since 2017

Олимпиада
по финансовой грамотности
и финансовой культуре

Экономический факультет МГУ имени М.В. Ломоносова

мои финансы

финансовая грамотность в вузах
Федеральный сетевой методический центр

Тематический вебинар

Как работать с данными социологических исследований

Виктор Аркадьевич Брызгалин
Экономический факультет
МГУ имени М.В. Ломоносова

fingramota.econ.msu.ru/olymp/

План

1. Введение в количественную социологию и опросы
2. Принципы построения индексов на основе социологических данных
3. Анализ устойчивости индексов

План

1. **Введение в количественную социологию и опросы**
2. Принципы построения индексов на основе социологических данных
3. Анализ устойчивости индексов

Где количественная социология используется в «экономических» целях?

1. Для получения данных об экономических явлениях, которые затруднительно получить другими способами
   Пример: уровень безработицы, структура расходов д/х, структура сбережений д/х и т.д.
2. Для определения возможных факторов поведения человека
   Пример: оценка ценностей и поведенческих установок людей
3. Маркетинг (предпочтения потребителей и т.д.)

Какие данные лучше: опросные или статистические?

Предположим, мы хотим ответить на вопрос об уровне преступности в определенной стране (как часто совершаются там преступления разного рода). Что лучше – провести опрос или воспользоваться данными о зарегистрированных преступлениях от полиции?

Аргументы «ЗА» опросных данных

- Могут быть проведены независимыми организациями
- Учитывают и те преступления, когда заявители на дошли до полиции
- Учитывают и те преступления, когда полиция не стала регистрировать обращение
- Могут дать информацию о любых видах преступлений (в том числе тех, по которым не ведется официальная статистика)

Аргументы «ЗА» полицейских данных

- Можно получить информацию о некоторых типов преступлений, которые недоступны опросными методами (убийства, преступления по отношению к детям и фирмам)
- Нет необходимости обеспечивать репрезентативность выборки
- Нет проблем с ошибками в датировке преступлений (и «забыванием» старых)
- Учитывают преступления против иностранных граждан
- Нет «завышения» количества преступлений, если жертвами стало несколько человек

Опросные данные могут дать то, что не может предоставить официальная статистика – но они же имеют и свои недостатки

