

Модуль 7: Анализ данных

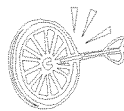
 © 2007. The World Bank Group. All rights reserved.



Цели обучения

В конце семинара участники должны освоить:

- основные концепции анализа данных
- взаимосвязь между типами данных, типами выборки и методами анализа данных
- способы использования результатов анализа данных в процессе мониторинга



2

World Bank Institute Evaluation Group



Стратегия анализа данных

Основной выбор:

- Количественный анализ
- Качественный анализ

3

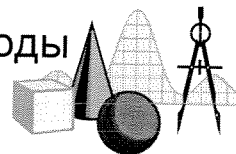
World Bank Institute Evaluation Group



Количественный анализ

Три основных вида

- Описательные методы
- Ассоциативные методы
- Детерминистические методы



4

World Bank Institute Evaluation Group



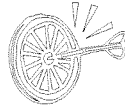
Переменные величины

Переменная величина – это характеристика или свойство, которое варьируется или меняется во времени или среди людей или групп

Примеры: возраст, пол, сельскохозяйственное производство, мили мощеных дорог, количество детей, которые не доедают, гектары национальных парков и т.д.

Независимая переменная величина: вмешательство или объяснимая переменная величина

Зависимая переменная величина: то, что мы ожидаем, изменится в результате изменений в независимой переменной величине



5

World Bank Institute Evaluation Group



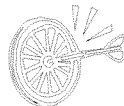
Примеры

Независимая величина: образование

Зависимая величина: доход

Независимая величина: доступ к услугам квалифицированных акушерок

Зависимая величина: материнская смертность



6

World Bank Institute Evaluation Group



Типы переменных величин

Дискретные и непрерывные:

- **дискретные** переменные величины измеряются единицами, которые не могут быть разделены, например, количество книг на моей полке, количество детей в школе
- **непрерывные** переменные величины измеряются единицами, которые могут быть разделены, например, температура, время

Номинальные и порядковые:

- **номинальные** (категориальные) переменные величины разделяют по категориям, например, мужчина, женщина, женат, разведен
- **порядковые** переменные величины приписывают названия каждой возможной категории ответа, но категории могут варьироваться, например, уровень удовлетворенности от тренинга (от неудовлетворительного до удовлетворительного)

Интервальные/коэффициентные переменные величины:

- **интервальная** шкала использует эквидистантное измерение, но в котором ноль не является значащим (например, цельсий)
- **коэффициентная**, где ноль означающая величина (например, ноль означает отсутствие того, что измеряется) например, доход, годы школы, уровень рождаемости, километры мощеной дороги

7

World Bank Institute Evaluation Group



Количественные описательные методы

Применяются к одной переменной величине

- **Частота/Процент Распределения**
 - График или таблица, которая показывает, как часто каждое значение или ряд значений переменной появляются в данных.
- **Центральная тенденция**
 - Измерение расположения середины или центра распределения.
 - Центральная тенденция может относиться к **Среднему значению, Медиане, Моде**
- **Дисперсия**
 - Описывает насколько наблюдения варьируются от центральной тенденции.
 - Разброс и Стандартное отклонение

8

World Bank Institute Evaluation Group



Частота Распределения

Сколько мужчин и женщин в программе?

Распределение респондентов

Пол	Число	%
Мужчина	100	33%
Женщина	200	67%
Итого	300	100%

Вывод: Из 300 человек в этой программе, 67 процентов женщин и 33 процента мужчин.

9

World Bank Institute Evaluation Group



Описание распределения

Центральная тенденция:

- Каковы типичные характеристики?
- Пример: Каков средний возраст выпускников?

Дисперсия :

- Насколько непохожими или концентрированными являются характеристики?
- Пример: Каково колебание в возрасте?

