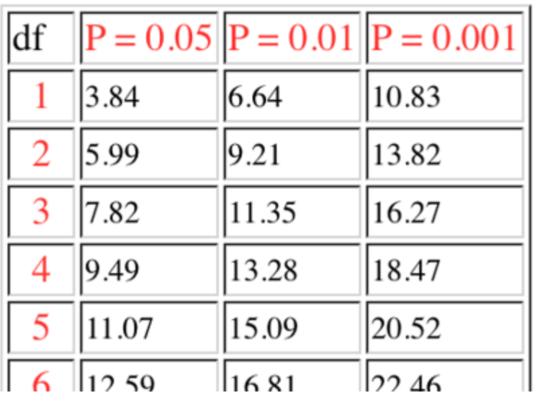
Workshop Week 11 - COMP20008 2020

1. Consider the following data set for a binary class problem:

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
F	F	-

We wish to select the feature that best predicts the class label using the χ^2 method.

- Write down the observed and expected contingency tables for feature A
- Calculate the $\chi^2(A, Class)$ value.
- Using the table below, conclude whether feature A is independent of the class label for $p = 0.05$.



df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46

- Repeat the process for feature B and decide which feature could be best used for predicting the class label.

Observed table:

A	A=T	A=F	Total
Class=+	4	0	4
Class=-	2	4	6
Total	6	4	10

Expected table:

A	A=T	A=F	Total
Class=+	2.4	1.6	4
Class=-	3.6	2.4	6
Total	6	4	10

$\chi^2(A, Class) = \frac{(4-2.4)^2}{2.4} + \frac{(0-1.6)^2}{1.6} + \frac{(2-3.6)^2}{3.6} + \frac{(4-2.4)^2}{2.4} = 4.44$

Degrees of freedom $= (2-1) \times (2-1) = 1$
Lookup value in table (3.84). Since our calculated χ^2 value is greater than the critical value in the table, conclude A is not independent of Class for $p = 0.05$

For feature B:
Observed table:

B	B=T	B=F	Total
Class=+	3	1	4
Class=-	1	5	6
Total	4	6	10

Expected table:

B	B=T	B=F	Total
Class=+	1.6	2.4	4
Class=-	2.4	3.6	6
Total	4	6	10

$\chi^2(B, Class) = \frac{(3-1.6)^2}{1.6} + \frac{(1-2.4)^2}{2.4} + \frac{(1-2.4)^2}{3.6} + \frac{(5-3.6)^2}{3.6} = 3.40$

Degrees of freedom $= (2-1) \times (2-1) = 1$
Lookup value in table (3.84). Since our calculated χ^2 value is less than the critical value in the table, conclude B is independent of Class for $p = 0.05$

Feature A best predicts class label.

2. Suppose you are conducting data linkage between two databases, one with m records and the other with n records (assume $m < n$). Under a basic approach, $m \times n$ record comparisons will be needed.

   - Assume there are no duplicates. What is the maximum number of record matches? What is the corresponding number of non-matching comparisons required in this circumstance? **Answer: maximum number of record matches is m, number of non-matches is m* (n-1)**

   Now suppose a blocking method is employed, where each record is assigned to exactly one block. Assume this method results in b number of blocks.

   - What is the smallest possible number of comparisons? **0 (all blocks contain only records from one of the two databases)**
   - What is the smallest possible number of comparisons that will return all matches? **(m)** What is the value of b? **(anything between m+ 1 and n)**

- What is the largest possible number of comparisons? ($m \times n$, **when** $b = 1$)
- If records are evenly allocated to b blocks, how many comparisons will be needed? $(m/b) \times (n/b) \times b = m \times n/b$
- What is the advantage with a large b? What is the advantage with a small b? **(fast, inaccurate vs slow, accurate)**
- In practice, a record is assigned to more than one block and records are not evenly allocated to blocks. How would this affect your analysis of large b vs small b?
  - **Dominant block size is the bottleneck for efficiency. So, even if b is large, the blocking can still lead to inefficient record linkage**
  - **Small b tends to correspond to inefficiency; but it does not automatically guarantee accuracy. If matches are all off blocks, accuracy can still be low.**
- What is the advantage of records being assigned to multiple blocks? **To reduce the chance of matched pairs not allocated to a common block. Matched pairs may not have the same parts of the records in common, sometimes, a blocking key may miss a pair but captured by another blocking key.**

3. One may evaluate the output of a data linkage system according to how many records are linked correctly and how many records are linked incorrectly.

   - What are the reasons two records could be linked incorrectly? **(They do not correspond to the same entity. E.g. two persons with accidentally the same name, but not the same date of birth)**
   - Suppose a false positive (FP) is two records that are linked by the system, which a human believes should not have been linked. Suppose a false negative (FN) is when two records are not linked by the system, which a human believes should have been linked. A true positive (TP) is two records linked by the system which a human believes should have been linked and a true negative (TN) is two records not linked by the system which a human believes should not have been linked.
     - What are the relative sizes of the categories TP, TN, FP, FN? In practice, how might one calculate these sizes?
       Construct the confusion matrix

	Actual:Link=true	**Actual:Link=false**
Prediction:link=true	**Small (TP)**	**Small (FP)**
Prediction: link=false	**Small (FN)**	**Very large (TN)**

     - It is desirable to minimise both FP and FN, but it may be difficult to minimise both simultaneously. Give an example application where minimising FP is more important than minimising FN. Give an example application where minimising FN is more important than minimising FP.

   More important to minimise FP: e.g. Centrelink - might want to avoid sending out debt recovery letters to people who don't actually owe a debt
   Minimise FN: e.g. Medical screening tests - Often used to flag people who may have a particular disease for further tests. Don't want a FN to

mean that people who actually do have the disease don't receive further tests and treatment. Also national security, don't want a FN to mean that a potential terrorist plot is not followed up on.
In practice there's often a trade-off involving the resources available for further investigations. We are willing to have a number of false positives to ensure we're avoiding false negatives, but don't want so many false positives that we can't manually screen and investigate them all.