Отчет по 2-му заданию

Масло Михаил, Лазарев Владимир, Подкина Александра

МГУ имени М. В. Ломоносова

Москва, 2017

Введение

В отличие от традиционной проверки гипотез, предназначенной для проверки априорных предположений, касающихся связей между переменными (например, "Имеется положительная корреляция между возрастом человека и его/ее нежеланием рисковать"), разведывательный анализ данных (РАД) применяется для нахождения связей между переменными в ситуациях, когда отсутствуют (или недостаточны) априорные представления о природе этих связей. Как правило, при разведочном анализе учитывается и сравнивается большое число переменных, а для поиска закономерностей используются самые разные методы.

Типичные техники, используемые в РАД

- Диаграмма размаха(Boxplot)
- Гистограмма
- Многовариантный график
- Диаграмма "стебель-листья"и др.

В качестве исходных данных мы имеем информацию о производстве оружия и количестве единиц сломанного оружия за каждый месяц каждым из кузнецов. Верхние строки этой таблицы выглядят так:

	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
7	1	1	7	0	11	harpy.co
8	1	2	2	105	0	harpy.co
9	1	2	3	0	2	harpy.co
10	1	2	4	0	2	harpy.co
11	1	2	5	0	2	harpy.co
12	1	2	6	0	13	harpy.co

В качестве показателя, на основании которого будет приниматься итоговое решение о выборе компании, возьмем процент дефектных орудий для каждой произведенной партии. Покажем как выглядят первые строки получившихся таблиц для двух компаний.

	V1	V2	V3	V4	V5	V6
average_harpy_defects	2.173776	3.005650	2.983568	14.02532	13.52404	11
average_harpy_defects	2.670400	3.218810	1.992718	14.00968	14.99522	10
average_harpy_defects	2.655395	2.576923	2.502439	14.33117	12.98058	9
average_harpy_defects	2.661927	2.425331	2.501182	15.40895	12.47867	10
average_harpy_defects	2.652997	2.403774	3.485849	17.71519	11.03774	10
average_harpy_defects	3.009524	3.007678	2.519139	14.03481	11.96190	11
average_harpy_defects	2.677316	2.584762	1.947619	14.00314	12.00000	9
average_harpy_defects	2.669316	2.804554	3.516432	15.01238	11.49541	15
average_harpy_defects	2.329635	3.607211	1.995249	16.67508	14.54286	10
average_harpy_defects	2.170846	2.802632	2.754098	16.00000	12.49767	11
average_harpy_defects	2.677994	3.222437	2.745721	15.65909	13.52174	12

Анализ
Таблица с процентами дефектных орудий

Каждый столбец отвечает за конкретный месяц после производства.
Каждая строка соответствует одному кузнецу.

