Выборка в социологическом исследовании

Лекция 6

Звоновский, к.с.н.

Основные понятия выборочного метода

Генеральная совокупность – совокупность всех единиц наблюдения. Почти всегда «объект» исследования и «генеральная совокупность» – это одно и то же.

Выборка (выборочная совокупность) - часть объектов генеральной совокупности, которые непосредственно подвергаются измерению.

Единицы выборки – однородные элементы генеральной совокупности, из которых формируется выборочная совокупность

Ошибка выборки – степень рассогласования между значением (долей или средним) признака выборочной совокупности и значением релевантного этому признаку генеральной совокупности

Гипотетическая совокупность

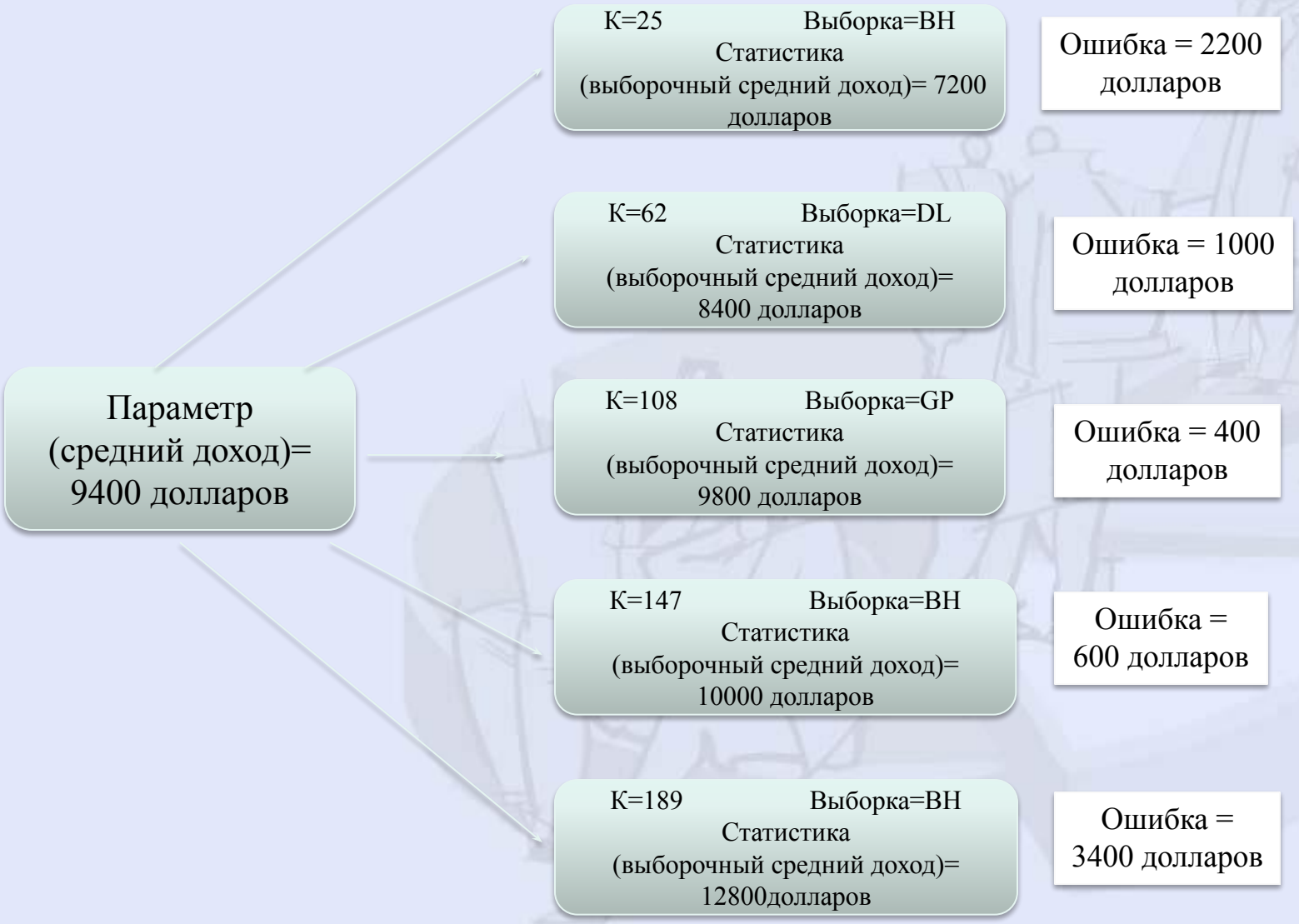
Элемент	Доход (в долларах)	Образование (лет)	Подписка на газету		Элемент	Доход (в долларах)	Образование (лет)	Подписка на газету
A	5600	8	X		K	9600	13	X
B	6000	9	Y		L	10000	13	Y
C	6400	11	X		M	10400	14	X
D	6800	11	Y		N	10800	14	Y
E	7200	11	X		O	11200	15	X
F	7600	12	Y		P	11600	16	Y
G	8000	12	X		Q	12000	16	X
H	8400	12	Y		R	12400	17	Y
I	8800	12	X		S	12800	18	X
J	9200	12	Y		T	13200	18	Y

Производная совокупность выборок объемом n=2

Выборка			Выборка			Выборка		
к	**пара**	**среднее**	**к**	**пара**	**среднее**	**к**	**пара**	**среднее**
1	AB	5800	61	DK	8200	122	HR	10400
2	AC	6000	**62**	**DL**	**8400**	123	HS	10600
3	AD	6200	63	DM	8600	124	HT	10800
25	**BH**	**7200**	85	ET	10200	145	JT	11200
26	BI	7400	86	FG	7800	146	KL	9800
27	BJ	7600	87	FH	8000	**147**	**KM**	**10000**
28	BK	7800	88	FI	8200	148	KN	10200
48	CN	8600	**108**	**GP**	**9800**	188	RS	12600
49	CO	8800	109	GQ	10000	**189**	**RT**	**12800**
50	CP	9000	110	GR	10200	190	ST	13000
51	CQ	9200	111	GS	10400			

Среднее средних – 9400 долларов

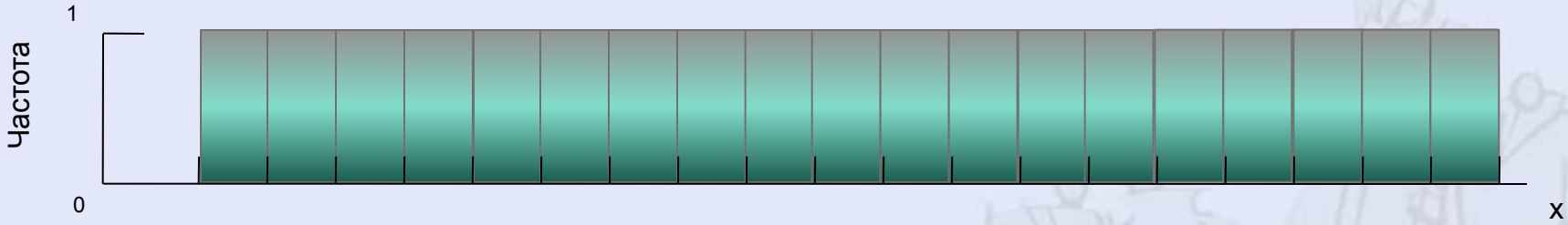
Примеры выборок и соответствующих ошибок