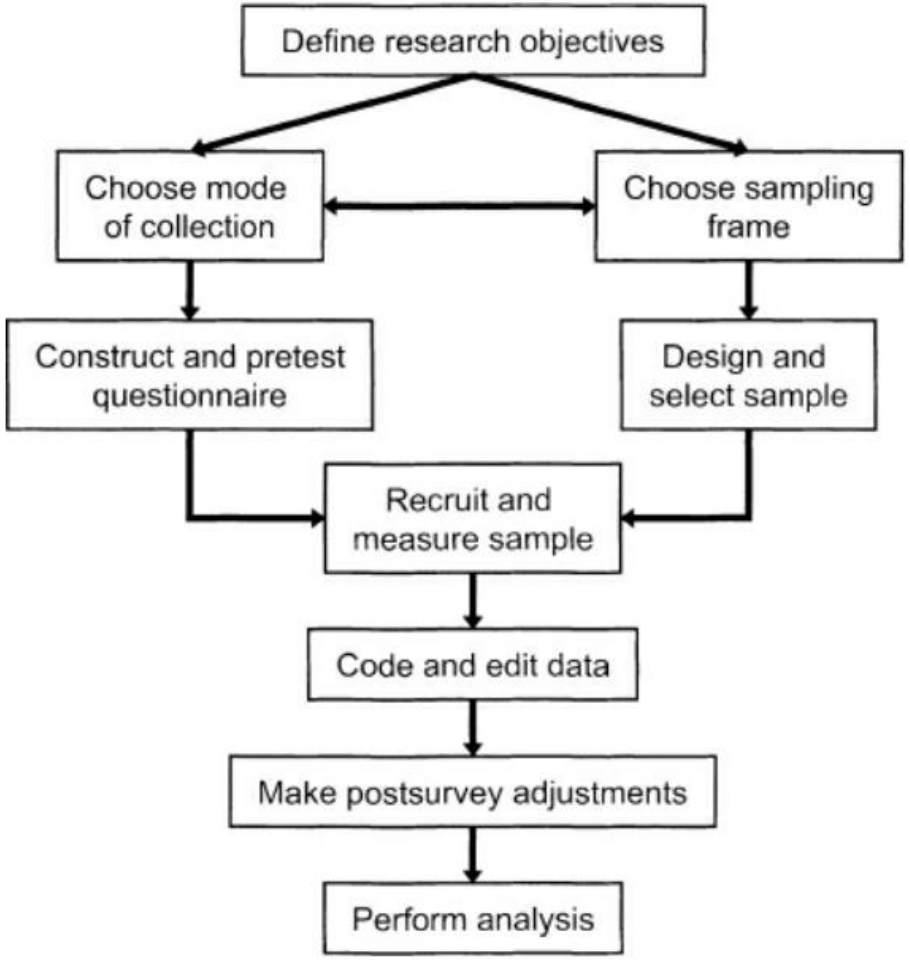
Виды опросов

- Кросс-секционные (однократные)
- Повторяемые кросс-секционные (регулярно повторяются на одной анкете, но на разных выборках)
- Лонгитюдные (одна и та же группа людей регулярно опрашивается)

Кросс-секционные vs лонгитюдные опросы

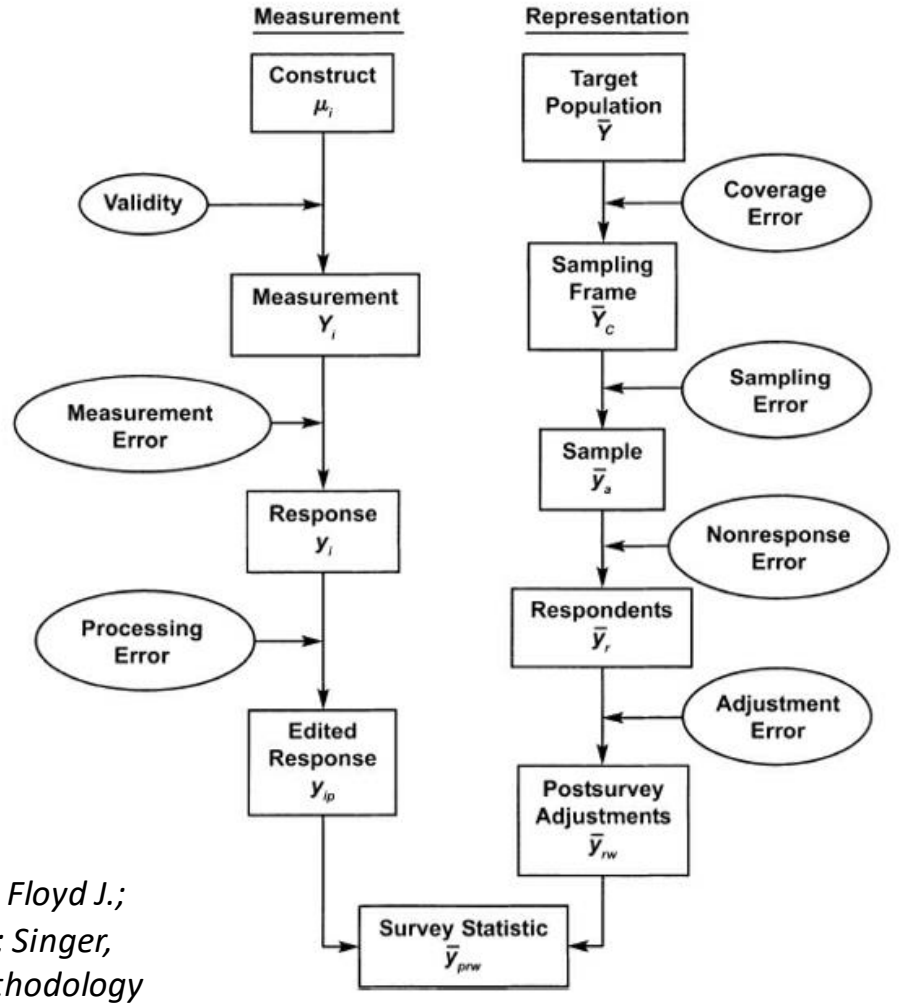
	Плюсы	Минусы
Кросс-секционные	• Учитывает временные изменения в структуре выборки • Большая легкость в сборе данных	• Меньше возможностей для каузального анализа • Влияние структуры выборки • Смешение возрастных и поколенческих эффектов
Лонгитюдные	• Контроль индивидуальных различий • Анализ динамики • Контроль поколенческих эффектов	• «Осыпание» панели • Старение панели • Трудность доступа

