<u>Question no 1</u>: Suppose you have used your favourite concept learning algorithm to learn a hypothesis h1 from some training data. You are interested in knowing the accuracy that the hypothesis can be expected to achieve on the underlying population. To assess this accuracy you apply the hypothesis to a test data set consisting of 45 instances that you had held back from the training data set. The error rate observed on the training data is 6.67%. Calculate the 95% confidence interval for the true error?

- **n** (number of instances) = 45
- **e** (error rate) = 6.67% = 0.0667

The 95% confidence interval for the true error 'e' is:

e = **e ± 1.96 * sqrt((e * (1 - e)) / n)**

```
Standard Error = sqrt((0.0667 * (1 - 0.0667)) / 45)
               = sqrt((0.0667 * 0.9333) / 45)
               = sqrt(0.06220011 / 45)
               = sqrt(0.0013822246)
               = 0.037178

Margin of Error = 1.96 * 0.037178
               = 0.07286888

Lower bound = 0.0667 - 0.07286888
            = -0.00616888

Upper bound = 0.0667 + 0.07286888
            = 0.13956888
```

Since an error rate can't be negative, we can say the lower bound is 0.

So, 95% confidence interval for the true error is approximately **0 to 0.1396** means **0% to 13.96%**.

<u>Question no 2</u>: You now decide to change a few parameters within the learning algorithm used in Question 7.1 and learn two more hypotheses, h2 and h3. The error rates for these new hypotheses observed on the test data set of 45 instances were 8.89% and 13.3%, respectively. To what degree can you be confident that h2 will perform worse than h1 on the underlying population. Is your confidence higher or lower for h3 performing worse than h1 on the underlying population?

- **h2:** 8.89%
- **h3:** 13.3%

For h1, the confidence interval is approximately **0% to 13.96%**.

For H2

For h2, with n = 45 and observed error rate e = 0.0889:

Standard Error = sqrt((0.0889 * (1 - 0.0889)) / 45)
= sqrt((0.0889 * 0.9111) / 45)
= sqrt(0.08109679 / 45)
= sqrt(0.00180215)
= 0.042452

Margin of Error (95%) = 1.96 * 0.042452
= 0.08320592

Lower bound = 0.0889 - 0.08320592
= 0.00569408

Upper bound = 0.0889 + 0.08320592
= 0.17210592

The 95% confidence interval for the true error of h2 is approximately **0.57% to 17.21%**.

For H3

For h3, with n = 45 and observed error rate e = 0.133:

Standard Error = sqrt((0.133 * (1 - 0.133)) / 45)
= sqrt((0.133 * 0.867) / 45)
= sqrt(0.115311 / 45)

= sqrt(0.00256246)
= 0.050621

Margin of Error (95%) = 1.96 * 0.050621
= 0.09921716

Lower bound = 0.133 - 0.09921716
= 0.03378284

Upper bound = 0.133 + 0.09921716
= 0.23221716

The 95% confidence interval for the true error of h3 is approximately **3.38% to 23.22%**.

Summary of the 95% confidence intervals:

H1 ----------⟶ **0% to 13.96%**
H2 ----------⟶ **0.57% to 17.21%**
H3 ----------⟶ **3.38% to 23.22%**

Conclusion

Based on the 95% confidence intervals :

- We cannot state with 95% confidence that h2 will perform worse than h1, as their confidence intervals overlap.
- Similarly, we cannot state with 95% confidence that h3 will perform worse than h1, as their confidence intervals also overlap.

<u>Question no 3</u>: You now decide to try out a decision tree induction algorithm to see if it can out perform your favourite concept learning algorithm. You decide to use 10 fold cross validation. The error rates for the 10 cross validation folds for the two algorithms are shown in Table 1.

CV Fold	Favourite Algorithm	Decision Tree Induction
1	8.89%	9.3%
2	9.52%	9.48%
3	8.13%	9.12%
4	9.48%	9.13%
5	10.12%	9.98%
6	10.23%	11.01%
7	8.56%	9.02%
8	9.12%	8.56%
9	9.23%	9.23%
10	9.11%	9.08%

1. **Calculate the difference in error rates for each fold:**
   Difference = Favourite Algorithm - Decision Tree Induction

   1 ----------⟶ (0.0889 - 0.093) = -0.0041
   2 ----------⟶ (0.0952 - 0.0.0948) = 0.0004
   3 ----------⟶ (0.0813 - 0.0.0912) = -0.0099
   4 ----------⟶ (0.0948 - 0.0913) = -0.0099
   5 ----------⟶ (0.1012 - 0.0998) = 0.0014
   6 ----------⟶ (0.1023 - 0.1101) = -0.0078
   7 ----------⟶ (0.0856 - 0.0902) = -0.0046
   8 ----------⟶ (0.0912 - 0.0856) = 0.0056
   9 ----------⟶ (0.0923 - 0.0923) = -0.0000
   10 ----------⟶ (0.0911 - 0.0908) = 0.0003

2. **Calculate the mean of these differences.**

Mean Difference (d): Sum of differences / Number of folds
(-0.0041 + 0.0004 - 0.0099 + 0.0035 + 0.0014 - 0.0078 - 0.0046 + 0.0056 - 0.0000 + 0.0003**)** / 10
= - 0.0152 / 10
= - 0.00152

The mean difference is negative, mean that the "Favourite Algorithm" has a slightly higher error rate on average than the "Decision Tree Induction" algorithm.

