ВВЕДЕНИЕ В СТАТАНАЛИЗ

О.А. Клиценко

СТАТИСТИКА – наука о сборе, представлении и анализе данных

БИОСТАТИСТИКА (биометрия) – статистика в приложении к демографии, эпидемиологии, клиническим исследованиям

Из теории информации

Данные - функциональные значения информационных кодов для действий аппарата их интерпретации, абстрагированные от природы симметричных взаимодействий лежащих в основе переноса этих кодов.

Диссертация – «информационный продукт»

Процесс исследования

1. Замысел, основная идея исследования (из предмета, целей, задач) (II):

- **есть эффект - нет эффекта;**
- **выше – ниже;**
- **связь есть – связи нет;**
- **причина – следствие.**

2. Дизайн исследования (план, схема работы):

a) **Единица исследования;**

b) **Ее характеристики;**

c) **Группы наблюдений, способы их формирования;**

d) **Этапы наблюдений и требования к ним (динамика).**

Окончательная детализация гипотез (III):

- ✔ **что предполагаем об отдельных параметрах в конкретных группах, подгруппах;**
- ✔ **что предполагаем о соотношениях.**

3. Выбор методов и методик исследования (целесообразность, возможность).

Процесс исследования

4. Информация:

- **состав, структура;**
- **способ фиксации;**
- **точность измерений;**
- **правила кодирования;**
- **объем выборок, размеры групп.**

5. Сбор данных.

6. Анализ.

7. Интерпретация результатов (возможен возврат до уровня предмета исследования).

Итог защиты – «признать выводы обоснованными»

Диссертация – описание процесса: актуальность проблемы → цель → задачи →информация →анализ →выводы

Информация – что, в каком объеме, как собираем + процедуры сбора

Гипотезы - задачи

1. Интерпретационная – что это?
2. Описательная – каков этот объект?
3. Систематизирующая – упорядоченность в описании, классификации, типологии, эмпирическом обобщении.
4. Объяснительная – почему?
5. Экстраполяционная – в какой степени это имеет значение для другого места, времени и объекта.
6. Методологическая – как это лучше изучать.

Цель, задачи

Обоснование

Выводы, практические рекомендации

Научная новизна

Виды клинических задач

1. Диагностика состояний. Верификация!!!!!!
2. Возникновение, течение болезни.
3. Этиология и патогенез. Возможности измерений.
4. Прогнозирование состояний. ЧТО??????
5. Оценка методов профилактики, лечения, реабилитации.

Массивы данных

Дизайн:
Тип исследования.
Конкретные группы: суть, размер, способ формирования.

Состав;
Структура;
Типы данных – правила фиксации, способы кодирования.

Требования к информации

1. **К структуре массива (**зависимые и независимые переменные**);**
2. **По типам данных (**max количественных**);**
3. **К правилам кодирования;**
4. **К точности измерений;**
5. **По способам фиксации сведений;**
6. **Независимые и связные выборки;**
7. **Объем выборок, размеры групп, допустимость пропусков**

Принципы формирования массива

1. Минимальная достаточность;
2. Что обеспечит новизну?;
3. Единая по одним и тем же объектам исследования таблица;
4. Набор показателей «под задачи»;
5. Показатель → набор его значений.
6. 1 показатель – 1 столбик;
7. Строка – все сведения одного и того же объекта;

Что может статистика?

- **Статистическое описание, оценивание**
- **Сравнение групп, этапов, проверка гипотез**
- **Статистическое моделирование**

- ***Придать исследованию, анализу наукообразность***

Что статистика не может?

