Machine Learning

Chapter 05

Evaluating Hypotheses

Evaluating Hypotheses

- Statistical methods are used for estimating hypothesis accuracy
- First, given the observed accuracy of a hypothesis over a limited sample of data, how well does this estimate its accuracy over additional examples?
- Second, given that one hypothesis outperforms another over some sample of data, how probable is it that this hypothesis is more accurate in general?
- Third, when data is limited what is the best way to use this data to both learn a hypothesis and estimate its accuracy?

Evaluating Hypotheses

- MOTIVATION: when we must learn a hypothesis and estimate its future accuracy given only a limited set of data, two key difficulties arise:
- **Bias in the estimate**. First, the observed accuracy of the learned hypothesis over the training examples is often a poor estimator of its accuracy over future examples. Because the learned hypothesis was derived from these examples, they will typically provide an biased estimate of hypothesis accuracy over future examples.
- **Variance in the estimate**. Second, 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.

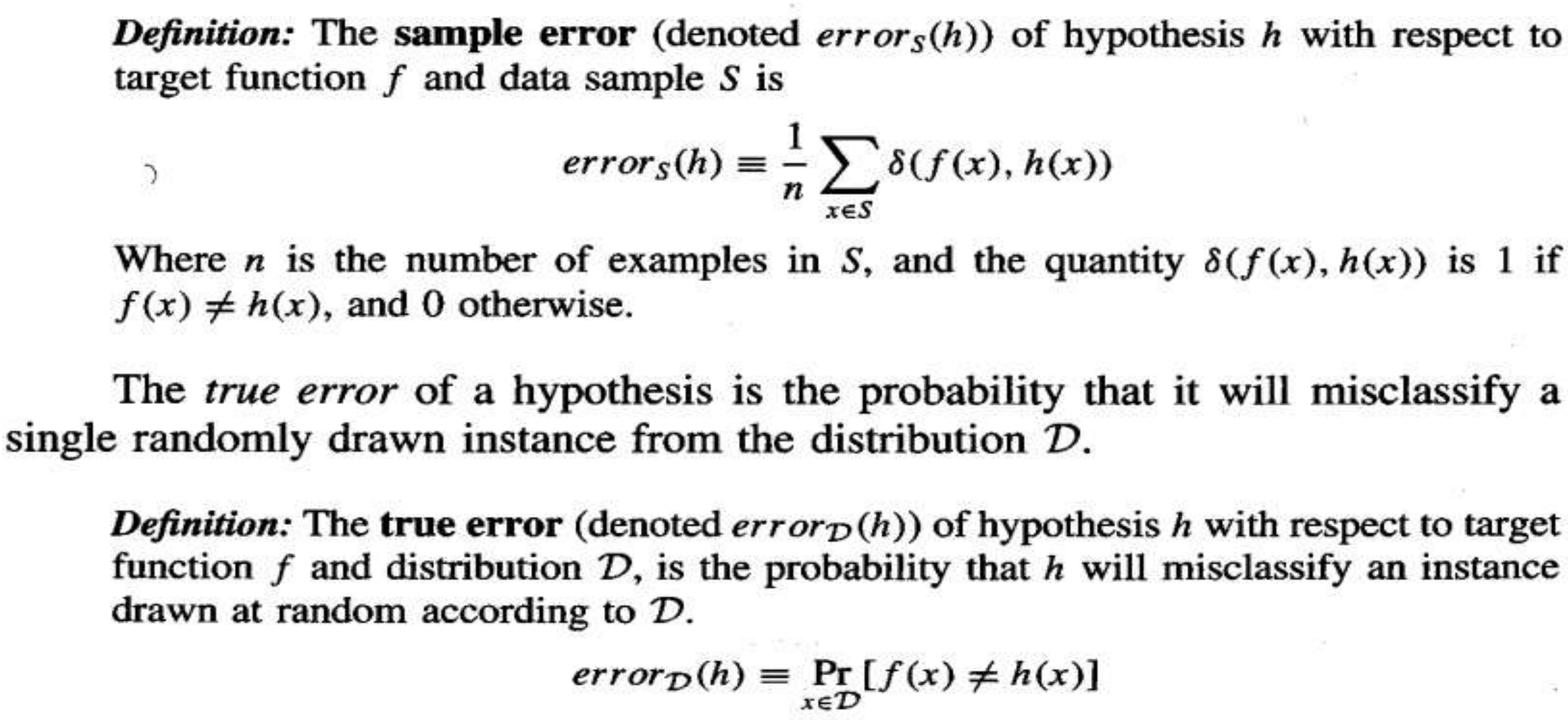
ESTIMATING HYPOTHESIS ACCURACY

- When evaluating a learned hypothesis we are interested in estimating the accuracy with which it will classify future instances.
- And also to know the probable error in the accuracy estimate.
- There is some space of possible instances X. We assume that different instances in X may be encountered with different frequencies. A convenient way to model this is to assume there is some unknown probability distribution D that defines the probability of encountering each instance in X.
- Training examples of the target function f are provided to the learner by a trainer who draws each instance independently, according to the distribution D, and who then forwards the instance x along with its correct target value f (x) to the learner.

- We are interested in the following two questions:
  - 1. Given a hypothesis h and a data sample containing n examples drawn at random according to the distribution D, what is the best estimate of the accuracy of h over future instances drawn from the same distribution?
  - 2. What is the probable error in this accuracy estimate?

Sample Error and True Error

- We need to distinguish between two notions of accuracy or, equivalently, error. One is the error rate of the hypothesis over the sample of data that is available. The other is the error rate of the hypothesis over the entire unknown distribution D of examples. We will call these the **sample error** and the **true error** respectively.
- The sample error of a hypothesis with respect to some sample S of instances drawn from X is the fraction of S that it misclassifies:



Definition: The **sample error** (denoted $error_S(h)$) of hypothesis h with respect to target function f and data sample S is

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

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.

The *true error* of a hypothesis is the probability that it will misclassify a single randomly drawn instance from the distribution $\mathcal{D}$.

Definition: The **true error** (denoted $error_{\mathcal{D}}(h)$) 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 $\mathcal{D}$.

