# Problem 7: Scotia Bank

*Alessandro Selvitella*

*August 16, 2016*

### The General Problem:

How to combine two relative rankings of credit risk into one ranking?

In the wholesale space, Banks rates each individual borrower and gives a credit rating in a way similar to rating agencies. The internal credit rating plays an important role in the credit risk management framework.

Credit rating is defined as the relative ranking of default risk for a borrower of the Bank. The assignment of a relative ranking is industry-specific.

Each industry uses a rating model internally built by the Bank. A rating from the model is the starting point of rating assessment. The model rating usually will be further adjusted based on experienced credit judgement after considering borrower's past performance and forecast of the future performance. When assigning the rating both borrower specific factors (e.g. level of debt, profitability and so on) and macro factors (e.g. Industry, geography and so on) are considered.

For an example, we rank borrowers in the Oil and Gas industry into 16 buckets and also rank borrowers in the Financial Services industry into 16 buckets. How can we combine these two relative rankings into one relative ranking scale so that the relative ranking is now considering borrowers in both industries?

## First Problem to Address:

Are the two distributions the same distributions?

The answer is basically obvious, but I think it should be proved right away and not think about it anymore.

Each industry has several companies and each company gets ranked monthly, so each company has its own time series. It is not completely easy to compare such type of data.

## First step: Static "oversimplified" Problem.

A possibility is find a common value for the time series and so give a unique rank for all the history of the time series.

The first possibility is to give the average rank, but one could choose, the median, the last rank and so on...

We will discuss the PROs and CONs of this at the end of the section.

In this way, we have two univariate samples, one for Oil and Gas and the other for Financial and they can be compared both parametrically and non parametrically.

We start with the non parametric tests.

**Answer: The Kolmogorov-Smirnov Test**

The Kolmogorov-Smirnov Test may be used to test whether two underlying one-dimensional probability distributions differ. The Kolmogorov-Smirnov statistic is

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively.

The null hypothesis is rejected at level $\alpha\alpha$ if

$$D_{n,n'} > c(\alpha)\sqrt{\frac{n+n'}{nn'}}..$$

Here $n$ and $n'$ are the sizes of first and second sample respectively.

The value of $c(\alpha)$ is given in the table below for each level of $\alpha$:

| $\alpha$ | $c(\alpha)$ |
|---|---|
| 0.10 | 1.22 |
| 0.05 | 1.36 |
| 0.025 | 1.48 |
| 0.01 | 1.63 |
| 0.005 | 1.73 |
| 0.001 | 1.95 |

Remark

Note that the two-sample test checks whether the two data samples come from the same distribution. This does not specify what that common distribution is (e.g. whether it's normal or not normal).

## The Data Analysis

In this section we use the Kolmogorov Smirnov test, to perform the statistical analysis described above.

We upload the OIL and GAS data

```
ScotiaBank_Data_OilGas<-read.csv("/Users/selvit/Desktop/IPSW 2016/Scotia Bank/ScotiaBank_Data_OilGas.cs

dim(ScotiaBank_Data_OilGas)
```

```
## [1] 1967  129
```

```
ScotiaBank_Data_OilGas<-data.matrix(ScotiaBank_Data_OilGas)
```

We compute the means of the OilGas scoring for each company.

```
means_ScotiaBank_Data_OilGas<-rep.int(0,length(ScotiaBank_Data_OilGas[,1]))
length(means_ScotiaBank_Data_OilGas)
```

```
## [1] 1967
```