- **Улучшить выборку**
- **Оценить неизвестные признаки**
- **Исправить ошибки в измерениях**
- **Дать интерпретацию результатов**

Этапы статистического анализа

I. **Постановка задачи**
II. **Подготовка данных к анализу**
III. **Проверка данных**
IV. **Обоснованный выбор методов статистического анализа**
V. **Анализ.**
VI. **Интерпретация результатов**
VII. **Представление результатов**

I. Постановка задачи

- **Garbage in, garbage out**
- **Никакая статистическая обработка данных не может устранить неизвестную систематическую ошибку**
- **Проверка гипотез (первичный анализ данных) или выдвижение гипотез (вторичный анализ - post hoc analysis - data dredging)**

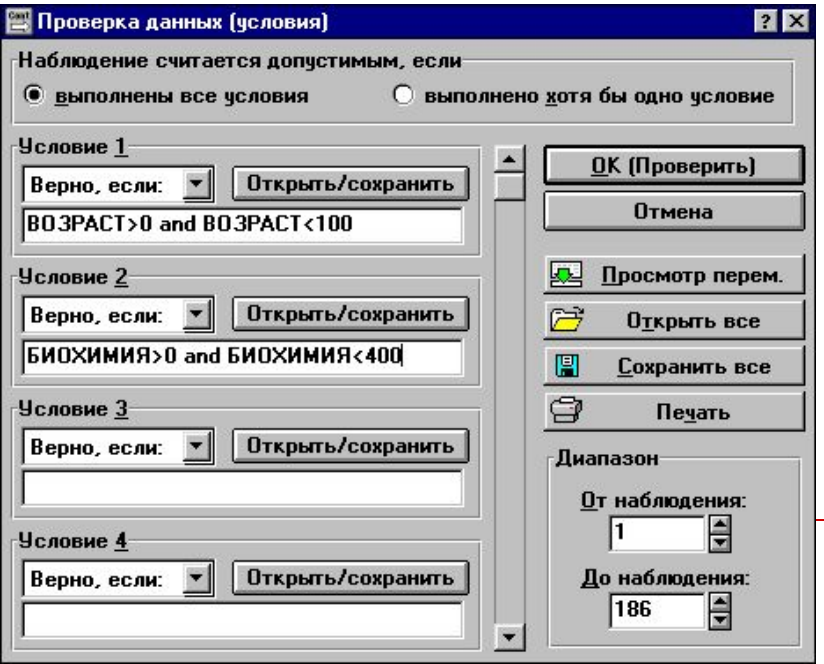
II. Подготовка данных

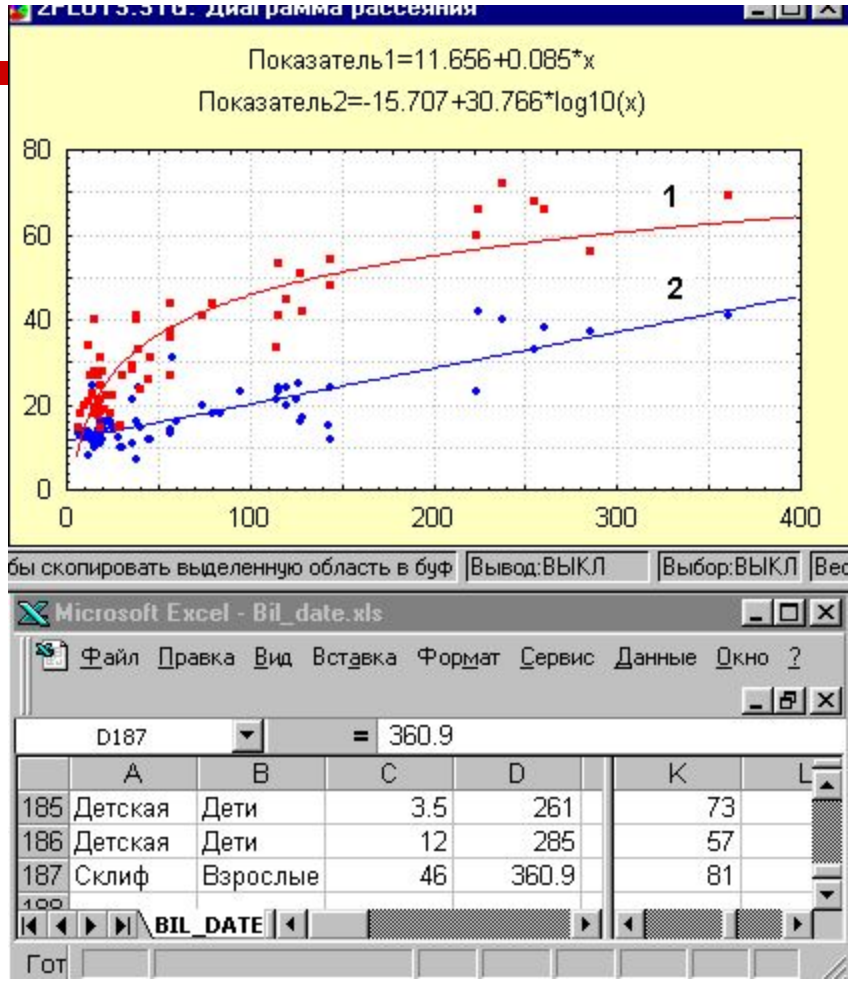
- **Разбиение области значений на интервалы, округление и точность**
- **Предварительные расчеты**
- **Использование стандартных шкал для клинических признаков**
- **Пропущенные значения**
- **Выбор объекта наблюдений**
- **Контрольные группы**
- **Интервал нормы**

