Technical Test for Research Associates(CAFRAL) Solutions

Suraj Kumar

January 29, 2018

1 Problem 1: True or False

a: FALSE
Let's assume we want to estimate a set of probability distribution parameters θ given a dataset D. The Bayes' Rule states

$$posterior = \frac{likelihood * prior}{evidence(dataset)} \tag{1}$$

$$p(\theta/D) = \frac{p(D/\theta) * p(\theta)}{p(D)} \tag{2}$$

where $p(D) = \int_\theta p(D/\theta) * p(\theta) d\theta$

Thus from above equations ([1],[2])the likelihood is probability of observing a given data given that parameters values is as stated by hypothesis H.

b: FALSE
The maximum likelihood estimate is not always the best guess of for the value of parameters. It may not exist or may not be unique. It doesn't even consider any prior belief. It treats $\frac{p(\theta)}{p(D)}$ as constant. MLE, $\hat{\theta}$ is a point estimate, not a random variable, which maximizes the likelihood $p(D/\theta)$. The Bayesian estimate may be better in case where we have prior information. Thus, MLE can only be best guess when we don't have any prior information and it exist and unique.

c: FALSE
If a text statistic lies inside the acceptance region, it implies that we don't reject the null hypothesis H_0. It does not imply that we accept the null hypothesis.

d: FALSE
The p-value is the smallest level of significance α_0 such that we would reject the null-hypothesis at level α_0 with the observed data.

2 Problem 2: Deal or no deal

Let H_i denote the hypothesis that the prize is in suitcase i where $i = A, B, C$..Let the prior probability of getting the prize is as shown below

$$p(\text{get prize}) = p(H_A)p(A) + p(H_B)p(B) + p(H_C)p(C) \tag{3}$$

Since the suitcases are equally likely to get selected imply $p(A) = p(B) = p(C) = 1/3$.

a)
The a-priori probabilities of H_i are $p(H_i) = 1/3$ for all , $i = A, B, C$

b)
Let E_i denotes that the person has chosen i. Let M_j denotes that the host open suitcase j which is empty. Let E_A and M_B occurs.

$$p(H_A/M_B) = \frac{p(M_B/H_A) * p(H_A)}{p(M_B)} = \frac{1/2 * 1/3}{1/2} = 1/3 \tag{4}$$

$$p(H_B/M_B) = \frac{p(M_B/H_B) * p(H_B)}{p(M_B)} = \frac{0 * 1/3}{1/2} = 0 \tag{5}$$

$$p(H_C/M_B) = \frac{p(M_B/H_C) * p(H_C)}{p(M_B)} = \frac{1 * 1/3}{1/2} = 2/3 \tag{6}$$

Here likelihood of H_A is $p(M_B/H_A) = 1/2$. The likelihood of H_C is $p(M_B/H_C) = 1$, if A is selected by the person and the prize is in C then the hose would always choose B. The likelihood of H_B is $p(M_B/H_B) = 0$, it is 0 because if B had a prize then the host would never open it.

c)
The a-posteriori probabilities of each H_i is $p(H_i/M_B)$ for all $i = A, B, C.$. Their values are shown in equation[4], equation[5] and equation[6] respectively for A, B and C.

d)
As the posterior probability of $p(H_C/M_B) = 2/3 > 1/3 = p(H_A/M_B)$, the person should switch to C.

3 Problem 3: Coin Tosses

a)
The null hypothesis is $H_0 : p = 1/2$ and the alternative hypothesis is $H_1 : p \neq 1/2$. Here p denotes the probability of getting head on a single coin toss.

b)
Let $q = h - t$ where h denotes the number of heads and t denotes the number of tails in n coin tosses. Let $q = x$ then $h = \frac{n+x}{2}$ and $t = \frac{n-x}{2}$. Hence the PMF of $q = x$ is similar to distribution of $h = \frac{n+x}{2}$ which follows the binomial distribution $bin(n, p)$. Thus the PMF is as follows:

$$P(q = x) = {}^nC_{\frac{n+x}{2}} * (p)^{\frac{n+x}{2}} * (1-p)^{\frac{n-x}{2}} \tag{7}$$

$$\text{Under } H_0 \Rightarrow p = 1/2 \Rightarrow P(q = x) = {}^nC_{\frac{n+x}{2}} * (\frac{1}{2})^n \tag{8}$$

