АНАЛИЗ ЭМПИРИЧЕСКИХ ДАННЫХ

Виды анализа данных

Эмпирические данные могут быть представлены в виде:

- совокупности чисел, характеризующих те или иные объекты;
- множества индикаторов определенных отношений между рассматриваемыми объектами;
- результатов попарных сравнений респондентами каких-либо объектов;
- совокупности определенных высказываний (например, при ответе респондентов на открытые вопросы);
- текстов документов;
- так или иначе зафиксированных результатов наблюдения за невербальным поведением каких-либо людей и т.п.

Группировка, табулирование и представление данных

- До начала анализа данные необходимо сгруппировать, упорядочить по одному признаку. Когда данные сгруппированы, по каждой группе устанавливается ее абсолютная частота (число наблюдений в данной выборке) и относительная частота (т.е. доля каждой группы в общей массе наблюдений). Результаты представляют в виде таблицы частотного распределения для каждой переменной.

Таблица 1

Пол	Количество (абсолютная частота)	% от общего кол-ва респондентов (относительная частота)
Мужской	160	40
Женский	240	60

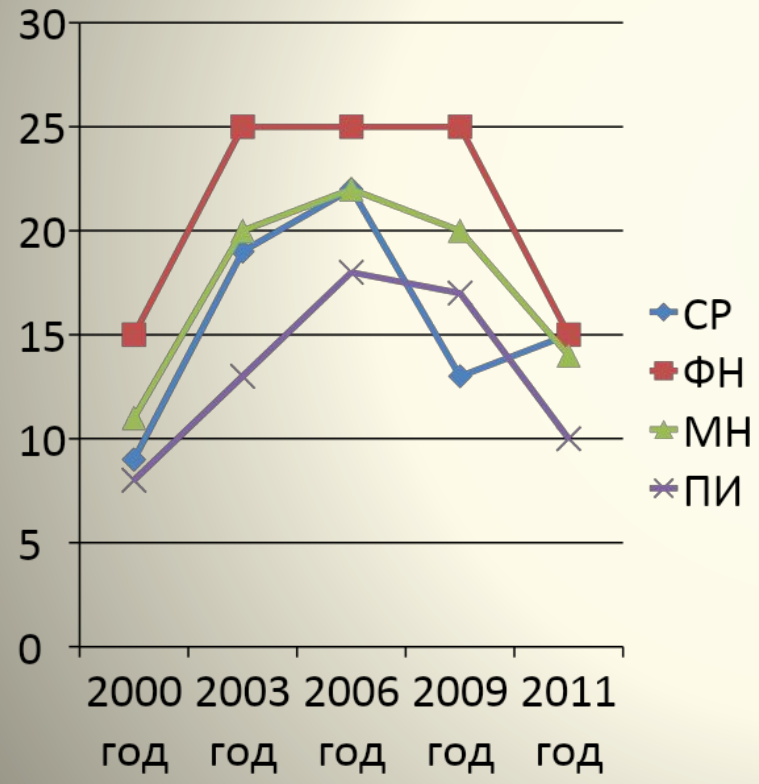
Графическое представление данных

Круговая диаграмма

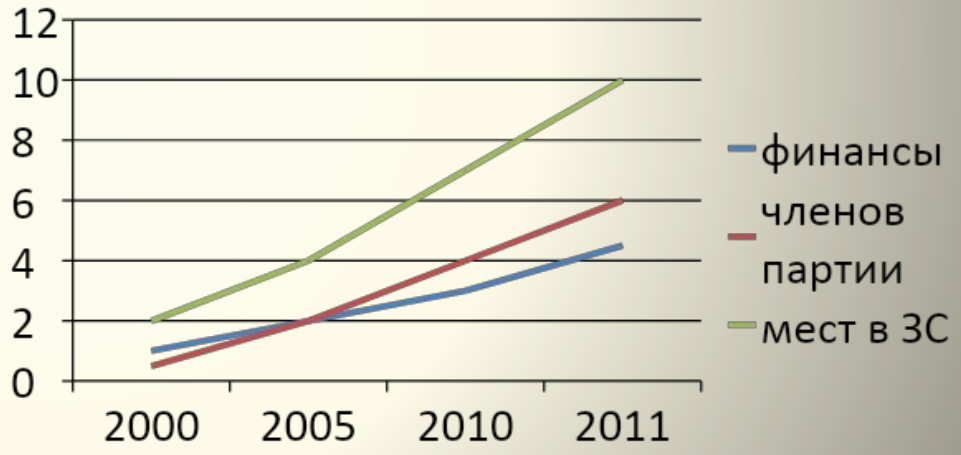
Гистограмма

Графики связи двух и более переменных

Динамика набора студентов НШФ ЮФУ



Корреляция роста численности и влияния партии с финансированием



Группировка

- В примере с двумя вариантами значений переменной (пол: либо мужской, либо женский) производить вычисления и строить частотные распределения достаточно легко. В случае же, если таких значений много (например, 200 респондентов указывают свой возраст в годах: количество вариантов может быть порядка 50-70, в зависимости от того, кто попал в выборку), для облегчения работы с частотными распределениями, используют группировку и обобщение, а также обобщающие числовые значения – **статистики**.
- Группировка:
- Сгруппировать данные о возрасте респондентов: 20,20,21,22,22,23,25,26,27,28,28,29,30,32,33,34,34,35,35,35,36,37,38,39,40,40,40,41,42,43,44,44,45,47,48,49: в две, три, шесть групп.

Статистики

Наибольшее значение имеют две группы статистик: меры центральной тенденции и меры изменчивости (разброса).