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

Confidence Intervals for Discrete-Valued Hypotheses

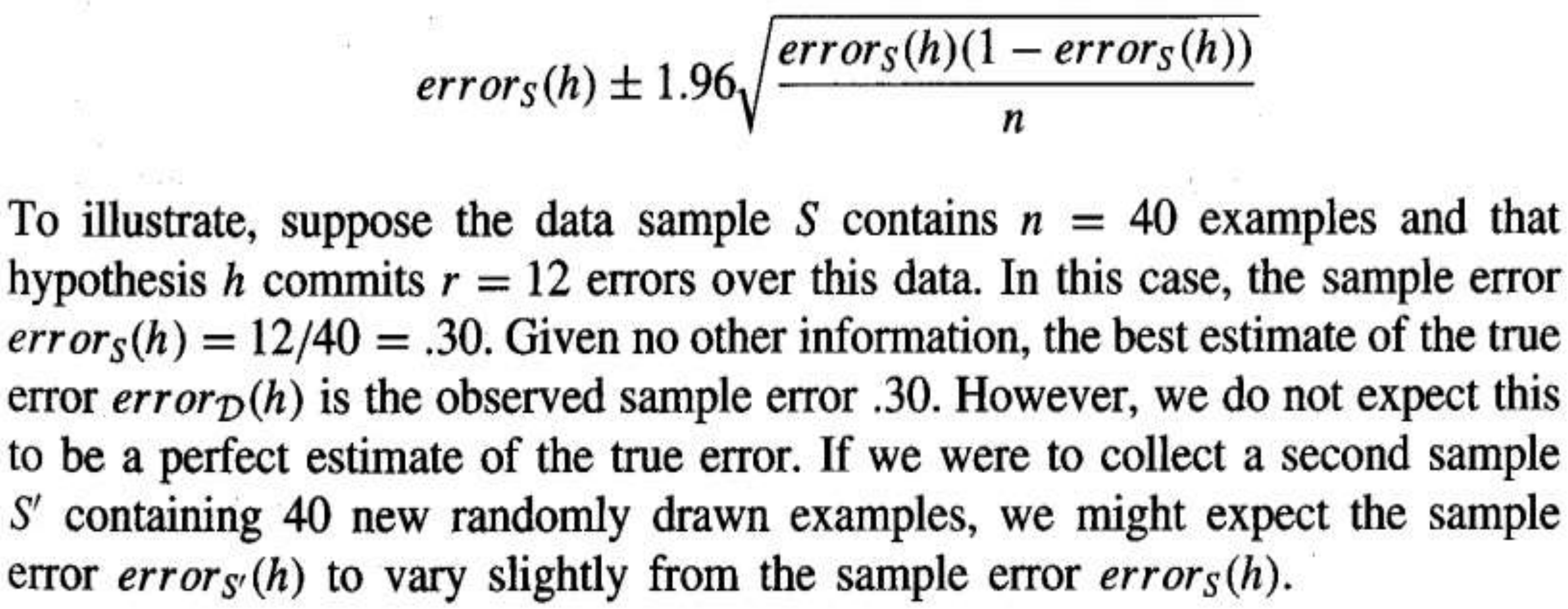
- "How good an estimate of $error_D(h)$ is provided by $error_s(h)$?"

  - for the case in which h is a discrete-valued hypothesis.

- 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 >= 30
  - hypothesis h commits r errors over these n examples $error_s(h) = r/n$

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

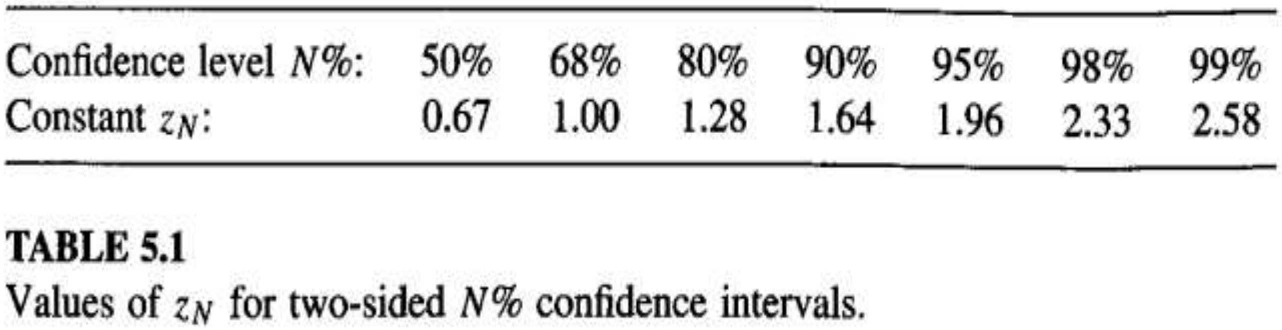
1. Given no other information, the most probable value of $\text{error}_D(h)$ is $\text{error}_s(h)$.
2. With approximately 95% probability, the true error $\text{error}_D(h)$ lies in the interval.



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

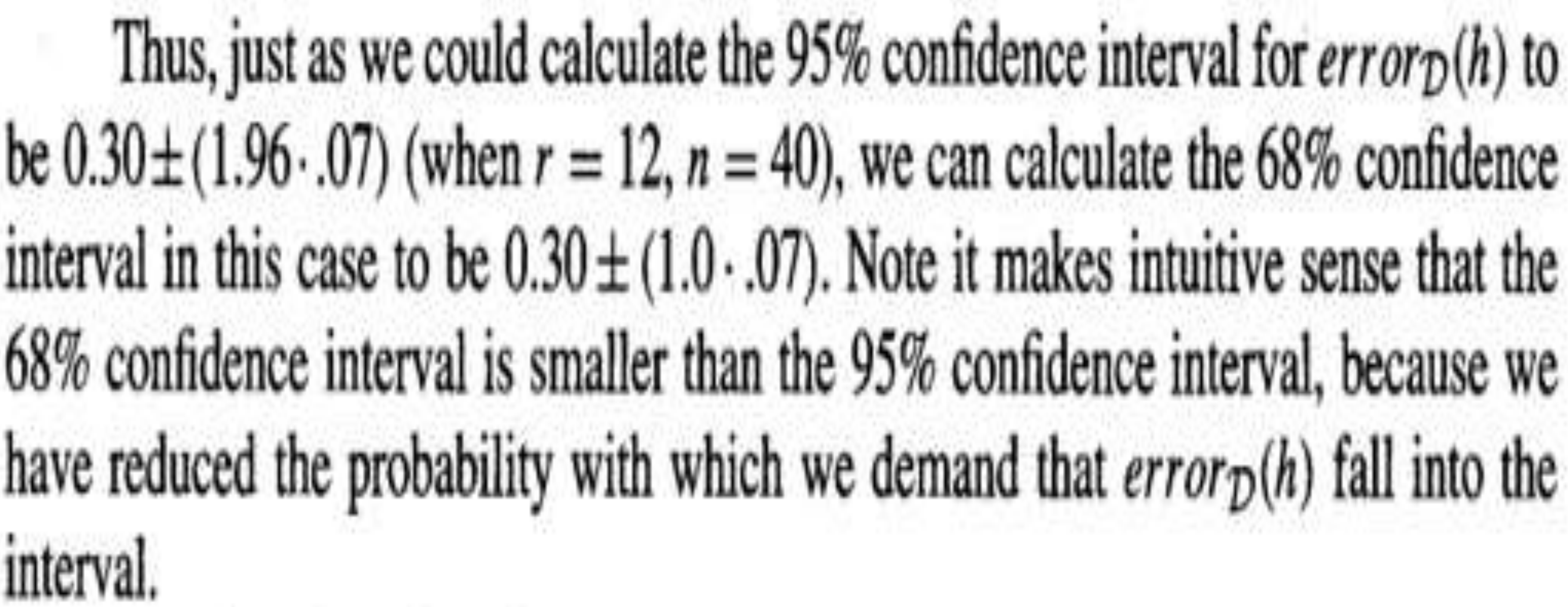
To illustrate, suppose the data sample S contains $n = 40$ examples and that hypothesis h commits $r = 12$ errors over this data. In this case, the sample error $error_S(h) = 12/40 = .30$. Given no other information, the best estimate of the true error $error_{\mathcal{D}}(h)$ is the observed sample error .30. However, we do not expect this to be a perfect estimate of the true error. If we were to collect a second sample S' containing 40 new randomly drawn examples, we might expect the sample error $error_{S'}(h)$ to vary slightly from the sample error $error_S(h)$.

