Проектная группа НИУ ВШЭ
«Визуализация событий жизненного пути»

Подготовка данных для анализа событий жизни

Митрофанова Екатерина Сергеевна

к.с.н.,
старший преподаватель кафедры демографии,
научный сотрудник Института демографии

Страница на сайте Вышки https://www.hse.ru/staff/emitrofanova

Алгоритм начала работы с данными

1. Изучить сопутствующую документацию (вопросники, гайды, инструкции, методические пояснения и т.д.)
2. Изучить основные переменные (frequences, crosstabs). Определить, есть ли несостыковки в данных, пропуски, выбросы
3. Разработать или найти алгоритм для исправления несостыковок и ошибок (консультация с коллегами, литературы)
4. Доработка данных:
    - склейка, если это панель
    - замена или удаление пропусков
    - исправление ошибок, удаление или замена скомпроментированных данных
    - расчет вспомогательных переменных (например, возрастов из дат наступления событий)
    - подготовка данных для применения к ним специальных методов (например, перевод из одного формата в другой)

Person-Level Data

	ID	дата_рождения	работа1	брак1	ребенок1	брак_обязателен
1	1	10.09.67	04.03.84	12.12.85	05.09.87	5
2	2	04.03.90	21.03.08	.	.	1
3	3	11.01.81	02.02.05	01.03.07	06.01.11	5
4	4	21.07.54	09.09.70	17.11.72	28.05.73	5

Person-Period Data

	ID	дата_рождения	работа1	брак1	ребенок1	брак_обязателен
1	1	10.09.67	04.03.84	12.12.85	05.09.87	4
2	1	10.09.67	04.03.84	12.12.85	05.09.87	4
3	1	10.09.67	04.03.84	12.12.85	05.09.87	5
4	2	04.03.89	.	.	.	2
5	2	04.03.89	.	.	.	2
6	2	04.03.90	21.03.08	.	.	1
7	3	11.01.81	.	.	.	2
8	3	11.01.81	02.02.05	01.03.07	.	5
9	3	11.01.81	02.02.05	01.03.07	06.01.11	5
10	4	21.07.54	09.09.70	17.11.72	28.05.73	5
11	4	21.07.54	09.09.70	17.11.72	28.05.73	5
12	4	21.07.54	09.09.70	17.11.72	28.05.73	5

Примеры

Подготовка данных

1. Чистка данных (особенно для панели)
2. «Возрастная дилемма»
3. Пропущенные данные

Стартовые события жизненного пути

демографические

- вступление в 1ое партнерство
- вступление в 1ый брак
- рождение 1го ребенка

социоэкономические

- 1ое отделение от родителей
- получение образования наивысшего уровня
- трудоустройство на 1ую работу

Данные

Родители и дети, мужчины и женщины
в семье и обществе (РиДМиЖ) –
часть проекта Generations and Gender Programme (GGP)