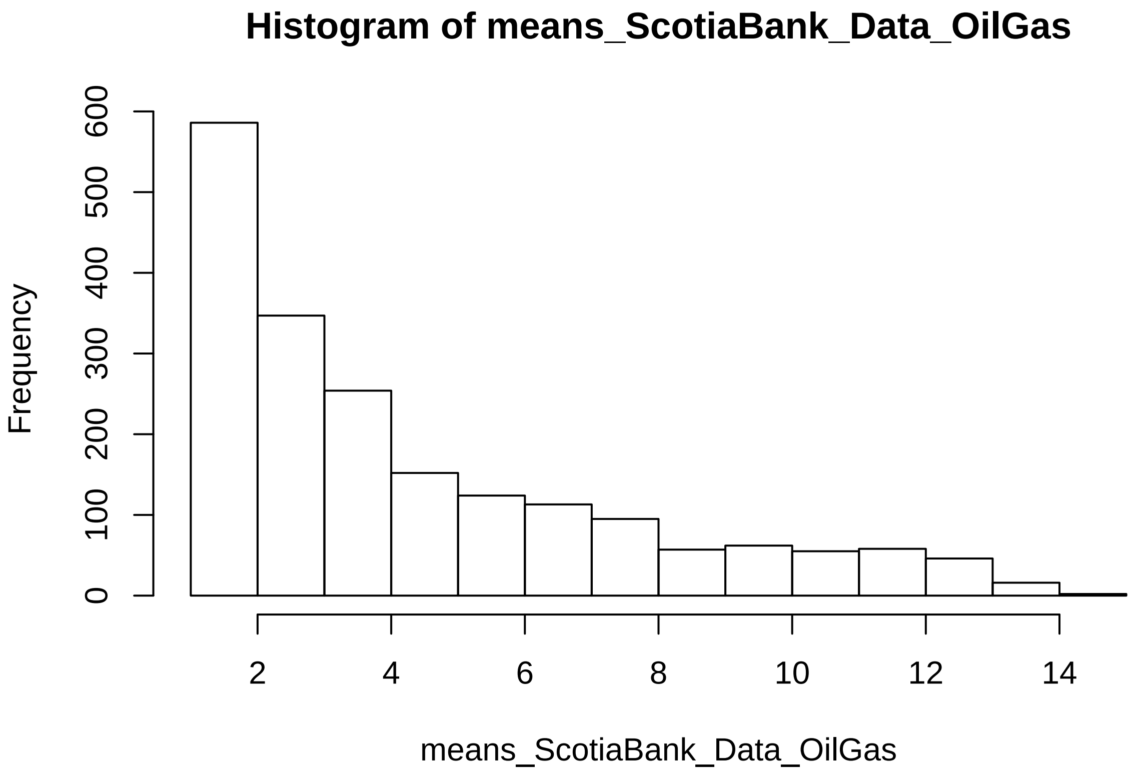
```
for (i in 1:length(ScotiaBank_Data_OilGas[,1])) {
  means_ScotiaBank_Data_OilGas[i]<- mean(ScotiaBank_Data_OilGas[i,3:129])
}
#means_ScotiaBank_Data_OilGas

length(means_ScotiaBank_Data_OilGas)
```

```
## [1] 1967
```

```
#View(means_ScotiaBank_Data_OilGas)
```

We plot a histogram with the means of the OilGas scoring for each company.

```
hist(means_ScotiaBank_Data_OilGas)
```



#Financial data

```
ScotiaBank_Data_Financial<-read.csv("/Users/selvit/Desktop/IPSW 2016/Scotia Bank/ScotiaBank_Data_Financ
#View(ScotiaBank_Data_Financial)
dim(ScotiaBank_Data_Financial)
```

```
## [1] 8848  129
```

```
ScotiaBank_Data_Financial<-data.matrix(ScotiaBank_Data_Financial)
```

We compute the means of the Financial scoring for each company.

```
means_ScotiaBank_Data_Financial<-rep.int(0,length(ScotiaBank_Data_Financial[,1]))
length(means_ScotiaBank_Data_Financial)
```

```
## [1] 8848
```

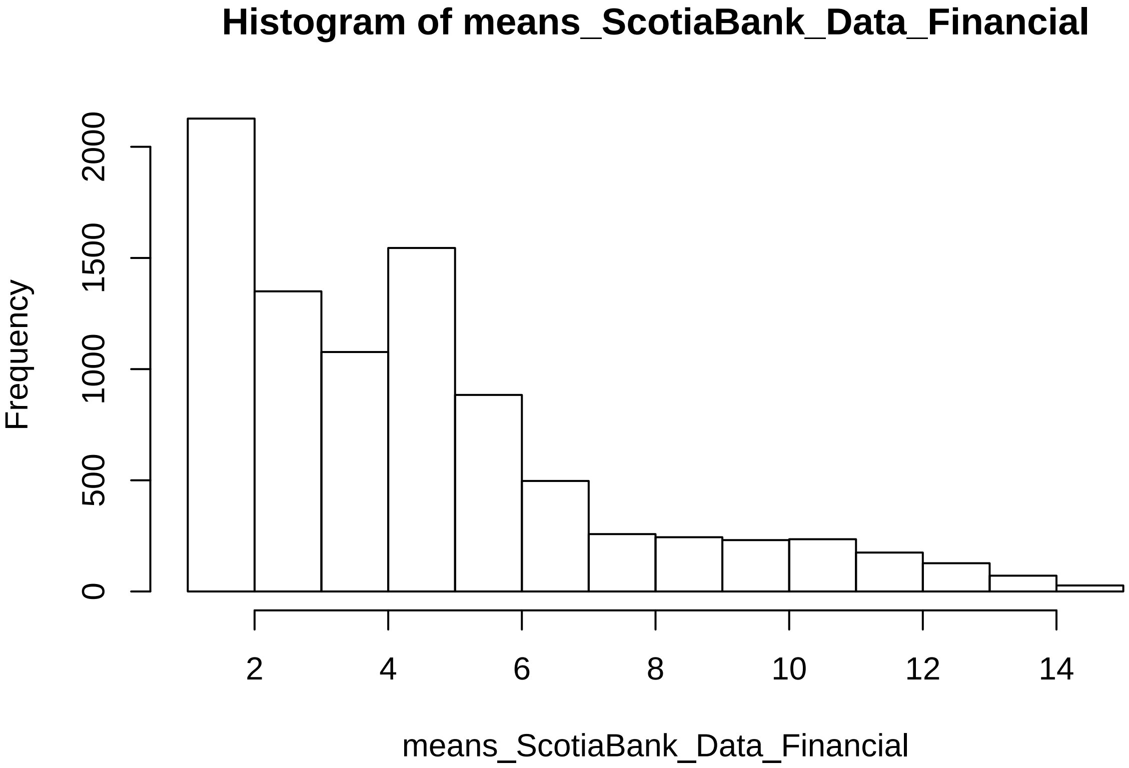
```
for (i in 1:length(ScotiaBank_Data_Financial[,1])) {
  means_ScotiaBank_Data_Financial[i]<- mean(ScotiaBank_Data_Financial[i,3:129])
}
#means_ScotiaBank_Data_Financial

length(means_ScotiaBank_Data_Financial)
```

```
## [1] 8848
```

```
#View(means_ScotiaBank_Data_Financial)
```