- If we repeat this experiment over and over, each time drawing a new sample Si containing 40 new examples, we would find that for approximately 95% of these experiments, the calculated interval would contain the true error.
- For this reason, we call this interval the 95% confidence interval estimate for $\text{error}_D(h)$. In the current example, where r = 12 and n = 40, the 95% confidence interval is, according to the above expression, $0.30 \pm (1.96 \cdot .07) = 0.30 \pm .14.$



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

TABLE 5.1
Values of z_N for two-sided $N\%$ confidence intervals.



Thus, just as we could calculate the 95% confidence interval for $error_{\mathcal{D}}(h)$ to be $0.30 \pm (1.96 \cdot .07)$ (when $r = 12$, $n = 40$), we can calculate the 68% confidence interval in this case to be $0.30 \pm (1.0 \cdot .07)$. Note it makes intuitive sense that the 68% confidence interval is smaller than the 95% confidence interval, because we have reduced the probability with which we demand that $error_{\mathcal{D}}(h)$ fall into the interval.

- A more accurate rule of thumb is that the above approximation works well when

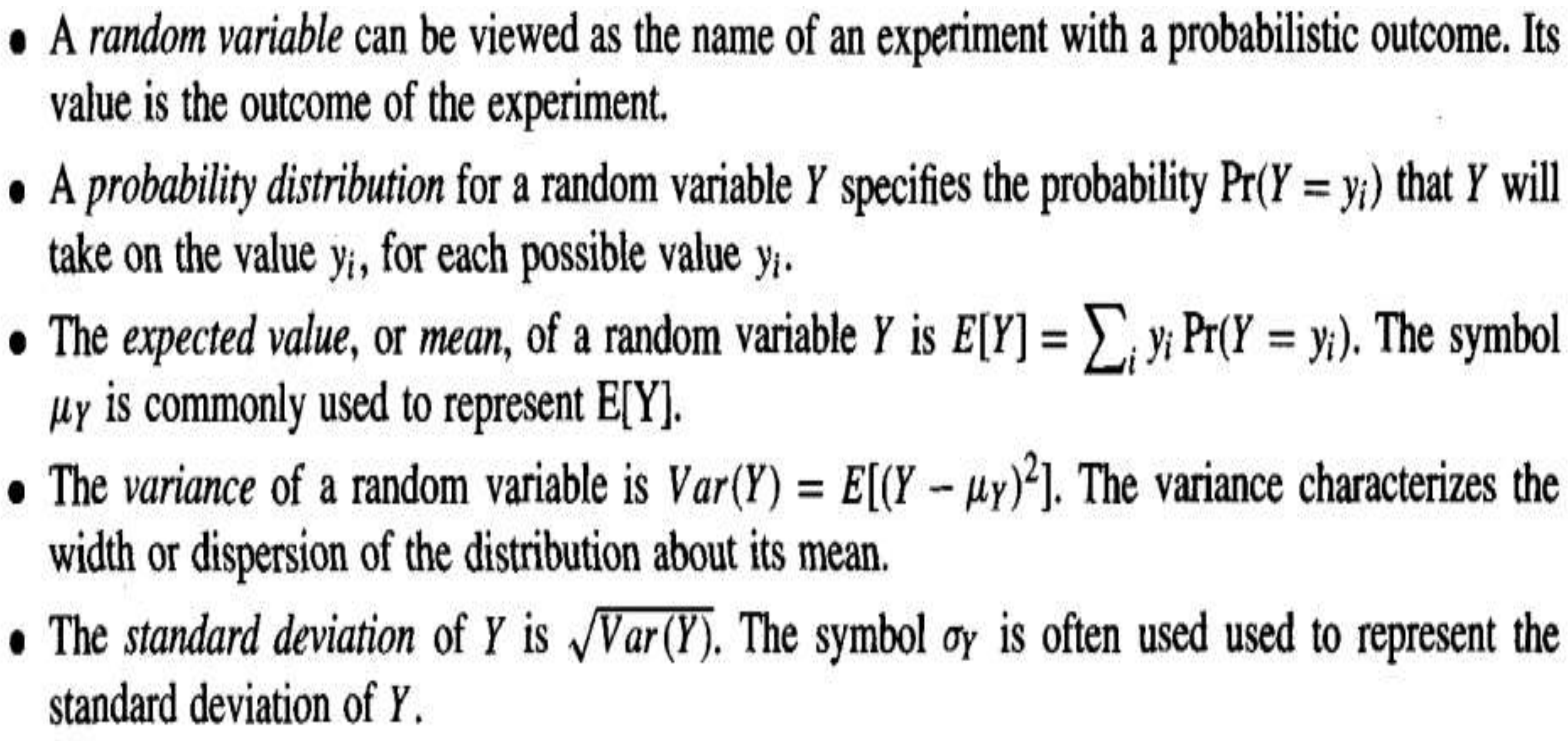
$$n\ error_S(h)(1 - error_S(h)) \geq 5$$

- Above we summarized the procedure for calculating confidence intervals for discrete-valued hypotheses.

- The following section presents the underlying statistical justification for this procedure.