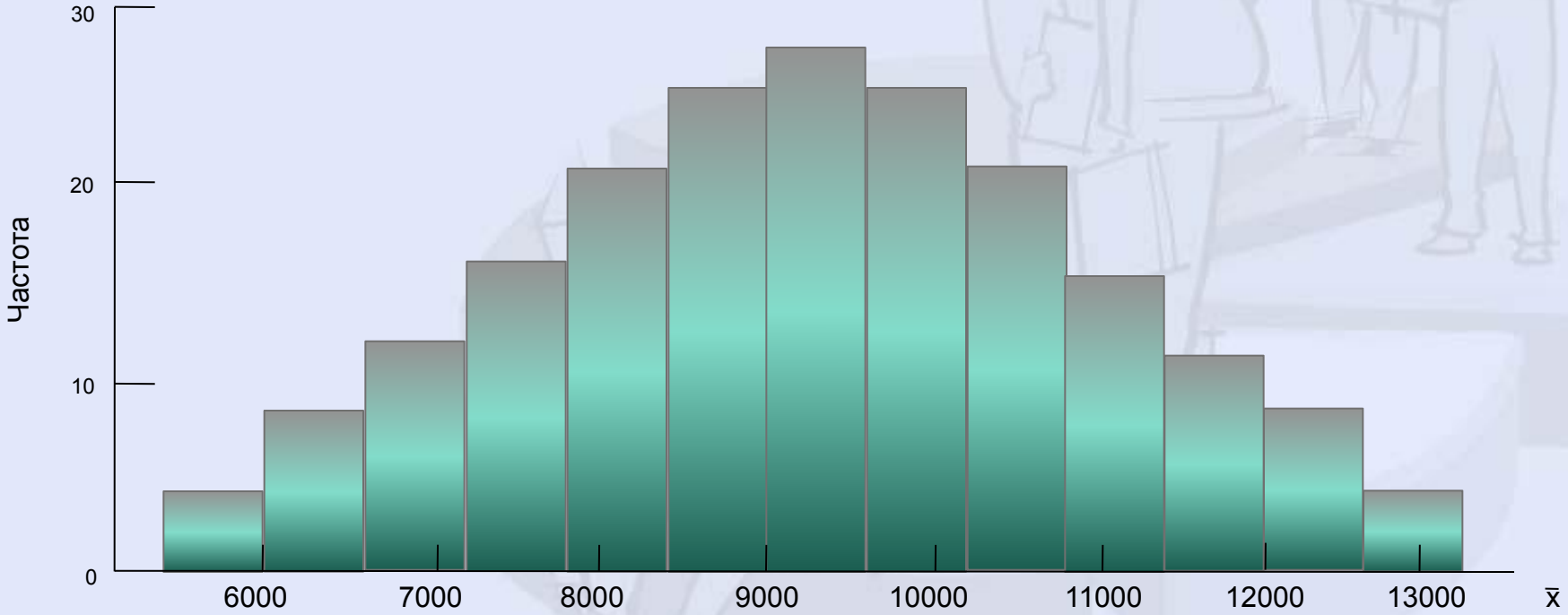
Распределение по числу выборок

выборочное среднее (рублей)	количество выборок
Не более 6100	**2**
от 6101 до 6600	**7**
От 6601 до 7200	**11**
От 7201 до 7800	**16**
От 7801 до 8400	**20**
От 8401 до 9000	**25**
От 9001 до 9600	**28**
От 9601 до 10200	**25**
От 10201 до 10800	**20**
От 11801 до 11400	**16**
От 11401 до 12000	**11**
От 12001 до 12600	**7**
12601 и более	**2**

Распределение количественного признака в генеральной совокупности и



Распределение оценок в производственной совокупности

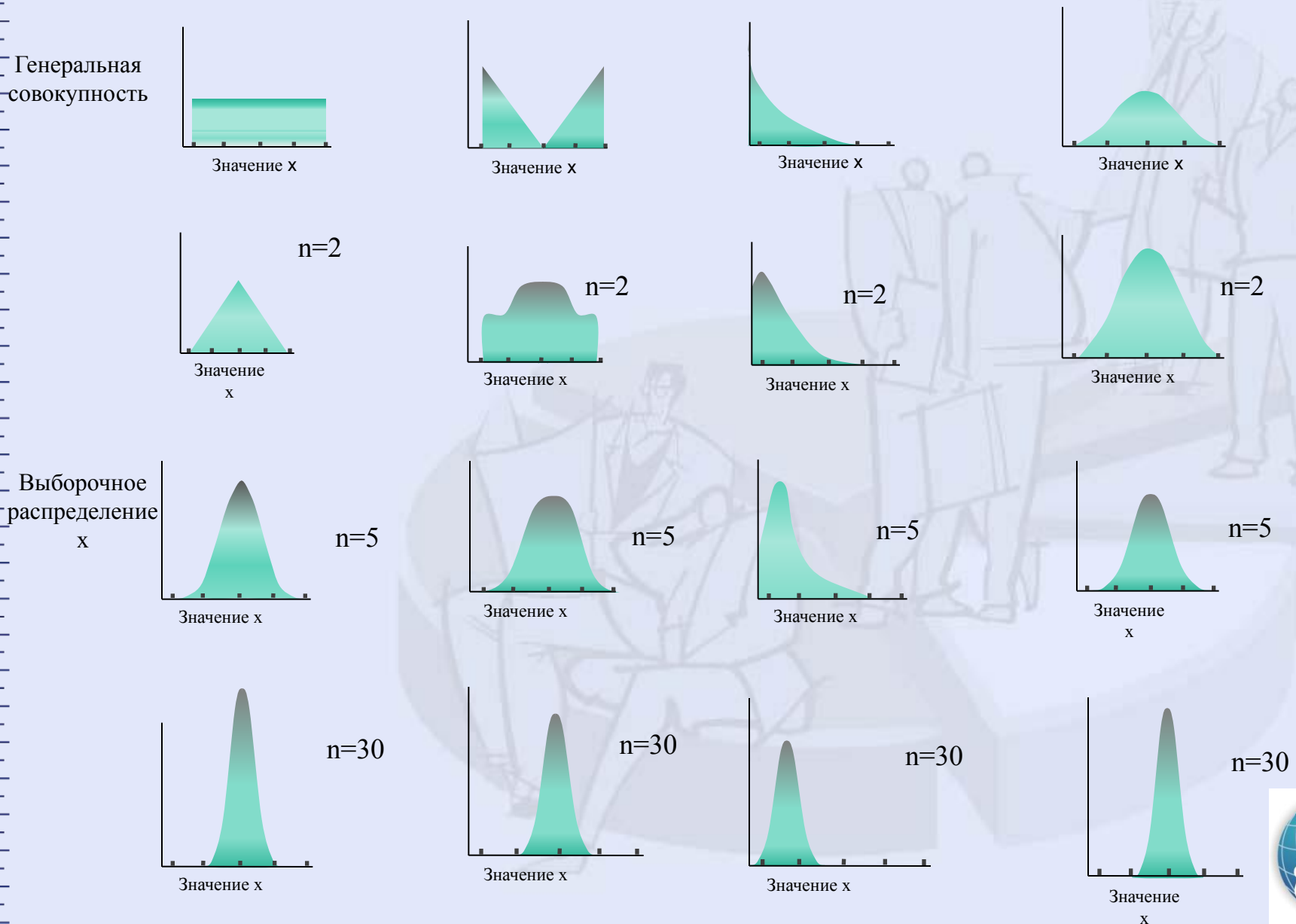


Центральная предельная теорема

Для простых случайных выборок объемом n, выделенных из генеральной совокупности с генеральным средним μ и дисперсией δ^2 , при больших n распределение выборочного среднего $\overline{x}$ приближается к нормальному с центром, равным μ, и с дисперсией δ^2 / n. Точность названного приближения возрастает с возрастанием n.

Простая случайная выборка объемом n имеет среднее, близкое к среднему генеральной совокупности, и степень этой близости возрастает с увеличением n.

Распределение выборочных средних для выборок различного объема и различных популяционных распределений



Доверительные интервалы

Доверительный интервал - интервал, который покрывает неизвестный параметр с заданной надёжностью.

68,26% выборочных средних отклоняются от генерального среднего не более, чем на ±1δ

95,45% выборочных средних отклоняются от генерального среднего не более, чем на ±2δ

99,73% выборочных средних отклоняются от генерального среднего не более, чем на ±3δ

$$\mu - z^*\delta \leq \bar{x} \geq \mu + z^*\delta$$

Доверительные интервалы (при μ=9200)

номер выборки	пара	среднее	нижний предел	верхний предел
1	AB	5800	2689	8911
2	AC	6000	2889	9111
3	AD	6200	3089	9311
4	AE	6400	3289	9511
5	AF	6600	3489	9711
6	AG	6800	3689	9911
7	AH	7000	3889	10111
8	AI	7200	4089	10311
9	AJ	7400	4289	10511
10	AK	7600	4489	10711

Репрезентативность

Репрезентативность – соответствие характеристик выборочной совокупности характеристикам генеральной. Репрезентативность определяет, насколько возможно обобщать результаты исследования с привлечением определённой выборки на всю генеральную совокупность.

Сбор данных на нерепрезентативных выборках всегда является результатом систематической ошибки.

