

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

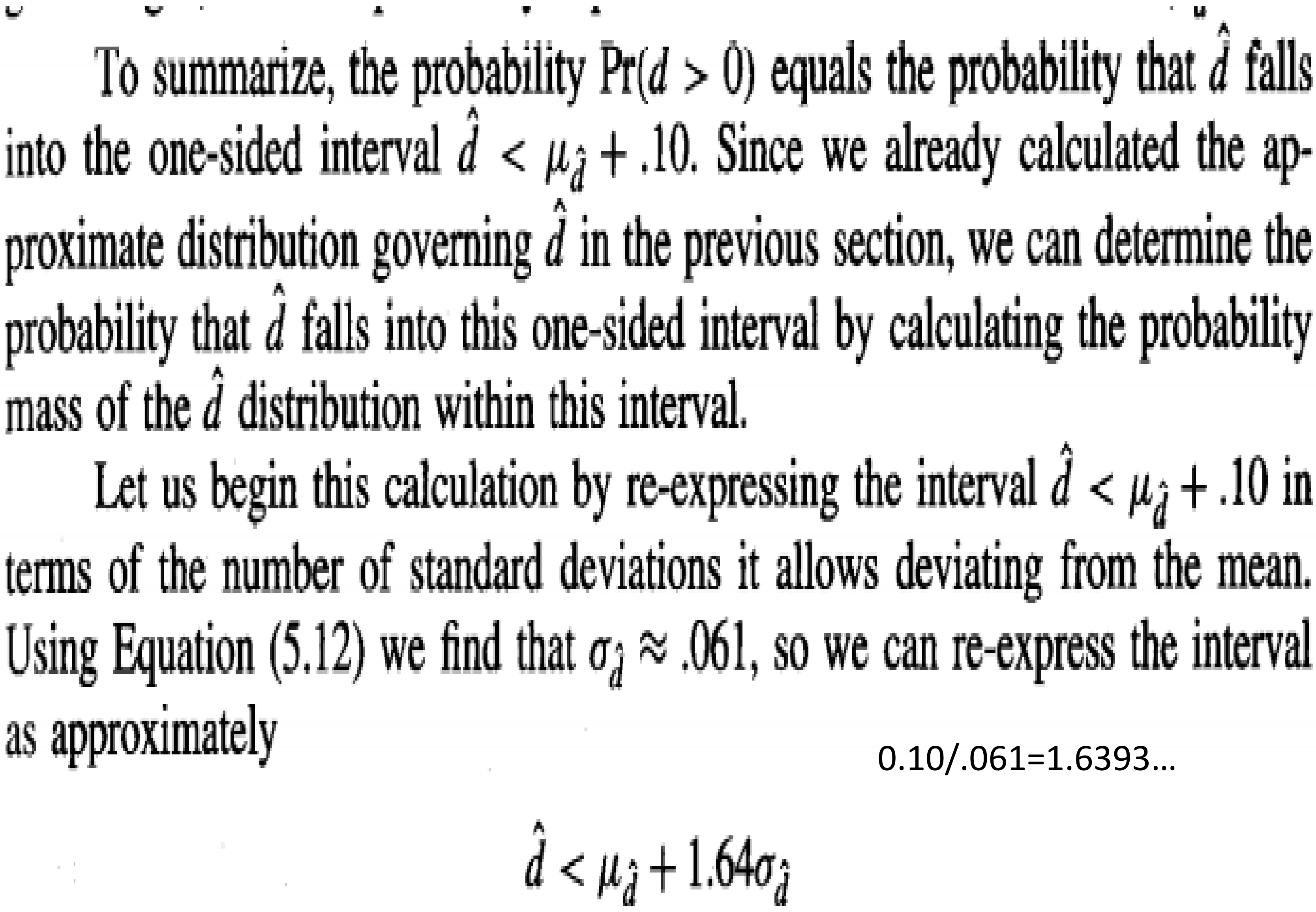
$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_S}{n_2}}$$

# Hypothesis Testing



In some cases we are interested in the probability that some specific conjecture is true, rather than in confidence intervals for some parameter. Suppose, for example, that we are interested in the question "what is the probability that $error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$?" Following the setting in the previous section, suppose we measure the sample errors for $h_1$ and $h_2$ using two independent samples $S_1$ and $S_2$ of size 100 and find that $error_{S_1}(h_1) = .30$ and $error_{S_2}(h_2) = .20$, hence the observed difference is $\hat{d} = .10$. Of course, due to random variation in the data sample, we might observe this difference in the sample errors even when $error_{\mathcal{D}}(h_1) \leq error_{\mathcal{D}}(h_2)$. What is the probability that $error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$, given the observed difference in sample errors $\hat{d} = .10$ in this case? Equivalently, what is the probability that $d > 0$, given that we observed $\hat{d} = .10$?

Note the probability $\Pr(d > 0)$ is equal to the probability that $\hat{d}$ has not overestimated $d$ by more than .10. Put another way, this is the probability that $\hat{d}$ falls into the one-sided interval $\hat{d} < d + .10$. Since $d$ is the mean of the distribution governing $\hat{d}$, we can equivalently express this one-sided interval as $\hat{d} < \mu_{\hat{d}} + .10$.



To summarize, the probability $\Pr(d > 0)$ equals the probability that $\hat{d}$ falls into the one-sided interval $\hat{d} < \mu_{\hat{d}} + .10$. Since we already calculated the approximate distribution governing $\hat{d}$ in the previous section, we can determine the probability that $\hat{d}$ falls into this one-sided interval by calculating the probability mass of the $\hat{d}$ distribution within this interval.

Let us begin this calculation by re-expressing the interval $\hat{d} < \mu_{\hat{d}} + .10$ in terms of the number of standard deviations it allows deviating from the mean. Using Equation (5.12) we find that $\sigma_{\hat{d}} \approx .061$, so we can re-express the interval as approximately

0.10/.061=1.6393…

$$\hat{d} < \mu_{\hat{d}} + 1.64\sigma_{\hat{d}}$$

What is the confidence level associated with this one-sided interval for a Normal distribution? Consulting Table 5.1, we find that 1.64 standard deviations about the mean corresponds to a two-sided interval with confidence level 90%. Therefore, the one-sided interval will have an associated confidence level of 95%.