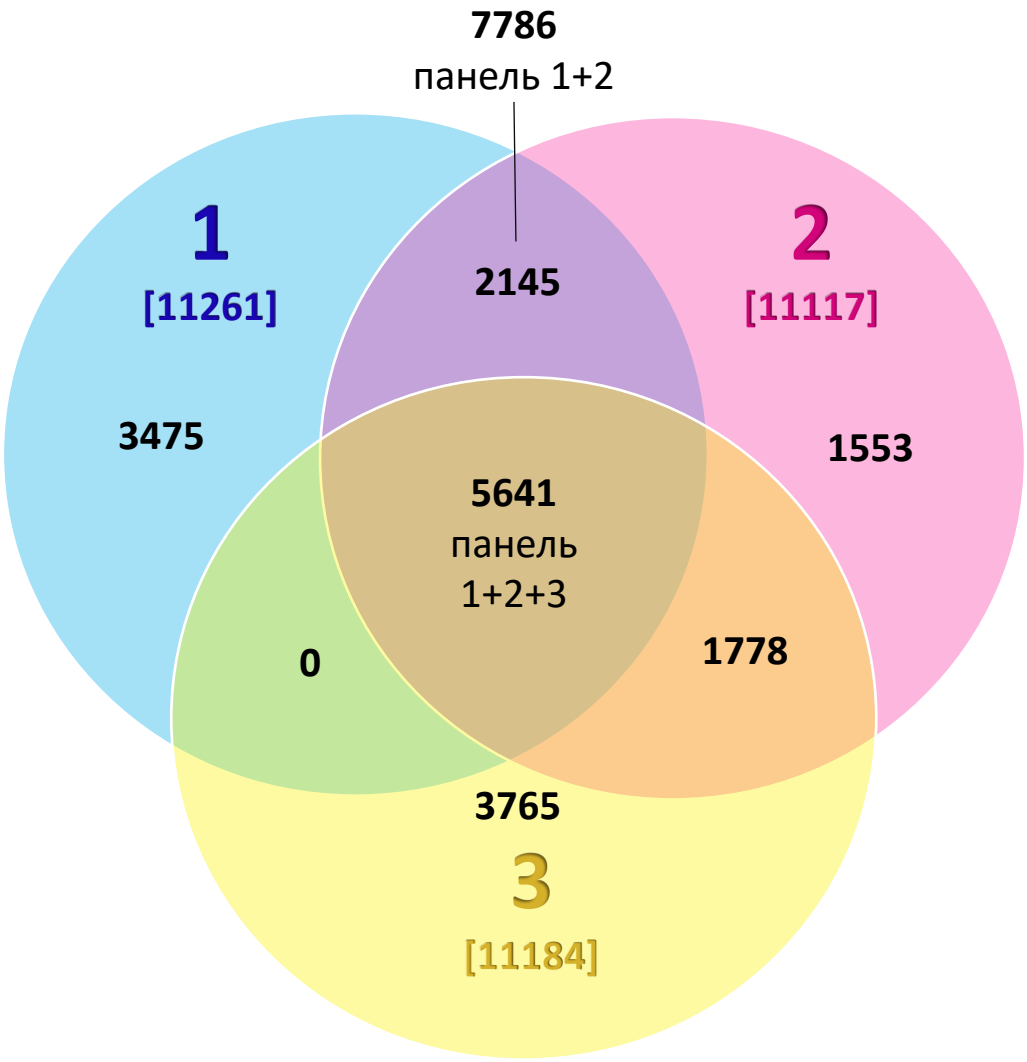
Три волны обследования:

- 2004 г.: 11261 чел.
- 2007 г.: 11117 чел.
- 2011 г.: 11184 чел.

Панель 3х волн: 5641 чел.

Возраст: 25-81 год
1930-1986 г.р.

Формирование панели РиДМиЖ



Блоки вопросов

Характеристики респондентов:

1. Пол
2. Дата рождения
3. Тип населенного пункта

Неповторяющиеся события:

1. Начало проживания вне родительского дома (минимум 3 месяца)
2. Трудоустройство на первую работу (минимум 6 месяцев)
3. Получение образования наивысшего уровня

Повторяющиеся события:

1. Партнерства (минимум 3 месяца) + расставания
2. Браки (официальная регистрация) + разводы
3. Дети

Характеристики респондентов

1. Пол:

- в случае расхождения данных – проверка по полу партнера из той же волны, по вопросам про беременность

2.1. Месяц рождения:

- миссинги, для которых нельзя найти замену, меняются на 6-ой месяц
- если два месяца одинаковые, а третий от них отличается, меняем третий
- если известно всего два месяца, и оба разные, выбираем тот, который стал известен раньше

2.2. Год рождения:

- если два года одинаковые, а третий от них отличается, меняем третий
- если известно всего два года, и оба разные, выбираем тот, который стал известен раньше

2.3. Объединение месяца и года в дату

Подсказка: в SPSS годы формата YY иногда преобразовываются при создании даты в годы 21го века, например, в 2045 вместо 1945. Эти настройки можно поменять в: Edit – Options – Data, ограничив End Year.