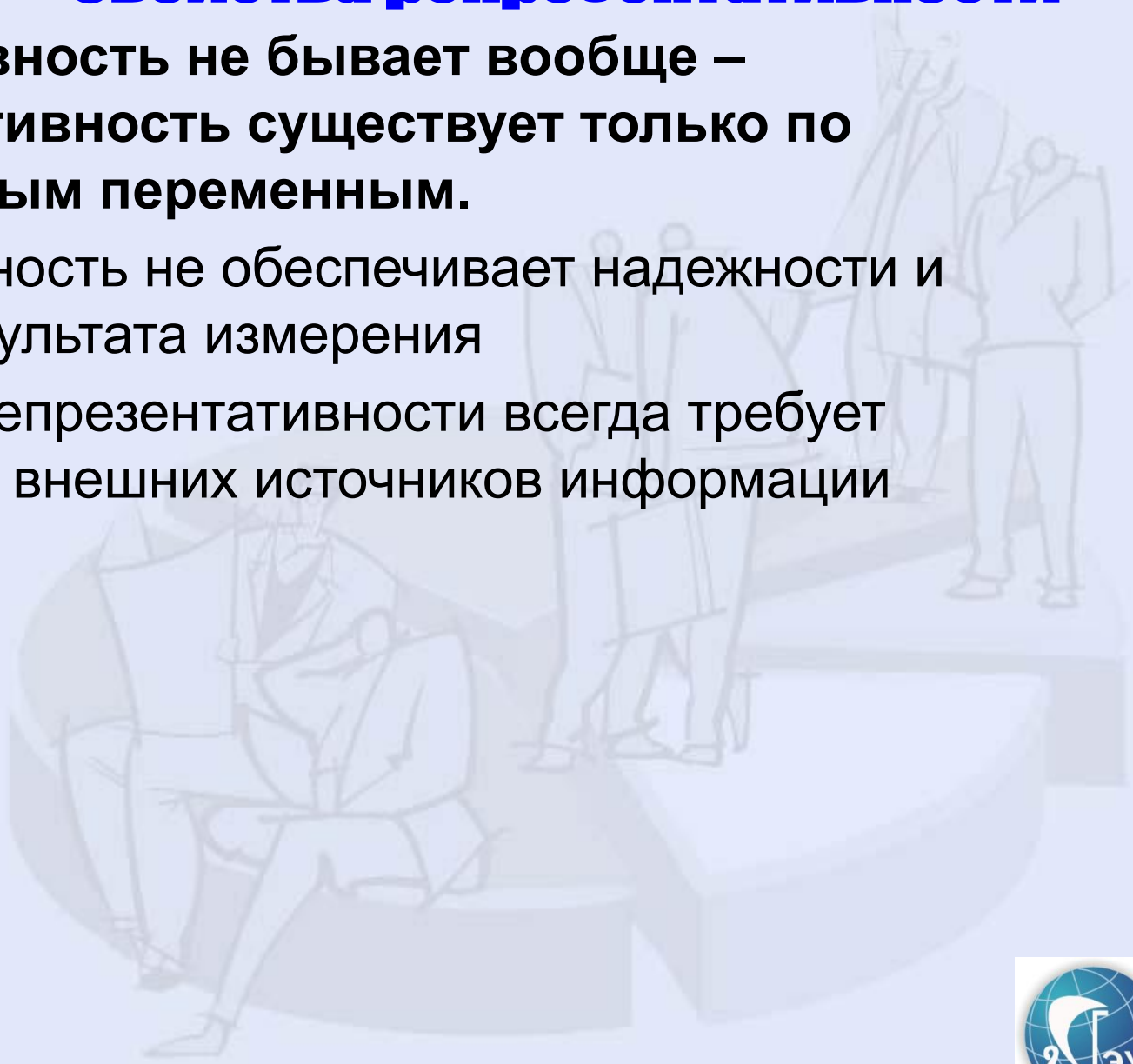
Случайные ошибки не делают выборку нерепрезентативной. Они лишь уменьшают точность измерения.

Свойства репрезентативности

Репрезентативность не бывает вообще – репрезентативность существует только по определенным переменным.

Репрезентативность не обеспечивает надежности и точности результата измерения

Утверждение репрезентативности всегда требует привлечения внешних источников информации



Типы выборки

Вероятностные

- Простая
- Систематическая
- Стратифицированная
- Кластерная

Невероятностные

- Квотная
- Метод снежного кома
- Направленная
- По удобству

Простая вероятностная

Выборка в которой каждый элемент генеральной совокупности имеет одинаковую, заданную и независимую вероятность попадания в выборочную совокупность.

Преимущества:

- простота понимания процедуры
- структура генеральной совокупности неизвестна
- репрезентирует генеральную совокупность

Недостатки:

- Сложность реализации процедуры
- Географическая дисперсия выборочной совокупности
- Невысокая точность

Систематическая

Выборка в которой сначала из генеральной совокупности N случайно выбирается первый элемент выборочной совокупности i_1, а затем с шагом k отбираются все остальные элементы выборочную совокупности i_k.

Например, в совокупности из 20 единиц нужно выбрать 5 единиц. Значит, шаг будет равен 4. Случайно выберем первый элемент выборки, Пусть это будет 2, тогда выборку дополнят 6, 10, 14 и 18-ый элементы.

Преимущества:

- простота реализации процедуры
- структура генеральной совокупности не имеет значения

Недостатки:

- Не снижает географическую дисперсию выборочной совокупности
- Не повышает точность

Стратифицированная

Двухэтапная выборка, при которой сначала генеральная совокупность делится на страты (слои), каждая из которых содержит максимально сходные между собой единицы отбора, а затем внутри каждой из страт формируется выборочная совокупность с помощью простой случайной выборки.

Преимущества:

- увеличивается точность измерения
- репрезентирует генеральную совокупность
- Позволяет формировать непропорциональные страты

Недостатки:

- Необходимость знания структуры выборки генеральной совокупности
- Географическая дисперсия выборочной совокупности

Гипотетическая совокупность

Элемент	Доход (в долларах)	Образование (лет)	Подписка на газету		Элемент	Доход (в долларах)	Образование (лет)	Подписка на газету
A	5600	8	X		K	9600	13	X
B	6000	9	Y		L	10000	13	Y
C	6400	11	X		M	10400	14	X
D	6800	11	Y		N	10800	14	Y
E	7200	11	X		O	11200	15	X
F	7600	12	Y		P	11600	16	Y
G	8000	12	X		Q	12000	16	X
H	8400	12	Y		R	12400	17	Y
I	8800	12	X		S	12800	18	X
J	9200	12	Y		T	13200	18	Y

Распределение по числу выборок

выборочное среднее (рублей)	количество выборок простая	количество выборок стратифицированная
Не более 6100	2	
от 6101 до 6600	7	
От 6601 до 7200	11	
От 7201 до 7800	16	3
От 7801 до 8400	20	12
От 8401 до 9000	25	21
От 9001 до 9600	28	28
От 9601 до 10200	25	21
От 10201 до 10800	20	12
От 11801 до 11400	16	3
От 11401 до 12000	11	
От 12001 до 12600	7	
12601 и более	2	

Определение средней и среднеквадратичной ошибки

1 страта		2 страта	
Элемент	Доход	Элемент	Доход
B	6000	N	10800
E	7200	S	12800
Среднее: $\bar{y}_1 = \frac{6000 + 7200}{2} = 6600$		$\bar{y}_2 = \frac{10800 + 12800}{2} = 11800$	
Дисперсия : $s_1^2 = \frac{(6000 - 6000)^2 + (7200 - 6000)^2}{2 - 1} = 72000$		$s_2^2 = \frac{(10800 - 11800)^2 + (12800 - 11800)^2}{2 - 1} = 2000000$	
Дисперсия оценки: $s_{\bar{y}_1}^2 = \frac{s_1^2}{n_1} = \frac{7200}{2} = 360000$		$s_{\bar{y}_2}^2 = \frac{s_2^2}{n_2} = \frac{2000000}{2} = 1000000$	
Полная выборка			
Среднее: $\bar{y} = \frac{10}{20}(6600) + \frac{10}{20}(11800) = 9200$			
Дисперсия оценки: $s_{\bar{y}}^2 = \left(\frac{10}{20}\right)^2 (360000) + \left(\frac{10}{20}\right)^2 (1000000) = 340000$			
Среднеквадратичная ошибка оценки: $s_{\bar{y}} = \sqrt{s_{\bar{y}}^2} = 583$			

Кластерная

Выборка в которой сначала генеральная совокупность делится на кластеры (гнезда), каждый из которых имеет примерно ту же степень разнообразия единиц, что и генеральная совокупность в целом. Затем производится случайная выборка кластеров и внутри каждого производится либо сплошной, либо выборочный сбор данных.

Кластер можно назвать уменьшенной копией генеральной совокупности. Кластеры – непересекающиеся и исчерпывающие генеральную совокупность подмножества.

Преимущества:

- Снижает географическую дисперсию выборочной совокупности

Недостатки:

- Не снижает, а часто увеличивает ошибки при одинаковом объеме выборки

Территориальная выборка

Кластерная выборка чаще всего используется в случаях, когда необходимо собрать данные в генеральной совокупности, распределенной по значительной территории. Например, среди населения в большом городе. При этом есть предположение, что степень разнообразия полученных данных внутри каждого кластера не будет меньше разнообразия по городу в целом.

В качестве кластера в городе можно использовать избирательные участки.

1. ИУ – локализованы на небольших территориях, имеют небольшую и примерно одинаковую численность избирателей (от 1500 до 2600).
2. Не пересекаются и исчерпывают генеральную совокупность подмножества.
3. Регулярно обновляются государственными органами власти и легко доступны.

Территориальная выборка

1 этап – генеральная совокупность разделена на непересекающиеся, исчерпывающие генеральную совокупность, сравнимые по объему друг с другом кластеры – избирательные участки.

2 этап – производится выборка из этих (ИУ) кластеров. Количество кластеров определяется количеством интервьюеров. Если есть 20 интервьюеров необходимого качества, то можно выбрать 20 участков. Тогда, для опроса 1000 респондентов в городе, на каждом из нужно выбрать 50 респондентов. Если на среднем участке зарегистрировано примерно 2200 избирателей, значит, необходимо опросить примерно каждого 44-ого жителя. А, учитывая, что в отдельном домохозяйстве проживает чуть менее трех человек, то респондент должен находиться в каждом пятнадцатом.

3 этап – отбор домохозяйства внутри каждого из кластеров (ИУ). Существует в тех случаях, когда необходимо произвести выборку домохозяйств. Если данный отбор реализуется с помощью вероятностных выборок, то результат будет также вероятностным.

По удобству

Выборка в которой выборочная совокупность формируется исходя из возможностей исследователя. Чаще всего, процесс выборки локализован в одном месте и в одно время.

Опросы студентов, учащихся, слушателей курсов и тренингов, участников собраний и конференций.

