Empirical Methods for Computer Science (CS 5970) Homework 1 Solutions

October 9, 2008

Question 1

1. (10pts) Suppose we have an **independent** variable that can take on one of two values (A, B), and a **dependent** variable that can take on one of three values (x, y, z). The following table gives the number of occurrences of each combination:

	x	y	z
A	17	45	32
B	28	72	54

Compute χ^2 for the hypothesis that the distribution of the dependent variable is the same given the independent variable (we refer to this as the *null* hypothesis – more on this later). Show your work.

```
expected =

   17.0565   44.3468   32.5968
   27.9435   72.6532   53.4032

chi2_ind =
```

```
    0.0002    0.0096    0.0109
    0.0001    0.0059    0.0067

chi2 =

    0.0334
```

2. (10pts) Suppose we have an independent variable that can take on one of two values (C, D), and a dependent variable that can take on one of three values (x, y, z). The following table gives the number of occurrences of each combination:

	x	y	z
C	15	36	8
D	21	43	16

Compute χ^2 for the hypothesis that the distribution of the dependent variable is the same given C and D. Show your work.

```
expected =

   15.2806   33.5324   10.1871
   20.7194   45.4676   13.8129

chi2_ind =

    0.0052    0.1816    0.4695
    0.0038    0.1339    0.3463

chi2 =

    1.1403
```

3. (10pts) What can you conclude (relatively) about these two different hypotheses?

   The χ^2 statistic for the 2nd case is higher. Because both contingency tables are the same size (and hence have the same number of DOFs), it is less likely that the null hypothesis will hold for the 2nd case.

4. (10pts) "dat1" is a matrix containing a set of paired samples of an independent variable (column 1) and a dependent variable (column 2). What is $p(v_1)$? What is $p(v_2|v_1)$? According to χ^2 is there a relationship between these two variables?

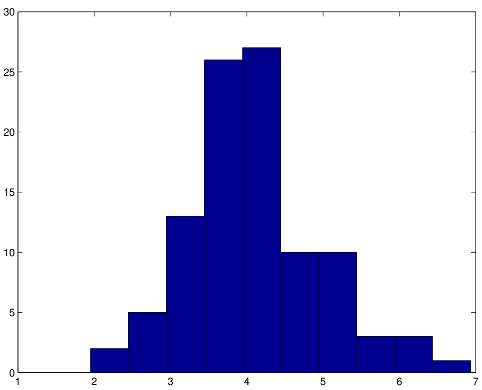
```
pv1 = 0.3120    0.3850    0.3030

prob =
    0.7821    0.0513    0.1667
    0.7455    0.1091    0.1455
    0.2673    0.6337    0.0990

(rows are the independent variable)

chi2 = 348.3774

p = 0
```

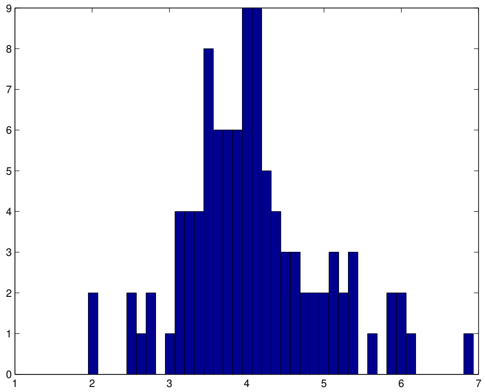
   Yes: there is a strong relationship between these two variables.

Question 2

1. (10pts) "dat2" contains several samples of a continuous random variable. Describe (in brief) the distribution of the data. Is this distribution monotonic? Is it Gaussian? (this is not intended to be a long answer; you do not need to do any hypothesis testing)

The distribution does see to have a single mode. Here are two histograms of the samples:



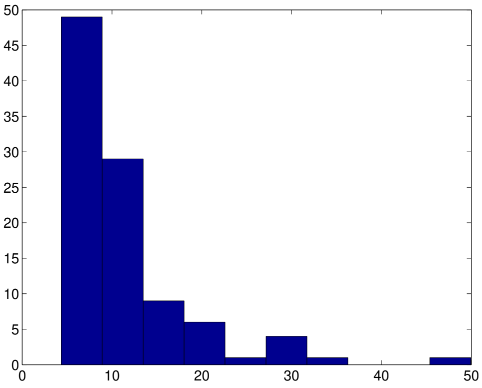


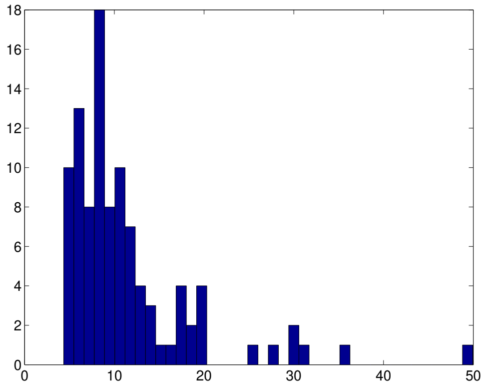
Other than a slight imbalance to the right, this does appear to be a Gaussian distribution.

