Итоги работы второго задания

Алияров Р., Байрамкулов А., Полушкин А.

МГУ имени М. В. Ломоносова

Москва, 2017

Основные положения разведовательного анализа данных

Разведовательный анализ данных направлен на выявление основных характеристик и суммирование их с целью построения некоторых гипотез. Для удобства чаще всего используют визуализацию, по которой достаточно просто построить гипотезы.

Методы, используемые для разведовательного анализа данных

- Использование пространственно-некогерентного света ртутной лампы;
- Применение RAW-конверторов, использование полного динамического диапазона цифровой фотокамеры;
- Усовершенствование линейных алгоритмов деконволюции;
- Многовариантные графики;

Этапы анализа

Первый этап - анализ первичных данных

Рассмотрим данные по производству:

```
  productionData <- read.csv('C:\\Users\\apolushkin\\Documents\\ppp-2017\\ППП\\tasks\\task2\\production-data.csv')
  productionDataHarpy <- subset(productionData, productionData$supplier == 'harpy.co')
  productionDataWesteros <- subset(productionData, productionData$supplier != 'harpy.co')
```

Harpy:

```
  head(productionDataHarpy)
```

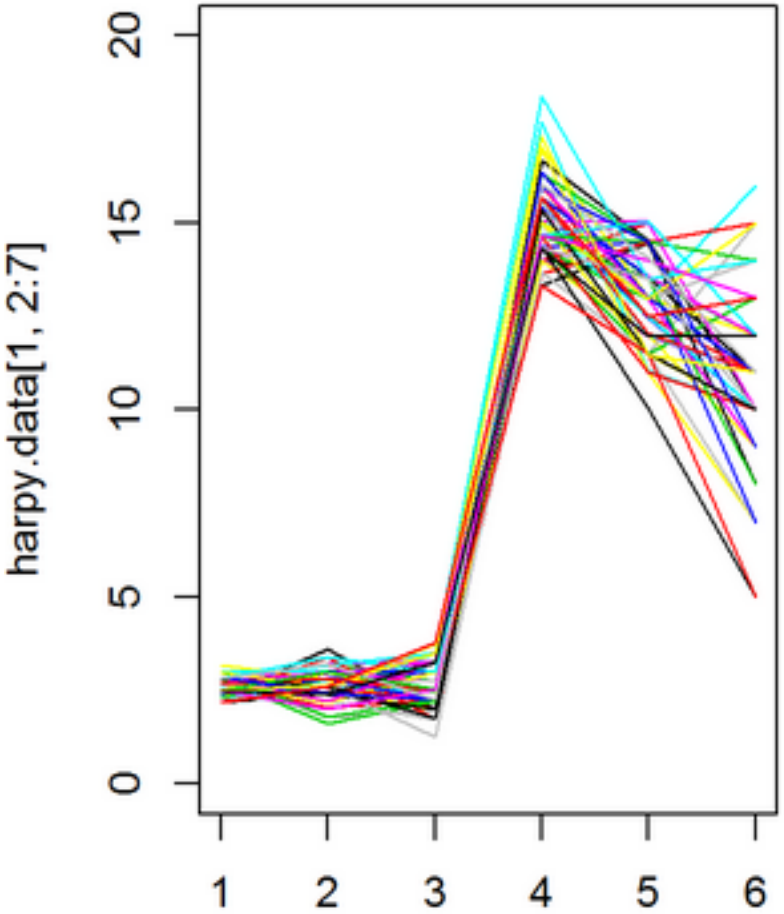
```
##   unsullen.id production.date report.date produced defects supplier
## 1           1               1           1      103       0 harpy.co
## 2           1               1           2        0       2 harpy.co
## 3           1               1           3        0       4 harpy.co
## 4           1               1           4        0       5 harpy.co
## 5           1               1           5        0      13 harpy.co
## 6           1               1           6        0      11 harpy.co
```