Опрос посетителей торговых центров без использования процедур отбора и фильтрации

Опрос читателей журнала, газеты

Опрос на каком-либо неопросном интернет-ресурсе

Преимущества:

- Невысокая стоимость
- Оперативность

Недостатки:

- Значительная систематическая ошибка

Направленный отбор

Выборка в которой выборочная совокупность из тех единиц генеральной, которые по мнению исследователя отвечают целям исследования. Отбор может происходить как на основе простых характеристик (социально-демографических), так и на основе сложным (политические и потребительские предпочтения, стиль жизни и пр.)

Преимущества:

- Низкая стоимость
- Небольшие требуемые гуманитарные ресурсы

Недостатки:

- Высокая субъективность отбора
- Возможность значительной систематической ошибки

Квотный отбор

Выборка в которой вначале выбираются критерии для отбора респондентов – пол, возраст, район проживания, партийные или потребительские предпочтения и пр. Исходя из представлений исследователя о долях имеющих такие характеристики в популяции (полученных, например, от органов государственного статистического учета) формируются квотные задания для интервьюеров. На втором этапе интервьюеры реализуют индивидуальные квотные задания любым из детерминированных способов отбора – по удобству, направленному или «снежным комом».

Преимущества:

- Низкая стоимость
- Небольшие требуемые гуманитарные ресурсы

Недостатки:

Квотный отбор

Если выбраны релевантные целям данного исследования и значимые характеристики, то результаты данного отбора будут формировать репрезентативную выборочную совокупность.

Преимущества:

- Низкая стоимость
- Высокая скорость сбора данных
- Невысокая стоимость

Недостатки:

- Высокая субъективность отбора (может быть компенсирована большим числом качественных интервьюеров)
- Возможность значительной систематической ошибки при неверном определении квотных параметров
- Требование определять всякий раз определять набор квотных параметров

Снежный ком

Этап формирования выборочной совокупности, который проводят после отбора респондентов по любой из схем вероятностного отбора (простой, систематический, стратифицированный или кластерный).

Чаще всего, используются тогда, когда целевая группа крайне немногочисленна, но когда ее члены лучше знакомы друг с другом, чем средний представитель жителей данного населенного пункта. Например, мамы маленьких детей лучше знакомы друг с другом, чем их же соседи.

Преимущества:

- Незаменим для узких целевых групп
- Сокращает время опроса

Недостатки:

- Нерепрезентативность
- Увеличивает систематическую ошибку

Реализация репрезентативной выборки в массовом опросе

Лекция 7

Звоновский, к.с.н.

Территориальный дизайн выборки

Формирование выборки

ИПН Самарской области строится **на основе данных опросов** общественного мнения, проводящихся **один раз в три месяца** Фондом социальных исследований.

Индекс потребительских настроений (ИПН) представляет собой количественный показатель, отражающий диспозицию населения к наиболее общим формам потребительского поведения в контексте оценок личного материального положения и экономической ситуации в целом. Данный индекс был предложен специалистами Университета Мичигана в 1946 году (Consumer sentiment index). В настоящее время это ведущий индекс США для прогнозирования потребительской активности населения.

Опрос производился по специально спроектированной **многоступенчатой выборке**, репрезентирующей взрослое (старше 18 лет) население Самарской области.

Выборка спроектирована для воспроизведения именно **потребительского поведения** населения области.

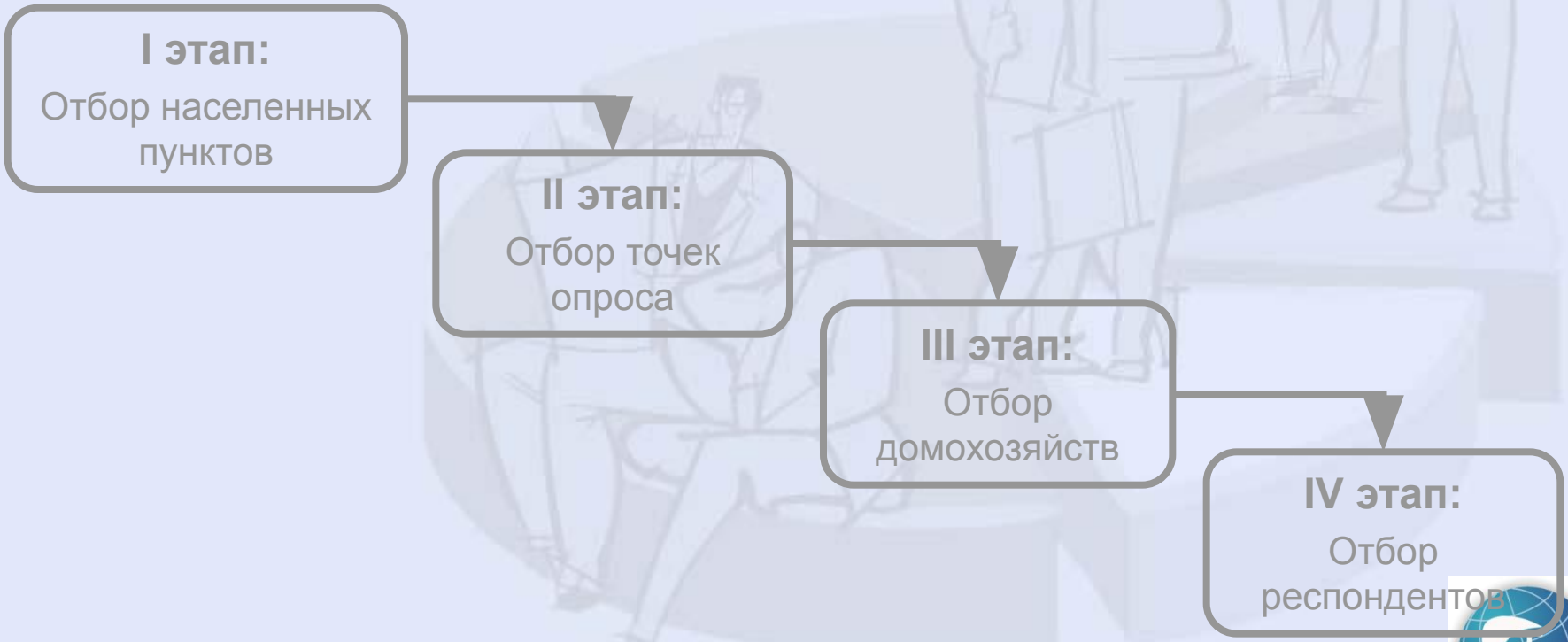
В марте было проведено тестовое измерение ИПН в г. Самаре (объем выборки – 544 респондента). Объем выборки в I и II волне ИПН – уже в рамках всей Самарской области – составлял 1202 и 1154 человек соответственно, в последней, III волне – 1024 человек.

Многоступенчатость отбора была призвана **обеспечить** необходимую **точность воспроизведения** структуры населения области. Она выразилась в применении методов **стратификации** и **кластеризации** по основным демографическим признакам: месту жительства, полу и возрастной группе.

Этапы формирования выборки

Отбор производился в **четыре этапа**. На **первом** этапе отбирались населенные пункты, где должен был проводиться опрос. На **втором** – точки опроса, представляющие собой избирательные участки. **Третий** этап включал в себя отбор домохозяйств. **Четвертый** этап – отбор конкретных респондентов.

В основу стратификации по месту жительства положены следующие критерии: **размер** населенного пункта и его **расположение** относительно областного центра и городов.