10

World Bank Institute Evaluation Group



Показатели центральной тенденции

Три показателя: Мода, медиана и среднее значение

Мода: наиболее часто встречающийся ответ

Медиана: срединная точка распределения

Среднее значение: среднее арифметическое

11

World Bank Institute Evaluation Group



Показатели центральной тенденции: количество транспортных средств в час

9	31	
17	34	Мода =
19	38	Медиана =
23	41	Средн. значение =
23	151	
28	Сумма = 414	

12

World Bank Institute Evaluation Group



Измерение расширения: разброс и стандартное отклонение

Разброс: разница между наибольшим и наименьшим значением

Стандартное отклонение: измеряет распределение значений – удаление от среднего показателя

- Небольшое стандартное отклонение: небольшое распределение; большинство данных или значений близки к среднему показателю
- Большое стандартное отклонение: большинство данных или значений далеки от среднего показателя

13

World Bank Institute Evaluation Group



Показатели дисперсии: сколько часов в месяц проводится перед телевизором

	11	3	
	16	4	Какое
	18	6	распределение
	19	12	имеет большее
	<u>21</u>	<u>60</u>	стандартное
Сумма =	85	85	отклонение?
Ср. Знач. =	17	17	
Медиана =	?	?	

14

World Bank Institute Evaluation Group



Стандартное (среднеквадратическое) отклонение

По мере увеличения отклонения увеличивается также дисперсия – потенциально на величину отличную от данных в распределении, которую мы исследуем.

Мы только возводим в квадрат отклонения, чтобы сумма не равнялась нулю (свойство среднего значения). Теперь когда мы имеем ненулевое число, извлекаем квадратный корень, чтобы получить статистическую величину, которая снова в метрической системе, с которой мы начинали:

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}, \text{ так что } s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

s – стандартное (среднеквадратическое) отклонение выборки.

15

World Bank Institute Evaluation Group



Дисперсия

Дисперсия: $(X_i - \bar{X})$

Один из способов измерения всех отклонений – сложить отклонения по каждому наблюдению: $\sum (X_i - \bar{X})$

Однако, мы знаем, что одним из особых свойств среднего значения есть то, что оно всегда равно нулю, так что оно не имеет особого практического значения для понимания дисперсии.

Возводя в квадрат каждое отклонение, мы можем получить величину, которая всегда будет положительной.

Если мы разделим сумму на количество наблюдений, мы получим *дисперсию*.

$$\frac{\sum (X_i - \mu)^2}{N} = \text{дисперсия распределения совокупности} = \sigma^2$$

$$\frac{\sum (X_i - \bar{X})^2}{n-1} = \text{дисперсия выборочного распределения} = s^2$$

Примечание: помните, что N – объем совокупности, а n – объем выборки. Позже мы более детально рассмотрим, почему $n-1$ в знаменателе выборочного среднеквадратического отклонения, а не среднеквадратического отклонения генеральной совокупности.

Рассматривайте дисперсию как среднеквадратическое отклонение от среднего значения.

16

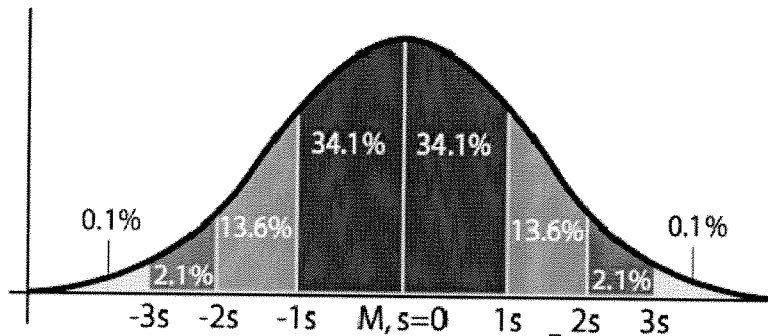
World Bank Institute Evaluation Group



Измерение дисперсии: среднеквадратическое отклонение

Нормальное распределение: колоколообразная кривая

- 68,26% вариации в пределах 1 среднеквадратического отклонения от среднего значения
- 95,44% вариации в пределах 2 среднеквадратических отклонений от среднего значения



Примечание:
масштабированию не подлежит

World Bank Institute Evaluation Group



Описательная статистика

Применяется к 2 или более переменным

- Сравнение средних значений
- Сводная таблица

18

World Bank Institute Evaluation Group



Сравнение средних значений

Зарабатывают ли мужчины больше женщин?