Подготовка данных

Непосредственный ввод

Импорт из баз данных, текстовых файлов или электронных таблиц.





Верификация данных

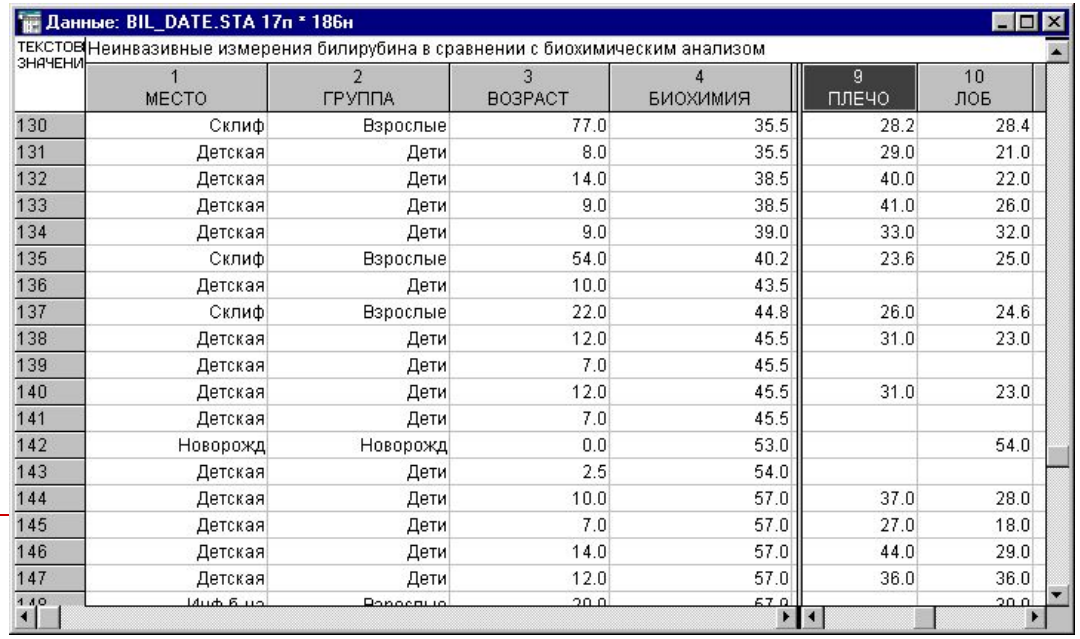
III. Проверка данных

- **Ошибки набора**
- **Артефакты**
- **Выпадающие значения**

Типы информации

- Массовые исследования (десятки тысяч наблюдений и сотни показателей).
- Результаты отдельных исследований (наблюдения за группами объектов).