The cumulative distribution function of the q cam be expressed as:

$$F(q \leq k) = \sum_{x=-n}^{k} {}^nC_{\frac{n+x}{2}} * (p)^{\frac{n+x}{2}} * (1-p)^{\frac{n-x}{2}} \tag{9}$$

$$\text{Under } H_0 \Rightarrow p = 1/2 \Rightarrow F(q \leq k) = \sum_{x=-n}^{k} {}^nC_{\frac{n+x}{2}} * (\frac{1}{2})^n \tag{10}$$

Since the distribution of $q = x$ is similar to distribution of $h = \frac{n+x}{2} = k$ which follows the $Bin(n, 1/2)$ with PMF as ${}^nC_k * p^k * (1-p)^(n-k)$. It can be argued that since for large enough n say $n > 20$ and p not too near 0 and 1 (say $0.05 < p < 0.95$),the binomial distribution approximately follows the Normal distribution. If $h \sim binomial(n, p)$, then h approximately follows the Normal distribution with mean $E(h) = np$ and $\sigma = \sqrt{var(h)} = \sqrt{np(1-p)}$. So $Z = \frac{h-np}{\sqrt{np(1-p)}}$ is approximately $N(0, 1)$. Thus the distribution of q, which is similar to h, can be approximated by normal distribution(Gaussian).

c)

$$\begin{aligned} E[q] &= E[h] - E[t] = n/2 - n/2 = 0 \\ Var(q) &= Var(h-t) = Var\Big[h - (n-h)\Big] = Var(2h-n) \\ &= 4 * Var(h) + Var(n) - 2 * Cov(2h, n) = 4 * Var(h) \\ &= 4 * n * p(1-p) \\ p = 1/2 \Rightarrow Var(q) &= 4 * n * (1/4) = n \Rightarrow sd(q) = \sqrt{n} \end{aligned} \tag{11}$$

Here $Var(n)$ and $Cov(2h, n)$ is 0 as n is constant which means $E(n) = n$

d)
As noted in the part b of this problem, the probability distribution function(CDF) of q will be roughly normal given $n > 20$ and $0.05 \leq p \leq 0.95$. As noted in the part c of this problem $E(q) = 0$ and $sd(q) = \sqrt{n}$. Thus $q \sim N(0, \sqrt{n})$ and distribution of q is symmetric around 0. As we have two-sided test, let us define a q^* such that