Процесс проведения опроса



Источник: Groves, Robert M.; Fowler, Floyd J.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor; Tourangeau, Roger. Survey Methodology

Виды ошибок, которые возникают при проведении опроса



Источник: Groves, Robert M.; Fowler, Floyd J.; Couper, Mick P.; Lepkowski, James M.; Singer, Eleanor; Tourangeau, Roger. Survey Methodology

Понятие репрезентативности

- **Репрезентативность** — соответствие характеристик выборки характеристикам популяции или генеральной совокупности в целом.
- Выборка репрезентативна, если в ходе исследования удалось свести к минимуму все соответствующие данному направлению ошибки (покрытия, выборки, отклика и корректировки)
- «Выборка репрезентативна по <название характеристики>» означает, что выборка отражает генеральную совокупность по указанным характеристикам

Зачастую термин «репрезентативна» используется для больших выборок или для выборок, в которых в нужной пропорции присутствуют целевые для исследования группы. Однако это не является признаками репрезентативности!

Направление «Репрезентативность»

Основная задача – как корректно сформировать выборку для опроса?

Основные этапы и ошибки:

- Этап 1: формирование фрейма выборки.
  Основная ошибка: ошибка покрытия (coverage error)
- Этап 2. формирование выборки.
  Основная ошибка: ошибка выборки (sampling error)
- Этап 3. проведение опроса.
  Основная ошибка: ошибка отклика (nonresponse error)
- Этап 4. корректировка данных («ремонт выборки»)
  Основная ошибка: ошибка корректировки (adjustment error)

Направление «Репрезентативность»

Основная задача – как корректно сформировать выборку для опроса?

Основные этапы и ошибки:

- Этап 1: формирование фрейма выборки.
  Основная ошибка: ошибка покрытия (coverage error)
- **Этап 2. формирование выборки.**
  Основная ошибка: ошибка выборки (sampling error)
- Этап 3. проведение опроса.
  Основная ошибка: ошибка отклика (nonresponse error)
- Этап 4. корректировка данных («ремонт выборки»)
  Основная ошибка: ошибка корректировки (adjustment error)

Ошибка выборки

- Два вида ошибки выборки: ***ошибка смещения*** и ***ошибка вариации***
- ***Ошибка смещения*** возникает, когда некоторые элементы выборочного фрейма не имели шанса попасть в выборку (или их шансы попасть были меньше, чем у других элементов)
- ***Ошибка вариации*** возникает из-за того, что есть вероятность, что взятая выборка случайным образом будет плохо представлять генеральную совокупность (из-за большого разброса)

Что такое ошибка выборки?

Ошибка выборки: максимальное значение, на которое результаты выборки (с определенной вероятностью) отличаются от результатов генеральной совокупности

Простой способ расчета:

$$\text{Ошибка выборки} = 1{,}95 * \sqrt{\frac{p*(1-p)}{N}}$$

Где p – ожидаемая пропорция , N – размер выборки. Так как ожидаемая пропорция заранее неизвестная, нужно ориентироваться на максимальный размер выборки, который достигается при p = 0,5.