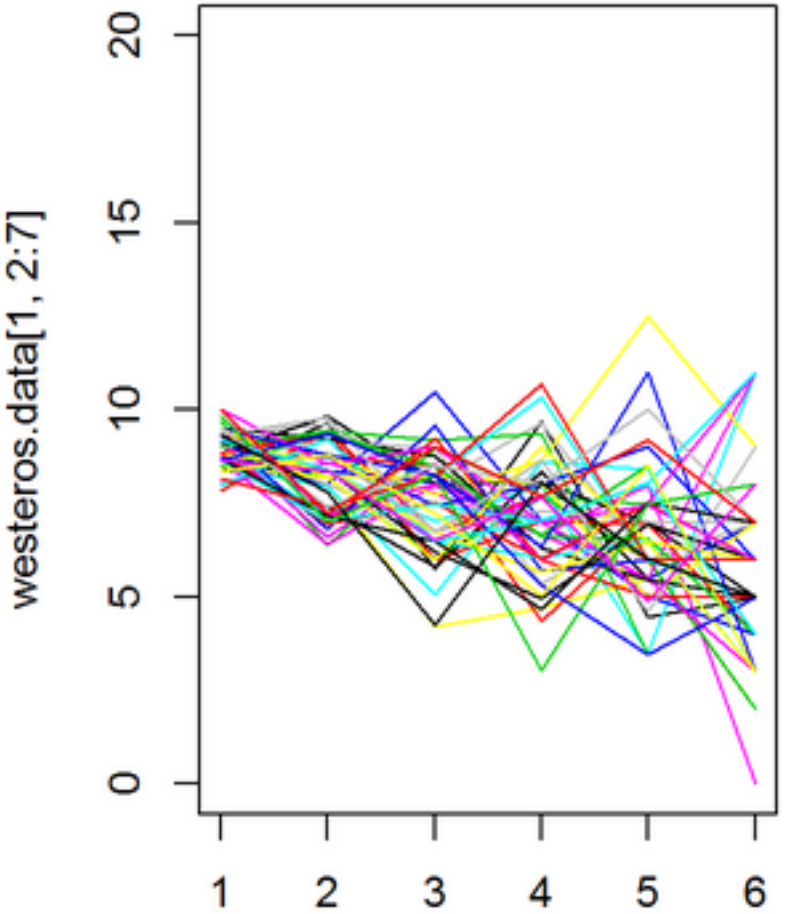
Westeros:

```
  head(productionDataWesteros)
```

```
##      unsullen.id production.date report.date produced defects     supplier
## 1351          51               1           1      106       0 westeros.inc
## 1352          51               1           2        0       8 westeros.inc
## 1353          51               1           3        0       9 westeros.inc
## 1354          51               1           4        0       8 westeros.inc
## 1355          51               1           5        0      12 westeros.inc
## 1356          51               1           6        0       0 westeros.inc
```

Все дефекты распределены по партиям, произведенным каждым кузнецом. Рассмотрим распределение дефектов по каждому кузнецу на первый месяц после производства, на второй и так далее.





Как видно из графиков harpy имеет большую дисперсию, а среднее имеет схожие черты с левой частью графика функции плотности нормального распределения. График westeros имеет меньшую дисперсию, а среднее похоже на убывающую линейную функцию $y = kx+b$.