BASICS OF SAMPLING THEORY

Basic notions from statistics and sampling theory



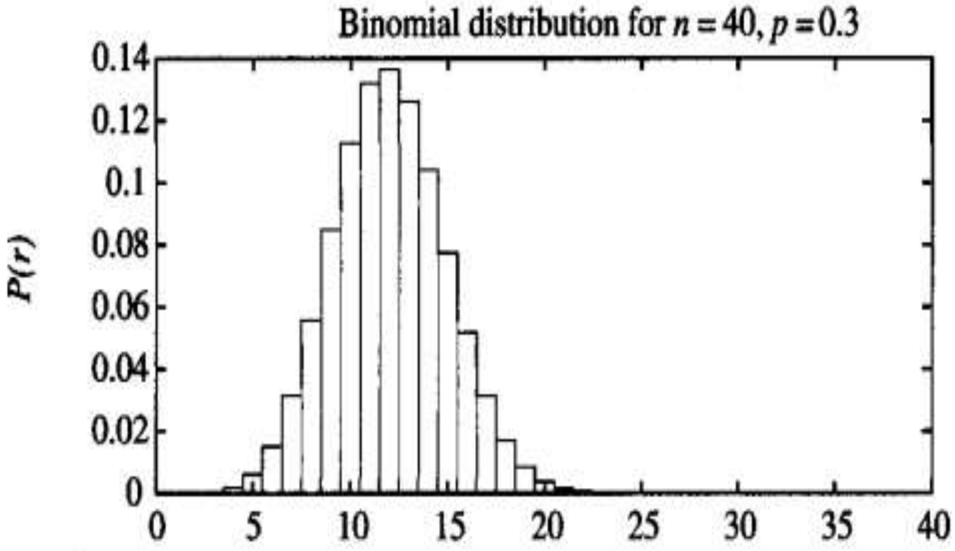
- A *random variable* can be viewed as the name of an experiment with a probabilistic outcome. Its value is the outcome of the experiment.
- A *probability distribution* for a random variable Y specifies the probability $\Pr(Y = y_i)$ that Y will take on the value y_i, for each possible value y_i.
- The *expected value*, or *mean*, of a random variable Y is $E[Y] = \sum_i y_i \Pr(Y = y_i)$. The symbol μ_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 σ_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%.

Error Estimation and Estimating Binomial Proportions

- We first collect a random sample S of n independently drawn instances from the distribution D, and then measure the sample error errors(h).
- If we were to repeat this experiment many times, each time drawing a different random sample Si of size n, we would expect to observe different values for the various errorsi(h), depending on random differences in the makeup of the various Si.
- We say in such cases that errorsi(h), the outcome of the ith such experiment, is a random variable.
- The value of the random variable is the observed outcome of the random experiment.

- Imagine that we were to run k such random experiments, measuring the random variables $error_{s1}(h)$, $error_{s2}(h)$ . . . $error_{sk}(h)$.
- Imagine further that we then plotted a histogram displaying the frequency with which we observed each possible error value.
- As we allowed k to grow, the histogram would approach the form of the distribution shown in Table 5.3.
- This table describes a particular probability distribution called the Binomial distribution.



A *Binomial distribution* gives the probability of observing r heads in a sample of n independent coin tosses, when the probability of heads on a single coin toss is p. It is defined by the probability function

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

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 $P(r)$
- The expected, or mean value of X, $E[X]$, is

$$E[X] = np$$

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

$$Var(X) = np(1-p)$$

- The standard deviation of X, σ_X, is

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

For sufficiently large values of n the Binomial distribution is closely approximated by a Normal distribution (see Table 5.4) with the same mean and variance. Most statisticians recommend using the Normal approximation only when $np(1-p) \geq 5$.

TABLE 5.3
The Binomial distribution.

The Binomial Distribution

- given a worn and bent coin and asked to estimate the probability that the coin will turn up heads when tossed. Let us call this unknown probability of heads p. You toss the coin n times and record the number of times r that it turns up heads. A estimate of p is 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, or underlying, experiment (e.g., toss of the coin) whose outcome can be described by a random variable, say 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 underlying 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 Y1, Y2, . . . , Yn. Let R denote the number of trials for which Yi = 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

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

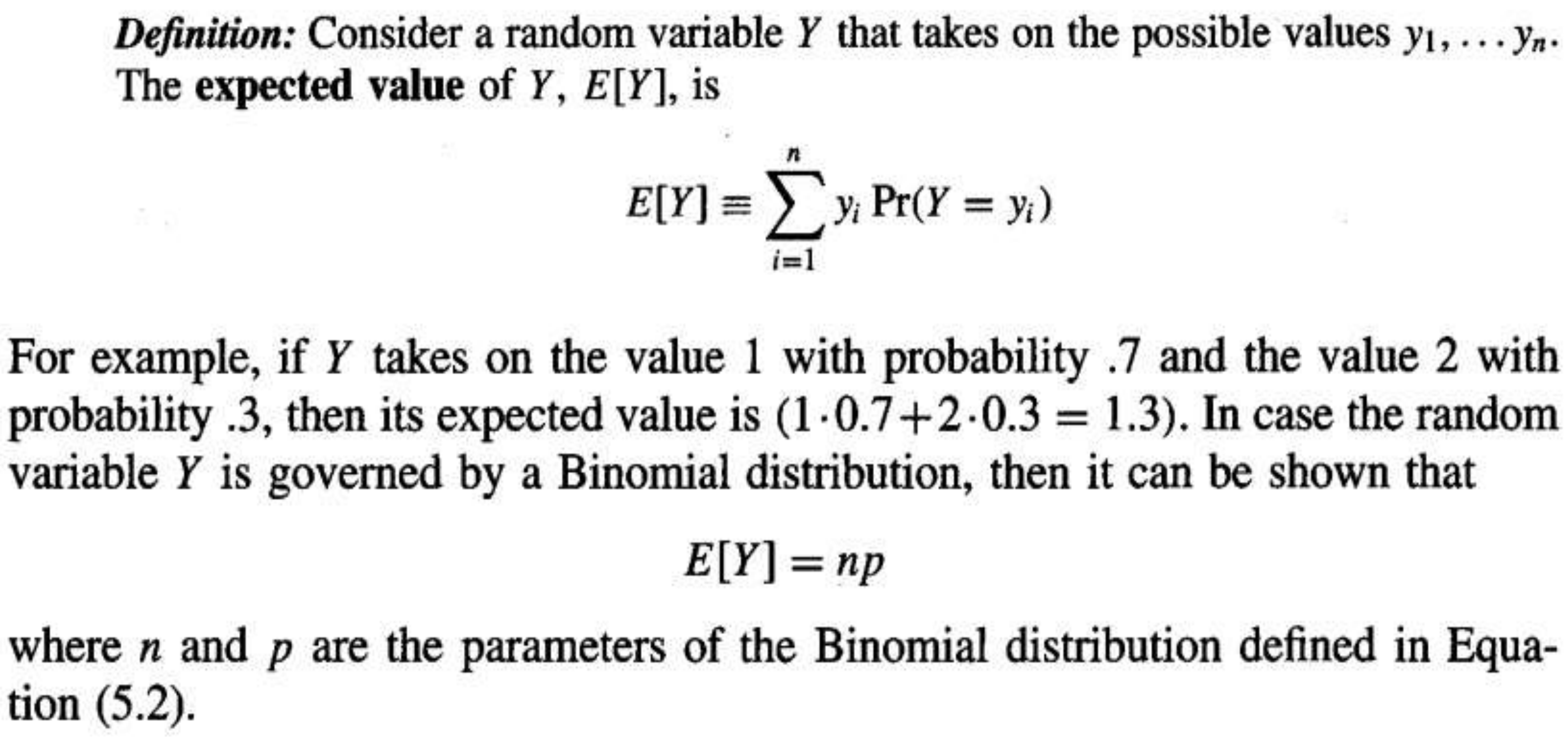
A plot of this probability distribution is shown in

Table 5.3.