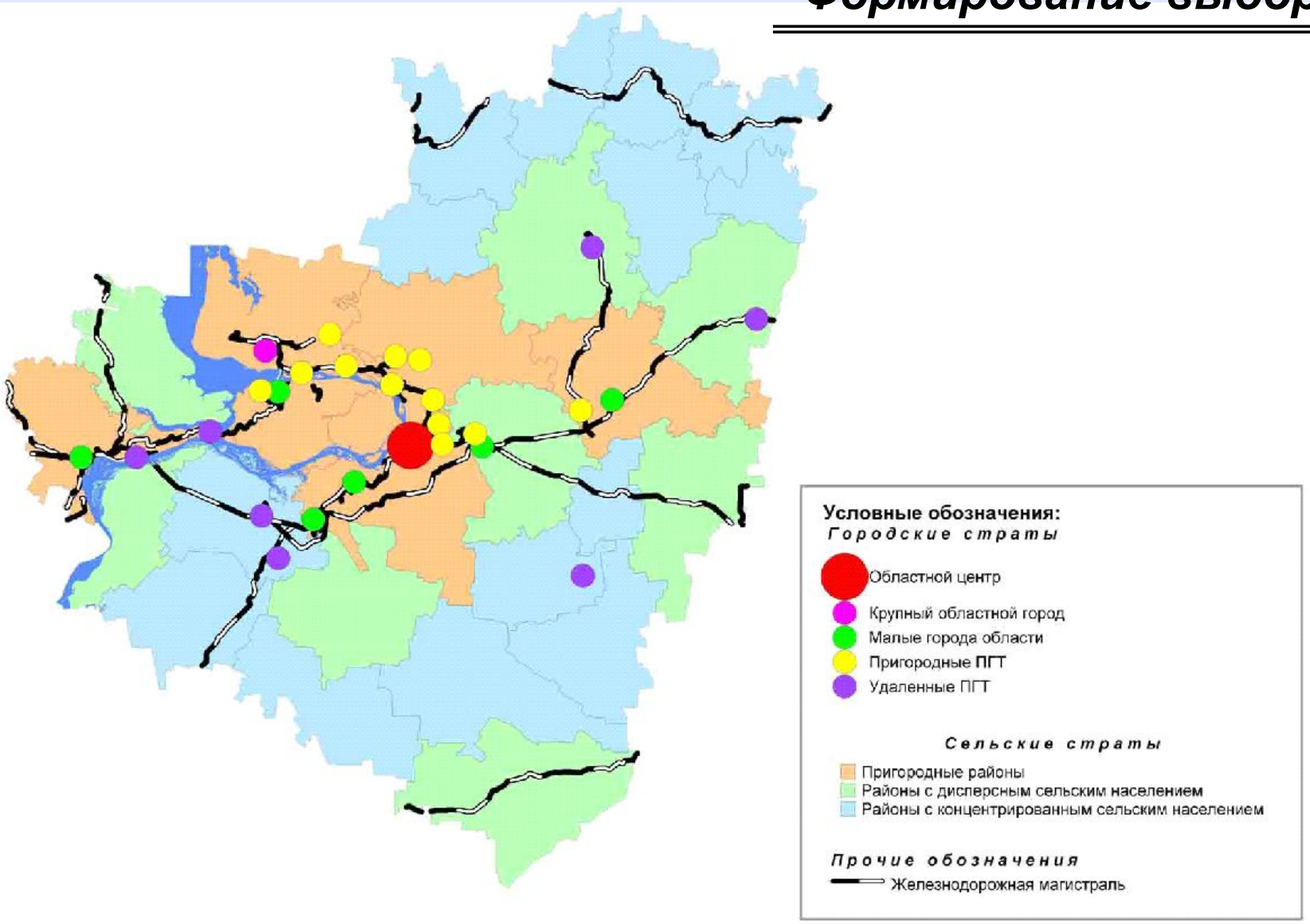
Стратификация области по месту жительства

На **первом** этапе все населенные пункты области были стратифицированы на восемь частей по типу поселения, исходя из приближенности к крупным локальным рынкам:

1. Областной центр (городское население Самары),
2. **Крупный областной город (городское население Тольятти),**
3. Малые города области (городское население Сызрани, Новокуйбышевска, Чапаевска, Отрадного, Жигулевска, Кинеля),
4. **Пригородные ПГТ (население крупных ПГТ, прилегающих к городам области, составляющим три первые страты),**
5. Удаленные ПГТ (городское население Октябрьска, Нефтегорска, Похвистнево, а также население крупных ПГТ, расположенных вне непосредственной близости к городам области, составляющим три первые страты),
6. **Пригородные районы (население сельских пунктов и малых ПГТ, прилегающих к городам области, составляющим три первые страты),**
7. Районы с дисперсным сельским населением (население сельских районов, которые насчитывают более одного крупного населенного пункта, расположенного на их территории),
8. **Районы с концентрированным сельским населением (население сельских районов, на территории которых расположен единственный крупный населенный пункт).**

Охват мелких поселений при реализации данной выборки определяется необходимостью учесть степень концентрации сельских населенных пунктов, влияющей на потребительское поведение их жителей

Формирование выборки



Формирование выборки

Далее **городские** страты были стратифицированы с целью максимально точного **воспроизведения** в выборочной совокупности **соотношения** населения в отдельных городах и городских районах населенных пунктов **первых трех** страт.

Наконец, в стратах была проведена кластеризация с целью **представить** в выборочной совокупности **доли** этих страт в генеральной совокупности. За **кластеры** принимались **избирательные участки** области.

Остальные страты (пгт и села) были кластеризованы с целью **представить** их в выборочной совокупности **пропорционально доле** этих страт в генеральной совокупности. За **кластеры** (единицы отбора) принимались **населенные пункты**, численность которых составляет **400 и более** человек всех возрастов.

Формирование выборки

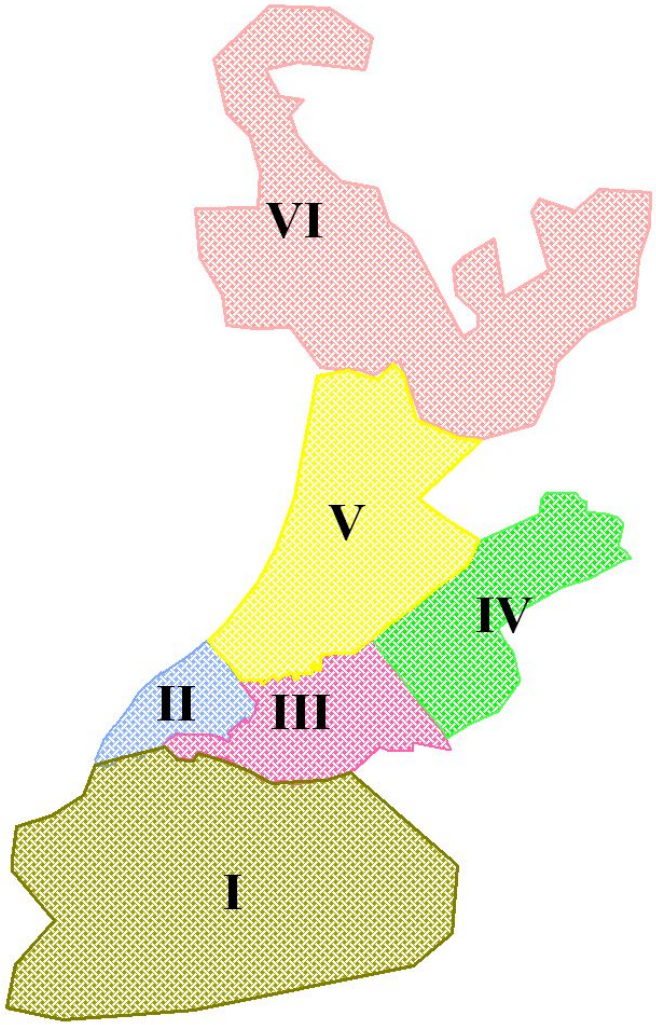
«Крупный областной город» **Тольятти** был стратифицирован по **административным районам** города с образованием **трех страт**: Автозаводской, Комсомольской и Центральной.

Исследовательский опыт показывает, что **деление** Самары по **административным районам** не всегда оправдано, поскольку различия в настроениях населения определяются другими, менее строгими границами.

Формирование выборки

Самара делится на **4+2 страты**: на **4** делится **основная часть** города, части примерно равны, границы частей проходят по границам избирательных участков 2003 года, оставшиеся **2 страты** – **удаленные части** города – Куйбышевский район с одной стороны, и Красноглинский район – с другой.

Стратам были приданы **веса** в соответствии с **долей населения** города, проживающего на данной территории.



Формирование выборки

Третий и **четвертый** этапы отбора (отбор домохозяйств) был **различным** для **Самары** и **Тольятти** (крупнейших городов области), с одной стороны, и **остальных населенных пунктов**, с другой.

Для **всех населенных пунктов**, кроме **Самары** и **Тольятти**, отбор **домохозяйств** (**третий** этап) проводился **по маршруту с заданным шагом**, то есть интервьюер получал описание избирательного участка и обходил его с самого начала по порядку. Порядок определялся интервьюером.

В домохозяйствах (**четвертый** этап) **респонденты** отбирались согласно **методике «ближайшего дня рождения»**.

Формирование выборки

В **Самаре** и **Тольятти** третий и четвертый этапы формирования выборки были реализованы иным способом. В этих городах была проведена предварительная работа – **восстановлена** (составлена) **полная база домохозяйств**, принадлежащих отобранным избирательным участкам.

Из этой базы с помощью специального программного обеспечения (модуль SPSS Complex Samples) **случайным образом** были отобраны **домохозяйства** для проведения интервью.

Четвертый этап (отбор **респондентов**) в Самаре и Тольятти **в разное время** осуществлялся по **двум разным** схемам: **адресной** и **именной**.

Адресная и именная схемы выборки

Адресная схема отбора респондентов

Принцип

При адресной схеме отбора каждый интервьюер должен был опросить на выданном ему избирательном участке определенное (также указанное руководителем работ) число респондентов (15 – 17 человек). С этой целью интервьюеру выдавался список адресов участка, число которых вдвое превосходило число требуемых законченных интервью.

В домохозяйствах респондент отбирался согласно методике ближайшего дня рождения.

Данная схема отбора респондентов использовалась в I волне ИПН (июнь).

Квотные ограничения

Помимо этого каждому интервьюеру выдавалось квотное задание, в котором было указано, сколько респондентов определенного пола и возраста должен опросить интервьюер на своем участке.

До тех пор, пока ни одна из квот не выбрана, интервьюеры отбирали и опрашивали респондентов «по ближайшему дню рождения».

После того, как любая первая квота была выбрана, интервьюер переставал опрашивать тех респондентов, которые должны были быть опрошены согласно отбору по ближайшему дню рождения, и мог опросить другого члена данного домохозяйства, если он не являлся представителем также выбранной квоты.

Если же все члены данного домохозяйства являлись представителями выбранных квот, то интервьюер переходил к другому адресу.