$$\Phi(\frac{q^*}{\sqrt{n}}) - \Phi(\frac{-q^*}{\sqrt{n}}) = 0.95 \tag{12}$$

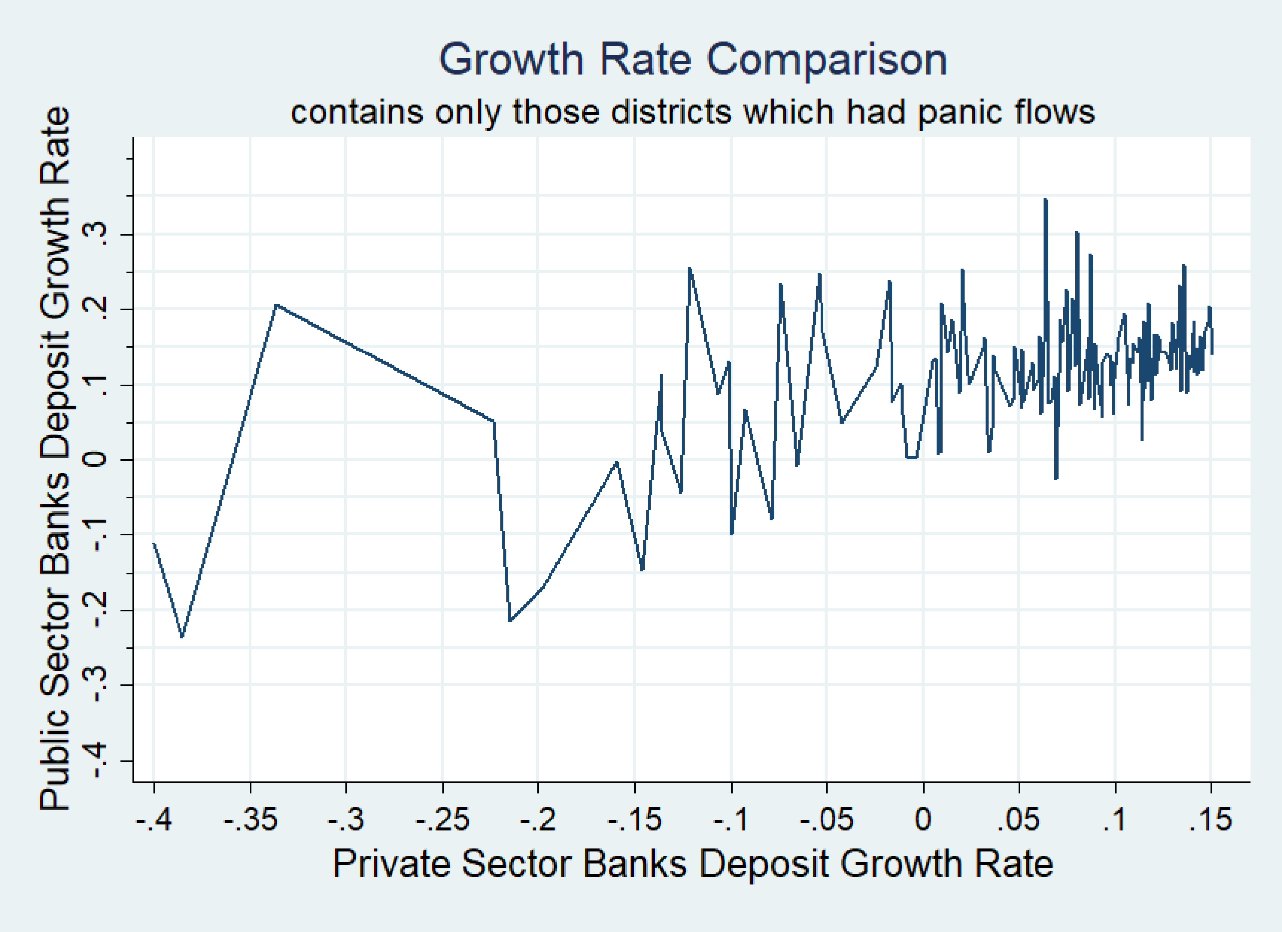
Here, $\Phi \sim N(0, 1)$ is standard normal distribution. Thus the critical region to reject null hypothesis is $q_{calc} \in (-\infty, -q^*) \cup (q^*, \infty)$. Here $q^* = 1.96 * \sqrt{n}$