The Binomial distribution characterizes the probability of observing r heads from n coin flip experiments, as well as the probability of observing r errors in a data sample containing n randomly drawn instances

Mean and Variance

- Two properties of a random variable that are often of interest are its expected value (also called its mean value) and its variance. The 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** of Y, $E[Y]$, is

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

For example, if Y takes on the value 1 with probability .7 and the value 2 with probability .3, then its expected value is $(1 \cdot 0.7 + 2 \cdot 0.3 = 1.3)$. In case the random variable Y is governed by a Binomial distribution, then it can be shown that

$$E[Y] = np$$

where n and p are the parameters of the Binomial distribution defined in Equation (5.2).

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 σ_Y.

Definition: The **standard deviation** of a random variable Y, σ_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 variance and standard deviation are given by

$$\begin{aligned} Var[Y] &= np(1-p) \\ \sigma_Y &= \sqrt{np(1-p)} \end{aligned}$$

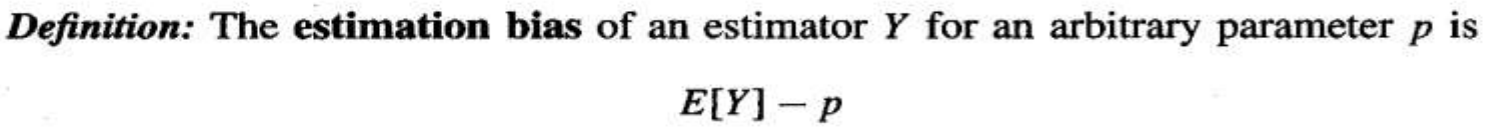
Estimators, Bias, and Variance

- Now we have shown that the random variable $error_S(h)$ obeys a Binomial distribution, What is the likely difference between $error_S(h)$ and the true error $error_D(h)$?
- Let us describe $error_S(h)$ and $error_D(h)$ using the terms in probability Distribution defining the Binomial distribution. We then have

$$error_S(h) = \frac{r}{n}$$

$$error_{\mathcal{D}}(h) = p$$

- where n is the number of instances in the sample S, r is the number of instances from S misclassified by h, and p is the probability of misclassifying a single instance drawn from D.
- We define the estimation bias to be the difference between the expected value of the estimator and the true value of the parameter.

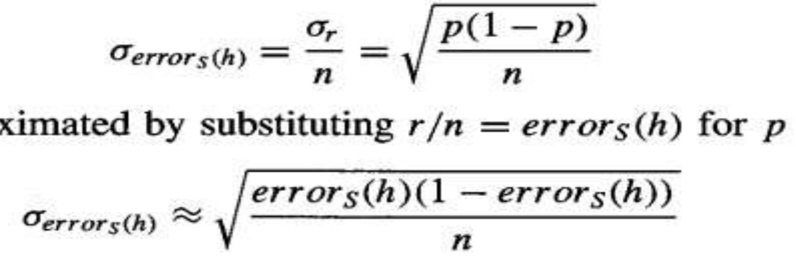


Definition: The **estimation bias** of an estimator Y for an arbitrary parameter p is

$$E[Y] - p$$

- If the estimation bias is zero, we say that Y is an unbiased estimator for p. 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.

- 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$_D$(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 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*0.3(1 - 0.3) = 8.4, or a corresponding standard deviation of √8.4=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.
- In general, given r errors in a sample of n independently drawn test examples, the standard deviation for error$_S$(h) is given by



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

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

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

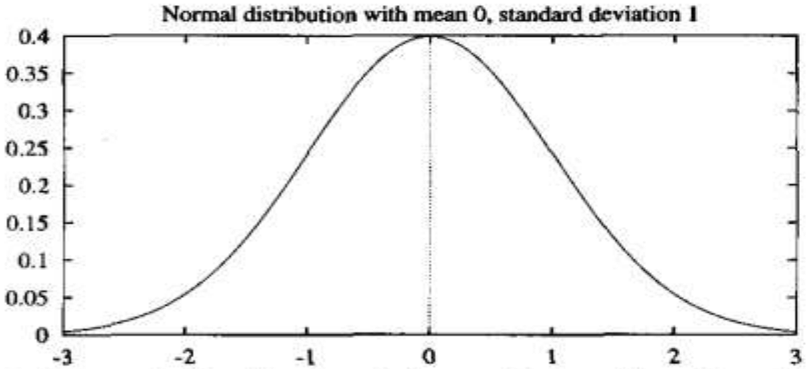
Eqn. (5.9)

Confidence Intervals

- 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 confidence interval estimates.**
- **Definition**: An N% confidence interval for some parameter p is an interval that is expected with probability N% to contain p.
- For example, if we observe r = 12 errors in a sample of n = 40 independently drawn examples, we can say with approximately 95% probability that the interval 0.30 ±0.14 contains the true error errorD(h).

- How can we derive confidence intervals for errorD(h)?
- we know the Binomial probability distribution governing the estimator errors(h).
- The mean value of this distribution is errorD(h), and the standard deviation is given by Equation (5.9).
- to derive a 95% confidence interval, we need to only find the interval centered around the mean value errorD(h), which is wide enough to contain 95% of the total probability under this distribution.
- This provides an interval surrounding errorD(h) into which errors(h) must fall 95% of the time.

- for the Binomial distribution this calculation are quite tedious.
- based on the fact that for large sample sizes the Binomial distribution can be approximated by the Normal distribution.



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: μ and σ.

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

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

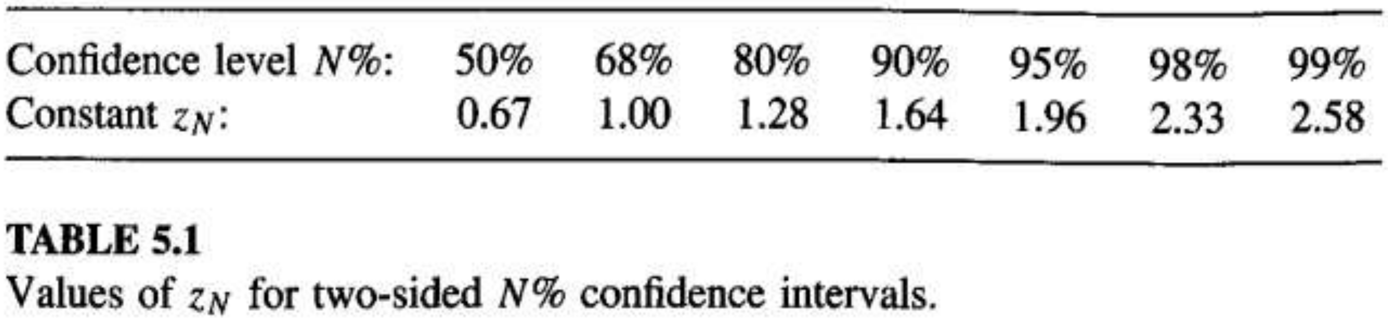
- The standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

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.

TABLE 5.4
The Normal or Gaussian distribution.

