1. Basic Statistics
   - Mean
   - SSD
   - Variance
   - Skew
   - Issues
   - Confidence
2. Sampling Distributions
   - Results
   - Comments
   - Conclusions
   - Issues
3. t-test
   - Results

Example 1 - Mean, Median, Variance, Sample Standard Deviation, and Skewness

- We want to make a generalized statement on the accuracy of a single classifier.
- The classifier is Naive Bayes in `Weka`.
- The data set is the UCI Labor database
- Run classifier on the data set 10 times
- Each run, train on randomly selected 80% of the examples and hold out the remaining 20% for testing
- For each run, compute the accuracy as "correctly classified percentage", i.e. produce 10 readings. These results are listed on page 3
- Compute the mean, median, variance, and standard deviation for the observed accuracies
- Interpret these results together

Table: Naive Bayes classifier on Labor data

Run	No. Train	No. Test	Accuracy
1	46	11	91.6667
2	45	12	100
3	46	11	91.6667
4	46	11	91.6667
5	46	11	91.6667
6	47	10	83.3333
7	47	10	83.3333
8	45	12	100
9	46	11	91.6667
10	45	12	100

Accuracy	Frequency
83.3333	2
91.6667	5
100	3

The Mean (average) & Median

- Use the formula:
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- In this case, the mean = 92.50001
- The Median is the middle accuracy value for the lower table on page 3
- In this case, Median = 91.6667
- The value of the median is not the same as that of the mean. This suggests that the distribution of accuracy values is skewed. This is very much apparent by having a shew value other than 0

Sample Standard Deviation

- Use the formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- In this case, the SSD or s = 6.148881761
- The standard deviation is a measure of how widely values are dispersed from the average value (the mean)

Variance

- Use the formula:

$$Var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- In this case, the variance = 37.80874691
- Variance for Naive Bayes accuracy on the Labor data set

Skewness

- Use the formula:

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$$

- In this case, the Skew = -0.165965671
- The distribution is negatively skewed
- Skewness characterizes the degree of asymmetry of a distribution around its mean.
- Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values (values higher than the mean)
- Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values (values lower than the mean)

Issues - Sample Size

- How many runs do we need to present results that are reliable? Well, there is no obvious answer, however we used a rule of thumb here. Our rule is to never use data points fewer than 10. This is based on a statistical advice to use a minimum number of n to be 10 or greater
- In this example, we applied this rule of thumb twice, once for the number of test examples (we randomly selected at least 10 examples to test) and we performed at least 10 random runs. Is this sufficient? We reckon this is the least.

Issues - Distribution

- What are our assumptions?
- We implicitly assumed that these values are normally distributed and we measured their mean, median, variance, and skewness.
- Is this really true? The answer at this point is "not sure" because we have no evidence to support this
- If we repeat this experiment a million times will we obtain a normal distribution? According to our results, it may not be the case, the distribution is skewed.
- However, based on central limit theorem and sampling distribution, we know that the distribution of sample means is approximately normal.
- Can we use this to address this issue? (later!)

Issues - Confidence

- What about confidence?
- How reliable are these numbers?
- We use a statistical method to compute confidence

Confidence Interval

- We address the issue of confidence abut the mean of observed accuracy from the previous example by computing the 95% confidence intervals of the observed mean
- When $\alpha = 0.05$ (or 5%), this means we compute the interval of possible mean values at the $1 - \alpha = 0.95$ (which is the 95%) level of confidence
- We use the formula:

$$\bar{x} \pm z \left(\frac{\sigma}{\sqrt{n}} \right)$$

  where σ is the population standard deviation and is estimate by s from the previous example.

- When $\alpha = 0.05$, the value of the z-statistic for standard normal distribution is 1.96 (normal distribution table)

Results

- the mean accuracy is $\bar{x} = 92.50001$
- and the 95% confidence interval for the mean accuracy is $[88.68896363, 96.31105637]$
- In our sample of 10 measurements of classifier accuracies, the average accuracy is 92.5% with a standard deviation of 6.15%. When $\alpha = 0.05$, the corresponding confidence interval is $92.5 \pm 3.81104637 = [88.68896363, 96.31105637]$.
- For any population mean, μ_0, in this interval, the probability of obtaining a sample mean further from μ_0 than 92.5% is more than 0.05. Likewise, for any population mean, μ_0, outside this interval, the probability of obtaining a sample mean further from μ_0 than 92.5% is less than 0.05.

Issues:

- We have not addressed the issue of sampling distributions to obtain a normally distributes sample of accuracy values.
- Our calculation of the confidence interval assumes a normally distributed population! (we used z statistic of a normal distribution!)
- The width of the confidence interval (approximately 7.622%). If we wish to increase the precision of our analysis, we need to make this interval more narrow, how?