3. Тип населенного пункта: сохранение переменных для каждой волны

Все события (неповторяющиеся и повторяющиеся)

Общие алгоритмы:

1. Для дат, где указан только год и не указан месяц, берется 6ой месяц
2. Если известен месяц и нет ни одной вспомогательной переменной с годом (соседней того же типа, похожего типа), удаляем дату
3. Если очевиден факт опечатки, меняем похожие по написанию цифры: 0 на 9… и наоборот
4. Если нарушена последовательность наступления событий, меняем их местами
5. Если даты соседних событий совпадают, удаляем то событие, которое стоит следующим по порядку
6. Если две даты/года/месяца одинаковые, а третья от них отличается, меняем третью
7. Если известно всего две даты/года/месяца, и обе разные, выбираем ту, которая стала известна раньше

Неповторяющиеся события: работа и образование

Работа

Информация о первом опыте работы есть только во 2 и 3 волне. Если у респондента нет даты завершения 1ой работы во 2 волне, то при наличии такой даты в 3 волне, указываем последнюю.

Образование

Если у респондента нет даты завершения 1го образования (edu_1.1_my), но есть дата получения образования самого высокого уровня из 2 волны (edu_high_2_my), то указываем эту дату.

Если есть только дата по 3 волне (edu_high_3_my), указываем ее.

Повторяющиеся события

Пример алгоритмов для союзов

1. Для пар дат «**Завершение союза^{n}**» и «**Начало союза$^{n+1}$**» при наличии одной из дат и отсутствии другой – отсутствующая приравнивается к той, которая имеется.
2. Для дат «**Начало совместного проживания**» и «**Регистрация брака**»:
   - при наличии даты завершения союза и отсутствии даты его начала, а также отсутствии информации о предыдущих союзах этого типа, старт союза приравнивается к началу союза другого типа (при наличии соответствующей информации и совпадении дат окончания союзов). Например, для сожительства используется дата начала брака, для брака – начало сожительства.
   - для 2 и 3 волны (2007, 2011 год): при наличии даты завершения союза и отсутствии информации о союзах другого типа и других очередностей в качестве старта союза берется дата проведения предыдущей волны, приравненная к 01.06.2004 и 01.06.2007.
3. Для дат «**Завершение отношений**» и «**Развод**»:
   - при наличии даты начала союза и отсутствии даты его завершения, а также отсутствии информации о последующих союзах этого типа, окончание союза приравнивается к завершению союза другого типа (при наличии соответствующей информации и совпадении дат начала союзов). Например, для завершения сожительства используется дата развода, для развода – завершение сожительства.
   - для 2 и 3 волны (2007, 2011 год): при наличии даты начала союза и отсутствии информации о союзах другого типа и других очередностей в качестве окончания союза берется дата проведения предыдущей волны, приравненная к 01.06.2004 и 01.06.2007.