- The constant Z_N given in Table 5.1 defines the width of the interval about the mean
- that includes *N%* of the total probability mass under the bell-shaped Normal distribution.
- More precisely, Z_N gives half the width of the interval (i.e., the distance from the mean in either direction) measured
- Figure 5.l(a) illustrates such an interval for z_{80}



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

TABLE 5.1
Values of z_N for two-sided $N\%$ confidence intervals.

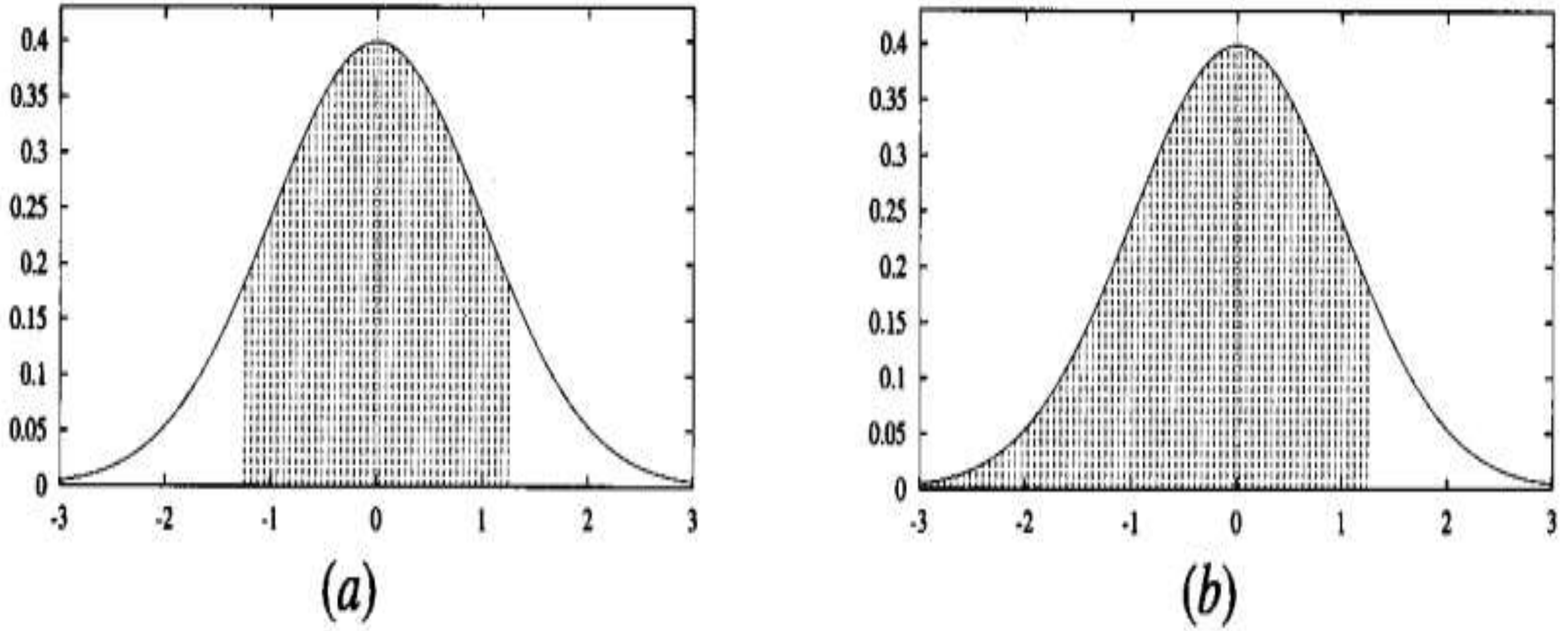


FIGURE 5.1

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 $[-\infty, 1.28]$.

- To summarize, if a random variable Y obeys a Normal distribution with mean μ and standard deviation σ, then the measured random value y of Y will fall into the following interval N% of the time

$$\mu \pm z_N \sigma \qquad \text{Eqn. (5.10)}$$

- Equivalently, the mean μ will fall into the following interval N% of the time

$$y \pm z_N \sigma \qquad \text{Eqn. (5.11)}$$

- To derive the general expression for N% confidence intervals for discrete-valued hypotheses given in Equation (5.1).
- First, we know that errors(h) follows a Binomial distribution with mean value $\text{error}_D(h)$ and standard deviation as given in Equation (5.9).
- Second, we know that for sufficiently large sample size n, this Binomial distribution is well approximated by a Normal distribution.
- Third, Equation (5.11) $y \pm z_N \sigma$ tells us how to find the N% confidence interval for estimating the mean value of a Normal distribution.

- Substituting the mean and standard deviation of errors(h) into Equation 5.11 yields the expression from Equation errorD(h) for N% confidence intervals for discrete-valued hypotheses.

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

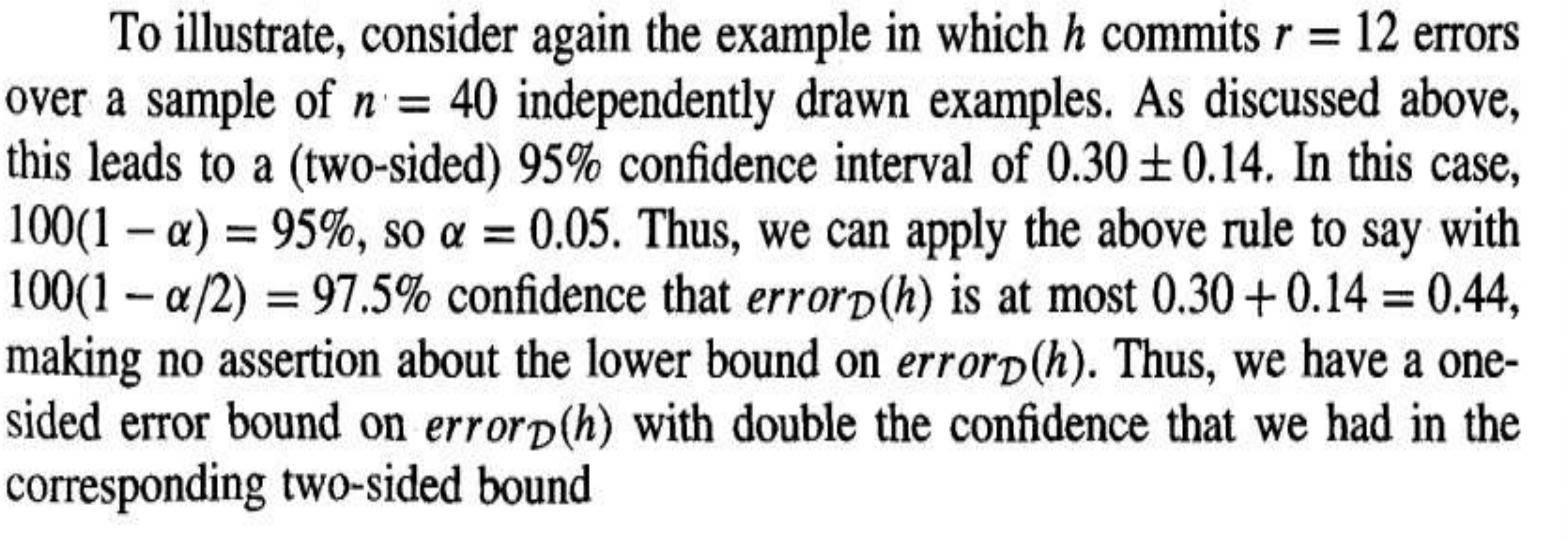
Two approximations were involved in deriving this expression, namely:

1. we have approximated errorD(h) by errors(h) and
2. The Binomial distribution has been approximated by the Normal distribution.