Или обуславливает ли пол отличия в уровне доходов?

Независимая переменная=пол

Зависимая переменная=доход

Пол	Средний доход
Мужчины	\$924
Женщины	\$798

19

World Bank Institute Evaluation Group



Сводная таблица

•Применяется при работе с номинальными и ординальными данными.

•Может использоваться с данными интервалов/отношений, которые были распределены по категориям.

Учителя	Внедрили новые методы обучения	Не внедрили новые методы обучения	Всего
Прошли 1-недельное обучение по применению современных педагогических методов в учебной работе	25%	75%	100%
Не участвовали в обучении, но получили пособие по применению современных педагогических методов в учебной работе	15%	85%	100%

Интерпретация:

■ Более вероятно, что учителя, которых обучили современным педагогическим методам (25%), будут внедрять современные педагогические методы в учебной работе по сравнению с теми учителями, которые только получили пособия по современным педагогическим методам (15%).

■ Похоже, что существует взаимосвязь, но насколько она сильна?

20

World Bank Institute Evaluation Group



Ассоциативная статистика

Сила и направление

Насколько сильной является взаимосвязь?

- Несколько вариантов показателей взаимосвязи
- Некоторые показатели связи лежат в пределах от 0 до 1
- Другие колеблются в пределах от -1 до +1

Взаимосвязь не доказывает наличие причинно-следственных отношений!

21

World Bank Institute Evaluation Group



Установление причинной связи

Причинная связь: В оценках влияния, наша основная задача определить **связь** причины и следствия в изучаемом явлении

Три фактора, необходимые для определения причинно-следственной связи:

1. Причина должна предшествовать результату. Изменения в независимых переменных величинах должны появляться до изменений в зависимых величинах.
2. Причина и результат действия должны быть взаимосвязанными (то есть, коррелированными).
3. Другие объяснения причинно-следственной связи должны быть исключены (то есть, исключите ложные или опровергающие факторы)

22

World Bank Institute Evaluation Group



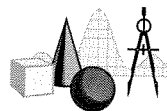
Детерминированные методы

Детерминированные статистические методы могут применяться для того, чтобы на основе описательных методов анализа исследовать причинные связи.

Типы... • Двумерная (простая) регрессия

• Многомерная (множественная) регрессия

• Подходы, использующие категориальные данные (н-р, логическая регрессия)



23

World Bank Institute Evaluation Group



Детерминированные методы

Модель двумерной регрессии:

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i, \quad i = 1, \dots, n$$

- X_1 , - независимая переменная (регрессор)
- β_0 = неизвестный отрезок совокупности
- β_1 = влияние изменения в X_1 на Y
- u = "вектор ошибок" (упущенные факторы)

Модель множественной регрессии совокупности:

Рассмотрим случай двух регрессоров:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, \dots, n$$

- β_1 = влияние на Y изменения в X_1 , сохраняя X_2 константу
- β_2 = влияние на Y изменения в X_2 , сохраняя константу X_1

24

World Bank Institute Evaluation Group



Анализ данных в планах мониторинга

- При анализе данных в ходе мониторинга больше используются основные методы
- Необходимо рассмотреть:
 - Использование подмножеств данных
 - Необходимость в сравнениях

25

World Bank Institute Evaluation Group



Определите подмножества данных в Планах мониторинга

Будьте осторожными, чтобы данные не были чрезмерно агрегированными!!!

Данные по каждому показателю конечного результата должны быть разбиты (дезагрегированы) для того, чтобы показать конечные результаты для различных подгрупп или частей.