Therefore, given the observed $\hat{d} = .10$, the probability that $error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$ is approximately .95. In the terminology of the statistics literature, we say that we accept the hypothesis that "$error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$" with confidence 0.95. Alternatively, we may state that we reject the opposite hypothesis (often called the null hypothesis) at a $(1 - 0.95) = .05$ level of significance.

# Comparing learning algorithms $L_A$ and $L_B$

What we'd like to estimate:

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

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
  $$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$$
- even better, repeat this many times and average the results (next slide)

# k-Fold Cross-Validation

1. Partition the available data $\boldsymbol{D_0}$ into $\boldsymbol{k}$ disjoint subsets $\boldsymbol{T_1, T_2, \ldots, T_k}$ of equal size, where this size is at least 30.

2. For $\boldsymbol{i}$ from **1** to $\boldsymbol{k}$, do

   use $\boldsymbol{T_i}$ for the test set, and the remaining data for training set $\boldsymbol{S_i}$

   - $\boldsymbol{S_i \leftarrow \{D_0 - T_i\}}$
   - $\boldsymbol{h_A \leftarrow L_A(S_i)}$
   - $\boldsymbol{h_B \leftarrow L_B(S_i)}$
   - $\boldsymbol{\delta i \leftarrow error_{Ti}(h_A) - error_{Ti}(h_B)}$

3. Return the value $\boldsymbol{avg(\delta)}$, where .

   $\mathbf{avg(\delta) = 1/k \sum_{i=1}^{k} \delta_i}$

# Comparing learning algorithms $L_A$ and $L_B$

1. Partition data $D_0$ into $k$ disjoint test sets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30.

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

   *use $T_i$ for the test set, and the remaining data for training set $S_i$*

   - $S_i \leftarrow \{D_0 - T_i\}$
   - $h_A \leftarrow L_A(S_i)$
   - $h_B \leftarrow L_B(S_i)$
   - $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$

3. Return the value $\bar{\delta}$, where

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

# Paired $t$ test to compare $h_A$,$h_B$

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

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

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

3. Return the value $\bar{\delta}$, where

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

$N$% confidence interval estimate for $d$:

$$\bar{\delta} \pm t_{N,k-1}\ s_{\bar{\delta}}$$

$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

*Note $\delta_i$ approximately Normally distributed*

# Comparing learning algorithms $L_A$ and $L_B$

Notice we'd like to use the paired $t$ test on $\bar{\delta}$ to obtain a confidence interval

but not really correct, because the training sets in this algorithm are not independent (they overlap!)

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

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

but even this approximation is better than no comparison

# Comparing Learning Algorithms

- Which of $\boldsymbol{L_A}$ and $\boldsymbol{L_B}$ is the better learning method on average for learning some particular target function $\boldsymbol{f}$ ?
- To answer this question, we wish to estimate the expected value of the difference in their errors
- $\boldsymbol{E_{S \subset D}[error_D(L_A(S))-error_D(L_B(S))]}$
- Of course, since we have only a limited sample $D_0$ we estimate this quantity by dividing $D_0$ into a ***training set*** $\mathbf{S_0}$ and a ***testing set*** $\boldsymbol{T_0}$ and measure: $\boldsymbol{error_{T0}(L_A(S_0))-error_{T0}(L_B(S_0))}$
- **Problem:** We are only measuring the difference in errors for one training set $S_0$ rather than the expected value of this difference over all samples $S$ drawn from $D$

**Solution:** ***k-fold Cross-Validation***

# Confidence of the k-fold Estimate

- The approximate N% confidence interval for estimating $\boldsymbol{E_{S\subset D0}[error_D(L_A(S))\text{-}error_D(L_B(S))]}$ using $\boldsymbol{avg(\delta)}$**,** is given by:

$$avg(\delta) \pm t_{N,k-1}\ s_{avg(\delta)}$$

where $\boldsymbol{t_{N,k-1}}$ is a constant similar to $\boldsymbol{Z_N}$ (See [Mitchell, Table 5.6]) and $\boldsymbol{s_{avg(\delta)}}$ is an estimate of the standard deviation of the distribution governing $\boldsymbol{avg(\delta)}$

$$s_{avg(\delta))}=\surd 1/k(k-1)\ \Sigma_{i=1}^{k}(\delta_i - avg(\delta))^2$$

The approximate $N\%$ confidence interval for estimating the quantity in Equa- ion (5.16) using $\bar{\delta}$ is given by

$$\bar{\delta} \pm t_{N,k-1}\, s_{\bar{\delta}} \tag{5.17}$$

vhere $t_{N,k-1}$ is a constant that plays a role analogous to that of $z_N$ in our ear- ier confidence interval expressions, and where $s_{\bar{\delta}}$ is an estimate of the standard eviation of the distribution governing $\bar{\delta}$. In particular, $s_{\bar{\delta}}$ is defined as

