Empirical Methods for Computer Science (CS 5453) Homework 1 Solutions

April 20, 2011

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	74	38	5
B	75	64	10

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.

```
prob =

    0.6325    0.3248    0.0427
    0.5034    0.4295    0.0671

expected =
```

```
   65.5376   44.8647    6.5977
   83.4624   57.1353    8.4023


chi2_part =

    1.0927    1.0503    0.3869
    0.8580    0.8248    0.3038

chi2 =

    4.5166

p =

    0.1045
```

Note: chi2cdf() is useful for computing the p-value.

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	23	88	178
D	42	150	301

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

```
prob =
```

```
    0.0796    0.3045    0.6159
    0.0852    0.3043    0.6105


expected =

   24.0217   87.9565  177.0217
   40.9783  150.0435  301.9783


chi2_part =

    0.0435    0.0000    0.0054
    0.0255    0.0000    0.0032

chi2 =

    0.0775

p =

    0.9620
```

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

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

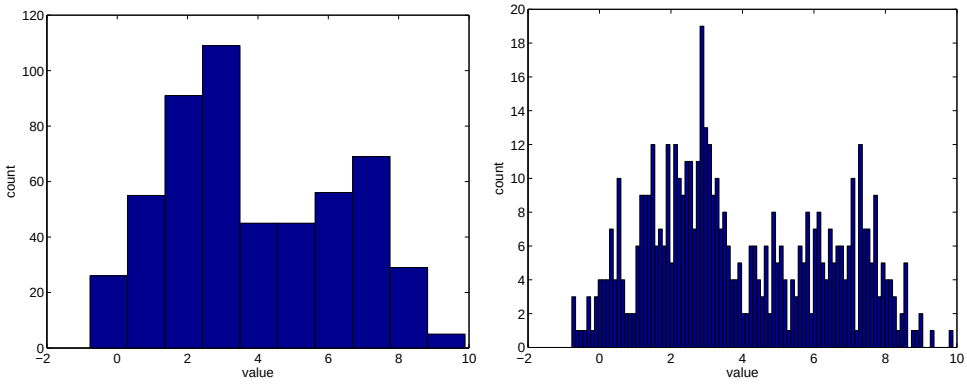
```
pv1 = 0.3850    0.1000    0.5150

prob =
    0.7766    0.0545    0.1688
    0.8900    0.0800    0.0300
    0.1961    0.4854    0.3184

(rows are the independent variable)

chi2 = 391.5355

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. Does the distribution have a single mode? Is it Gaussian? This is not intended to be a long answer (you do not need to do any hypothesis testing). But - do some visualization.

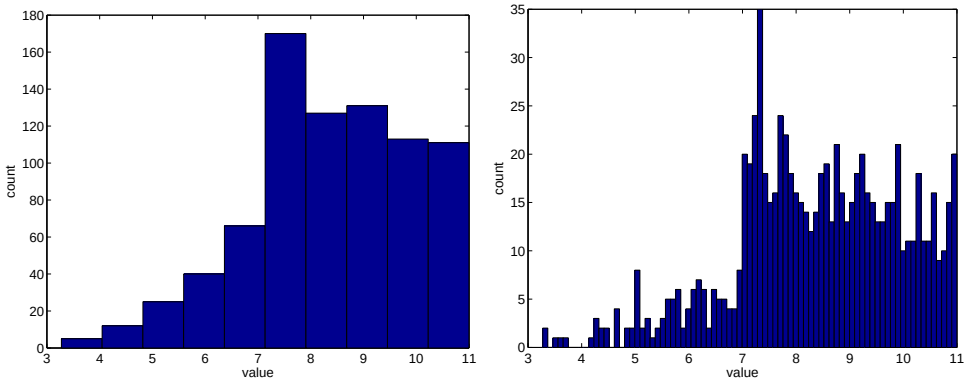
The distribution does have two modes, but you have to look at the right resolution. Here are two histograms of the samples:



The reality: this is a mixture of two Gaussian distributions ($\mu = 6.533$ and $\sigma = 1.51$ and $\mu = 2.233$ and $\sigma = 1.75$).