Меры центральной тенденции	Меры изменчивости
• Характеризуют расположение типичного значения признака, вокруг которого «сгруппированы» остальные наблюдения:	• Характеризуют степень неоднородности, несхожести данных, отклонения от типичного значения:
• Мода • Медиана • Среднее арифметическое • Отклонение от среднего	• Размах • Коэффициент вариации • Дисперсия • Стандартное отклонение

Меры центральной тенденции

1. **Мода (Mo) -** значение наблюдений, которое встречается наиболее часто. Например, группа респондентов указала свой возраст: 35, 24, 25, 28, 27, 30, 25, 33, 36 лет. В данном случае Mo = 25 лет (встречается дважды). В распределении могут быть две и более моды, либо мода может отсутствовать.
2. **Медиана (Md)** - это значение, которое делит упорядоченное множество данных пополам, так что одна половина наблюдений оказывается не меньше медианы, а другая – не больше. Для нахождения медианного значения для небольшого массива наблюдений упорядочивают наблюдения от меньших значений переменной к большим: то значение, которое оказывается в центре, и является медианой.

Для примера с возрастом получим:24, 25, 25, 27, 28, 30, 33, 35, 36 Md = 28 (т.е. 4 значения (24, 25, 25, 27) – не больше 28 и 4 значения (30, 33, 35, 36) - не меньше). Если число значений в группе наблюдений четное, то медианой будет среднее двух центральных значений упорядоченной совокупности.

3. **Среднее арифметическое** (X). Высчитывается путем деления суммы всех значений наблюдений на число наблюдений. Средний возраст респондентов (из приведенного примера) составит: (24 + 25 + 25 + 27 + 28 + 30 + 33 + 35 + 36)/ 9 = 29,2 года
4. **Отклонением от среднего** называется разность между значением отдельного признака совокупности и средним для данной совокупности

Меры изменчивости, разброса

1. **Размах** - описывает диапазон изменчивости значений. Так, в примере с возрастом размах = 36-24=12 лет
2. **Коэффициент вариации** (V) - процент наблюдений, не совпадающих с модальным значением. В нашем примере от модального отличаются 78 % значений, значит *V*=78 % (или *V*=0,78).
3. **Дисперсия.** Представительную информацию о вариации совокупности значений относительно среднего дают отклонения от среднего. Однако, поскольку сумма всех значений отклонения равна нулю (основное свойство средних), то в данном случае используют **квадраты отклонений,** вернее, их сумму. (Если данные однородны, то сумма квадратов отклонений будет маленькой, и, наоборот, когда данные неоднородны – большой). Для возможности сравнения сумм квадратов отклонений выборок разного размера каждую из них делят на *N,* где *N* - объем выборки. Полученная величина называется **дисперсией (*S*2)**. (***методичка с.45)***
4. **Стандартное отклонение.** Величина, равная квадратному корню из дисперсии, называется ***стандартным отклонением (s)***. Для небольших выборок *(N*< 100) лучше делить на *(N*-1). ***(методичка с.45)***

Самое главное значение стандартного отклонения - продемонстрировать «типичность» среднего: чем оно меньше, тем с большей вероятностью можно говорить, что среднее представительно для данной совокупности наблюдений.

Практическое задание

- Рассчитать статистики для данных о возрасте:

20,20,21,22,22,23,25,26,27,28,28,29,30,32,33,34,34,35,35,35,36,37,38,39,40,40,40,41,42,43,44,44,45,47,48,49:

- Мода
- Медиана
- Среднее арифметическое
- Отклонение от среднего для 20 и 49 (25, 35 и 45)
- Размах
- Коэффициент вариации

Проверка результатов

20,20,21,22,22,23,25,26,27,28,28,29,30,32,33,34,34,35,35,35,36,37,38,39,40,40,40,41,42,43,44,44,45,47,48,49:

- Мода – 35 и 40
- Медиана - 35
- Среднее арифметическое – 33,86 (39)
- Отклонение от среднего 1й – 13,86 / 36й – 15,14
- Размах - 29
- Коэффициент вариации – 83,6 %

Рассчитать статистики, в том числе дисперсию и стандартное отклонение

- За 2011 год Ира посетила драматический театр 2 раза, Света 3 раза, Юля 5 раз, Лена 6 раз.
- Мода
- Медиана
- Среднее арифметическое
- Размах
- Коэффициент вариации
- Дисперсия
- Стандартное отклонение

Анализ связи между двумя переменными

- При всей важности одномерного анализа в исследованиях основное внимание обычно уделяется анализу связей между переменными, поскольку именно это позволяет делать выводы о причинно-следственных связях, подтверждать или опровергать выдвинутые гипотезы. Самым распространенным является анализ взаимосвязи (сопряженности) двух переменных.
- Первым этапом этого процесса является перекрестная классификация, или построение таблицы сопряженности признаков (т.е. исследователю необходимо проследить информацию о совместном появлении переменных).
- См. методичку с. 88-89-90.

Сопряженность признаков

Рейтинг партии	Финансирование
1 Единая Россия	5000000
2 КПРФ	3000000
3 ЛДПР	2000000
4 Справедливая Россия	3000000
5 Яблоко	1000000

Пришел бы снова	Перешел бы на другой факультет		
Пришел бы	Нет	Затрудняюсь ответить	Да
Да	a	b	f
Затрудняюсь ответить	b	c	d
Нет	f	d	e

Рейтинг партии ЕР	Финансирование
8 (2000 г.)	1000000
6 (2001 г.)	2000000
6 (2002 г.)	1800000
4 (2003 г.)	3000000
3 (2004 г.)	3400000
2 (2005 г.)	4000000
1 (2006 г.)	5000000

Рейтинг партии	Частота появления в СМИ в месяц
1 Единая Россия	50
2 КПРФ	40
3 ЛДПР	40
4 Справедливая Россия	20