Проверка дат в биографиях

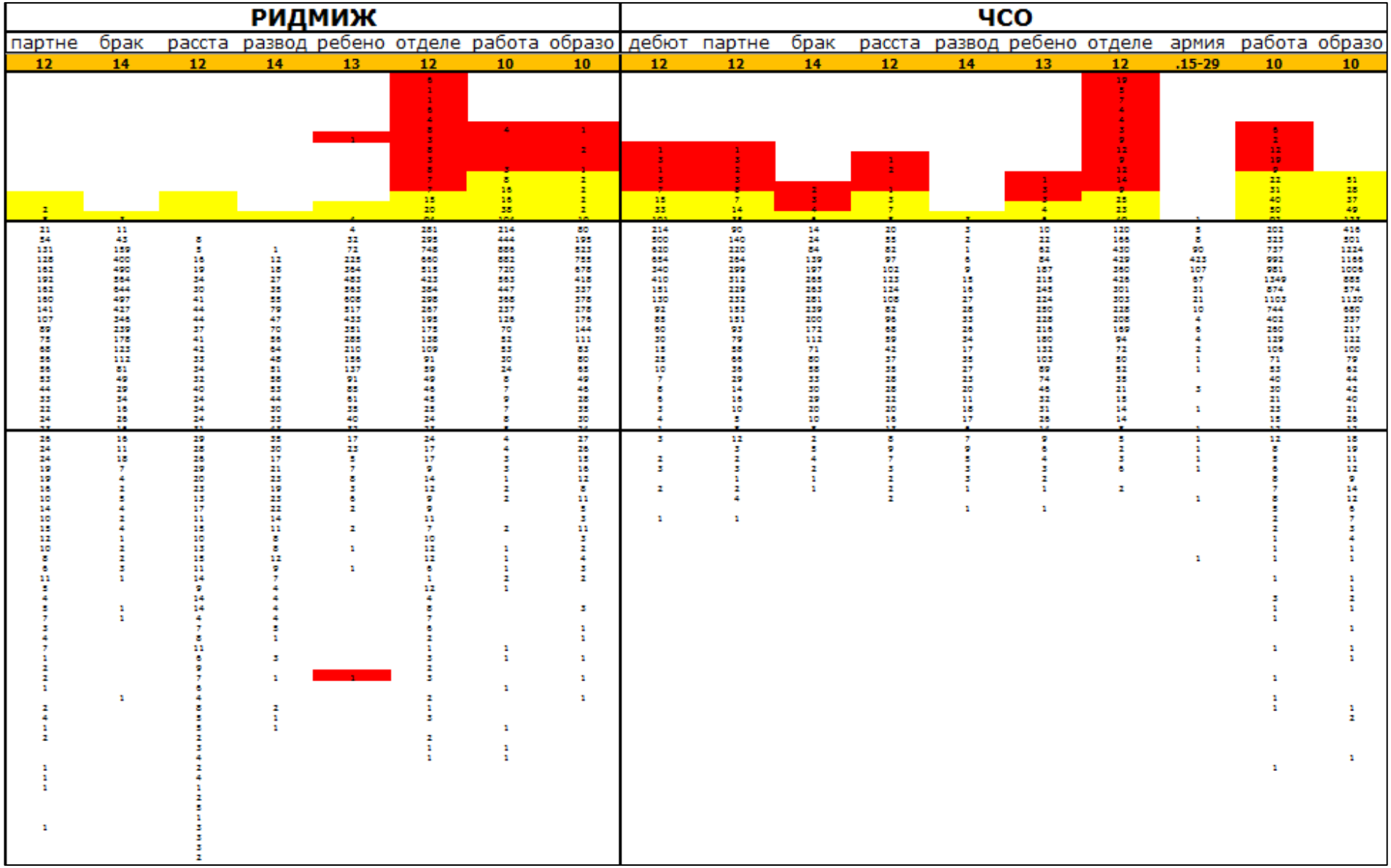
Все события:

Возрасты наступления событий на момент прохождения опроса должны быть как минимум больше нуля >> возрастная дилемма (след. слайд)

Повторяющиеся события:

1. «Расстояние» между союзами одного типа должно быть больше или равно нулю: разница между событием^{n} и событием$^{n+1}$
2. Длительность союзов (брак-развод, партнерство-расставание) должна быть больше 3 месяцев (в соответствии с тем, как был задан вопрос в анкете)

Возрастная «дилемма»



Возрастная «дилемма»