$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2} \tag{5.18}$$

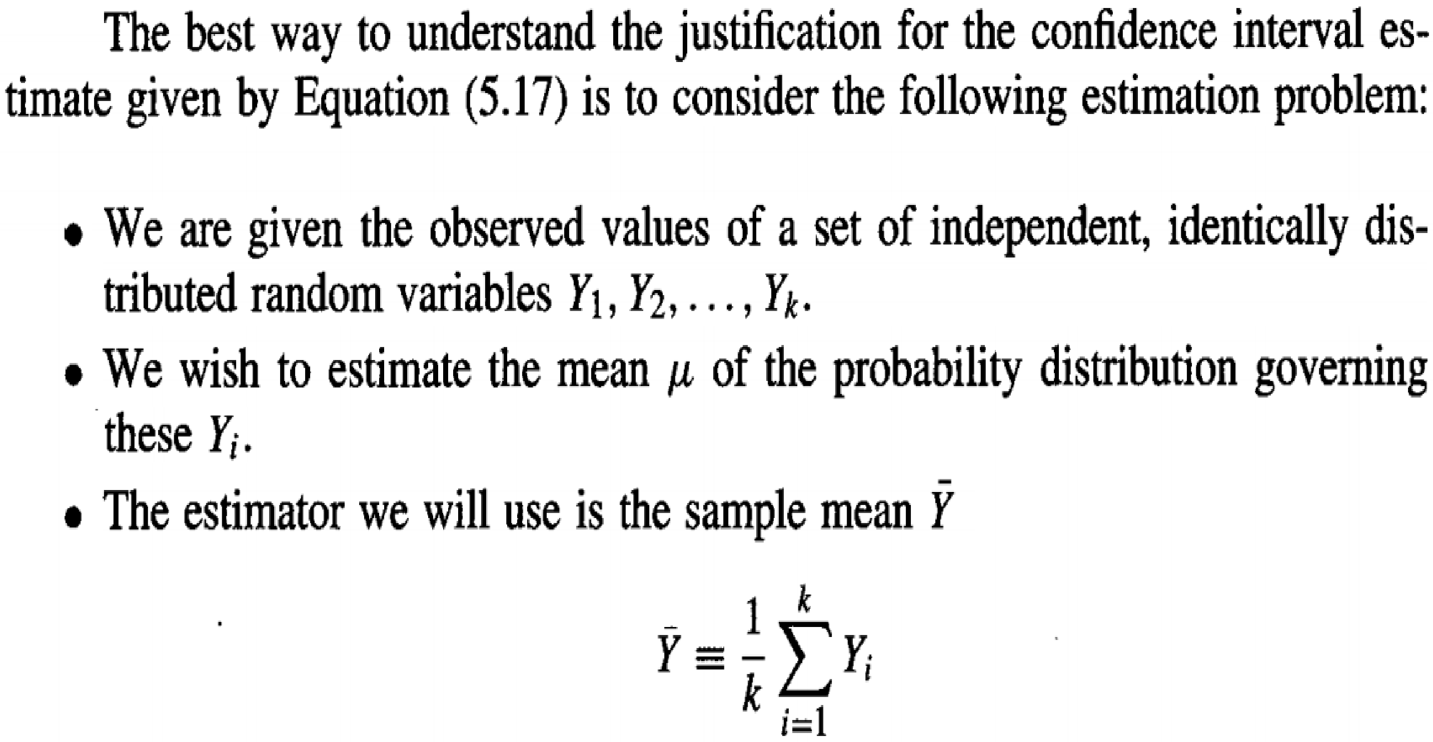
|  | Confidence level $N$ |  |  |  |
|---|---|---|---|---|
|  | 90% | 95% | 98% | 99% |
| $\nu = 2$ | 2.92 | 4.30 | 6.96 | 9.92 |
| $\nu = 5$ | 2.02 | 2.57 | 3.36 | 4.03 |
| $\nu = 10$ | 1.81 | 2.23 | 2.76 | 3.17 |
| $\nu = 20$ | 1.72 | 2.09 | 2.53 | 2.84 |
| $\nu = 30$ | 1.70 | 2.04 | 2.46 | 2.75 |
| $\nu = 120$ | 1.66 | 1.98 | 2.36 | 2.62 |
| $\nu = \infty$ | 1.64 | 1.96 | 2.33 | 2.58 |

**TABLE 5.6**

Values of $t_{N,\nu}$ for two-sided confidence intervals. As $\nu \to \infty$, $t_{N,\nu}$ approaches $z_N$.

## Paired t Tests

we described one procedure for comparing two learning methods given a fixed set of data. This section discusses the statistical justification for this procedure, and for the confidence interval defined



The best way to understand the justification for the confidence interval estimate given by Equation (5.17) is to consider the following estimation problem:

- We are given the observed values of a set of independent, identically distributed random variables $Y_1, Y_2, \ldots, Y_k$.
- We wish to estimate the mean $\mu$ of the probability distribution governing these $Y_i$.
- The estimator we will use is the sample mean $\bar{Y}$

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

This problem of estimating the distribution mean $\mu$ based on the sample mean $\bar{Y}$ is quite general. For example, it covers the problem discussed earlier of using $error_S(h)$ to estimate $error_{\mathcal{D}}(h)$. (In that problem, the $Y_i$ are 1 or 0 to indicate whether $h$ commits an error on an individual example from $S$, and $error_{\mathcal{D}}(h)$ is the mean $\mu$ of the underlying distribution.) The $t$ test, described by Equations (5.17) and (5.18), applies to a special case of this problem—the case in which the individual $Y_i$ follow a Normal distribution.

Now consider the following idealization of the method in Table 5.5 for comparing learning methods. Assume that instead of having a fixed sample of data $D_0$, we can request new training examples drawn according to the underlying instance distribution. In particular, in this idealized method we modify the procedure of Table 5.5 so that on each iteration through the loop it generates a new random training set $S_i$ and new random test set $T_i$ by drawing from this underlying instance distribution instead of drawing from the fixed sample $D_0$. This idealized method perfectly fits the form of the above estimation problem. In particular, the $\delta_i$ measured by the procedure now correspond to the independent, identically distributed random variables $Y_i$. The mean $\mu$ of their distribution corresponds to the expected difference in error between the two learning methods [i.e., Equation (5.14)]. The sample mean $\bar{Y}$ is the quantity $\bar{\delta}$ computed by this idealized version of the method. We wish to answer the question "how good an estimate of $\mu$ is provided by $\bar{\delta}$?"

First, note that the size of the test sets $T_i$ has been chosen to contain at least 30 examples. Because of this, the individual $\delta_i$ will each follow an approximately Normal distribution (due to the Central Limit Theorem). Hence, we have a special case in which the $Y_i$ are governed by an approximately Normal distribution. It can be shown in general that when the individual $Y_i$ each follow a Normal distribution, then the sample mean $\bar{Y}$ follows a Normal distribution as well. Given that $\bar{Y}$ is Normally distributed, we might consider using the earlier expression for confidence intervals (Equation [5.11]) that applies to estimators governed by Normal distributions. Unfortunately, that equation requires that we know the standard deviation of this distribution, which we do not.