Two-sided and One-sided Bounds

- the above confidence interval is a two-sided bound; it bounds the estimated quantity from above and from below. In some cases, we will be interested only in a one-sided bound.
- Any two-sided confidence interval based on a Normal distribution can be converted to a corresponding one-sided interval
- 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).

Example



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 ± 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

COMPARING LEARNING ALGORITHMS

- We wish to determine which of LA and LB is the better learning method on average for learning some particular target function f.
- consider the relative performance of these two algorithms averaged over all the training sets of size n that ae drawn from the instance distribution D.
- we wish to estimate the expected value of the difference in their errors

$$\underset{S \subset \mathcal{D}}{E}\left[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))\right] \qquad \text{-----5.14}$$

- where L(S) denotes the hypothesis output by learning method L when given the sample S of training data

- in practice we have only a limited sample Do of data when comparing learning methods.
- approach is to divide Do into a training set So and a disjoint test set To.
- The training data can be used to train both LA and LB, and the test data can be used to compare the accuracy of the two hypotheses.

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0)) \qquad \text{-----5.15}$$

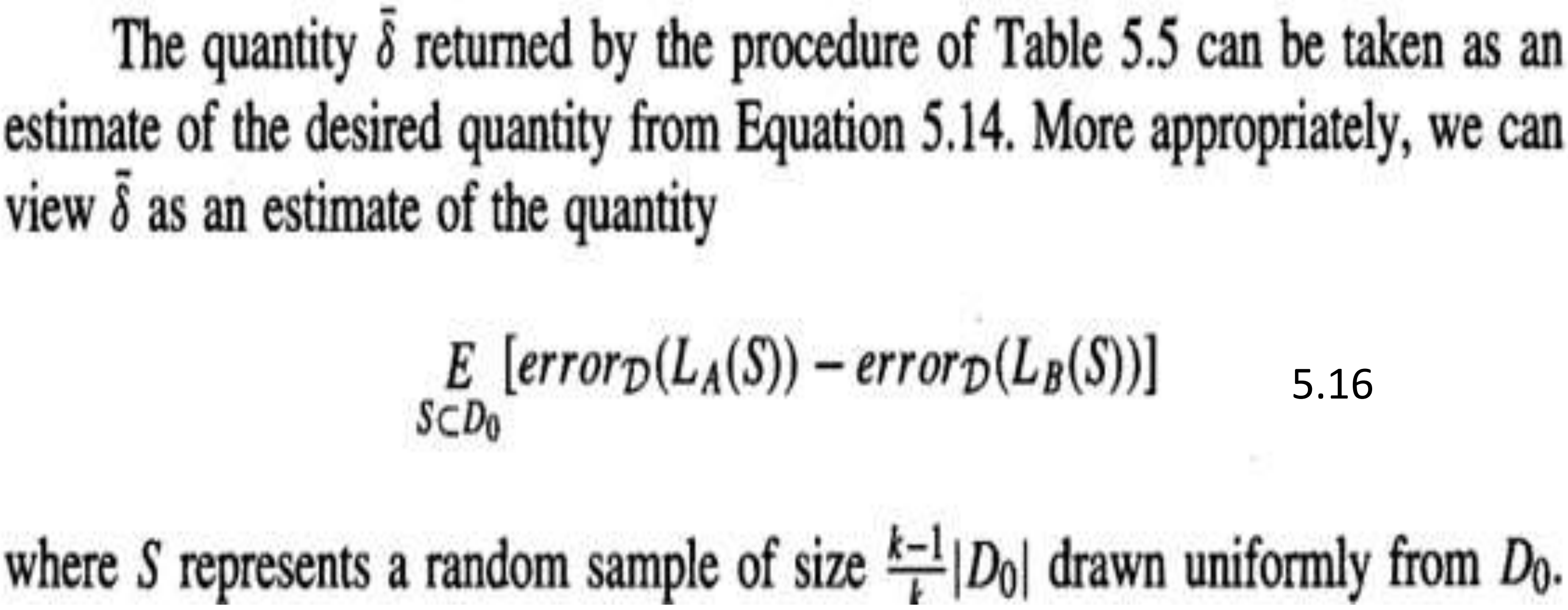
- two key differences between this estimator and the quantity   in Equation(5.14) is
- First, we are using error$_{To}$(h) to approximate error$_{D}$(h).
- Second, we are only measuring the difference in errors for one training set S0 rather than taking the expected value of this difference over all samples S that might be drawn from the distribution D.

- to improve the estimator given by Equation (5.15)
- repeatedly partition the data Do into disjoint training and test sets
- and take the mean of the test set errors for these different experiments.

```
1. Partition the available data $D_0$ into $k$ disjoint subsets $T_1, T_2, \ldots, T_k$ 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
       $\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$        (T5.1)
```

TABLE 5.5

A procedure to estimate the difference in error between two learning methods L_A and L_B. Approximate confidence intervals for this estimate are given in the text.



The quantity $\bar{\delta}$ returned by the procedure of Table 5.5 can be taken as an estimate of the desired quantity from Equation 5.14. More appropriately, we can view $\bar{\delta}$ as an estimate of the quantity

$$\mathop{E}_{S\subset D_0}\left[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))\right] \qquad 5.16$$

where S represents a random sample of size $\frac{k-1}{k}|D_0|$ drawn uniformly from D_0.