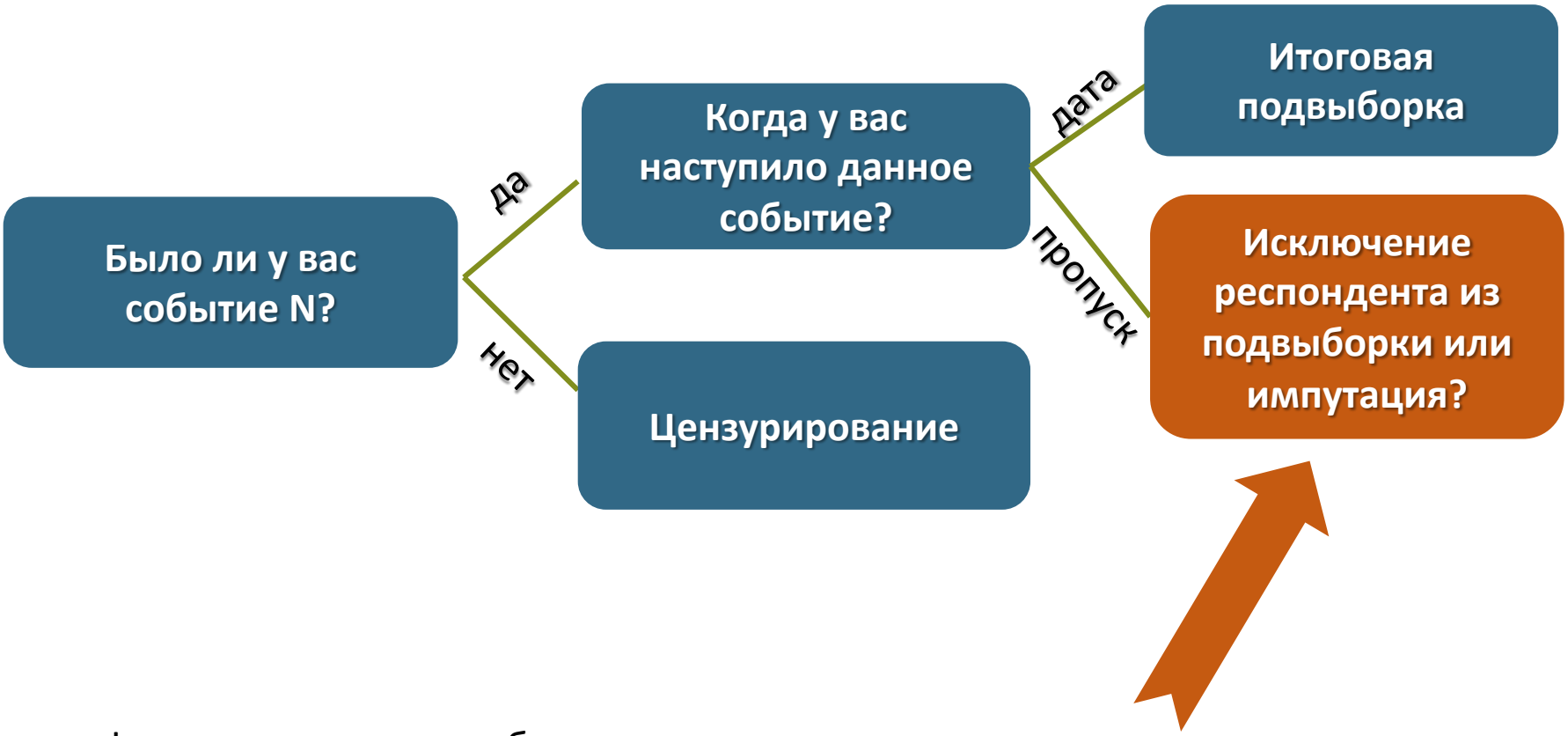
	РИДМИЖ								ЧСО									
	партне	брак	расста	развод	ребено	отделе	работа	образо	дебют	партне	брак	расста	развод	ребено	отделе	армия	работа	образо
	12	14	12	14	13	12	10	10	12	12	14	12	14	13	12	.15-29	10	10
0						6									19			
1						1									5			
2						1									7			
3						6									4			
4						4									4			
5						8	4	1							3		6	
6					1	3									9		2	
7						8		2	1	1					12		12	
8						3			3	3		1			9		19	
9						8	3	1	1	2		2			12		9	
10						7	8	2	3	3				1	14		22	51
11						7	16	2	7	8	2	1		3	9		31	28
12						15	16	2	15	7	3	3		3	25		40	37
13	2					20	38	2	33	14	4	7		4	23		50	49
14	8	3			4	94	104	10	101	35	6	8	3	6	49	1	92	123

. . . . . .