Sampling Distributions

- We want to estimate the expected accuracy for Naive Bayes classifier on labor data.
- We run cross-validation of 10 folds (data contains 57 examples, approximately 10% is the size of hold out set)
- we run 10 times of the above (i.e. 10 runs of 10-fold cross-validation to avoid cross-validation problems!)
- For each run of 10 folds, we compute:
  - average size of the training set
  - average size of the test set
  - average accuracy
  - standard deviation
  - and distribution skew.
- We compare all results obtained so far

Table: Results of running Naive Bayes on Labor data averaged over 10 folds of cross-validation for each of the 10 runs, e.g. each row is an average for all 10 folds except for SD and Skew, they are computed for all 10 fold of that particular run

Run	Train	Test	Mean Accuracy	Median Accuracy	SD Accuracy	Skew
1	51.3	5.7	90	100	14.06	-1.00
2	51.3	5.7	94.67	100	8.64	-1.08
3	51.3	5.7	95	100	8.05	-1.04
4	51.3	5.7	94.67	100	8.64	-1.08
5	51.3	5.7	95	100	8.05	-1.04
6	51.3	5.7	91.67	100	11.79	-1.19
7	51.3	5.7	94.67	100	11.68	-2.09
8	51.3	5.7	94.67	100	11.68	-2.09
9	51.3	5.7	94.33	100	9.17	-1.071
10	51.3	5.7	91	100	12.38	-0.96

Table: Analysis of accuracy of Naive Bayes on Labor data so far

	previous 10 random runs	all 100 (10×10 folds)	10 means (each 10 fold)
median	91.67	100	94.67
mean	92.50	93.57	93.57
STDEV	6.15	10.27	1.90
Skew	-0.17	-1.30	-1.17
95%-CI($\bar{x}$)	[88.69, 96.31]	[91.55, 95.58]	[92.39, 94.74]

Observations

- The median over all 100 runs is the highest.
- The median of 10 random runs is lowest at 91.67%, while the median of medians (94.67) is closest to the mean values.
- The mean of means is the same as the mean of the 100 runs (makes sense) and both are higher but close to the mean of the 10 random runs.
- The standard deviation of means is significantly lower than the other two means (reliable because of sampling distributions)
- Skewness is significantly higher in this last 100 runs.
- The 95%-confidence interval is very wide for the 10 random runs, it tightens a bit for the 100 runs, and is significantly tighter for the 10 means!

Conclusions

- At this point, we can safely conclude that the expected value for accuracy is very close to the value of the mean of means at 93.57% with standard deviation of 1.90% and 95%-confidence interval of $[92.39, 94.74]$.
- This conclusion is more reliable than previous measurement and is more valid since the assumption of normality holds.

Issues

- The standard deviation of the mean values (over the 10 runs) is not the standard deviation of the population. Calculating the confidence intervals assumes that the standard deviation of the population is known. Is it true that the measures standard deviation of the means of 1.90% is the same for the population? Can we possible use another method to verify this value? (Hint: hypothesis testing!). (Note: this could be a good practice problem!).
- Consider the average size of the training and testing data folds. The testing remains well below the size of 10. Is it possible that we could obtain better results by repeating these 10 runs of fold bigger than 10 examples of testing such that we can guarantee reliable results? (Note: again, good practice problem!).

The problem

- We are interested in knowing whether or not the difference in accuracy obtained by Naive bayes classifier to that obtained by a decision tree (J48) is statistically significant.
- We run 20 times using random split of training and testing
- We record accuracis for both classifiers and obtain their difference for each run

Table of Contents
Basic Statistics
Sampling Distributions
t-test
Results

Table: Results of running Naive Bayes and J48 on labor data using random split

Run	Train NB	Train J48	Test NB	Test J48	Acc NB	Acc J48	Difference
1	38	38	19	19	94.74	84.21	10.53
2	37	37	20	20	95	85	10
3	37	37	20	20	100	70	30
4	37	37	20	20	95	65	30
5	37	37	20	20	80	80	0
6	39	39	18	18	88.89	88.89	0
7	37	37	20	20	90	90	0
8	38	38	19	19	94.74	78.95	15.79
9	39	39	18	18	94.44	77.78	16.67
10	37	37	20	20	100	80	20
11	37	37	20	20	85	70	15
12	37	37	20	20	100	85	15
13	37	37	20	20	95	90	5
14	38	38	19	19	94.74	68.42	26.32
15	38	38	19	19	89.47	84.21	5.26
16	39	39	18	18	88.89	83.33	5.56
17	38	38	19	19	84.21	73.68	10.53
18	38	38	19	19	100	78.95	21.05
19	38	38	19	19	100	84.21	15.79
20	37	37	20	20	85	70	15

Table: Analysis of the difference in accuracy between Naive Bayes and J48 on labor data

Statistic	Naive Bayes Accuracy	J48 Decision Tree Accuracy	Difference in their Accuracy
Median	94.74	80.00	15.00
Mean	92.76	79.38	13.37
St. Dev.	6.05	7.58	9.22
Skew	-0.49	-0.40	0.25
SE			2.06
Paired t-test (two tailed)			6.49
CI Lower			9.06
CI Upper			17.69