Для расчета доверительных интервалов по конкретным группам/вопросам можно воспользоваться стандартной формулой:

$$\bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{(n)}}$$

Направление «Репрезентативность»

Основная задача – как корректно сформировать выборку для опроса?

Основные этапы и ошибки:

- Этап 1: формирование фрейма выборки.
  Основная ошибка: ошибка покрытия (coverage error)
- Этап 2. формирование выборки.
  Основная ошибка: ошибка выборки (sampling error)
- Этап 3. проведение опроса.
  Основная ошибка: ошибка отклика (nonresponse error)
- **Этап 4. корректировка данных («ремонт выборки»)**
  Основная ошибка: ошибка корректировки (adjustment error)

Взвешивание («ремонт выборки»)

Процедура взвешивания («ремонта выборки») – увеличение роли недопредставленных групп и уменьшение – перепредставленных в выборке.

$$\text{Вес группы} = \frac{\text{Доля группы в генеральной совокупности}}{\text{Доля группы в выборочной совокупности}}$$

- Если целевые группы, которые необходимо взвесить, построены на пересечении нескольких признаков (пол, возраст, образование и т.д.) и определение их доли в генеральной совокупности невозможно, используются специальные процедуры подбора весов (например, rim weighting)
- Обычно исследовательские базы содержат переменную, в которой уже находятся рассчитанные веса для взвешивания
- Пакет в R, который позволяет делать расчет значений с учетом взвешивания – svydesign.

Направление «Измерение»

- **Этап 1. Разработка анкеты**
  Основная ошибка: ошибка измерения (measurement error)
- Этап 2. Обработка данных
  Основная ошибка: ошибка обработки (processing error)

Выбор метода опроса: плюсы и минусы

	Личные интервью (face to face)	Телефонник (CATI)	Онлайн-опросы
Плюсы	• Высокий охват населения • Возможность опросить по длинной и сложной анкете	• Почти 100%-ный охват населения • Относительная дешевизна	• Дешевизна • Возможность опросить по длинной и сложной анкете
Минусы	• Дороговизна • Отсутствие охвата труднодоступных групп	• Ограничение по времени опроса • Риски смещения из-за отказов	• Не полный охват населения • Серьезные риски самоотбора

Направление «Измерение»

- Этап 1. Разработка анкеты
  Основная ошибка: ошибка измерения (measurement error)
- **Этап 2. Обработка данных**
  Основная ошибка: ошибка обработки (processing error)

Проверка корректности заполненных анкет

- Отсутствие «дублирующихся» строк (наблюдений с одинаковыми ответами);
- Отсутствие строк без дисперсии в разных вопросах (например, на серии вопросов со шкалой, везде стоит одна цифра);
- Отсутствуют наблюдения с экстремальными значениями (например, возраст 150 лет, высшее образование в возрасте 18 лет);
- Наличие вопросов «ловушек»: схожих по смыслу и вариантам ответа вопросов, расположенных в разных частях анкеты.

Работа с пропущенными значениями

Варианты работы с пропущенными значениями:

1. Удаление наблюдений с пропущенными значениями.
2. Подстановка среднего значения
3. Подстановка модального значения
4. Подстановка прогнозных значений (на основе регрессионного анализа)
5. Исключение из анализа вопроса с большим количеством пропущенных значений

Главный критерий при выборе метода работы с пропущенными значениями – минимизация смещения

Вызовы интерпретации ответов в социологических исследованиях

1. Специфика выборки (ошибка покрытия, ошибка выборки, ошибка отклика, ошибка корректировка)
2. Влияние типа опроса и анкеты (ограниченная сопоставимость разных типов опросов)
3. Влияние формулировки вопросов
4. Фальсификация предпочтений
5. Страновая несопоставимость
   - Корректность перевода
   - Разное восприятие шкалы респондентами из разных стран
   - Разные способы формирования выборок и методов опроса при проведении исследований
6. Считать – не значит делать

<u>Только репликация результатов на альтернативных выборках подтвердит корректность полученных результатов!</u>

План

1. Введение в количественную социологию и опросы
2. **Принципы построения индексов на основе социологических данных**
3. Анализ устойчивости индексов

Агрегирование данных в индексы

Развилка – 1. Вопросы от данных или от теории?