The $t$ test applies to precisely these situations, in which the task is to estimate the sample mean of a collection of independent, identically and Normally distributed random variables. In this case, we can use the confidence interval given by Equations (5.17) and (5.18), which can be restated using our current notation as

$$\mu = \bar{Y} \pm t_{N,k-1}\, s_{\bar{Y}}$$

where $s_{\bar{Y}}$ is the estimated standard deviation of the sample mean

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

and where $t_{N,k-1}$ is a constant analogous to our earlier $z_N$. In fact, the constant $t_{N,k-1}$ characterizes the area under a probability distribution known as the $t$ distribution, just as the constant $z_N$ characterizes the area under a Normal distribution. The $t$ distribution is a bell-shaped distribution similar to the Normal distribution, but wider and shorter to reflect the greater variance introduced by using $s_{\bar{Y}}$ to approximate the true standard deviation $\sigma_{\bar{Y}}$. The $t$ distribution approaches the Normal distribution (and therefore $t_{N,k-1}$ approaches $z_N$) as $k$ approaches infinity. This is intuitively satisfying because we expect $s_{\bar{Y}}$ to converge toward the true standard deviation $\sigma_{\bar{Y}}$ as the sample size $k$ grows, and because we can use $z_N$ when the standard deviation is known exactly.

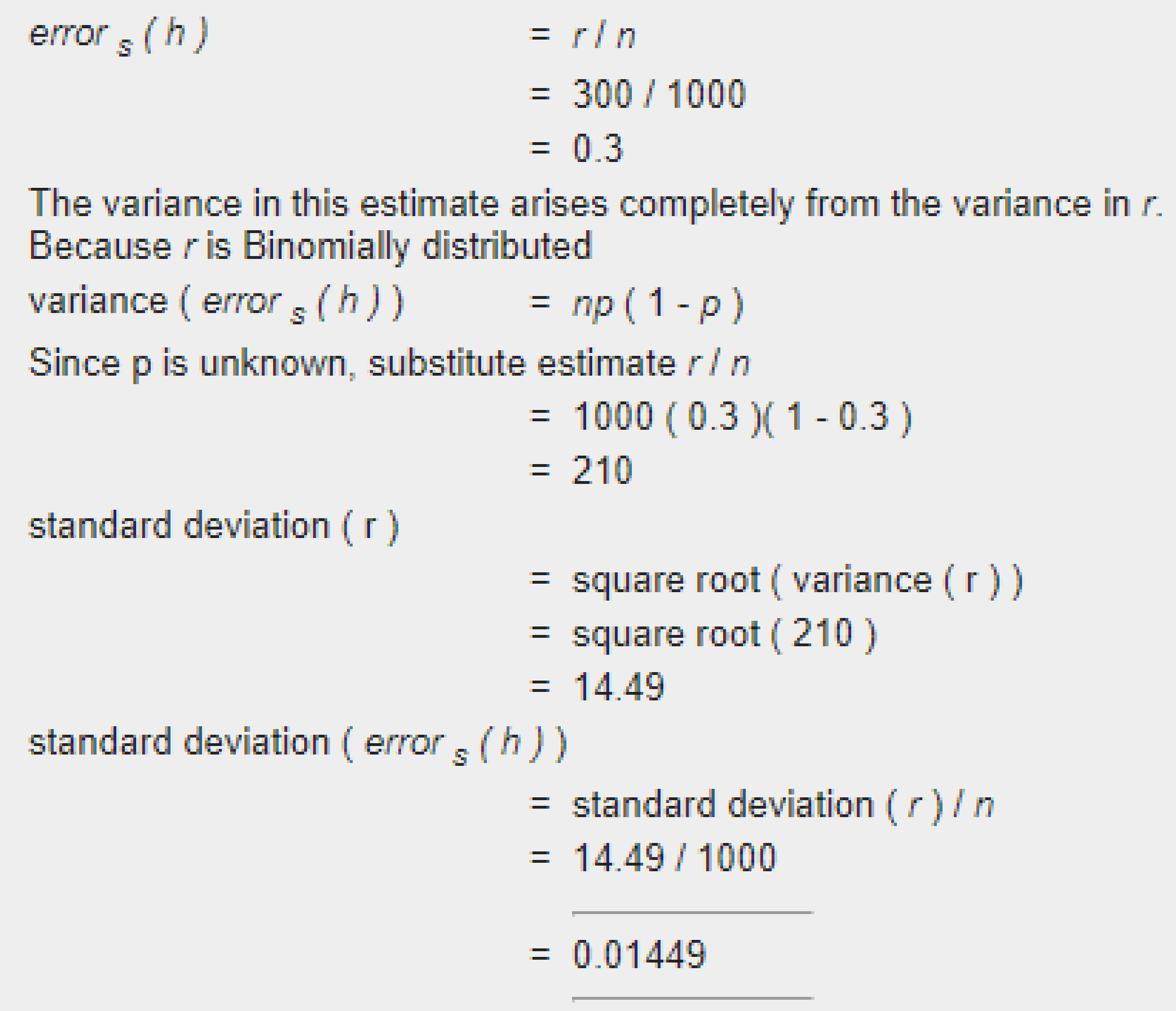
Apply the candidate elimination (CE) algorithm to the sequence of training examples specified in the table and name the contents of the sets $S$ and $G$ after each step.

| Training Example | N (running nose) | C (coughing) | R (reddened skin) | **Classification** |
|---|---|---|---|---|
| $d_1$ | $+$ | $+$ | $+$ | positive (ill) |
| $d_2$ | $+$ | $+$ | $-$ | positive (ill) |
| $d_3$ | $+$ | $-$ | $+$ | negative (healthy) |
| $d_4$ | $-$ | $+$ | $+$ | negative (healthy) |