58	1		6	3		3	1	1										1
59	2		9			2												
60	2		7	1	1	3		1									1	
61	1		6				1											
62		1	4			2		1									1	
63	2		8	2		1											1	1
64	4		5	1		3												2
65	1		5	1			1											
66	2		2			2												
67			3			1	1											
68			4			1	1											1
69	1		2														1	
70	1		4															
71	1		1															
72			2															
73			5															
74			1															
75	1		3															
76			3															
77			3															
78			2															
79			1															

Пропущенные данные

Что делать с пропущенными данными?



Фактически данное событие произошло, но мы не можем разместить его на временной оси и сопоставить с остальными событиями жизни данного индивида, не внеся искажения в последовательности.

Что делать? Удалять респондентов, восстанавливать пропущенные события или оставлять как есть?

Вид данных

Характеристики респондента					Событие		
Пол	Дата рождения	Поколе-ние	Образова-ние	Тип населенного пункта	Факт	Дата	Возраст
М	10.09.1947	2	1	1	1	01.10.1973	26
Ж	30.05.1984	6	2	.	0	.	.
М	05.06.1972	5	3	2	1	.	.
Ж	07.01.1951	3	.	3	1	01.05.1971	20
Ж	17.04.1966	4	2	1	.	.	.

Точками обозначены пропущенные данные

Варианты пропущенных данных (миссингов):

1. отсутствие информации, характеризирующей респондента из-за его отказа отвечать или ошибки интервьюера, или ошибки кодировщика
2. отсутствие даты события при наличии факта наступления события – самый проблемный случай
3. отсутствие даты события вследствие отсутствия факта события (не нужно ни на что заменять!)

Вариант 1. Отсутствие информации, характеризирующей респондента вследствие отказа или ошибки

Примеры:

Год рождения (ECC)

У 23 респондентов отсутствуют годы рождения. По возрасту на момент опроса годы восстановить не удалось, т.к. там тоже пропуски значений. В других вариантах базы данных пропуски у тех же респондентов.

Можно восстановить по дополнительным переменным.
Также можно использовать возрасты стартовых событий, но после того, как они будут почищены.

ID	Пол	Год рождения	Число внуков	Год рождения 1го внука	Есть ли правнуки	last 7 days: retired	last 7 days: paid work	last 7 days: education	last 7 days: unemployed, actively looking for job	last 7 days: unemployed, not looking for job	last 7 days: permanently sick/disabled	last 7 days: community or military service	last 7 days: housework, looking after children	Год выхода на пенсию /инвалидности
139	2		66	6666	6	0	1	0	0	0	0	0	0	6666
295	2		0	6666	6	0	1	0	0	0	0	0	0	6666
502	1		66	6666	6	0	1	0	0	0	0	0	0	6666
957	1		66	6666	6	0	0	0	0	1	0	0	0	6666
959	2		0	6666	6	0	0	0	0	1	0	0	0	6666
1433	1		66	6666	6	0	0	1	0	0	0	0	0	6666
1701	1		0	6666	6	0	1	0	0	0	0	0	0	6666
2080	1		66	6666	6	0	1	0	0	0	0	0	0	6666
2257	2		66	6666	6	0	1	0	0	0	0	0	0	6666
2263	2		66	6666	6	0	0	1	0	0	0	0	0	6666
2266	1		66	6666	6	0	1	0	0	0	0	0	0	6666
2267	1		66	6666	6	0	0	1	0	0	0	0	0	6666
2472	2		66	6666	6	0	0	1	0	0	0	0	0	6666
2473	1		66	6666	6	0	0	0	0	0	0	1	0	6666
2522	2		66	6666	6	0	0	0	0	0	0	0	1	6666
2590	1		66	6666	6	0	0	0	0	1	0	0	0	6666
2775	2		66	6666	6	0	1	0	0	0	0	0	1	6666
2910	1		66	6666	6	0	0	1	0	0	0	0	0	6666
2915	1		0	6666	6	0	1	0	0	0	0	0	0	6666
3197	1		66	6666	6	0	1	0	0	0	0	0	0	6666
3242	1		66	6666	6	0	1	0	0	0	0	0	0	6666
3360	2		66	6666	6	0	1	0	0	0	0	0	1	6666
3368	2		0	6666	6	0	1	0	0	0	0	0	0	6666

Судя по дополнительным переменным, респонденты с пропущенными годами рождений являются трудоспособными: ни у кого нет внуков, никто не является пенсионером.

