

Recap

1. Assessing the error of a single hypothesis h

A. Two-sided confidence interval:

$$\begin{aligned} & error_S(h) \pm z_N \sigma_S \\ &= error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \end{aligned}$$

“With $N\%$ confidence, true error is in this interval”

Example

Suppose hypothesis h is tested on a set of 100 examples, and it classifies 83 correctly. What is the 95% confidence interval for the true error rate?

On-line statistics tables: <http://surfstat.anu.edu.au/surfstat-home/tables/normal.php>

B. One-sided confidence interval:

If the two-sided confidence interval is for $N = 100(1 - \alpha)\%$ with lower bound L and upper bound U , this implies a one-sided confidence interval of $N_U = 100(1 - \alpha/2)\%$.

“With $N_U\%$ confidence, the true error is smaller than U .”

Example

Suppose hypothesis h commits $r = 10$ errors over a sample of $n = 65$ independently drawn examples. What is the 90% (two-sided) confidence interval for the true error rate? What is the 95% one-sided interval?

2. Comparing the error of two hypotheses, h_1 and h_2

A. Confidence interval around observed difference in error rate:

N% confidence interval around

$$\hat{d} = error_{s_1}(h_1) - error_{s_2}(h_2)$$

is

$$\hat{d} \pm z_N \sqrt{\frac{error_{s_1}(h_1)(1 - error_{s_1}(h_1))}{n_1} + \frac{error_{s_2}(h_2)(1 - error_{s_2}(h_2))}{n_2}}$$

“With confidence N%, the true difference in error rate d is in this interval.

Hypothesis testing

Given $\hat{d} = error_{s_1}(h_1) - error_{s_2}(h_2)$

what is the probability that $d > 0$ (i.e., h_2 is truly more accurate than h_1)?

Example: Suppose

$error_{s_1}(h_1) = .3$ and $error_{s_2}(h_2) = .2$ on samples with $n = 100$.

It appears that h_1 makes more errors:

$$\hat{d} = .1$$

Then, $Probability(d > 0) = Probability(\hat{d} < d + .1)$

Since d is the mean of the distribution of \hat{d} , can write this as:

$$\hat{d} < \mu_{\hat{d}} + .1 \text{ or}$$

$$\hat{d} < \mu_{\hat{d}} + \gamma\sigma_{\hat{d}} \text{ where } \gamma\sigma_{\hat{d}} = .1$$

We have :

$$\sigma_{\hat{d}} = \sqrt{\frac{(.3)(.7)}{100} + \frac{(.2)(.8)}{100}} \approx .061$$

so $\gamma \approx 1.64$.

Thus,

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

From our table, 1.64 standard deviations corresponds to a two-sided interval with confidence 90%. The associated one-sided confidence level is 95%.

Therefore:

$$Probability(error_D(h_1) > error_D(h_2)) \approx .95.$$

“We accept the hypothesis that $error_D(h_1) > error_D(h_2)$ with confidence .95”

“We reject the hypothesis that $error_D(h_1) > error_D(h_2)$ at a .05 level of significance.

Summary of Hypothesis testing

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

what is the probability that $d > 0$?

$$\begin{aligned} P(d > 0) &= P(\hat{d} < d + \hat{d}) \\ &= P\left(\hat{d} < \mu_{\hat{d}} + \frac{\hat{d}}{\sigma_{\hat{d}}}\sigma_{\hat{d}}\right) \end{aligned}$$

Look up $\frac{\hat{d}}{\sigma_{\hat{d}}}$ in z_n table to get N .

Find associated one - sided confidence interval N_U .

“With confidence $N_U\%$, we accept the hypothesis that $error(h_1) > error(h_2)$.”

Comparing Learning Algorithms

- Let L_1 and L_2 be two learning algorithms.
- How to determine if difference in performance of L_1 and L_2 is statistically significant?
- There are many approaches to this. One method: **paired t test** .

Defining “difference in performance”

- First, what does “difference in performance” mean?

Let

$$d = error_D(L_1(S)) - error_D(L_2(S))$$

where $L(S)$ is the hypothesis output by L on training set S .

We want the expected difference in performance over all training sets S of size n drawn from distribution D :

$$E[d]$$

S drawn from D

Estimating expected difference in performance

- In practice, can't get all samples S of size n over D . So need an estimator.
- Suppose we have sample D_0 of data, and divide it into S_0 (training set) and T_0 (test set). Then our estimator for the difference in performance between L_1 and L_2 is:

$$\hat{d} = \text{error}_{T_0}(L_1(S_0)) - \text{error}_{T_0}(L_2(S_0))$$

Improving estimator using k -fold cross-validation

1. Partition data D_0 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
 - use T_i for the test set, and the remaining data for training set S_i*
 - $S_i \leftarrow \{D_0 - T_i\}$
 - $h_A \leftarrow L_A(S_i)$
 - $h_B \leftarrow L_B(S_i)$
 - $\delta_i \leftarrow \text{error}_{T_i}(h_A) - \text{error}_{T_i}(h_B)$
3. Return the value $\bar{\delta}$, where

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

Note that $\bar{\delta}$ is the mean of samples from random variables δ_i .

$\bar{\delta}$ is called a “sample mean”. It is an unbiased estimator of the population mean $E[d]$
 S drawn from D

The sample standard deviation is:

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

But the true standard deviation of the sample mean, $\sigma_{\bar{\delta}}$, is unknown.

It is estimated by

$$S_{\bar{\delta}} = \frac{\sigma_{sample}}{\sqrt{n}}$$

This is also called “standard error”. It estimates the standard deviation of the distribution of samples. It reflects how much sampling fluctuation σ_{sample} has.

So we have:

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

Approximate N % confidence interval:

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

Here $t_{N, k-1}$ is a constant that corresponds to z_N .

When $k \rightarrow \infty$, $t_{N, k-1} \rightarrow z_N$.

Determining this confidence interval is called the “paired t test”.

$t_{N, k}$ table

Confidence level N

	90%	95%	98%	99%
$k = 2$	2.92	4.30	6.96	9.92
$k = 5$	2.02	2.57	3.36	4.03
$k = 10$	1.81	2.23	2.76	3.17
$k = 20$	1.72	2.09	2.53	2.84
$k = 30$	1.70	2.04	2.46	2.75
$k = \infty$	1.64	1.96	2.33	2.58