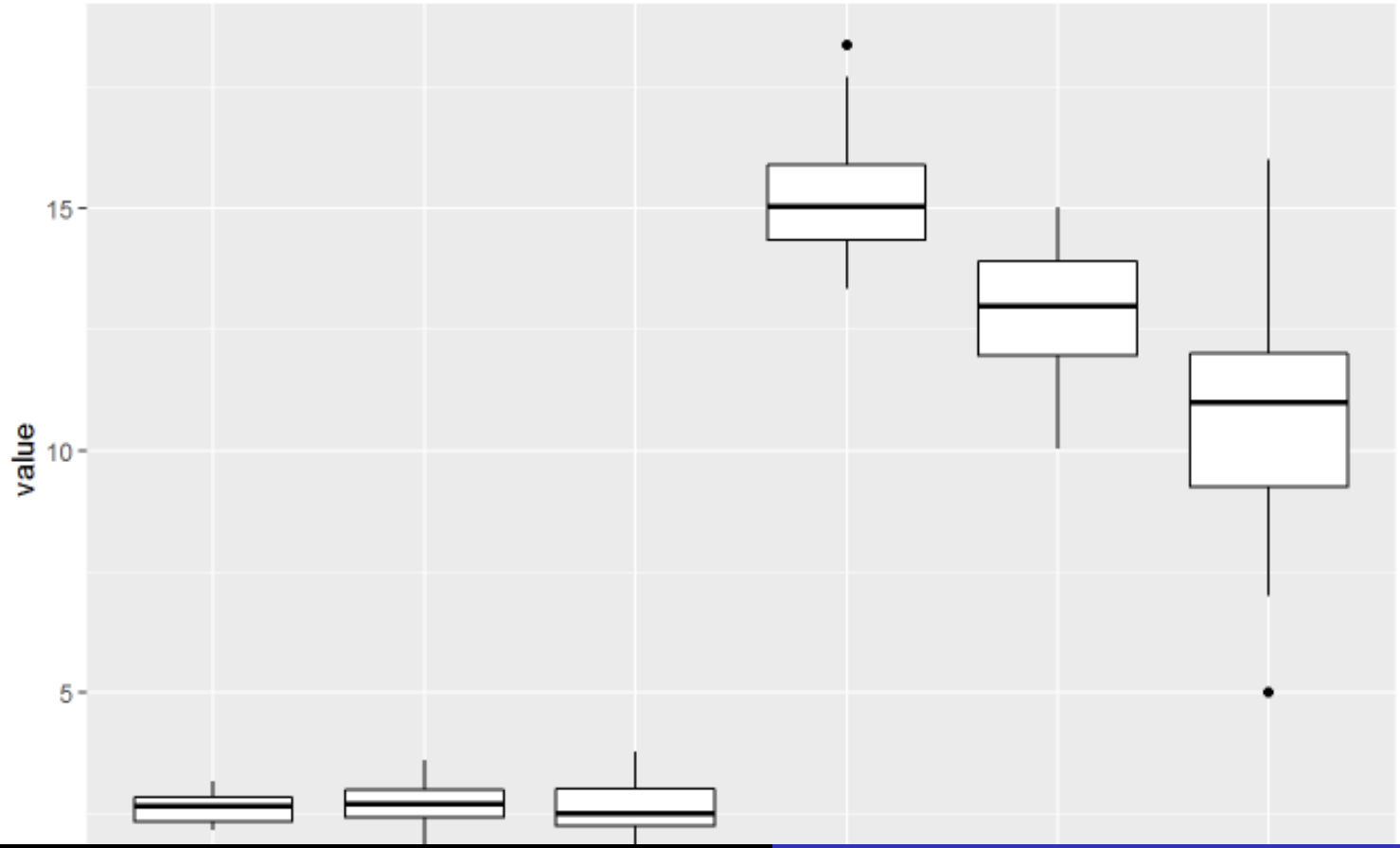
	V1	V2	V3	V4	V5	V6
average_westeros_defects	8.324238	9.639692	5.755396	9.689873	4.457143	5
average_westeros_defects	7.824089	9.423664	6.725962	7.680511	5.990291	2
average_westeros_defects	8.675969	8.202980	6.009346	7.658307	7.471963	6
average_westeros_defects	8.345313	8.388679	8.035377	6.990566	5.502326	7
average_westeros_defects	9.338608	8.388994	7.004717	7.965300	3.478873	11
average_westeros_defects	8.004724	8.589888	7.764706	6.655063	7.990521	0
average_westeros_defects	9.331190	7.988417	4.214458	4.690096	5.480952	3
average_westeros_defects	9.362205	9.621013	8.000000	9.663551	6.004739	7
average_westeros_defects	8.851911	9.848197	8.007075	8.101587	6.990291	5
average_westeros_defects	9.343849	8.782197	8.498812	4.344828	7.000000	6
average_westeros_defects	9.692429	6.396584	8.988124	7.679365	6.014634	2
average_westeros_defects	8.656786	8.025926	10.488372	7.028037	5.041475	4
average_westeros_defects	8.505547	7.962264	5.056075	8.623457	8.401869	3

Наш анализ будет строиться на основе диаграммы размаха(boxplot). Дадим формализацию этого понятия.