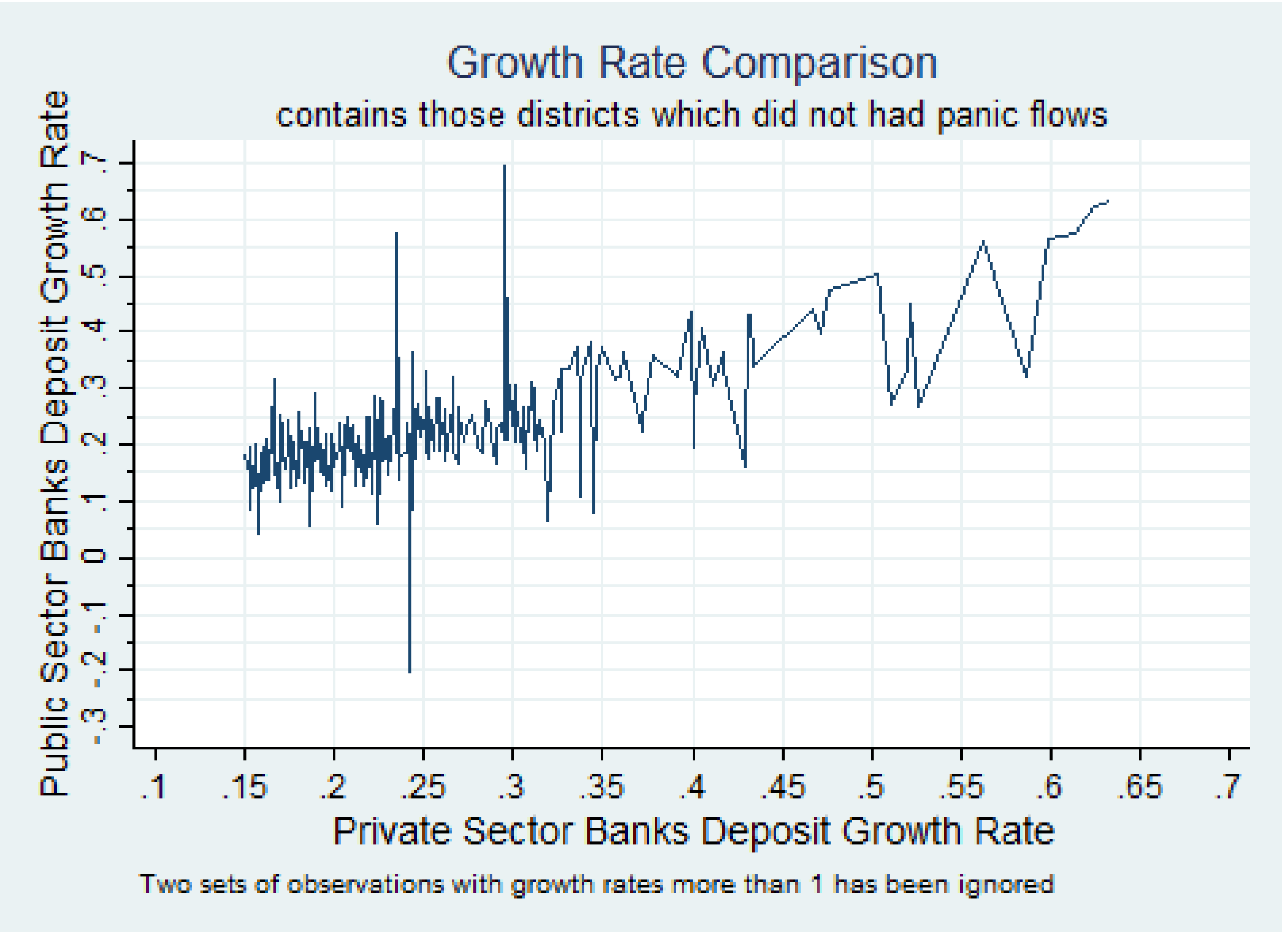
e)
Using the results of the part c and part d of the problem. It follows $q = 63 - 37 = 26$ and thus $q \sim N(0, 10)$. As $q_{calc} = 26 > q^* = 1.96 * 10 = 19.6$, the null hypothesis $H_0 : p = 1/2$ can be rejected. Thus the uncle can be accused of cheating.

4 Problem 4: Programming(STATA)

c)
The mean deposit growth for public sector banks in districts that had panic flows are .11866. The corresponding rate was .22919 for districts that had no panic flows.

d)
The results are shown in fig[1] and fig[2]


Growth Rate Comparison
contains only those districts which had panic flows
Public Sector Banks Deposit Growth Rate
-.4
-.3
-.2
-.1
0
.1
.2
.3
-.4
-.35
-.3
-.25
-.2
-.15
-.1
-.05
0
.05
.1
.15
Private Sector Banks Deposit Growth Rate


Growth Rate Comparison
contains those districts which did not had panic flows
Public Sector Banks Deposit Growth Rate
-.3
-.2
-.1
0
.1
.2
.3
.4
.5
.6
.7
.1
.15
.2
.25
.3
.35
.4
.45
.5
.55
.6
.65
.7
Private Sector Banks Deposit Growth Rate
Two sets of observations with growth rates more than 1 has been ignored

e)

This part of the problem set could not be done. As I was not able to understand the 3D coordinates of heat maps. I tried with some coordinates iteration such as $x =$ growth rate, $y = district$ and $z = district$, but then STATA showed crazy output or went into infinite loop.

f)

The results for paired t-test for panic and no panic flows has been shown in fig[3] and fig[4]. Here $panic_flow = 1$ means that panic flow is there and 0 otherwise. In both the cases the null hypothesis of equality in the group means can be rejected.

```
. ttest deposit_private_growth_rate == deposit_public_growth_rate if panic_flow == 1

Paired t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
deposi..	173	.0587428	.0075123	.0988086	.0439147	.0735709
deposi..	173	.1186577	.0062746	.0825299	.1062725	.1310429
diff	173	-.0599148	.0067829	.0892152	-.0733033	-.0465264