Тип населенного пункта (ECC)

ID	birth_y	sex	edu_type	edufld	eduyrs	nbthcld	hhmmb	estsz	regionru	stfsdlv	pplahlp	flclpla	mbltph	inttph	Наши предполож	Импутации	Наш выбор
	Год рожд	Пол	Тип обр	Спец-ть	Лет обр	Число детей	Число членов дх	Число сотрудн	Регион	Удовл усл жилья	Помощь людей	Другие близко	Моб. телеф.	Интер нет			
1319	1940	1	1	инжен	19	2	5	5	Центр	2	0	4	1	2	1	1,0	1
5041	1942	2	2	инжен	10	1	2	8	Волга	3	4	4	2	5	1/2	2,2	2
3187	1950	2	3	общее	11	2	2	1	Зап. Сибирь	7		3	2	5	3	2,4	3
2256	1954	2	1	эконом	16	2	3	4	Сев кав	3	3	4	1	5	1/2	1,6	1
3188	1954	1	3	общее	11	1	2	8	Зап. Сибирь	0		3	2	5	2	2,2	2
3186	1956	1	2	инжен	14	2	3	4	Зап. Сибирь	6	4		2	5	2	2,4	2
5286	1957	1	2	инжен	12	2	4	3	Зап. Сибирь	3	2	2	1	2	1/2	1,4	1
1138	1974	1	2	инжен	11	3	1	1	Север	6	1	3	1	5	1/2	2,4	2
2264	1980	2	2	соц раб	12	66	2	6	Сев. Кав.	0	0	5	1	5	2	2,4	2
2471	1983	2	1	общее	18	66	6	1	Сев. Кав.	7	5	4	1	2	1	1,2	1
3185	1983	2	2	соц раб	13	1	3	1	Зап. Сибирь	0			1	5	3	2,0	3
3275	1991	2	3	общее	8	66	4	6	Вост. Сибирь	10	5	2	1	1	1	1,6	1

1 - Большой город/областной центр

2 - Город/ПГТ

3 - Село

Множественная импутация: 5 итераций
Сначала сделан анализ частот, и отобраны только наиболее значимые переменные (поколение, пол, тип образования, чувствует близость с людьми в своей местности, наличие интернета)
Недостаток импутации: завышенные результаты
Общая проблема: миссинги во вспомогательных переменных (без цвета)

Импутация в СПСС

Среднее ряда. Заменяет пропущенные значения средним для всего ряда.

Среднее близлежащих точек. Заменяет пропущенные значения средним из валидных окружающих значений. Интервал ближайших точек здесь — количество точек, предшествующих текущей и следующих за ней, которые используются при вычислении среднего.

Медиана близлежащих точек. Заменяет пропущенные значения медианой из валидных окружающих значений. Интервал ближайших точек здесь — количество точек, предшествующих текущей и следующих за ней, которые используются при вычислении медианы.