Развилка – 2. Вопросы объединяются с равными весами или с разными?

(если с разными)

Развилка – 2.1. Веса определяются экспертно или расчетно?

Развилка – 3. Внутри вопросов есть вариация в баллах или нет?

Развилка – 4. Использовать исходные значения индекса или отмасштабированные?

Факторный анализ

Исследовательская задача: можно ли объединить несколько переменных, близких по смыслу, в одну мета-переменную?
Пример: какое поведение/установки отражает финансовую ответственность, доверие и долгосрочные горизонты соответственно?

- Фактор – латентная (скрытая) переменная (например, финансовая культура), состоящая из коррелированных отдельных переменных
- С помощью факторного анализа можно попробовать понять, сколько «латентных» переменных скрывается в наборе данных, и что туда входит
- «Латентных» переменных, во-первых, меньше, во-вторых, они могут иметь общую глубинную природу
- Для поиска латентных переменных используется **исследовательский (Exploratory) факторный анализ.** Зачастую под ним понимается **метод главных компонент** (principal component analysis, PCA)

Примеры определения весов

- **Пример 1.** Компонента А состоит из двух вопросов. За вопрос 1 дается 1 балл, за вопрос 2 – 2 балла. Значит, вес вопроса 2 в компоненте А выше, чем вес вопроса 1.
- **Пример 2.** Компонента А состоит из двух вопросов по 1 баллу, компонента Б – из трех вопросов по 1 баллу. Значит, в итоговом индексе вес компоненты Б больше, чем компоненты А.
- **Пример 3.** Вопрос 1 является бинарным (0 если нет, 1 если да), вопрос 2 является множественным, максимум – 10 вариантов ответа (за каждый выбранный дается 0,1 балла). Значит, конкретное действие в вопросе А имеет больший вес, чем каждое по отдельности действие в вопросе 2.

Агрегирование данных в индексы

Варианты расчета итогового индекса:

1. Вариант 1. Сохранение исходного значения
2. Вариант 2. Масштабирование (например, от 0 до 100) в соответствие с теоретическими минимумами и максимумами
3. Вариант 3. Масштабирование (например, от 0 до 100) в соответствие с фактическими минимумами и максимумами

Формула для линейного масштабирования:

$$X_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

План

1. Введение в количественную социологию и опросы
2. Принципы построения индексов на основе социологических данных
3. **Анализ устойчивости индексов**

Проверка качества индекса

Технические способы

- Корреляция вопросов в индексе друг с другом
- Альфа Кронбаха
- McDonald’s ω (omega)
- PCA (доля объясненной дисперсии компонентой + uniqueness каждого вопроса в индексе)
- Подтверждающий (confirmatory) факторный анализ

Содержательные способы

- Test–retest reliability
- Content validity
- Convergent validity
- Discriminant validity
- Nomological validity

Кронбах-Альфа: контекст

- В социологии и психологии факторный анализ обычно проводится для выявления латентных переменных (конструктов), которые не наблюдаются напрямую
- Для их выявления обычно используют опросы, в которых разные вопросы позволяют «пощупать» с разных сторон интересующий исследователь конструкт
- Есть две ключевые меры надежности выделенных факторов (и одновременно – инструментария, который позволил эти факторы выделить):
  - **Повторяемость во времени (test-retest reliability).** Фактор должен воспроизводится, если после повторного исследования тех же объектов по тому же инструментарию результат повторился.
  - **Устойчивость при делении (split-half reliability).** Т.е., если случайным образом вынуть отдельные компоненты фактора, значение фактора по конкретному объекту наблюдения не должно измениться
- Наиболее популярный способ имплементации второго подхода – Кронбах-Альф