We plot a histogram with the means of the Financial scoring for each company.

```
hist(means_ScotiaBank_Data_Financial)
```



## Kolmogorov Smirnov Test for two samples

```
ks.test(means_ScotiaBank_Data_OilGas,means_ScotiaBank_Data_Financial)
```

```
## Warning in ks.test(means_ScotiaBank_Data_OilGas,
## means_ScotiaBank_Data_Financial): p-value will be approximate in the
## presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  means_ScotiaBank_Data_OilGas and means_ScotiaBank_Data_Financial
## D = 0.093024, p-value = 1.605e-12
## alternative hypothesis: two-sided
```

The two sided Kolmogorov Smirnov Test rejects the null hypotheses and so the two distributions are different.

But: who has higher scores significantly?

## Signed Kolmogorov Smirnov: statistical significance not good

The reference:

“The signed Kolmogorov-Smirnov test: why it should not be used” by Guillaume J Filion Gigascience. 2015; 4: 9.

explains that (from the abstract) “he two-sample Kolmogorov-Smirnov (KS) test is often used to decide whether two random samples have the same statistical distribution. A popular modification of the KS test is to use a signed version of the KS statistic to infer whether the values of one sample are statistically larger than the values of the other. The underlying hypotheses of the KS test are intrinsically incompatible with this approach and the test can produce false positives supported by extremely low p-values. This potentially makes the signed KS test a tool of p-hacking, which should be discouraged by replacing it with standard tests such as the t-test and by providing confidence intervals instead of p-values.”

## Mann–Whitney U test: statistical significance

There are several other tests, non parametric that can say if the distributions are the same or not and test where the mean is shifted. As an example, we use the Mann-Whitney test.

```
wilcox.test(means_ScotiaBank_Data_OilGas,means_ScotiaBank_Data_Financial)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  means_ScotiaBank_Data_OilGas and means_ScotiaBank_Data_Financial
## W = 8167700, p-value = 1.986e-05
## alternative hypothesis: true location shift is not equal to 0
```

## Mann–Whitney U test: statistical bigger, TRUE

```
wilcox.test(means_ScotiaBank_Data_OilGas,means_ScotiaBank_Data_Financial, alternative="greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  means_ScotiaBank_Data_OilGas and means_ScotiaBank_Data_Financial
## W = 8167700, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

## Mann–Whitney U test: statistical smallerer, FALSE

```
wilcox.test(means_ScotiaBank_Data_OilGas,means_ScotiaBank_Data_Financial, alternative="less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  means_ScotiaBank_Data_OilGas and means_ScotiaBank_Data_Financial
## W = 8167700, p-value = 9.932e-06
## alternative hypothesis: true location shift is less than 0
```

These tests say that Oil and Gas tend to have higher rank than Financial.

## PROs and CONs of this approach:

- +does the job of comparing the two samples
- +simple
- +rank tend to be constant in time
- +test statistics already computed for us with respective p-values
- -does not take care of the evolution of the data
- -does not do anything more than comparing the history mean of the two samples non-parametrically and say if the distributions are the same, not that much. It's almost obvious.

## What to do next?

- We have done Non-parametric: K-S, sK-S, M-W U Test
- Parametric: Beta Distribution? Exponential, Truncated Exponential?
- Static–>Dynamic, more complicated... Let's see
- Here I haven't done prediction.