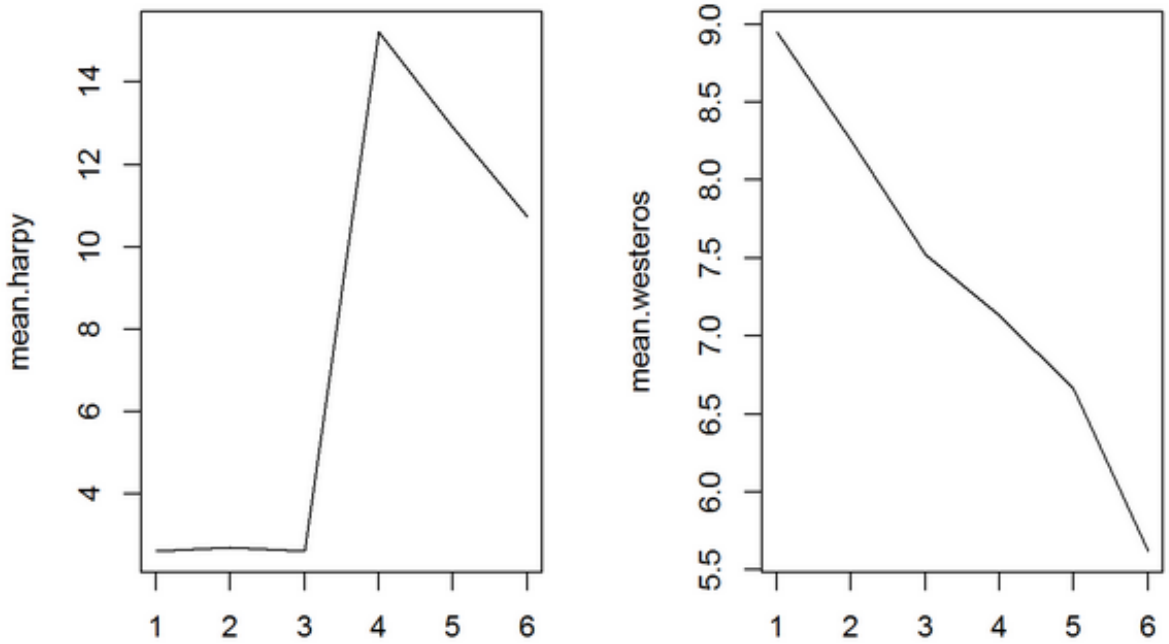
Этапы анализа

Второй этап - оценки для анализа поставок стали каждой компанией

1. Построим график среднего для полученных выше распределений:

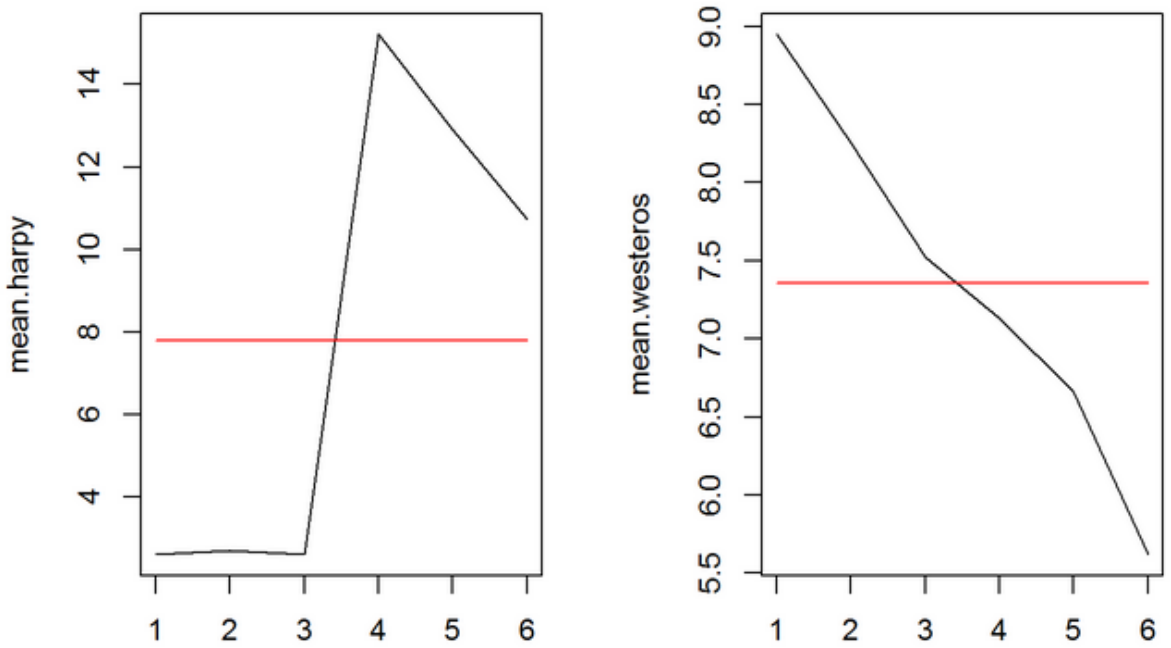
```
mean.harpy <- c()
mean.westeros <- c()
for (i in 1:6) {
  mean.harpy[i] <- mean(harpy.data[, i + 1])
  mean.westeros[i] <- mean(westeros.data[, i + 1])
}

par(mfrow=c(1,2))
plot(x = 1:6, y = mean.harpy, type = "l")
plot(x = 1:6, y = mean.westeros, type = "l")
```