Since the mean difference is negative, the favorite algorithm does not appear to outperform the decision tree algorithm. Therefore, the confidence level for this is very low.

<u>Question no 4</u>: Using the Euclidean distance on an ROC plot from the "perfect classifier" as the metric, choose the best classifier
• assuming equal costs for false positives and false negatives
• assuming that false positives cost 4 times as much as false negatives

h1 Confusion Matrix

Actual	Positive	Negative	Marginal Sum
Predicted Positive	29	1	30
Negative	2	13	15
Marginal Sum	31	14	45

h2 Confusion Matrix

Actual	Positive	Negative	Marginal Sum
Predicted Positive	29	3	32
Negative	1	12	13
Marginal Sum	30	15	45

h3 Confusion Matrix

Actual	Positive	Negative	Marginal Sum
Predicted Positive	27	3	30
Negative	3	12	15
Marginal Sum	30	15	45

To choose the best classifier, we will calculate the True Positive Rate (TPR), False Positive Rate (FPR), and then the Euclidean distance from the "perfect classifier" (0,1) on an ROC plot.

Formulas

True Positive Rate (TPR) / Sensitivity: TP / (TP + FN)

False Positive Rate (FPR): FP / (FP + TN)
Euclidean Distance (d): sqrt((FPR - 0)^2 + (TPR - 1)^2)

Hypothesis h1

- TP = 29, FN = 2
- FP = 1, TN = 13

- TPR = 29 / (29 + 2)
  = 29 / 31
  = 0.9355

- FPR = 1 / (1 + 13)
  = 1 / 14
  = 0.0714

- Euclidean Distance = sqrt((0.0714 - 0)^2 + (0.9355 - 1)^2)
  = sqrt(0.0051 + 0.0042)
  = sqrt(0.0093)
  = 0.0964

Hypothesis h2

- TP = 29, FN = 1
- FP = 3, TN = 12

- TPR = 29 / (29 + 1)
  = 29 / 30
  = 0.9667

- FPR = 3 / (3 + 12)
  = 3 / 15
  = 0.2000

- Euclidean Distance = sqrt((0.2000 - 0)^2 + (0.9667 - 1)^2)
  = sqrt(0.0400 + 0.0011)
  = sqrt(0.0411)
  = 0.2027

Hypothesis h3

- TP = 27, FN = 3
- FP = 3, TN = 12

- TPR = 27 / (27 + 3)
  = 27 / 30
  = 0.9000

- FPR = 3 / (3 + 12)
  = 3 / 15

= 0.2000

- Euclidean Distance = sqrt((0.2000 - 0)^2 + (0.9000 - 1)^2)
  = sqrt(0.0400 + 0.0100)
  = sqrt(0.0500)
  = 0.2236

Summary of Euclidean Distances (Equal Costs)

Hypothesis	TPR	FPR	Euclidean Distance from Perfect Classifier
h1	0.9355	0.0714	0.0964
h2	0.9667	0.2000	0.2027
h3	0.9000	0.2000	0.2236

Based on the calculations, **h1** has the smallest Euclidean distance (0.0964). Therefore, **h1 is the best classifier assuming equal costs for false positives and false negatives.**

b) Assuming false positives cost 4 times as much as false negatives:

In this case, we modify the distance formula for false positives more heavily:

Distance= $\text{sqrt}((4\times FPR)^2+(1-TPR)^2)$

h1 : Distance = $\text{sqrt}((4\times0.071)^2+(1-0.935)^2)$
$=\text{sqrt}(0.284^2+0.065^2)$
$=\text{sqrt}(0.080656+0.004225)$
$=\text{sqrt}(0.084881)$
$=0.291$

h2 : Distance = $\text{sqrt}((4\times0.2)^2+(1-0.906)^2)$
$=\text{sqrt}(0.8^2+0.094^2)$
$=\text{sqrt}(0.64+0.008836)$
$=\text{sqrt}(0.648836)$
$=0.805$

h3 : Distance = $\text{sqrt}((4\times0.2)^2+(1-0.9)^2)$
$=\text{sqrt}(0.8^2+0.1^2)$
$=\text{sqrt}(0.64+0.01)$
$=\text{sqrt}(0.65)$
$=0.806$

Even with the higher penalty on false positives, **h1** is still the best classifier because its FPR is significantly lower than h2 and h3.
Therefore, **h1 is also the best classifier assuming that false positives cost 4 times as much as false negatives**.