Именная схема отбора респондентов

Принцип

Из полной базы респондентов по Самаре и Тольятти с помощью специального программного обеспечения (модуль SPSS Complex Samples) случайным образом были отобраны конкретные респонденты для проведения интервью.

Интервьюер для опроса получал список из адресов, количество которых превышало необходимое количество законченных интервью в n раз – коэффициент запаса.

Когда интервьюер достигал респондента, прежде чем проводить опрос, необходимо было сверить правильность написания его/ее имени, даты рождения и адреса с указанными в бланке.

Данная схема отбора респондентов использовалась в мартовской (тестовой и проводившейся только в Самаре), II (сентябрь) и III (декабрь) волне ИПН.

Тестовая волна показала значимое смещение половозрастной структуры выборочной совокупности относительно генеральной.

Квотные ограничения

Перед проведением сентябрьской волны интервьюерам выдавались квотные задания.

Реализация квотных ограничений состояла в том, что, когда в списке планируемых респондентов с запасом n квота старших возрастов была выбрана, интервьюер не мог в целях достижения количественного плана опроса (15 или 17 респондентов) опрашивать пожилых респондентов, и должен был либо работать с имеющимся списком, либо запрашивать у руководителя работ новый список потенциальных респондентов.

Данная техника не является квотированием выборки в чистом виде. Тем не менее, с целью реализации случайной выборки респондентов в чистом виде, в мартовской и декабрьской волнах исследования была использована исключительно методика случайного отбора без коррекции ее квотными заданиями.

Преимущества и недостатки адресной и именной выборок

Преимущества	Недостатки
Адресная выборка	
Нет ограничений на присутствие в домохозяйстве конкретного респондента	Необходимо строго соблюдать методику опроса и (следующая из этого) сложность контроля выполнения методики отбора
Меньше дефект базы (высокая валидность базы адресов)	Большое количество отказов от интервью
	Неконтролируемое смещение выборки в сторону менее «мобильных» респондентов
Именная выборка	
Проще осуществлять контроль над соблюдением методики	Больше дефект базы (низкая валидность базы жителей)
Небольшое количество отказов от интервью	Смещение выборки в сторону менее «мобильных» респондентов

Адресная vs. именная выборка

Результат контакта	именная март 2005	адресная июнь 2005	именная сентябрь 2005	именная декабрь 2005
оконченных интервью	32%	28%	32%	34%
респондента нет дома	6%	1%	5%	4%
никого нет дома	21%	19%	16%	8%
отказ от интервью	10%	34%	14%	20%
дефект базы	26%	2%	18%	24%
число контактов	2 526	3 568	3 265	3 126
число контактных адресов	1 669	2 360	2 180	1 662

Именная выборка позволяет существенно **увеличить** долю законченных интервью и **уменьшить** долю отказов.

Многократное посещение

Многократное посещение

Результат контакта	июнь 2005			сентябрь 2005			декабрь 2005		
	1 посещ.	2 посещ.	3 посещ.	1 посещ.	2 посещ.	3 посещ.	1 посещ.	2 посещ.	3 посещ.
оконченных интервью	19%	18%	16%	20%	23%	27%	19%	18%	17%
респондента нет дома	2%	3%	1%	9%	15%	11%	14%	16%	21%
никого нет дома	39%	51%	50%	29%	36%	45%	23%	32%	37%
отказ от интервью	25%	17%	17%	11%	9%	8%	12%	12%	9%
дефект базы	2%	0%	0%	17%	7%	0%	19%	8%	6%

const

Увеличение числа посещений увеличивает долю несостоявшихся контактов.
При этом доля законченных интервью остается примерно такой же.

Трехкратное посещение и мобильность молодежи

Количество посещений увеличивает охват мобильных респондентов.

Доли возрастных групп в числе поменявших место жительство респондентов

Возрастная группа	сентябрь 2005	декабрь 2005
18-35 лет	52%	47%
36-55 лет	34%	39%
56 лет и старше	14%	14%

возрастная группа	март 2005			сентябрь 2005			декабрь 2005		
	1 посещ.	2 посещ.	3 посещ.	1 посещ.	2 посещ.	3 посещ.	1 посещ.	2 посещ.	3 посещ.
18-35 лет	27%	35%	45%	32%	32%	48%	27%	36%	39%
36-55 лет	31%	36%	20%	38%	32%	38%	32%	41%	28%
56 лет и старше	42%	29%	35%	30%	36%	14%	41%	23%	33%

Увеличение числа посещений также увеличивает долю молодежи в выборке и приближает ее к доле молодежи в генеральной совокупности.

Шестикратное посещение

Однако даже шестикратное посещение не восстанавливает долю молодежи в генеральной совокупности.

возрастная группа	**декабрь 2005**				**генеральная совокупность**
	1 посещение	2 посещение	3 посещение	4+5+6 посещения	
18-35 лет	27%	36%	39%	42%	35%
36-55 лет	32%	41%	28%	33%	38%
56 лет и старше	41%	23%	33%	25%	27%

Возрастные группы респондентов, которых сначала не заставали дома, а затем все-таки опросили

Возрастная группа	адресная июнь 2005	именная сентябрь 2005	именная декабрь 2005	генеральная совокупность
18-35 лет	40%	36%	38%	35%
36-55 лет	37%	39%	38%	38%
56 лет и старше	23%	25%	24%	27%

В результате повторное посещение и при адресном, и при именном отборе, хотя и смещает выборку ближе к генеральной совокупности, не решает проблему репрезентации уже на полевом этапе исследований.

Качество базы жителей города и статистических данных

Возрастная группа	адресная июнь 2005	именная сентябрь 2005	именная декабрь 2005	генеральная совокупность	исходная база жителей
18-35 лет	35%	34%	32%	35%	33%
36-55 лет	38%	38%	36%	38%	34%
56 лет и старше	27%	28%	32%	27%	33%

Серьезной проблемой является вопрос о расхождении (причем, значимом при больших выборках) между возрастной структурой населения, предоставляемой органами государственной статистики и базами данными, чаще всего, представляющими собой базы данных паспортных столов, входящих в систему МВД.

Причем, приоритетным при решении этой проблемы является вопрос о том, какой из источников статистической информации является верным.

Поскольку на данный момент точного ответа на этот вопрос нет, будет корректным считать оба источника верными и неверными в равной степени. Поэтому следует усреднить данные о долях различных возрастных групп в генеральной совокупности и именно полученные в результате такой процедуры данные считать целевыми для коррекции выборки.

Возможное решение проблемы нехватки молодежи

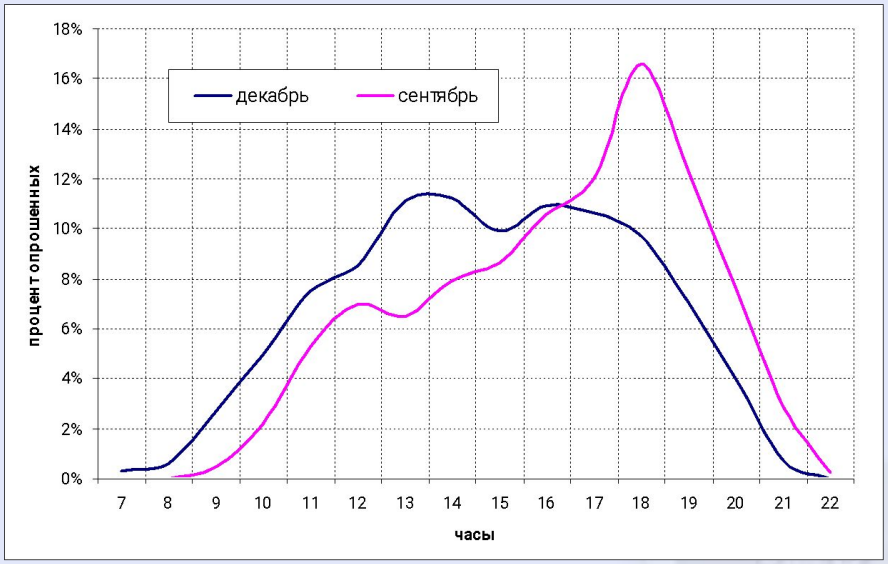
Поскольку задача репрезентации молодежи в выборочной совокупности даже после шестикратного посещения осталась нерешенной, необходимо устранить возникшее смещение в сторону респондентов среднего и пожилого возраста уже после окончания полевой части исследования. В принципе, для решения этой проблемы существует два пути.

Во-первых, можно искусственно увеличить объем выборки, т.е. сначала дополнительно взять столько интервью у всех возрастных групп, сколько требуется для достижения планового числа молодежной группы, а затем «отремонтировать» выборку, т.е. случайным образом исключить из выборочной совокупности излишние анкеты средней и старшей возрастных групп.

Второй путь – это перевзвешивание полученного массива по полу, возрасту и месту проживания. Несмотря на все недостатки «перевзвешивания» данных, именно этот способ коррекции финальной выборки представляется наиболее предпочтительным, поскольку позволяет сохранить в том или ином виде все собранные валидные материалы полевого этапа.