The reality: it is a Gaussian distribution ($\mu = 4.1$ and $\sigma = 0.74$).

2. (10pts) "dat3" contains another set of samples from a random variable. Describe (in brief) the distribution of the data. Is this distribution monotonic? Is it a Gaussian?

This distribution is probably monotonic. However, it is definitely not Gaussian:

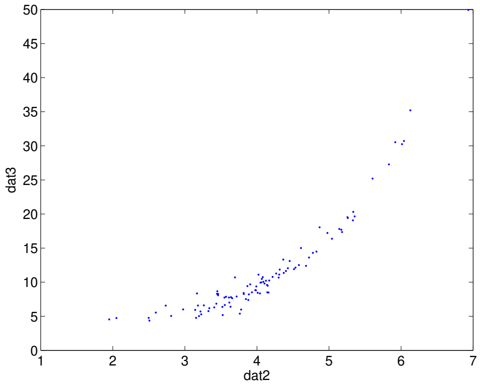




What is particularly strange is that the set of samples is abruptly cut off at 4.35.

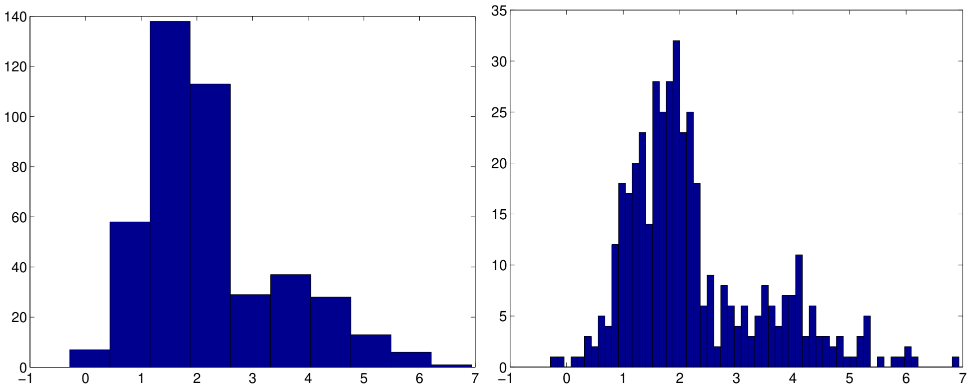
3. (10pts) Assume that dat2 and dat3 are paired tuples. Briefly describe the relationship between these two variables.

   The correlation between these two data sets is high: $R = .9$. A scatter plot reveals that there is a nonlinear relationship between these two variables:

4. (10pts) "dat4" contains yet another set of samples from a random variable. Describe (in brief) the distribution of the data. Is this distribution monotonic? Is it a Gaussian?

   This data set is not monotonic: there are two maxima in the histogram:

   The reality: it is a mixture of two Gaussian distributions: $\mu = 4.1$, $\sigma = .74$; $\mu = 1.667$, $\sigma = .349$.