```
     mean(diff) = mean(deposit_privat~e - deposit_public~e)        t =  -8.8332
 Ho: mean(diff) = 0                               degrees of freedom =      172

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

. ttest deposit_private_growth_rate == deposit_public_growth_rate if panic_flow == 0

Paired t test
```

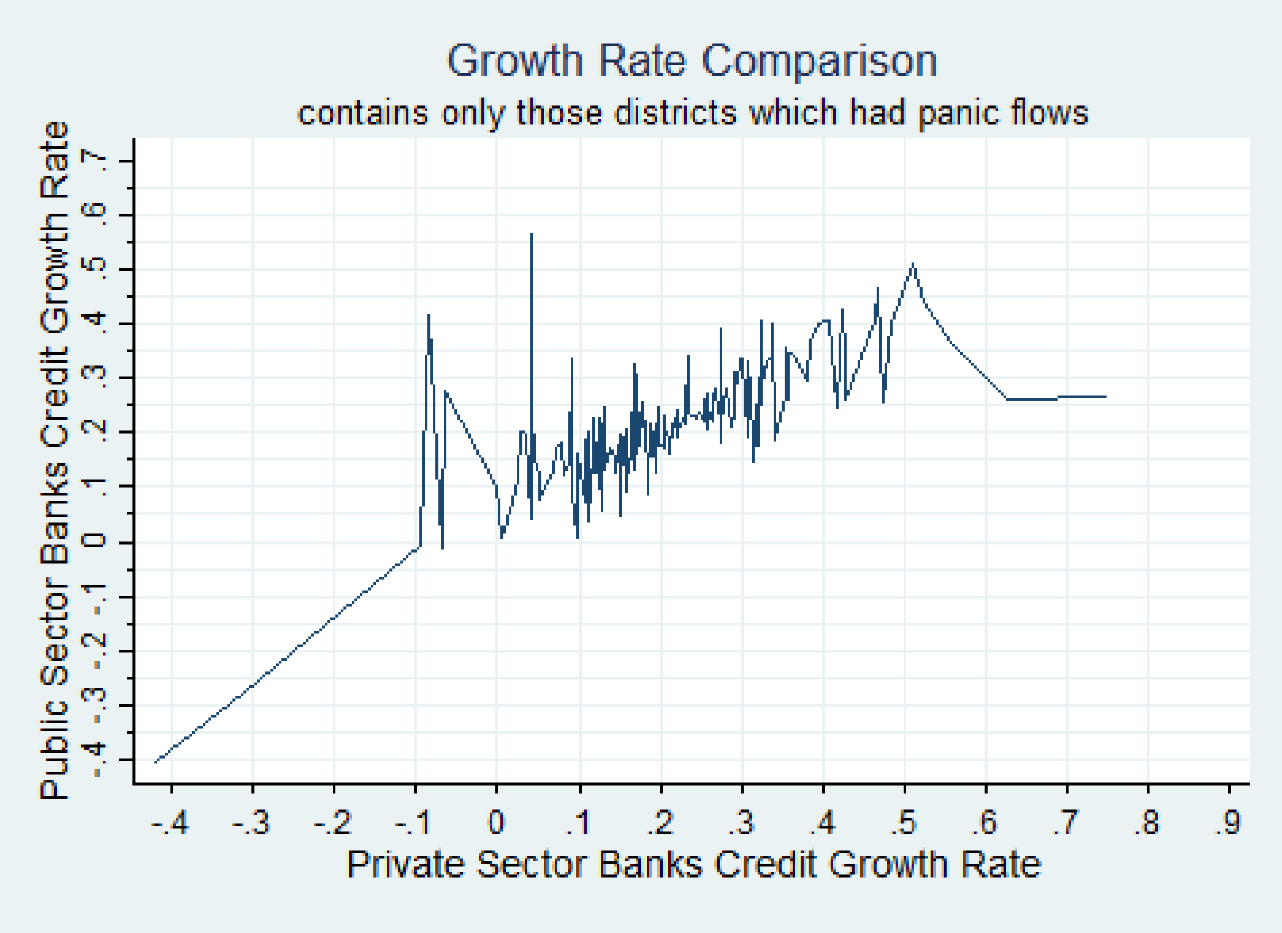
Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