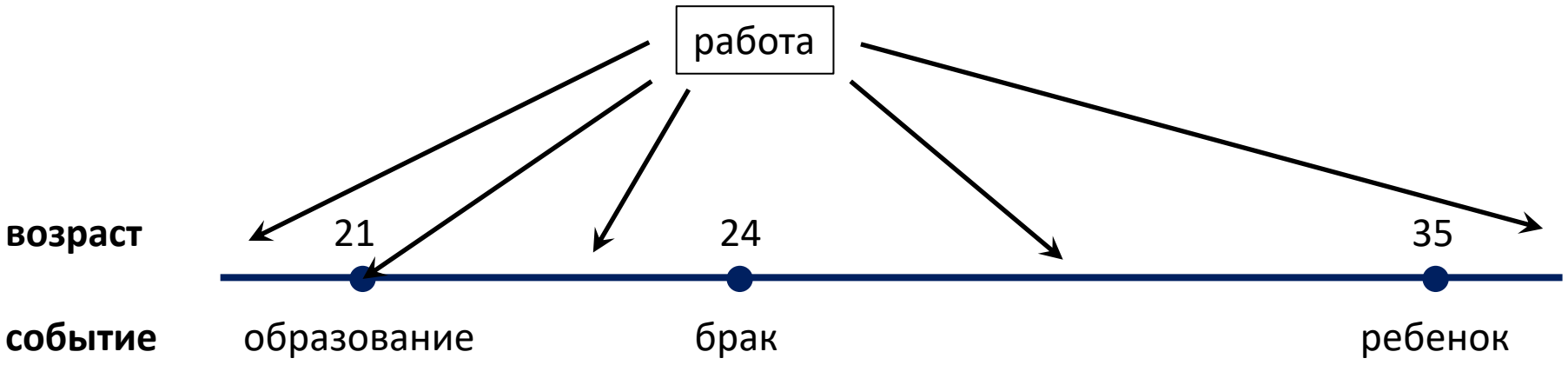
Линейная интерполяция. Заменяет пропущенные значения с помощью линейной интерполяции. Для интерполяции используются последнее валидное (непропущенное) значение перед пропущенным и первое валидное значение после пропущенного. Если пропущенное значение является первым или последним значением ряда, то такое значение не заменяется.

Линейный тренд в точке. Заменяет пропущенные значения линейным трендом для этой точки. Строится регрессия существующего временного ряда на индексную переменную со значениями от 1 до n. Пропущенные значения заменяются предсказанными значениями.

Множественная импутация. Метод выбирается автоматически на основе сканирования данных. Доступно: линейная регрессия, метод согласования предсказанного среднего.

Вариант 2. Отсутствие даты события при наличии факта наступления события

Известно, что человек работает, но не известно, когда он трудоустроился



Способы решения проблемы миссингов 2го типа

1. **Удаляем респондента из выборки** => искажаем выборку. ***Можно***: проанализировать, как изменяются ключевые переменные и иметь это в виду при анализе.
2. **Восстанавливаем дату/возраст** => искажаем последовательность, т.к. не знаем, между какими событиями произошло событие. ***Можно***: проанализировать похожих респондентов (использовать импутацию с большим кол-вом переменных). ***Проблема***: миссинги во вспомогательных переменных.
3. **Не восстанавливаем дату наступления события** => искажаем последовательность. ***Можно***:
   а) принять, что респондент соврал, и если **нет даты, то не было и факта** => серьезное допущение, особенно если таких миссингов много + рискуем восстанавливать пропущенные данные по восстановленным данным (слишком много абстракции). Можно принять эти риски. ***Обоснование***: на некоторые вопросы трудно дать однозначный ответ, т.к. ситуации могут быть разными. Например, в школе отучился, но нет аттестата - и респондент решил не указывать, что получил образование. Или с партнером не понятно: вроде и был, но жили не вместе и т.д. Поэтому если мы где-то огрубляем, то это справедливо, т.к. сами люди могли и не иметь однозначного ответа.

   б) всем событиям с неизвестными датами **присвоить какой-то определенный возраст**, например, 0 или 10 лет. Нам важно сохранить эти события для последовательностей. Можно таким событиями приписать **отдельную букву**, и получится отдельная траектория: такое-то событие есть, но не ясно где. ***Проблема***: это сильно увеличит кол-во последовательностей.

   в) создать дополнительные переменные, маркирующие миссинги (для каждого события или для всех сразу), и сравнить хронограммы для тех, у кого нет миссингов с теми, у кого есть. Если различия незначительные – оставлять все как есть, если значительные, то возвращаемся к п.1 и п.2. ***Проблема***: если строить переменную для каждого события, то хронограмм будет слишком много.