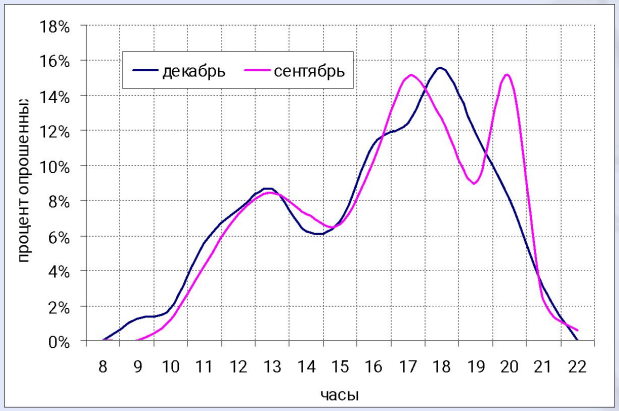
Суточная динамика результатов опроса

Суточная динамика результатов опроса

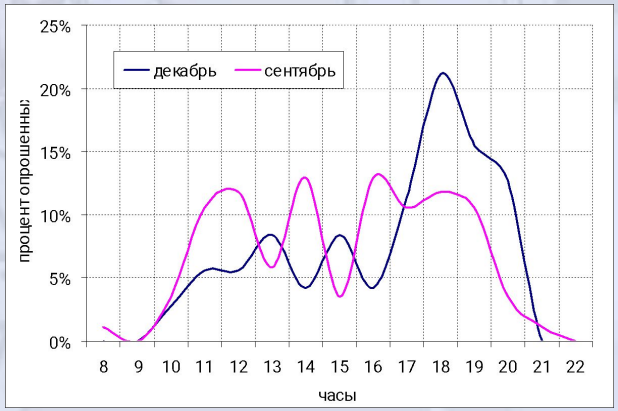


Хотя в декабре первое посещение интервьюер делал в более ранние часы, во второе и в третье посещения суточная динамика приближалась к сентябрьской.

Суточная динамика момента первого посещения в сентябре и декабре 2005 г.



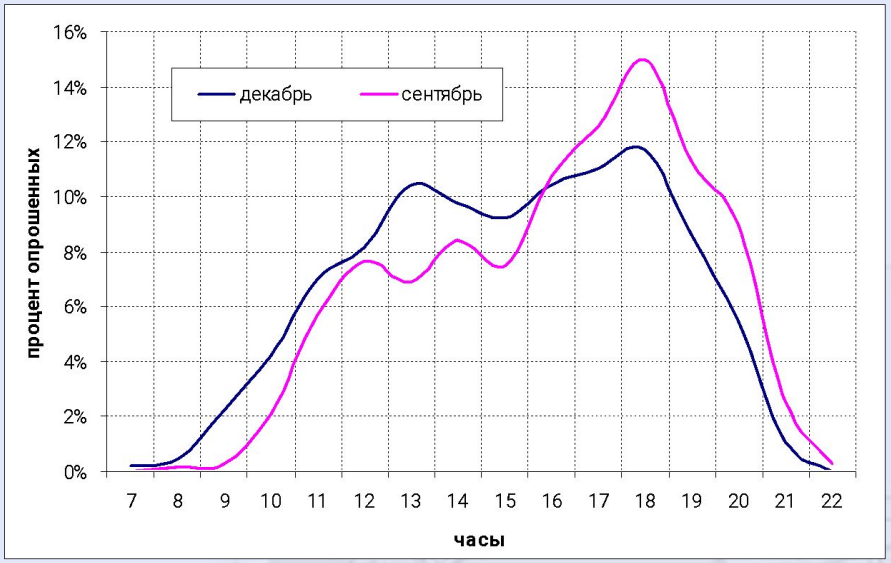
а)



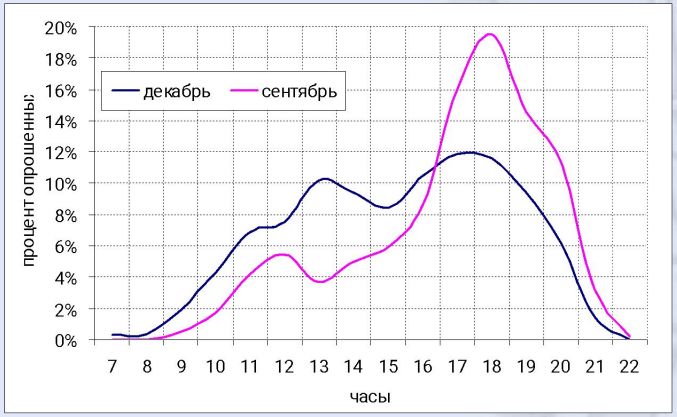
б)

Суточная динамика момента второго (а) и третьего (б) посещений в сентябре и декабре 2005 г.

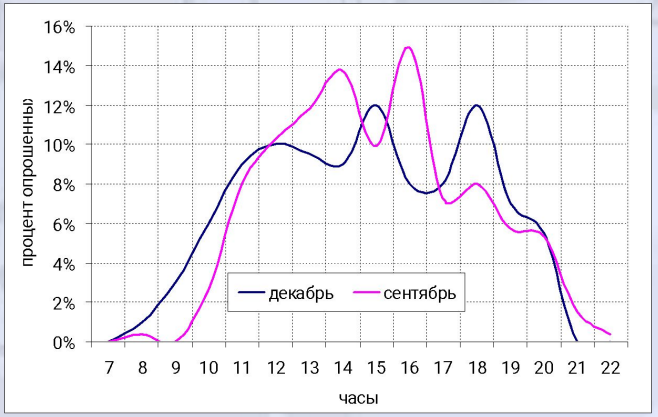
Суточная динамика результатов опроса



Суточная динамика момента опроса за все три посещения в сентябре и декабре 2005 г.



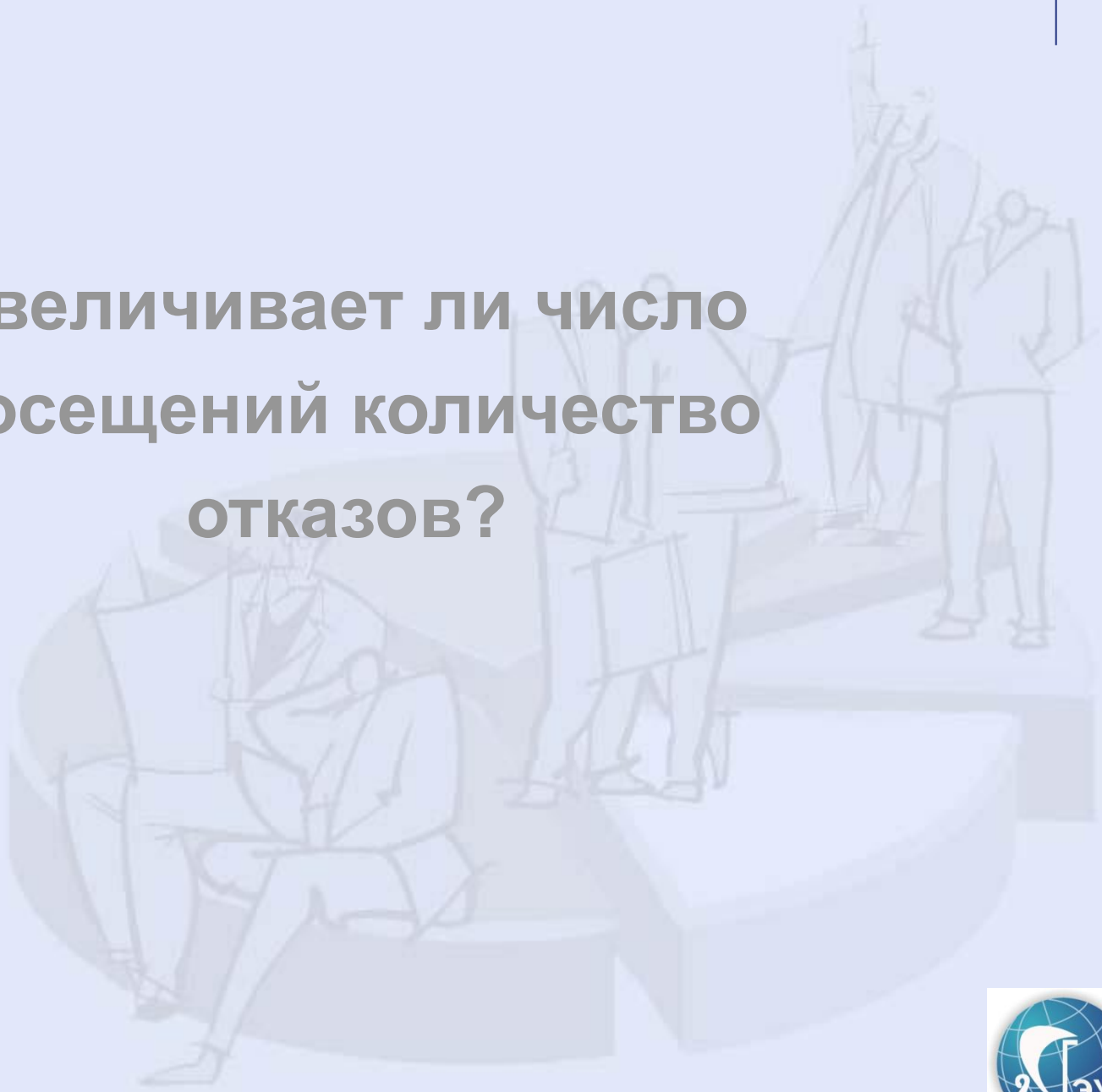
а)



б)

Суточная динамика опроса в будние (а) и в выходные (б) дни
в сентябре и декабре 2005 г.