deposi..	346	.2703212	.0149231	.277586	.2409695	.299673
deposi..	346	.2291859	.0069659	.1295733	.2154849	.2428869
diff	346	.0411353	.0144323	.2684569	.0127489	.0695217

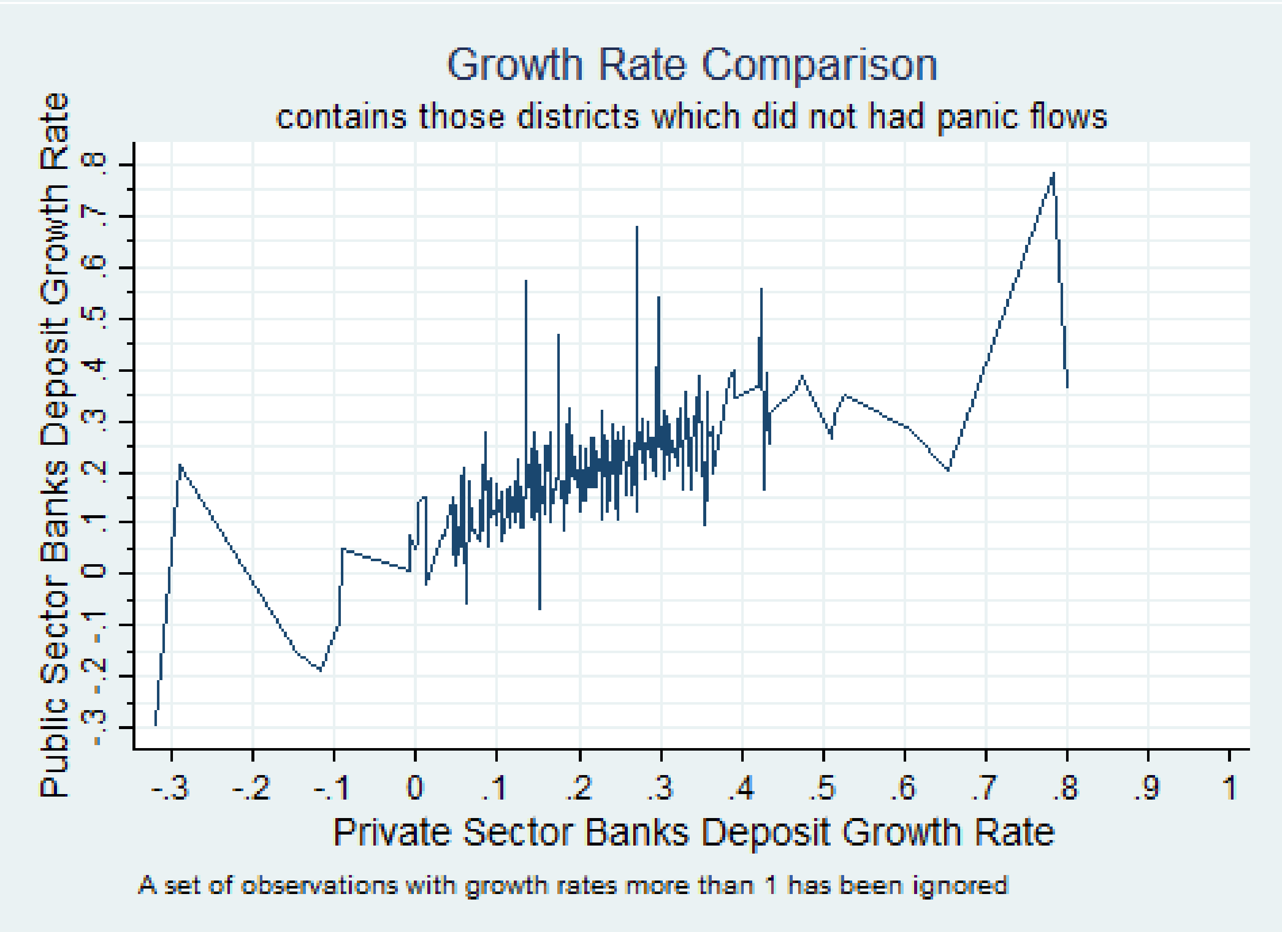
```
     mean(diff) = mean(deposit_privat~e - deposit_public~e)        t =   2.8502
 Ho: mean(diff) = 0                               degrees of freedom =      345