2. Добавим оценку математического ожидания для полученого выше среднего:

```
par(mfrow=c(1,2))
plot(x = 1:6, y = mean.harpy, type = "l")
lines(x = 1:6, y = rep(mean(mean.harpy), 6), type = "l", col = "red")
plot(x = 1:6, y = mean.westeros, type = "l")
lines(x = 1:6, y = rep(mean(mean.westeros), 6), type = "l", col = "red")
```



Из графиков видно, что оценка математического ожидания дефектов для компании Harpy незначительно выше, чем для компании Westeros:

```
mean(mean.harpy)
```

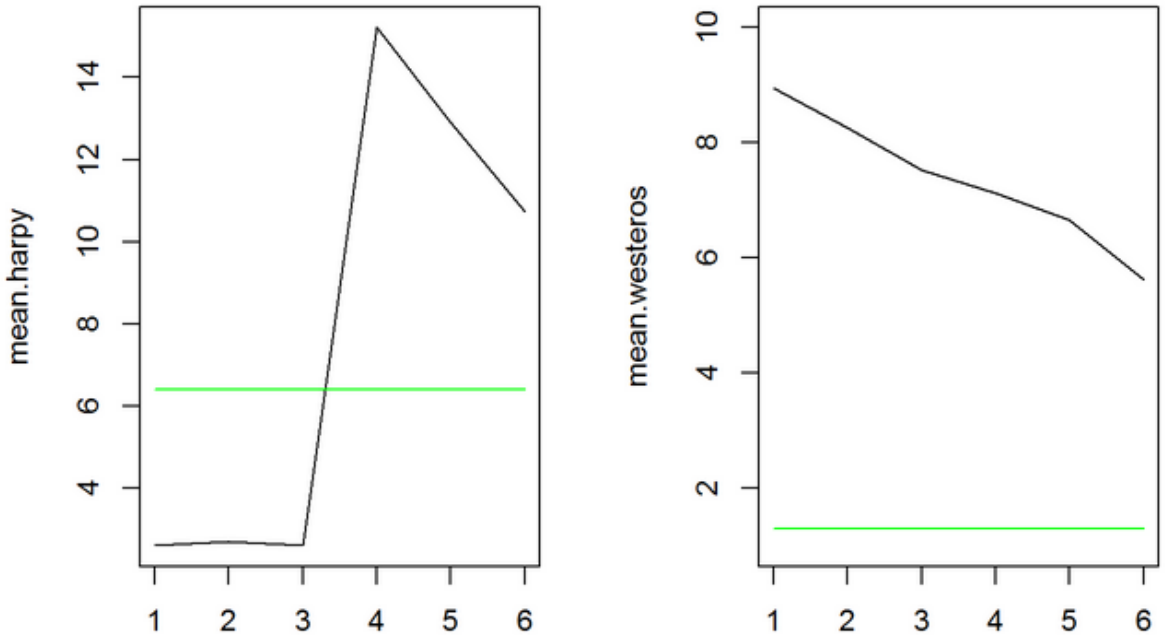
```
## [1] 7.791396
```

```
mean(mean.westeros)
```

```
## [1] 7.357553
```

3. Рассмотрим в качестве оценки показатель “выборочная дисперсия” - это оценка теоретической дисперсии распределения, рассчитанная на основе данных выборки. Для сохранения размерности, возьмем корень из дисперсии, то есть среднеквадратичное отклонение.

```
par(mfrow=c(1,2))
plot(x = 1:6, y = mean.harpy, type = "l")
lines(x = 1:6, y = rep(sd(mean.harpy) * (6 / 5) ^ 0.5, 6), type = "l", col = "green")
plot(x = 1:6, y = mean.westeros, type = "l", ylim = c(1, 10))
lines(x = 1:6, y = rep(sd(mean.westeros) * (6 / 5) ^ 0.5, 6), type = "l", col = "green")
```



Выборочная дисперсия

Пусть $X_1, \ldots, X_n, \ldots$ - выборка из распределения вероятности. Тогда

- выборочная дисперсия — это случайная величина $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$ где символ $\bar{X}$ обозначает выборочное среднее;
- несмещённая (исправленная) дисперсия — это случайная величина $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Очевидно, $S^2 = \frac{n}{n-1}S_n^2$