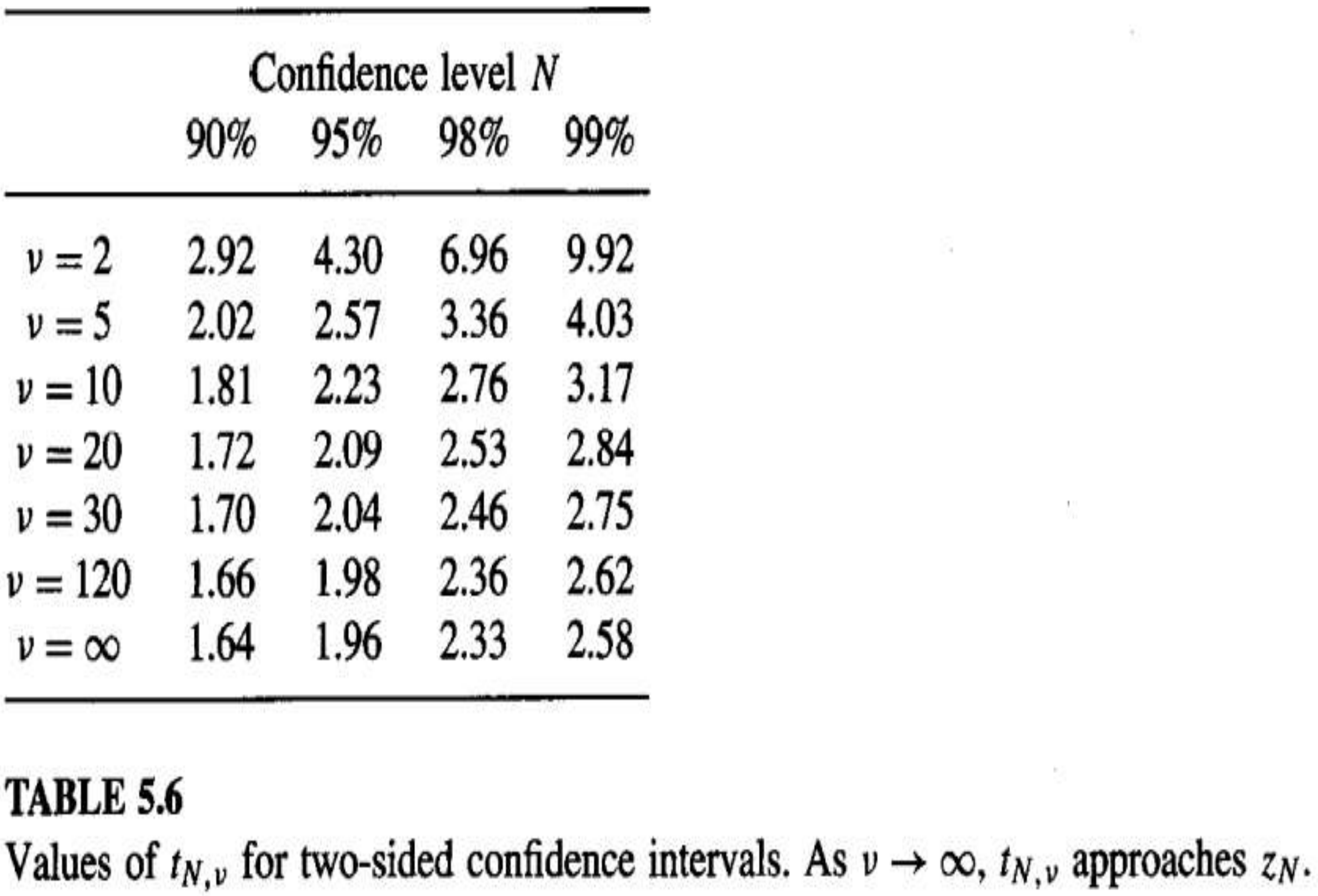
The approximate $N\%$ confidence interval for estimating the quantity in Equation (5.16) using $\bar{\delta}$ is given by

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

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

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

Notice the constant $t_{N,k-1}$ in Equation (5.17) has two subscripts. The first specifies the desired confidence level, as it did for our earlier constant z_N. The second parameter, called the number of *degrees of freedom* and usually denoted by ν, is related to the number of independent random events that go into producing the value for the random variable $\bar{\delta}$. In the current setting, the number of degrees of freedom is $k - 1$. Selected values for the parameter t are given in Table 5.6. Notice that as $k \to \infty$, the value of $t_{N,k-1}$ approaches the constant z_N.



	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.

- Tests where the hypotheses are evaluated over identical samples are called paired tests.
- Paired tests typically produce tighter confidence intervals because any differences in observed errors are due to differences between the hypotheses.

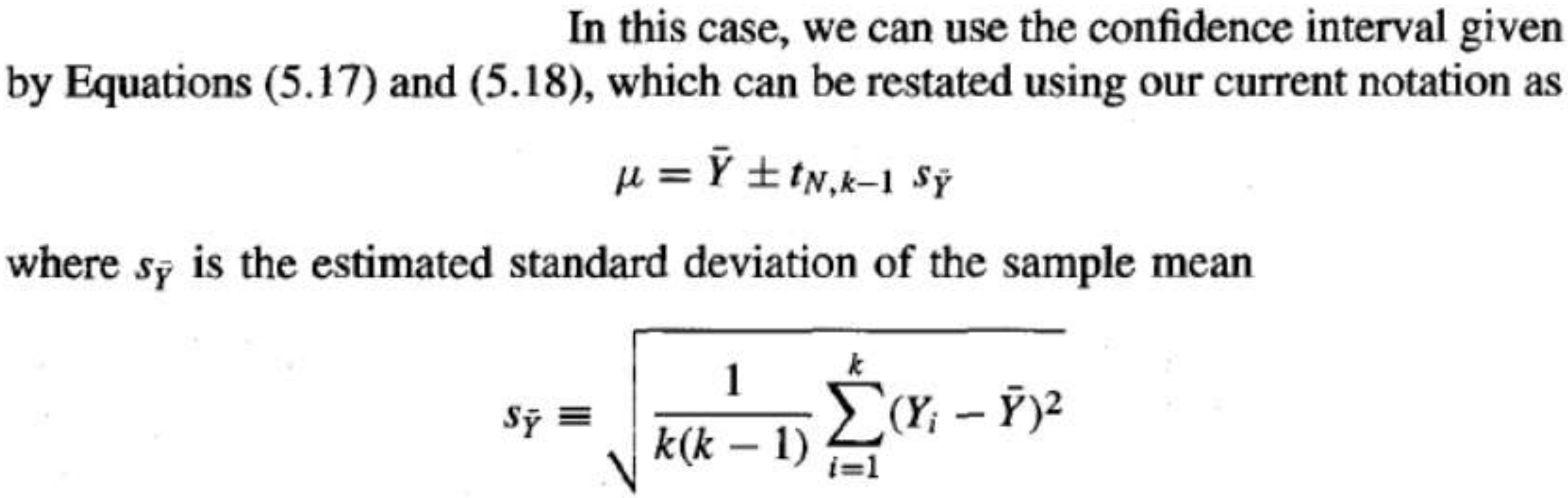
Paired t Tests

- 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 Y1, Y2, . . . , Yk.
  - We wish to estimate the mean μ of the probability distribution governing these Yi.
  - The estimator we will use is the sample mean $\bar{Y}$

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

- the individual Yi follow a Normal distribution.

- in this idealized method we modify the procedure of Table 5.5
- on each iteration through the loop it generates a new random training set *Si* and new random test set *Ti* by drawing from this underlying instance distribution instead of drawing from the fixed sample *Do.*
- In particular, the δi measured by the procedure now correspond to the independent, identically distributed random variables *Yi.*
- The mean μ 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 $\overline{\delta}$ computed by this idealized version of the method.



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

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

- 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 when k approaches infinity.