Увеличивает ли число посещений количество отказов?



Увеличивает ли число посещений количество отказов?

Здравый смысл подсказывает, что повторные посещения могут привести к увеличению доли отказов.

Результат контакта	именная март 2005	адресная июнь 2005	именная сентябрь 2005	именная декабрь 2005
респондента нет дома	6%	1%	5%	4%
никого нет дома	21%	19%	16%	8%
отказ от интервью	10%	34%	14%	20%

Очень важным является контроль не только проведенных интервью, но и полученных ими отказов.

Рост отказов происходит, в основном, за счет достижения определенных групп населения по мере роста числа посещений.

Возрастные группы респондентов, отказавшихся от интервью после первого посещения

Возрастная группа	адресная июнь 2005	именная сентябрь 2005	именная декабрь 2005
18-35 лет	19%	32%	28%
36-55 лет	40%	41%	35%
56 лет и старше	41%	26%	37%

Возрастные группы респондентов, которых сначала не заставали, а потом они отказались от интервью

Возрастная группа	адресная июнь 2005	именная сентябрь 2005	именная декабрь 2005	Самара
18-35 лет	39%	43%	36%	35%
36-55 лет	49%	33%	36%	37%
56 лет и старше	12%	23%	28%	28%

Определение объема выборки

Лекция 8

Звоновский, к.с.н.

Расчет объема выборки

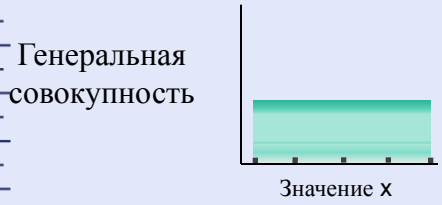
Выборочное измерение проводят с целью получить значение одного из количественных параметров генеральной совокупности

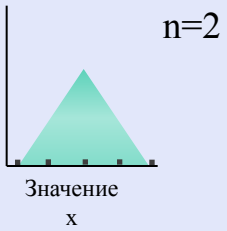
Поскольку мы имеем дело со статистической оценкой, то измерение имеет определенную точность и достоверность.

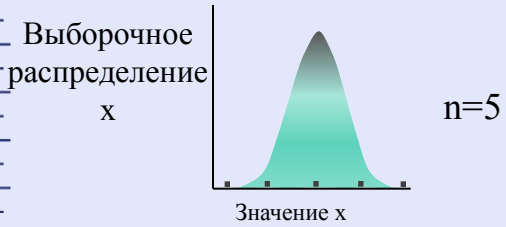
Точность – степень возможного отклонения выборочного среднего от генерального среднего. Определяется величиной доверительного интервала

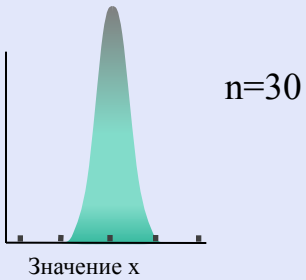
Достоверность – вероятность возможного выхода значения генерального среднего за пределы доверительного интервала, рассчитанного на данной выборочной совокупности.

Дисперсия оценки выборочного среднего определяет объем выборки









Центральная предельная теорема позволяет получить среднее значение уже при небольших объеме выборки.

Увеличение объема выборки дает возможность увеличить точность (уменьшить доверительный интервал) и увеличить достоверность измерения .

Увеличение размера выборки позволяет уменьшить величину средне квадратичной ошибки

$$\sigma_x = \sigma / \sqrt{n}$$

Объем выборки для оценки среднего

Случай когда выборочная дисперсия известна

Пусть выборочная оценка (результат измерения) не должна отклоняться от генерального среднего более, на ±25 рублей (доверительный интервал). Такова требуемая **точность.**

Пусть вероятность возможного выхода значения средней генеральной совокупности составит за пределы указанного интервала (**достоверность** измерения) составит 95%.

Поскольку $\mu - z^*\sigma_x = \dot{X}$. Тогда $\mu - \dot{X}$ (точность) $= z^*\sigma_x$.

А поскольку - $\sigma_x = \sigma / \sqrt{n}$, то $H = z^*\sigma / \sqrt{n}$, или

$$n = \sigma^2 * z^2 / H^2$$

Объем выборки для оценки среднего

Случай когда выборочная дисперсия известна

Пример. Необходимо определить объем выборки для оценки размера среднего чека в магазине с точностью ±250 рублей и достоверностью 95%. При этом дисперсия генерального среднего 1000 рублей.

Тогда размер выборки:

$$n=\sigma^2*z^2/H^2$$

$$n=1000^2*2^2/250^2=64$$

Очевидно, что если **точность** уменьшить вдвое, то требуемую выборку придется увеличить вчетверо.

Объем выборки также возрастет, если мы увеличим **достоверность**.

Объем выборки для оценки среднего

Случай когда выборочная дисперсия неизвестна

При первом расчете выборки мы оцениваем дисперсию генеральной совокупности.

При повторении расчета выборки (при имеющейся выборки) мы принимаем за дисперсию генеральной дисперсию выборочной совокупности.

Как можно оценить дисперсию генеральной совокупности?

1. На основе данных переписи.
2. На основе предыдущих исследований.
3. На основе косвенных данных.
4. На основе нормального закона распределения выборочной совокупности.

Объем выборки для оценки среднего

Случай когда выборочная дисперсия неизвестна

Оценка дисперсии: 15 посещений магазина в месяц и 300 рублей примерный средний чек в день. Итого 4500 рублей в месяц. Можно предположить, что дисперсия 4500/6=750 рублей. Тогда, планируемый объем выборки – 36 единиц.

Предположим, что в результате измерения выборочное среднее - $\dot{X}$=350 рублей, а дисперсия – 600 рублей.

Тогда доверительный интервал: $\dot{X} \pm 2*\sigma/\sqrt{n}$

$$350 \pm 2*600 / \sqrt{36}$$

$$n= 350 \pm 200$$

Интервал уже, чем предполагался.

Объем выборки в случае конечной генеральной совокупности

В случае, если объем **выборочной совокупности составляет значимую долю генеральной** (5% и более) необходимо делать поправку на объем выборки:

$$\sigma_x = \sigma / \sqrt{n} * \sqrt{(N-n)/(N-1)}$$

Объем выборки для оценки доли

Распределение выборочных долей при небольших объемах выборки (n=30) является биноминальным. Но при больших объемах выборки его можно аппроксимировать нормальным.

Среднеквадратичная ошибка доли

$$H=\sqrt{\rho(1-\rho) / n}$$

А объем выборки

$$n= \rho(1-\rho)*z^2/H^2$$

Пример: Требуется получить оценку доли жителей микрорайона вокруг магазина среди покупателей магазина с точностью ±2% и доверительном уровне 95% (z = 2).

$$n = 2^2/ (0,02)^2 * \rho(1-\rho)$$

Коррекция объема выборки

Коррекция на **инцидентность (проникновение).** В случае, если в выборочной совокупности доля целевой подгруппы составляет менее 100%, необходимо увеличивать объем выборки для того, чтобы представители целевой подгруппы в нее попали в необходимом количестве.

Пусть расчет показал, что нам необходимо опросить **1000** респондентов, но опросу подлежат лишь женщины от 20 до 55 лет, а таких в городе **33%**. Тогда расчетную выборку необходимо увеличить в 1000/0,33 = **3** раза.

Коррекция объема выборки

Коррекция **на неполное заполнение.** В случае, если анкеты заполнены не полностью, необходимо увеличить объем собранных данных по целевым и вспомогательным переменным, чтобы в финальном массиве данных было минимально необходимое число данных в требуемом объеме.

После сбора всех данных оказывается, что отдельные части анкет остаются незаполненными. По этой причине следует увеличить выборку на какое-то число записей для восполнения очевидного недостатка.

Например, исследователь решает что для целей исследования необходимо, чтобы анкета была заполнена на 90%. Из предыдущих измерений он знает, что анкет с меньшей заполненностью будет не более 5%. Тогда он должен увеличить начальный объем выборки в 1/0,95=1,05 раза.

Коррекция объема выборки

Объем выборки следует увеличивать в случае измерения параметра в перекрестных группах. Например, доли сторонников кандидата необходимо измерить среди мужчин с доходом от 20 до 40 т.р. с аналогичной долей среди женщин с двумя и более детьми.

В этом случае необходимо рассчитать тот объем выборки, который будет достаточен для измерения искомого параметра в данной целевой подгруппе. Если эта группа составит, например, 15% от общей выборки, значит, всю выборку необходимо увеличить в 1/0,15= 6,7 раз.

Коррекция объема выборки

Мы нигде не указывали фактор цены и себестоимости измерения, хотя при определении типа отбора и расчете объем выборки, он часто имеет решающее значение.

Есть формулы, которые учитывают фактор цены, но они имеют лишь приблизительное значение и применимы лишь к узкой группе случаев и чаще всего к одному региону или небольшой их группе.