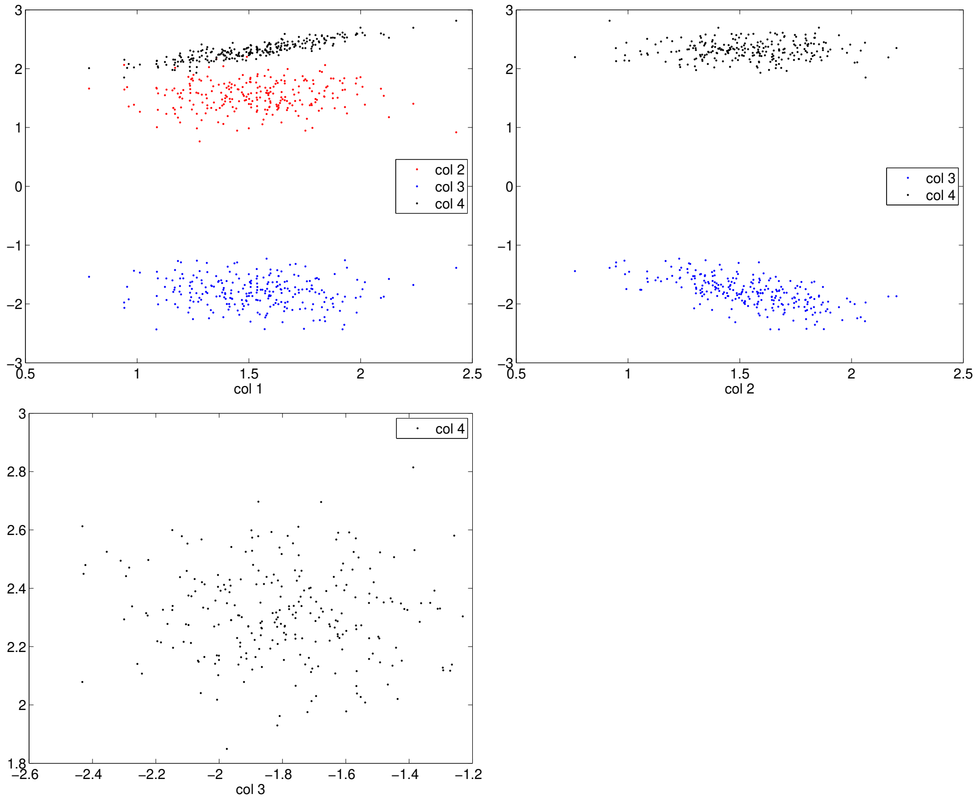
Question 3

(20pts) "dat5" contains a set of 4-tuple observations (the data are represented as a single matrix). Describe the relationship (if any) between the four variables and show the process by which you came to these conclusions.

The correlation matrix tells much of the story:

1.0000	0.0165	-0.0996	0.8799
0.0165	1.0000	-0.6294	0.0169
-0.0996	-0.6294	1.0000	-0.0775
0.8799	0.0169	-0.0775	1.0000

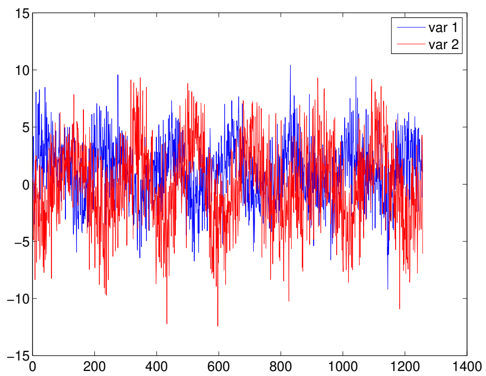
Column 1 is related to column 4, and columns 2 and 3 are related weakly. These are confirmed with scatter plots:



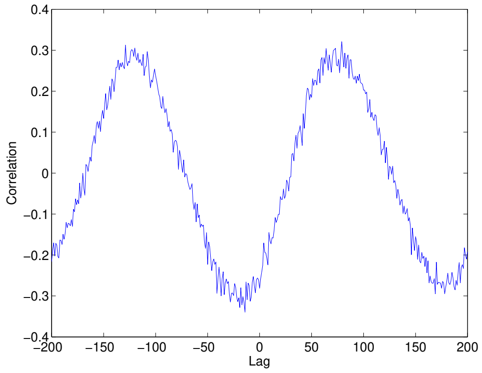
Question 4

(20pts) “dat6” contains two time series (represented as a single matrix). Describe the relationship between these variables and show the process by which you arrived at this conclusion.

The original data:

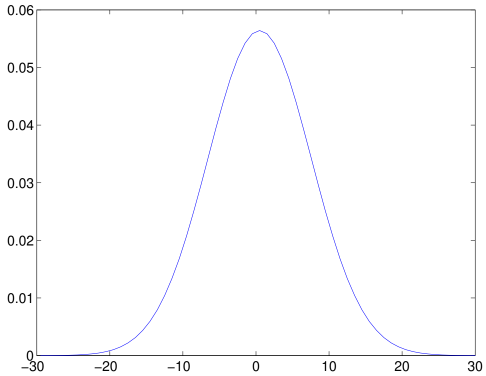


Cross-correlation between these two variables:

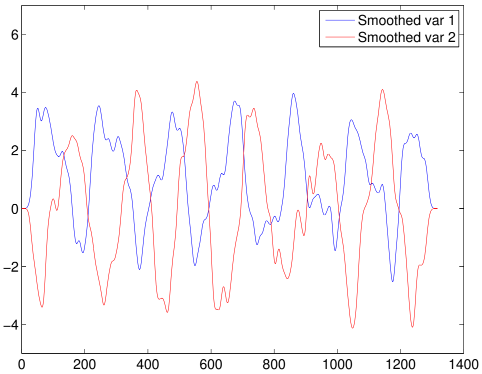


Note that the peak correlation is $R = -.3394$ *at* $t = -14$.

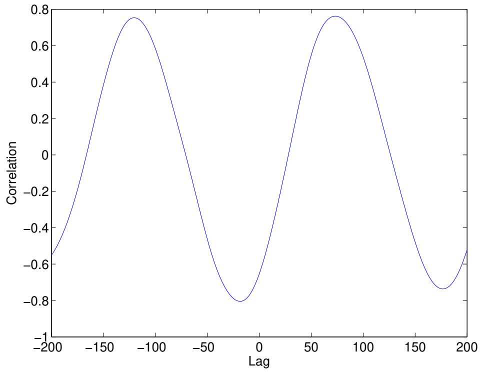
The original data was filtered using a pretty wide mask:



This cleaned up much of the high-frequency variation:



Cross-correlation between these two smoothed variables:



The peak correlation is $R = -.8045$ *at* $t = -18$. *This magnitude is substantially higher than the original time series. We also feel much more confident about the relative timing of these two signals.*