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9977         Pr(|T| > |t|) = 0.0046          Pr(T > t) = 0.0023
```

Regarding the part c of this subproblem. I am little confused in it. As subparts of the problem are bullets, not a and b. Also its not clear what is the meaning of the difference in the problem. However by guess it seems the empirical strategy is like a difference-in-difference approach.

g)

The answer to part d for this subproblem is as follows


Growth Rate Comparison
contains only those districts which had panic flows
Public Sector Banks Credit Growth Rate
-.4 -.3 -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7
-.4 -.3 -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7 .8 .9
Private Sector Banks Credit Growth Rate


Growth Rate Comparison
contains those districts which did not had panic flows
Public Sector Banks Deposit Growth Rate
-.3 -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7 .8
-.3 -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1
Private Sector Banks Deposit Growth Rate
A set of observations with growth rates more than 1 has been ignored

Due to reason given in part(e) of this sub-problem, the heatmaps could not be produced.

Regression Model

Assumptions:

1. Since the variable that is modelled is Credit Growth over the entire horizon of 5 Quarters. It is assumed that this problem is of cross-sectional type rather than panel data or time series type.
2. Controls are assumed to be of the state in which the district belongs to, bank type (PSB or Private) and panic flow.
3. Since the private bank data are not availble for 112 districts. They have been ignored from the analysis.
4. The effect quarter-wise variable has been ignored as the varible that is predicted is cumulative growth of entire 5 quarters. So variables such as deposit-private-2008-9Q4 or deposit-public-2009-10Q4 effect has been ignored. It can also be be argued that growth-rate-deposit contains the effect all these quarter wise variable

The Regression model is as follows:

$$g_{credit} = \beta_0 + \beta_1 * g_{deposit} + \beta_2 * banktype + \beta_3 * state + \beta_4 * PF + \epsilon \tag{13}$$

The results with robustness has been shown as follows

```
. reg growth_rate_credit_ growth_rate_deposit_ banktype state panic_flow, robust

Linear regression                               Number of obs     =      1,038
                                                F(4, 1033)        =       5.83
                                                Prob > F          =     0.0001
                                                R-squared         =     0.4974
                                                Root MSE          =     .20728

--------------------------------------------------------------------------------------
                     |               Robust
 growth_rate_credit_ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------+----------------------------------------------------------------
growth_rate_deposit_ |   1.103328   .4592264     2.40   0.016     .2022049    2.004451
            banktype |  -.0116212   .0115096    -1.01   0.313    -.0342061    .0109638
               state |    -.00078    .000879    -0.89   0.375    -.0025049    .0009449
          panic_flow |   .1737974   .0708557     2.45   0.014     .0347599    .3128349
               _cons |  -.0309936   .1173615    -0.26   0.792    -.2612877    .1993006
--------------------------------------------------------------------------------------
```

Analysis:

1. growth-rate-deposit strongly explains the credit growth rate. It's effect is as expected. It is significant at 5% level of significance.
2. panic-flow(PF) is significant meaning that for districts with $PF = 0$ are qualitatively different from districts with $PF = 1$. Due to inclusion of PF, the *state* becomes insignificant. Bur regressions without PF made *state* significant , it implies some degree of correlation between PF and *state*

```
. reg growth_rate_credit_ growth_rate_deposit_ banktype state , robust

Linear regression                               Number of obs     =      1,038
                                                F(3, 1034)        =       6.27
                                                Prob > F          =     0.0003
                                                R-squared         =     0.4319
                                                Root MSE          =     .22026
```

growth_rate_credit_	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
growth_rate_deposit_	.9492039	.4685279	2.03	0.043	.0298299	1.868578
banktype	-.0127697	.0120239	-1.06	0.288	-.0363637	.0108243
state	-.0023815	.0006148	-3.87	0.000	-.0035879	-.0011751
_cons	.0849955	.0838928	1.01	0.311	-.0796241	.249615

3. *banktype* is not significant in any regression form. It means that the impact of bank type is not important in explaining the credit growth rate. It may be case that after all the effect of *banktype* got diluted after 5 quarters. It can also be argued that initial distrust of private bank is counter balanced by government commitment to regulating economy or by inherit better ability of private banks to manage their funds as compared to public banks.