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

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

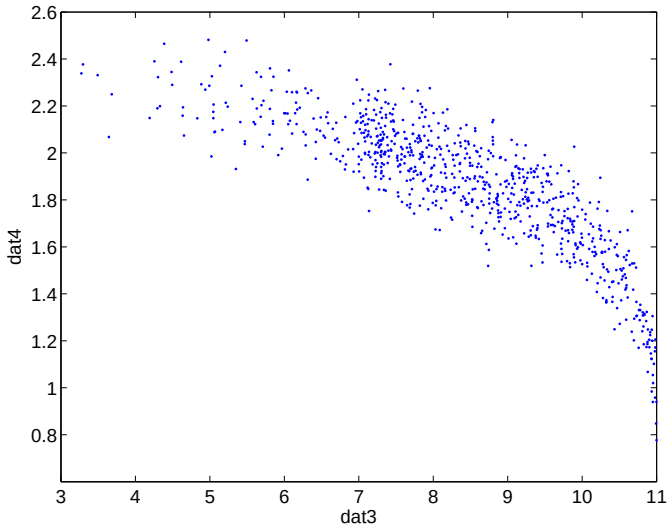


   What is particularly strange is that the set of samples is abruptly cut off around 11.

   The reality: this is a mixture of one Gaussian and one uniform distribution ($\mu = 6.533$ and $\sigma = 1.51$; and $[7..11]$).

3. (10pts) Assume that dat3 and dat4 are paired tuples. Briefly describe the relationship between these two variables.

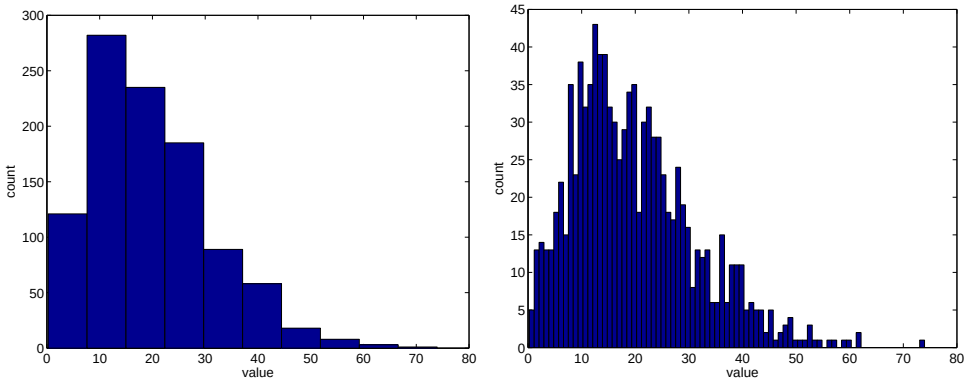
   The anti-correlation between these two data sets is high: $R = -.84$. A scatter plot reveals that there is a nonlinear relationship between these two variables:



Note: corrcoef() or corr() are very useful for this.

4. (10pts) "dat5" contains yet another set of samples from a random variable. Describe (in brief) the distribution of the data. Is this distribution unimodal? Is it a Gaussian?

This data set is unimodal:



The shape is not Gaussian – it appears more like a Poisson distribution.

The reality: this distribution is derived from a Gaussian. Each sample from the Gaussian is raised to the power of two, which yields the longer

tail on the right (in fact, this is a common transformation to get us from a Gaussian-like shape to a Poisson-like shape, although we tend to go in the opposite direction).

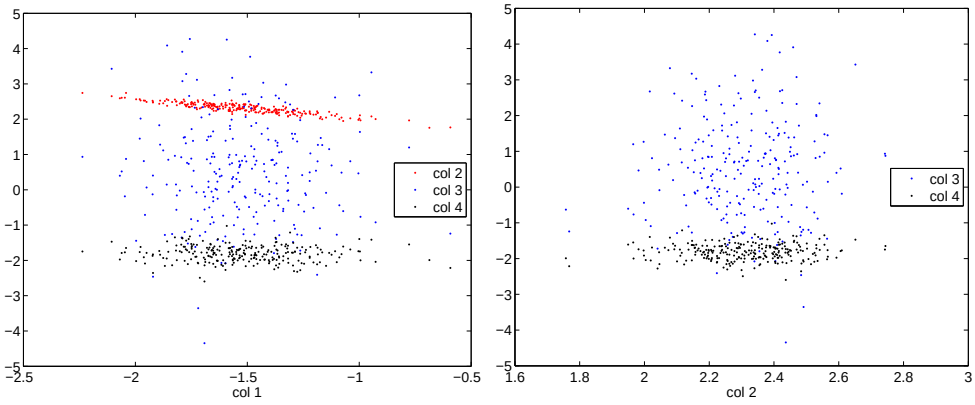
Question 3

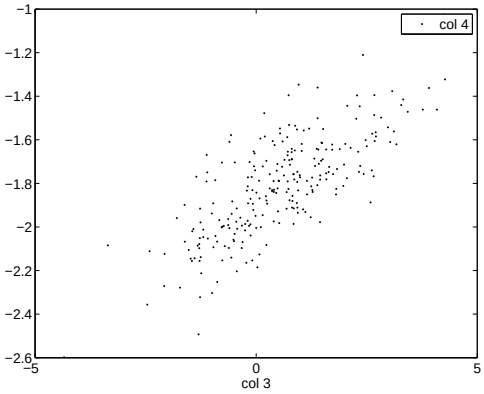
(20pts) "dat6" contains a set of 4-tuple observations (the data are represented as a single matrix). Describe the relationship (if any) between the four variables and show the process by which you came to these conclusions.

The correlation matrix tells much of the story:

```
 1.0000   -0.8955   -0.0399   -0.0707
-0.8955    1.0000    0.0132    0.0474
-0.0399    0.0132    1.0000    0.7604
-0.0707    0.0474    0.7604    1.0000
```

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





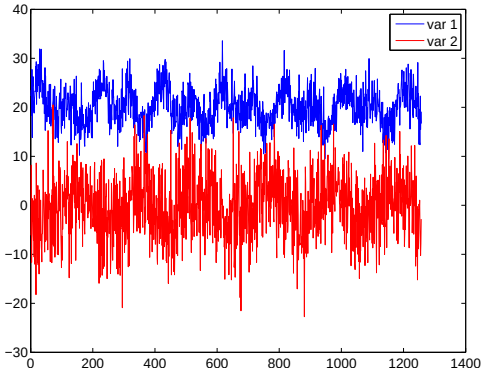
Note that col1 and col2 have a high anti-correlation. However, the range of the scatter plot masks this (compare red vs black in the first scatter plot).

Question 4

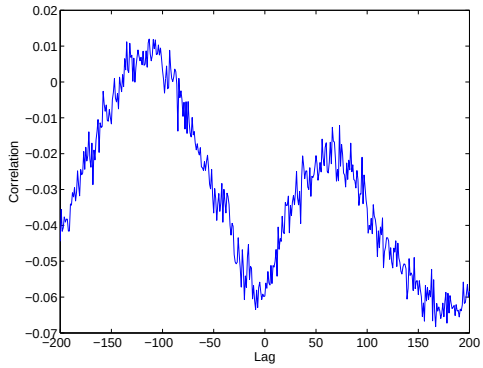
(20pts) "dat7" contains two time series (represented as a single matrix). Describe the relationship between these variables and show the process by which you arrived at this conclusion.

Note: I used xcorr() for computing these statistics. It took me a while, but I convinced myself that the "coeff" option gives us a Pearson's R.

The original data:

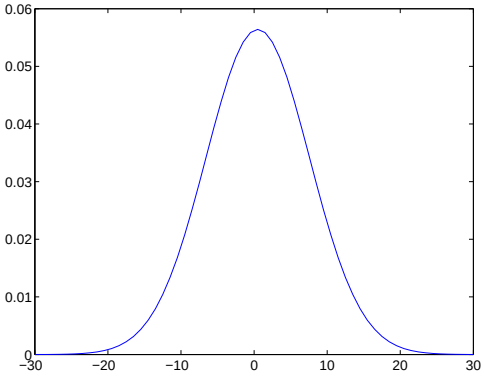


Cross-correlation between these two variables:

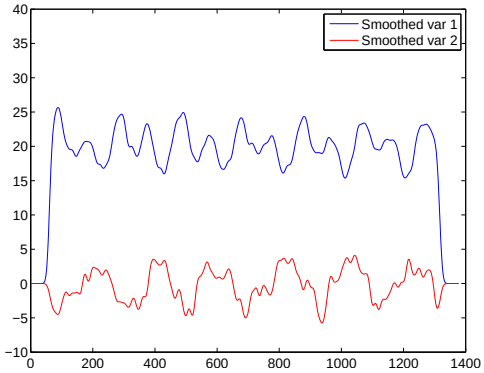


Note that the peak correlation is $R = -.00683$ *at* $t = +168$*. However, this is a really small correlation value.*

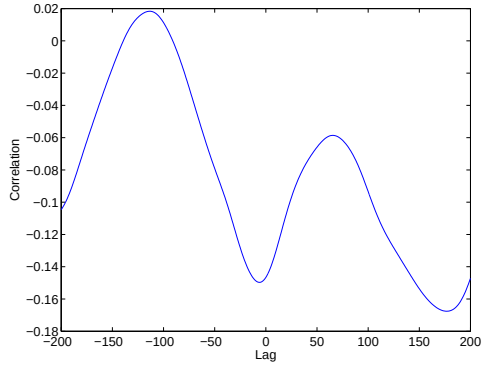
The original data was filtered using a pretty wide mask:



This cleaned up much of the high-frequency variation:



Cross-correlation between these two smoothed variables:



The peak correlation is $R = -.1676$ *at* $t = +177$. *This magnitude is substantially higher than the original time series. However, this* R *is still rather small. So, it is hard to be convinced that these two timerseries are related to one-another.*

The reality is: they are, but the relationship is nonlinear and a function of time, so cross-correllation has a hard time pulling out a relationship (remember that it is a linear operation).