Из графиков видно, что оценка среднеквадратичного отклонения распределения дефектов для компании Harpy значительно выше, чем для компании Westeros:

```
sd(mean.harpy) * (6 / 5) ^ 0.5
```

```
## [1] 6.385113
```

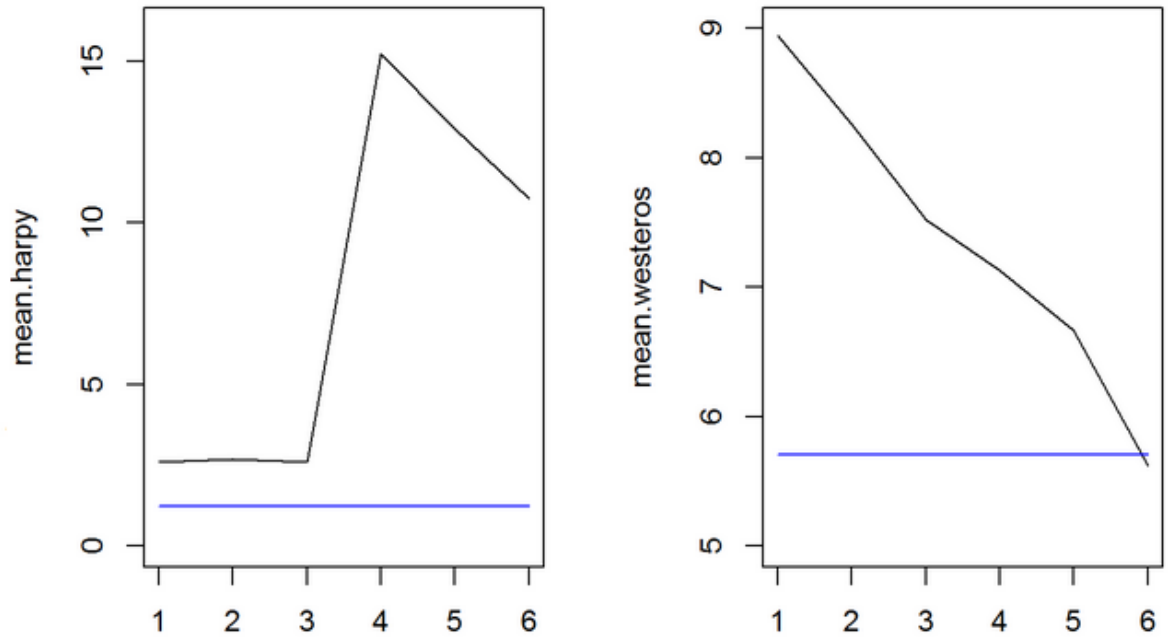
```
sd(mean.westeros) * (6 / 5) ^ 0.5
```

```
## [1] 1.288848
```

4. Итоговым показателем рассмотрим коэффициент Шарпа:

```
par(mfrow=c(1,2))

plot(x = 1:6, y = mean.harpy, type = "l", ylim = c(0, 16))
lines(x = 1:6, y = rep(mean(mean.harpy) / (sd(mean.harpy) * (6 / 5) ^ 0.5), 6), type = "l", col = "blue")
plot(x = 1:6, y = mean.westeros, type = "l",ylim = c(5, 9))
lines(x = 1:6, y = rep(mean(mean.westeros) / (sd(mean.westeros) * (6 / 5) ^ 0.5), 6), type = "l", col = "blue")
```



Коэффициент Шарпа

Показатель эффективности инвестиционного портфеля (актива), который вычисляется как отношение средней премии за риск к среднему отклонению портфеля.

Из графиков видно, что коэффициент Шарпа распределения дефектов для компании Harpy значительно ниже, чем для компании Westeros:

```
mean(mean.harpy) / (sd(mean.harpy) * (6 / 5) ^ 0.5)
```

```
## [1] 1.220244
```

```
mean(mean.westeros) / (sd(mean.westeros) * (6 / 5) ^ 0.5)
```

```
## [1] 5.708628
```

Вывод

Таким образом, компания Westeros лидирует по всем трем оценкам, причем столь большая разница по оценке среднеквадратичного отклонения и коэффициенту Шарпа с довольно большой вероятностью не приблизится к 0 за 11 месяцев. Поэтому по результатам разведывательного анализа следует выбрать компанию Westeros.