- Start (init): $G = \{\langle * * * \rangle\}$, $S = \{\langle \emptyset \emptyset \emptyset \rangle\}$
- **foreach** $d \in D$ **do**
  - $d_1 = [\langle + + + \rangle, pos] \Rightarrow G = \{\langle * * * \rangle\}$, $S = \{\langle + + + \rangle\}$
  - $d_2 = [\langle + + - \rangle, pos] \Rightarrow G = \{\langle * * * \rangle\}$, $S = \{\langle + + * \rangle\}$
  - $d_3 = [\langle + - + \rangle, neg]$
    - no change to $S$: $S = \{\langle + + * \rangle\}$
    - specializations of $G$: $G = \{\langle - * * \rangle, \langle * + * \rangle, \langle * * - \rangle\}$
    - there is no element in $S$ that is more specific than the first and third element of $G$
      $\rightarrow$ remove them from $G \Rightarrow G = \{\langle * + * \rangle\}$

- **foreach** $d \in D$ **do**
  - *loop continued ...*
  - so far we have $S = \{\langle + + *\rangle\}$ and $G = \{\langle * + *\rangle\}$
  - $d_4 = [\langle - + +\rangle, neg]$
    - no change to $S$: $S = \{\langle + + *\rangle\}$
    - specializations of $G$: $G = \{\langle + + *\rangle, \langle * + -\rangle\}$
    - there is no element in $S$ that is more specific than the second element of $G$
      $\rightarrow$ remove it from $G \Rightarrow G = \{\langle + + *\rangle\}$
  - Note:
    - At this point, the algorithm might be stopped, since $S = G$ and no further changes to $S$ and $G$ are to be expected.
    - However, by continuing we might detect inconsistencies in the training data.

Suppose you test a hypothesis *h* and find that it commits *r* = 300 errors on a sample *S* of *n* = 1000 randomly drawn test examples. What is the standard deviation in $error_s(h)$?



$error_s(h)$ = $r / n$

= 300 / 1000

= 0.3

The variance in this estimate arises completely from the variance in *r*. Because *r* is Binomially distributed

variance ( $error_s(h)$ ) = $np(1 - p)$

Since p is unknown, substitute estimate $r / n$

= 1000 ( 0.3 )( 1 - 0.3 )

= 210

standard deviation ( r )

= square root ( variance ( r ) )

= square root ( 210 )

= 14.49

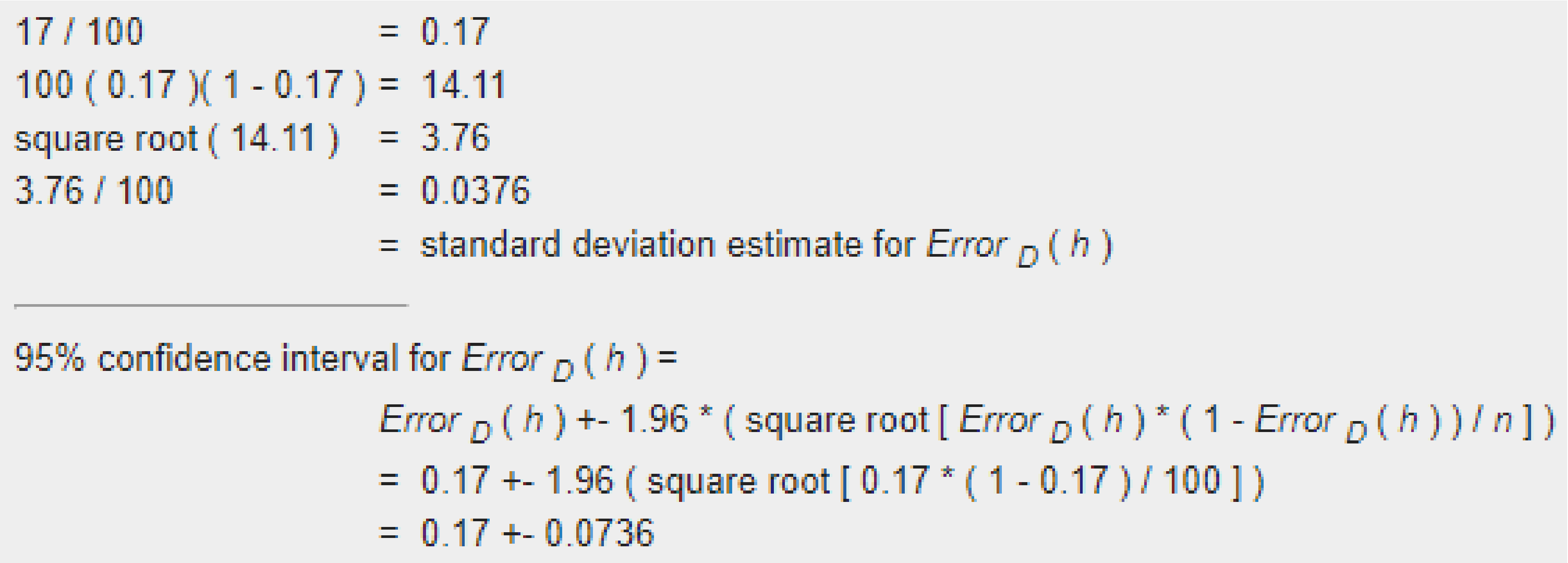
standard deviation ( $error_s(h)$ )