Диаграмма размаха

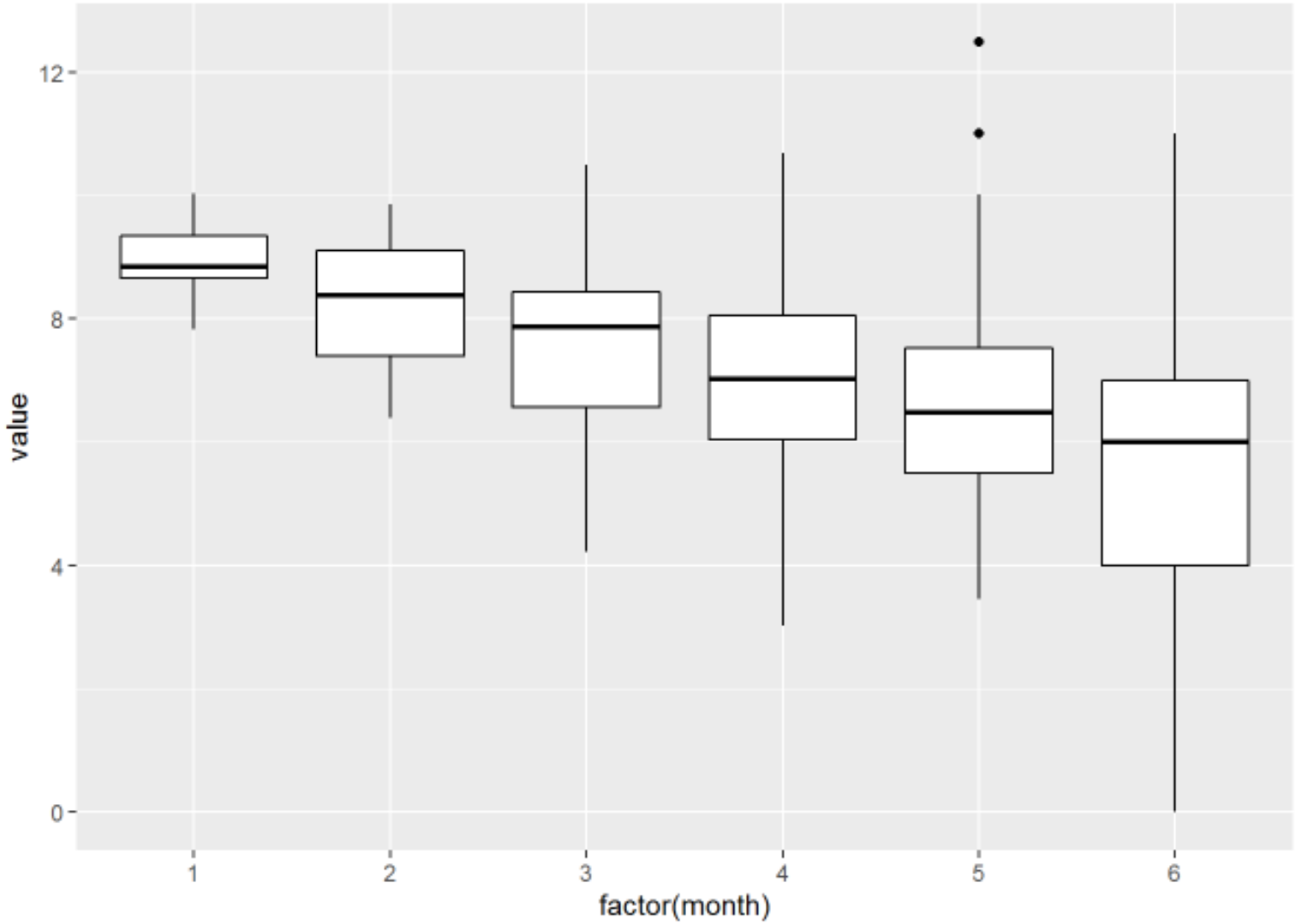
Диаграмма размаха - график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей. Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких диаграмм можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

Изобразим наши получившиеся данные в виде boxplot графиков, чтобы более наглядно увидеть различия в компаниях. Сначала изобразим boxplot для Harpy и дадим некоторые комментарии увиденному.



Первые 3 месяца показывают относительно идентичную динамику, т.е. дефекты более менее одинаковые(различие наблюдаются в разнице между high и low значениями, high - low имеет линейную динамику роста). Но на 4-й месяц происходит резкое изменение, что говорит о том, что в 4-м месяце достигается глобальный максимум поломки орудий. Далее, наблюдается линейное убывание дефектов по среднему значению.

Анализ
Диаграмма размаха Westeros



Здесь мы видим следующее. Дефекты напоминают линейную функцию, убывающую с ростом месяца. В данном случае и среднее значение, и дисперсия показывает линейную динамику, что является очень привлекательным фактом для прогноза процентов дефекта на будущее. Простая линейная регрессия вполне может дать валидные результаты, чего нельзя сказать с уверенностью про компанию Harpy.

Обе компании имеют как положительные, так и отрицательные стороны. Компания Harpy показывает гораздо более низкую дисперсию по проценту дефекта в первые 3 месяца по сравнению с компанией Westeros. Это позволяет нам лучше прогнозировать поломки орудий на первые 3 месяца после производства партии. Но наш горизонт планирования составляет 11 месяцев, т.е. значительно больше 3 месяцев. На этом временном промежутке преимущество по дисперсии стирается, так как на 4-6 месяцы, судя по имеющимся результатам, дисперсия возрастает и становится вполне сравнимой с компанией Westeros. Компания Westeros в свою очередь имеет линейную динамику как по среднему значению, так и по дисперсии, что свидетельствует о большей вероятности сохранения “тренда”. Это хорошее подспорье для валидности прогноза на наш горизонт планирования. В качестве минуса Westeros можно отметить большое расстояние между high и low значениями в каждый месяц после производства. Несмотря на это привлекательность компании Westeros выше Harpy, поэтому контракт выгоднее заключить с Westeros.