Кронбах - Альфа

Показатель Кронбах-Альфа показывает внутреннюю гомогенность и целостность фактора, лежащего за переменными.

$$\alpha = \frac{N^2 \overline{Cov}}{\sum s^2 + \sum Cov},$$

Где N – количество переменных, Cov – средняя ковариация между переменными, s – вариация переменных

- Рассчитывается как Кронбах-альфа в целом, так и для каждой переменной. Кронбах-альфа для каждой переменной – это Кронбах-альфа по выборке, из которой *исключена* переменная (alpha if item deleted)
- Если удаление пункта повышает α, он кандидат на исключение
- Значение больше 0,7 – ок, 0,8–0,90 – идеально, меньше 0,5 – прям совсем плохо

Кронбах – Альфа: NB

- Важные замечания насчет показателя:
  - **Кронбах-альфа считается внутри фактора.** Если после проведения факторного анализа были выделены компоненты, расчет Кронбах-альфа должен осуществляться внутри данных компонент.
  - **Вопросы должны быть сонаправлены.** «Перевернутые» вопросы значительно ухудшают Кронбах-альфа и делают возможным даже его отрицательное значение.
  - **Значение чувствительно к количеству переменных.** Чем больше количество переменных, тем выше показатель.
- Оценивается через функцию alpha (пакет psych)
- По понятным причинам, для факторов из одной переменной показатель не работает

Кронбах - Альфа

Обычный Кронбах-Альфа

Стандартизированный Кронбах-Альфа

```
Reliability analysis
Call: alpha(x = Normal_SC)

  raw_alpha std.alpha G6(smc) average_r S/N  ase mean   sd median_r
      0.83      0.87    0.85      0.62 6.5 0.04 0.32 0.12     0.62

    95% confidence boundaries
         lower alpha upper
Feldt     0.71  0.83  0.91
Duhachek  0.75  0.83  0.91

 Reliability if an item is dropped:
             raw_alpha std.alpha G6(smc) average_r S/N alpha se   var.r med.r
Друзья            0.75      0.80    0.74      0.58 4.1    0.056 0.00809  0.61
Коллеги           0.79      0.87    0.82      0.68 6.4    0.049 0.00729  0.64
Религиозн         0.77      0.83    0.79      0.61 4.8    0.054 0.02402  0.59
Спорт_и_культ     0.82      0.82    0.76      0.61 4.6    0.053 0.00041  0.61

 Item statistics
               n raw.r std.r r.cor r.drop mean    sd
Друзья        34  0.88  0.88  0.85   0.74 0.48 0.172
Коллеги       34  0.80  0.79  0.68   0.65 0.34 0.142
Религиозн     34  0.88  0.85  0.77   0.72 0.32 0.189
Спорт_и_культ 34  0.81  0.86  0.82   0.75 0.13 0.074
```

Кронбах-Альфа по отдельным переменным

Корреляция переменной с общей суммой переменных (без данной переменной). Если значение меньше 0,3 – повод серьезно насторожиться.