= standard deviation ( *r* ) / *n*

= 14.49 / 1000

= 0.01449

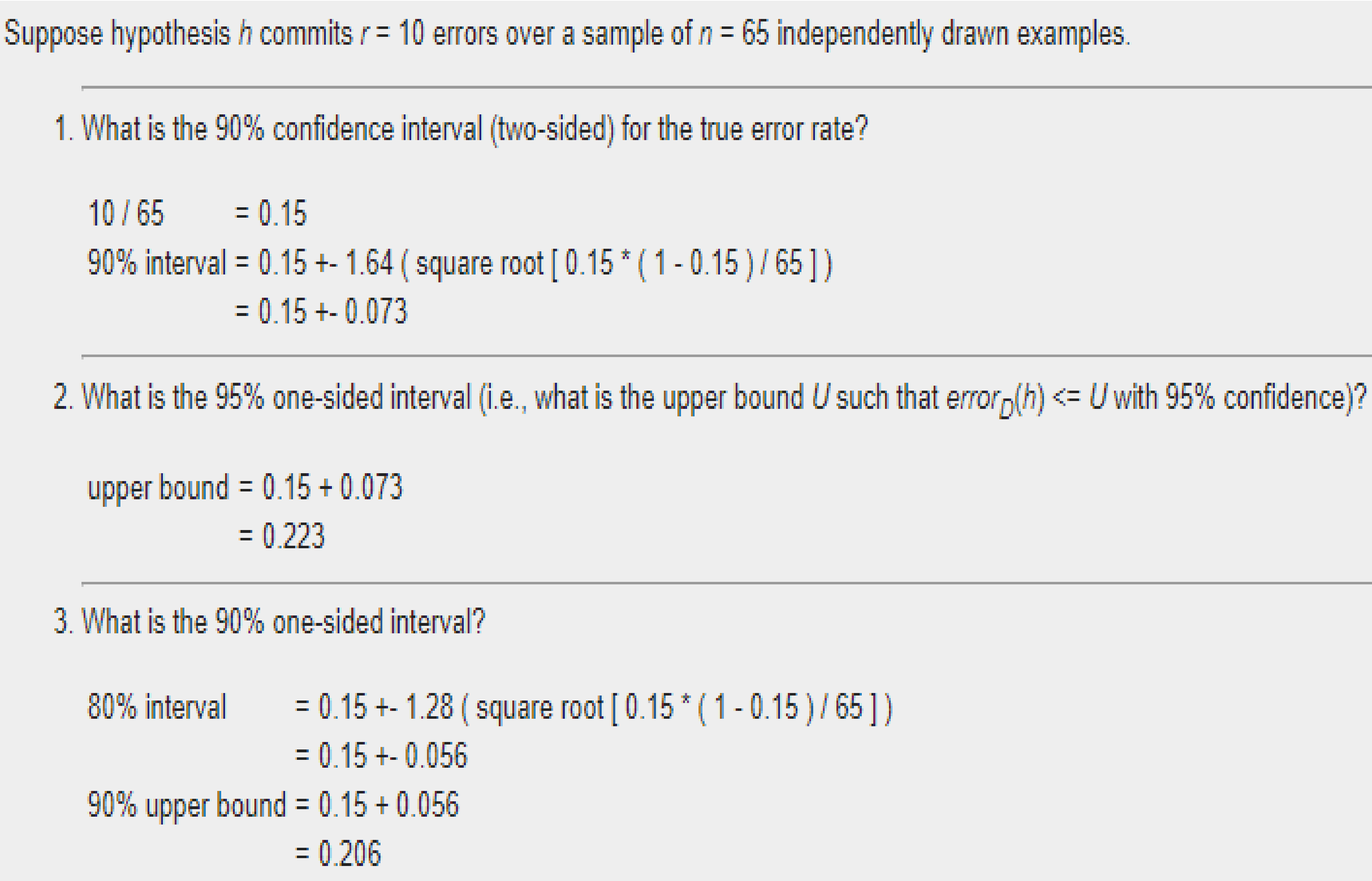
Consider a learned hypothesis, *h*, for some boolean concept. When *h* is tested on a set of 100 examples, it classifies 83 correctly. What is the standard deviation and the 95% confidence interval for the true error rate for $Error_D(h)$?



17 / 100 = 0.17
100 ( 0.17 )( 1 - 0.17 ) = 14.11
square root ( 14.11 ) = 3.76
3.76 / 100 = 0.0376
= standard deviation estimate for $Error_D(h)$

---

95% confidence interval for $Error_D(h)$ =
$Error_D(h)$ +- 1.96 * ( square root [ $Error_D(h)$ * ( 1 - $Error_D(h)$ ) / $n$ ] )
= 0.17 +- 1.96 ( square root [ 0.17 * ( 1 - 0.17 ) / 100 ] )
= 0.17 +- 0.0736



Suppose hypothesis $h$ commits $r = 10$ errors over a sample of $n = 65$ independently drawn examples.

1. What is the 90% confidence interval (two-sided) for the true error rate?

   10 / 65 = 0.15
   90% interval = 0.15 +- 1.64 ( square root [ 0.15 * ( 1 - 0.15 ) / 65 ] )
   = 0.15 +- 0.073

2. What is the 95% one-sided interval (i.e., what is the upper bound $U$ such that $error_D(h)$ <= $U$ with 95% confidence)?

   upper bound = 0.15 + 0.073
   = 0.223

3. What is the 90% one-sided interval?

   80% interval = 0.15 +- 1.28 ( square root [ 0.15 * ( 1 - 0.15 ) / 65 ] )
   = 0.15 +- 0.056
   90% upper bound = 0.15 + 0.056
   = 0.206

You are about to test a hypothesis $h$ whose true error rate $error_{\mathcal{D}}(h)$ is known to be in the range between 0.2 and 0.4. Table listed below should help you.

1. What is the minimum number of examples ($n$) you must collect to assure that the width of the two-sided 95% confidence interval will be smaller than 0.1?
2. What will be the change in the solution for $n$ if the range is between 0.2 and 0.6 instead of 0.2 and 0.4?

| $\alpha$ | 0.100 | 0.050 | 0.025 | 0.001 |
|---|---|---|---|---|
| $1-\alpha$ | 0.900 | 0.950 | 0.975 | 0.999 |
| $z_{1-\alpha}$ | 1.28 | 1.64 | 1.96 | 3.09 |

Table 1: $z$ is a value of quantile function of standard normal distribution for a random variable $X$; e.g. $Pr(X > 3.09) = 0.001$.