Количественные и качественные признаки. Группирующие переменные.



Данные: BIL_DATE.STA 17п * 186н

ТЕКСТОВЫЕ ЗНАЧЕНИЯ | Неинвазивные измерения билирубина в сравнении с биохимическим анализом

	1 МЕСТО	2 ГРУППА	3 ВОЗРАСТ	4 БИОХИМИЯ	9 ПЛЕЧО	10 ЛОБ
130	Склиф	Взрослые	77.0	35.5	28.2	28.4
131	Детская	Дети	8.0	35.5	29.0	21.0
132	Детская	Дети	14.0	38.5	40.0	22.0
133	Детская	Дети	9.0	38.5	41.0	26.0
134	Детская	Дети	9.0	39.0	33.0	32.0
135	Склиф	Взрослые	54.0	40.2	23.6	25.0
136	Детская	Дети	10.0	43.5		
137	Склиф	Взрослые	22.0	44.8	26.0	24.6
138	Детская	Дети	12.0	45.5	31.0	23.0
139	Детская	Дети	7.0	45.5		
140	Детская	Дети	12.0	45.5	31.0	23.0
141	Детская	Дети	7.0	45.5		
142	Новорожд	Новорожд	0.0	53.0		54.0
143	Детская	Дети	2.5	54.0		
144	Детская	Дети	10.0	57.0	37.0	28.0
145	Детская	Дети	7.0	57.0	27.0	18.0
146	Детская	Дети	14.0	57.0	44.0	29.0
147	Детская	Дети	12.0	57.0	36.0	36.0

IV. Обоснованный выбор методов статистического анализа

- **Типы данных**
- **Вид распределения**
- **Одно- и двусторонние тесты**
- **Связанные и несвязанные выборки**
- **Проблема множественных сравнений (алгоритмы, выбор уровня P)**
- **Хи-квадрат или ТКФ**
- **Корреляция или регрессия**

VI. Интерпретация результатов

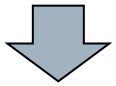
- **Отсутствие достоверных результатов не является подтверждением нулевой гипотезы**
- **Корреляционная связь – не причинно-следственная**
- **Валидизация многомерных моделей**
- **Data dredging (post hoc analysis)**
- **Соотношение статистической и клинической, эпидемиологической и другой предметной значимости**
- **Очень большие и очень маленькие выборки**
- **Суррогатные исходы и конечные точки**

VII. Представление результатов

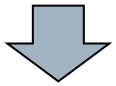
- **«Единые требования к статьям, представляемым в международные биомедицинские журналы» (Межд. журнал мед. практики, 1997, N 5, с. 53-64)**
- **Число наблюдений для каждого признака**
- **Описательная статистика -**
- **$M \pm SD$, Me (LQ;UQ), % (n/N)**
- **Точность результатов (оценки, P)**
- **ДИ (для основных результатов исследования) и P**
- **Указание на использованные стат. методы**
- **Указание на использованный стат. пакет**

V. Основные этапы анализа данных

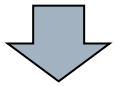
Подготовка данных: заполнение таблиц, импорт, проверка и сортировка.



Разведочный анализ: сопоставимость групп!!!, описательные статистики, графические методы.



Сравнение групп, **оценка динамики**: параметрические и непараметрические методы.



Выявление связей: корреляционный, факторный анализ.

Анализ зависимостей. Построение линейных и нелинейных моделей.

Разведочный анализ

Сопоставимость групп: по полу, возрасту, особенностям патологии. Определяется дизайном работы

Определение характера распределений переменных, визуальный анализ зависимостей и идентификация возможных выбросов.