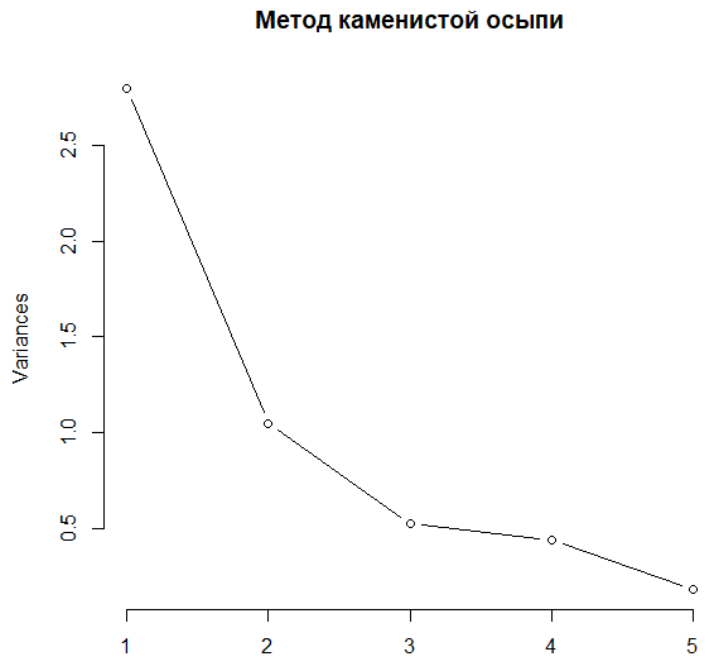
McDonald’s ω (omega)

- McDonald’s ω (омега) — это коэффициент надёжности шкалы, основанный на факторной модели, а не на средних корреляциях между пунктами, как Cronbach’s α
- В отличие от альфа Кронбаха:
  - Не требует одинаковость факторной нагрузки от каждой переменной
  - Не растет при добавлении нерелевантных пунктов
  - Строится на основе факторного моделирования
  - Более устойчив при разных шкалах
- Сейчас считается более надежным, чем альфа Кронбаха, особенно для категориальных данных, но обычно считают оба показателя
- Пороги аналогичны альфа Кронбаха

PCA

Расчет критерия Казейра в R: fit$sd^2

1	2	3	4	5
2.7936280	1.0485645	0.5274649	0.4436206	0.1867220




```
Родственники      Друзья      Коллеги     Религиозн Спорт_и_культ
     0.96519      0.76144      0.64815      0.70726      0.78633
```

Сумма факторных общностей: 0.77367

Построение каменистой осыпи в R: Plot(fit)

Расчет uniqueness и суммарная факторная мощность в R

Факторный анализ, шаг 3

- Для расчета факторной общности для разного количества факторов можно использовать функцию principal (пакет psych)
- Объект содержит лист «uniqueness». Она отражает долю уникальной, необъясненной общим фактором дисперсии показателя. Факторная общность – обратный показатель для уникальности
- Средняя факторная общность рассчитывается как сумма отдельных факторных общностей, деленная на количество переменных

Родственники	Друзья	Коллеги	Религиозн	Спорт_и_культ
0.96519	0.76144	0.64815	0.70726	0.78633

Сумма факторных общностей: 0.77367

Средняя факторная общность = кумулятивная объясненная дисперсия!

Confirmatory factor analysis (CFA)

- Методы обычно относятся к «нормальному» факторному анализу, но:
  - Направлены на достижение той же цели, что и reliability analysis (проверка качества модели)
  - Могут быть частично использованы и для проверки результатов PCA
- CFA предпочтительно проводить в случае, если количество переменных больше либо равно 3
- NB: CFA более требователен к размеру выборки. Для данного вида анализа рекомендуется соотношение 20 к 1: т.е. на одну переменную должно приходиться 20 наблюдений
- CFA использует похожие методы с SEM (structural equation modelling)
- При категориальных переменных используются разновидность – CFA с WLSMV (Weighted Least Squares Mean and Variance adjusted)

Методы оценки моделей в рамках CFA

1. **Хи квадрат (chi-square)**
2. **Comparative Fit Index (CFI).**
    1. Варьируется от 0 до 1, чем больше – тем лучше
    2. Rule of Thumb: значение должно быть больше 0,95
3. **Tucker Lewis Index (TLI).**
    1. Варьируется от 0 до 1, чем больше – тем лучше
    2. Rule of Thumb: значение должно быть больше 0,9
4. **Root mean square error of approximation (RMSEA).**
    1. Чем меньше RMSEA, тем лучше
    2. Rule of Thumb: значение должно быть меньше 0,1
    3. Пакет lavaan в R также проверяет гипотезу, что RMSEA < 0.05. Если нулевая гипотеза отвергается, то модель не очень хорошая CFI > 0.95, TLI > 0.90 and RMSEA < 0.10