26

World Bank Institute Evaluation Group



Возможные подмножества данных для мониторинга: демографические характеристики

Демографические характеристики

- По уровню дохода домохозяйства (или его замещающему показателю)
- По полу
- По возрастной группе
- По расовой/этнической принадлежности
- По географическому расположению, как например, село в сравнении с городом, по району, по муниципальному делению

27

World Bank Institute Evaluation Group



Возможные подмножества данных для мониторинга: характеристики услуг

Характеристики услуг

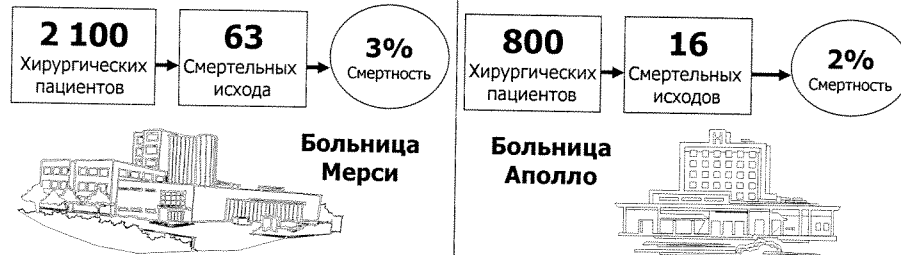
- По организационной единице, если услуга оказывается более чем в одном объекте (как, например, разных медицинских учреждениях, школах, парках, водных объектах или районах)
- По типу процедуры, используемой поставщиком услуги
- Объему или уровню услуг
- По потребностям потребителя

28

World Bank Institute Evaluation Group



Обсуждение: какую больницу вы бы выбрали?



29

World Bank Institute Evaluation Group



Обсуждение: Какую больницу вы бы выбрали?



Но...



30

World Bank Institute Evaluation Group



Сравнения для интерпретации данных конечного результата в целях мониторинга

- В разное время (прошлый год в сравнении с текущим годом или текущий месяц в сравнении с этим же месяцем прошлого года)
- В сравнении с целями, установленными ведомством
- По разным демографическим характеристикам потребителей
- По разным характеристикам оказания услуги
- С другими подобными программами
- С другими городами, странами или регионами

31

World Bank Institute Evaluation Group



Качественный анализ данных

- Данные из отчетов,
- Нерегламентированных интервью, фокус групп,
- Неструктурированных наблюдений
- Методы анализа
 - ✓ Индуктивный анализ
 - ✓ Логический анализ
 - ✓ Синтез

32

World Bank Institute Evaluation Group



Качественный анализ данных

- Данные из отчетов, нерегламентированных интервью, фокус-групп, неструктурированных наблюдений
- Определите общие слова, идеи, темы
- Составьте таблицу или делайте записи на карточках
- Отмечайте, где они размещаются
- Подберите «подходящие для цитирования цитаты»

33

World Bank Institute Evaluation Group



Качественный анализ данных

Главный риск: смещение

- Сложно признать то, что вы не ожидаете

Пусть анализ проведет кто-то еще

- Сравните результаты
- Выясните, чем обусловлены различия

34

World Bank Institute Evaluation Group



Качественный анализ данных

Запись результатов

- Отрадите основные темы
- «Некоторые участники заявили...»
- Подчеркните интересные точки зрения, даже если их выразили один или два человека
- Не пытайтесь указывать количество или процентные показатели

35

World Bank Institute Evaluation Group



План оценки

Конкретные проекты

- Для каждого вопроса или подвопроса в Плане оценки, определите оптимальные стратегии анализа для ответа на каждый вопрос или подвопрос.
- Включите информацию в графу «Анализ данных»

36

World Bank Institute Evaluation Group



План оценки

Общие вопросы	Конкретные под-вопросы	Вид/тип вопроса	Вид/тип дизайна	Показатели и измерения	Источник и данные	Сбор данных и выборка	

37

World Bank Institute Evaluation Group



Групповое задание

Ваш проект – для оценки

- Для каждого вопроса или подвопроса, определите оптимальную стратегию анализа

38

World Bank Institute Evaluation Group