1. We know the range of the confidence interval ($L = 0,2$, $U = 0,4$), so we know the midpoint $M = 0,3$. The midpoint $M$ is an estimation of error rate. Hence $p = 0,3$ and $(1-p) = 0,7$. Formulas:

$$M = p$$

$$L = p - z\sqrt{\frac{p(1-p)}{n}}$$

$$U = p + z\sqrt{\frac{p(1-p)}{n}}$$

   Confidence interval width is $U - L = 0,1$.

$$U - L = 2z\sqrt{\frac{p(1-p)}{n}}$$

   95% two-sided confidence interval means that we have to use $z_{0,975} = 1,96$.

$$n_1 \geq \frac{4z_{0,975}^2 p(1-p)}{(U-L)^2} = \frac{4.1,96^2.0,3.0,7}{0,1^2}$$

$$n_1 \geq 322,7$$

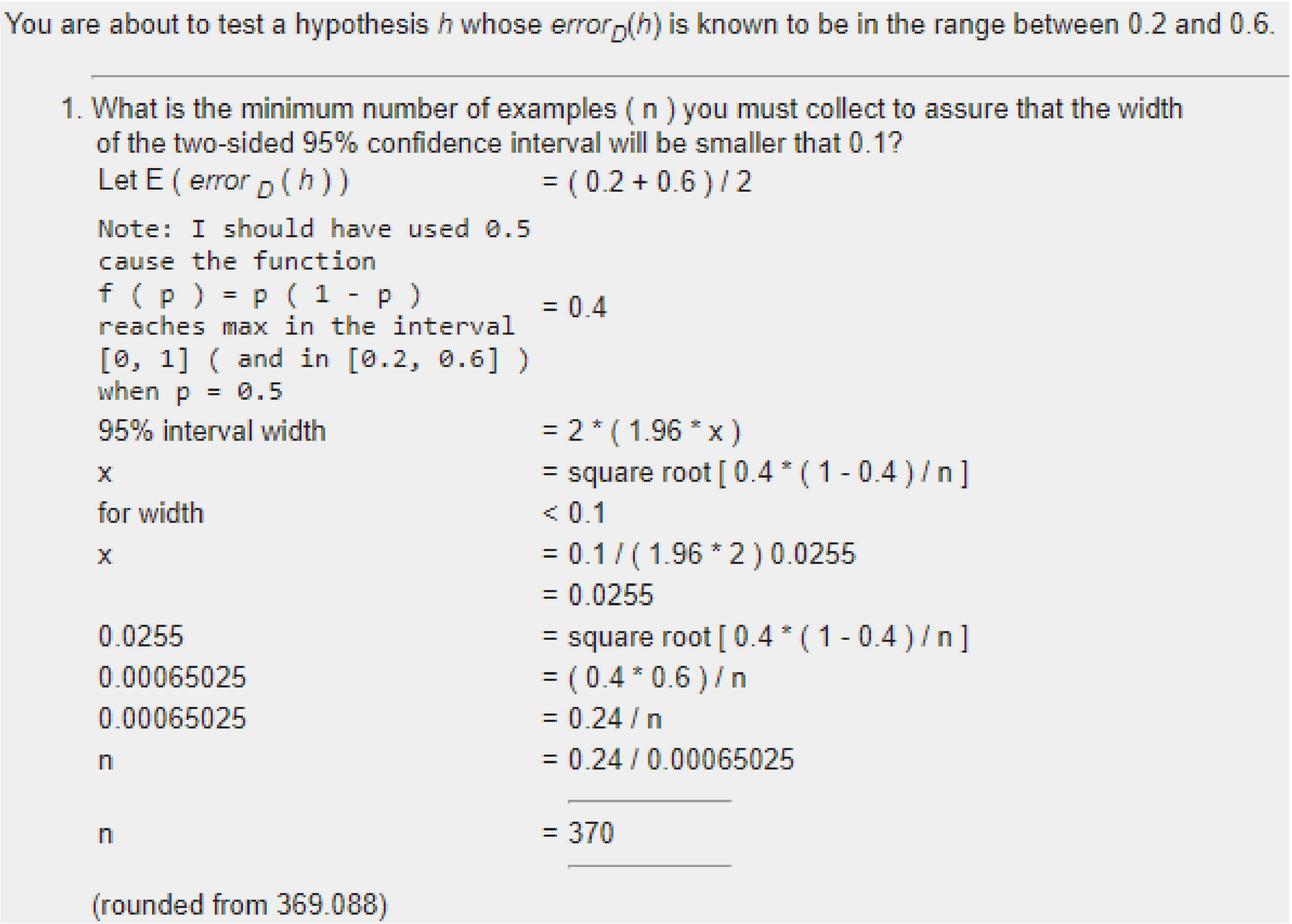
$$n_1 = 323$$

2. Midpoint $M$ is shifted. Hence $p = 0,4$ and $(1-p) = 0,6$.

$$n_2 \geq \frac{4.1,96^2.0,4.0,6}{0,1^2}$$

$$n_2 \geq 368,8$$

$$n_2 = 369$$



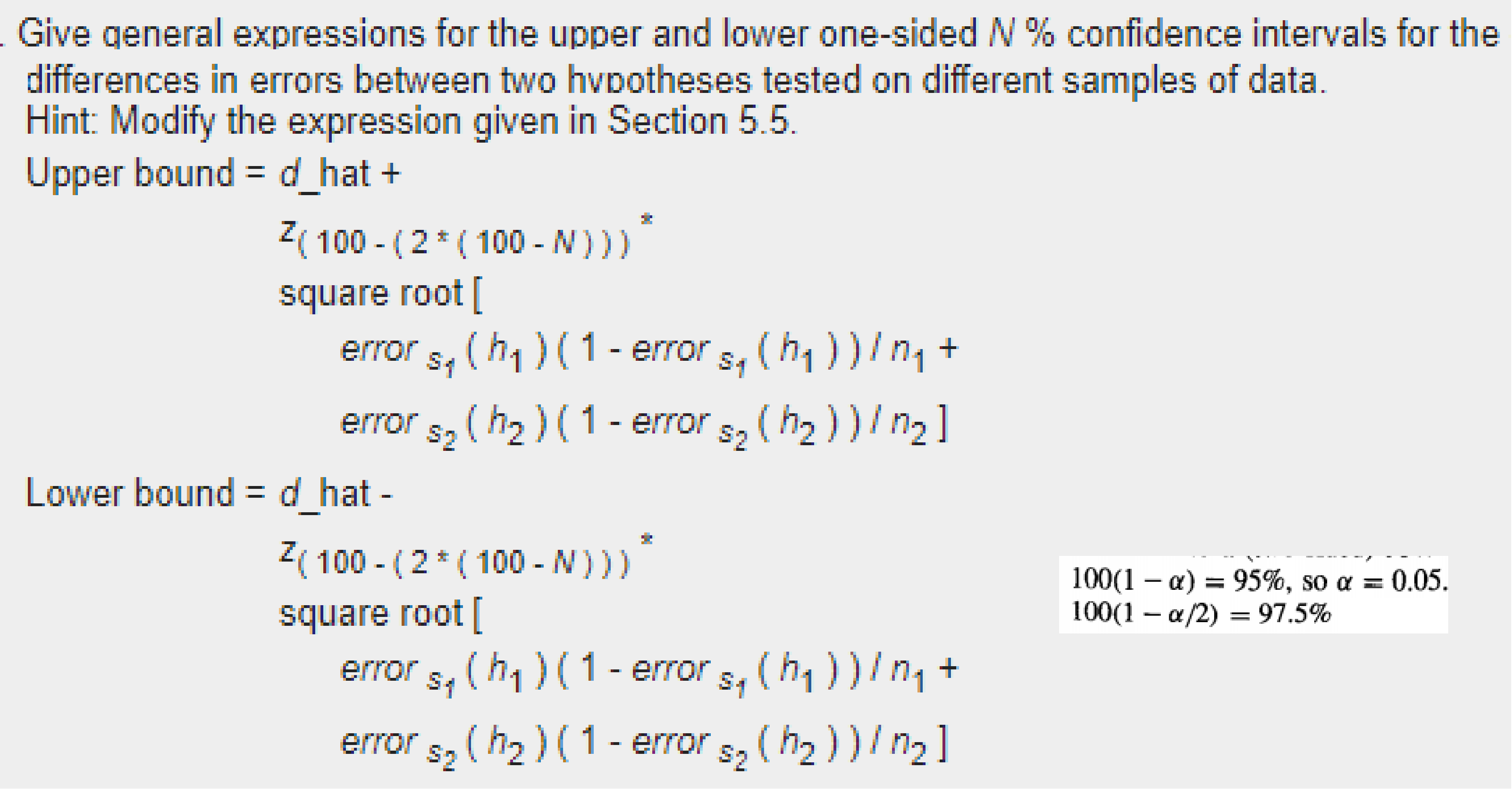
You are about to test a hypothesis *h* whose *error*$_D$(*h*) is known to be in the range between 0.2 and 0.6.

---

1. What is the minimum number of examples ( n ) you must collect to assure that the width of the two-sided 95% confidence interval will be smaller that 0.1?

| | |
|---|---|
| Let E ( *error* $_D$ ( *h* ) ) | = ( 0.2 + 0.6 ) / 2 |
| Note: I should have used 0.5<br>cause the function<br>f ( p ) = p ( 1 - p )<br>reaches max in the interval<br>[0, 1] ( and in [0.2, 0.6] )<br>when p = 0.5 | = 0.4 |
| 95% interval width | = 2 * ( 1.96 * x ) |
| x | = square root [ 0.4 * ( 1 - 0.4 ) / n ] |
| for width | < 0.1 |
| x | = 0.1 / ( 1.96 * 2 ) 0.0255 |
| | = 0.0255 |
| 0.0255 | = square root [ 0.4 * ( 1 - 0.4 ) / n ] |