Содержательные способы проверки качества индекса

1. **Test–retest reliability:** воспроизводимость индекса во времени
   - Нужно повторить технические способы (Кронбах-альфа и т.д.) на выборке за другой промежуток времени. Если показатели качества будут схожие – значит, индекс хороший
2. **Content validity:** четкое определение границ конструкта
3. **Convergent validity:** коррелированность с похожими характеристиками
4. **Discriminant validity:** не коррелированность с нерелевантными характеристиками
5. **Nomological validity:** поведение индекса в соответствие с теорией
   - Например, построение регрессий для проверки наличия связей с переменными, которые должны быть связаны с конструктом

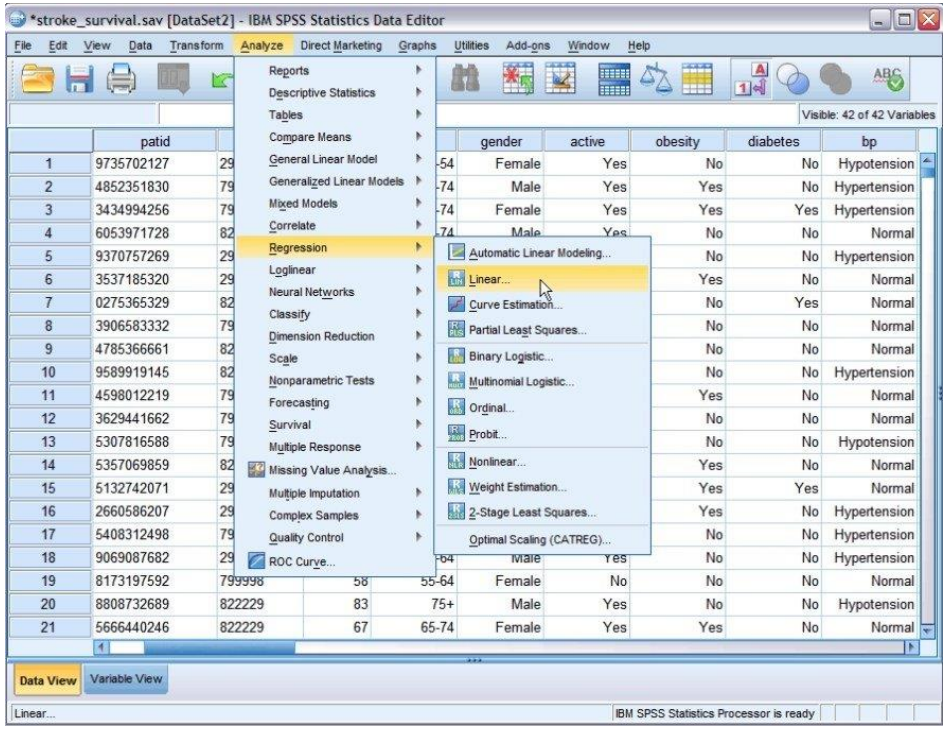
«Любимая программа» социологов

SPSS, преимущества:

- Удобный, интуитивно понятный интерфейс
- Возможно быстро делать базовые операции с опросными данными (таблицы сопряженности, средние и т.д.)

SPSS, недостатки:

- Негибкая визуализация
- Ограниченное количество методов анализа данных
- Неудобный синтаксис (затруднено совершение повторяемых операций)



since 2017

мои финансы

финансовая грамотность в вузах
Федеральный сетевой методический центр

Экономический факультет МГУ имени М.В. Ломоносова

Олимпиада

по финансовой грамотности и финансовой культуре

для студентов

Регистрация до 03.03.2026

fingramota.econ.msu.ru/olymp/