| 0.00065025 | = ( 0.4 * 0.6 ) / n |
| 0.00065025 | = 0.24 / n |
| n | = 0.24 / 0.00065025 |
| n | = 370 |

(rounded from 369.088)



. Give general expressions for the upper and lower one-sided *N* % confidence intervals for the differences in errors between two hypotheses tested on different samples of data.
Hint: Modify the expression given in Section 5.5.

Upper bound = *d*_hat +

$z_{(100 - (2 * (100 - N)))}$ *

square root [

$error_{s_1}(h_1)(1 - error_{s_1}(h_1)) / n_1$ +

$error_{s_2}(h_2)(1 - error_{s_2}(h_2)) / n_2$ ]

Lower bound = *d*_hat -

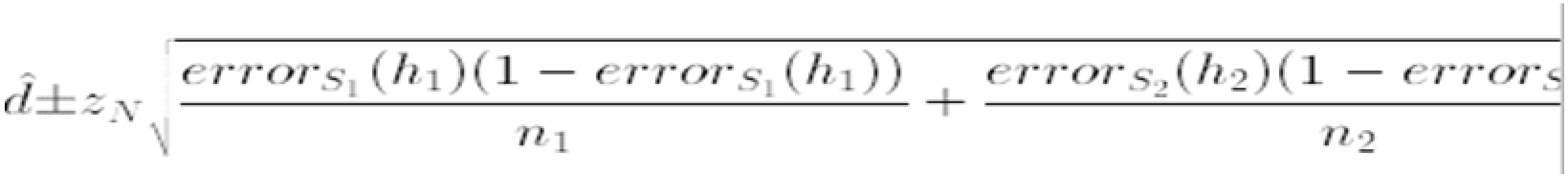
$z_{(100 - (2 * (100 - N)))}$ *

square root [

$error_{s_1}(h_1)(1 - error_{s_1}(h_1)) / n_1$ +

$error_{s_2}(h_2)(1 - error_{s_2}(h_2)) / n_2$ ]

$100(1-\alpha) = 95\%$, so $\alpha = 0.05$.
$100(1-\alpha/2) = 97.5\%$



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

Explain why the confidence interval estimate given in Equation (5.17) applies to estimating the quantity in Equation (5.16), and not the quantity in Equation (5.14).

This is so because the set of available data in Eq. (5.16) to take sub sets from is limited. The data available in Eq. (5.14) include all possible examples from the distribution. *k* in Eq. (5.17) is the number of sub sets taken from the available distribution and is used to determine the degrees of freedom in estimating the quantity in Eq. (5.16). As k approaches infinity (or our available data becomes closer to the full set), the value of $t_{N, k-1}$ approaches that of $z_N$, maybe…