Нормальное

Можно применять стандартные методы: t-критерии и дисперсионный анализ.

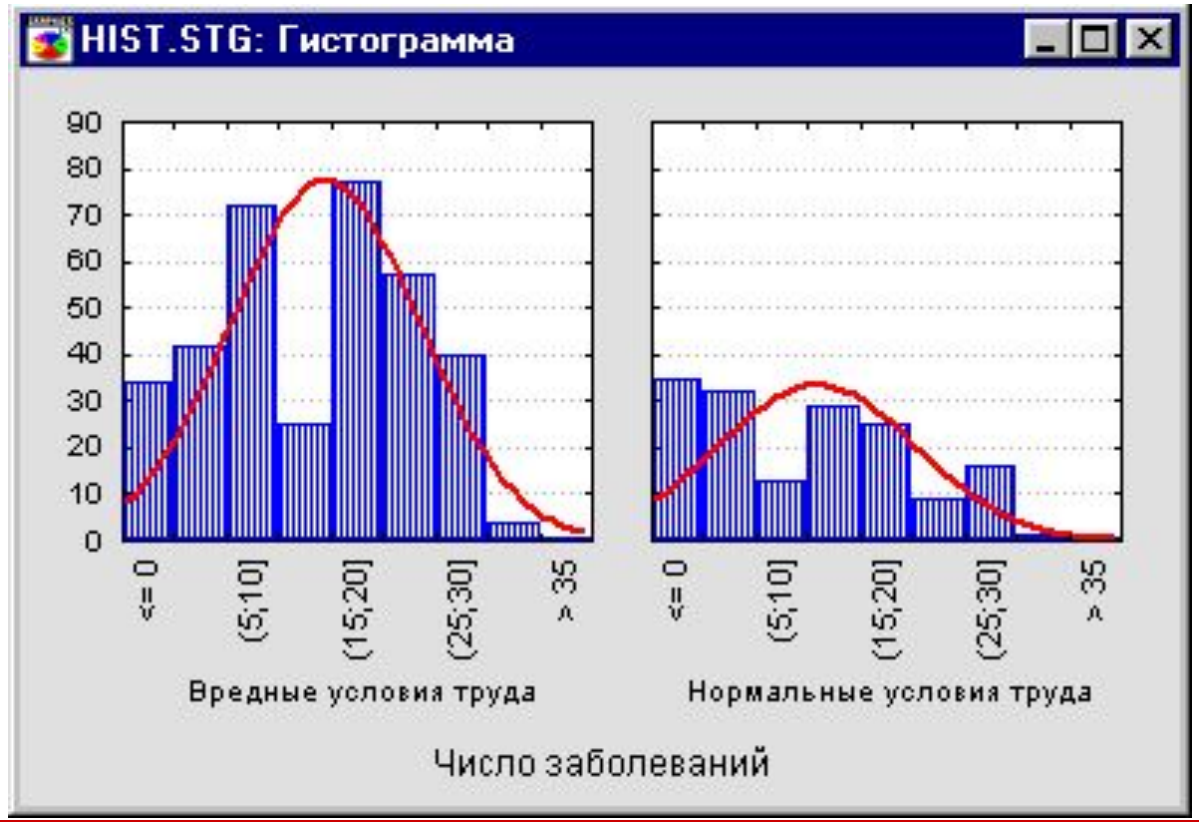
Отличное от нормального (или малая выборка)

Необходимо использовать непараметрические критерии.

Описание данных

- **Основные дескриптивные статистики.**
- **Дескриптивные статистики для группированных данных.**
- **Графики для дескриптивных статистик.**

Описание данных



Описание данных

Описание данных

Возраст Stem-and-Leaf Plot (диаграмма ветвей и листьев)

Frequency	Stem &	Leaf
6,00	3 .	677999
7,00	4 .	0223333
14,00	4 .	66677788888999
23,00	5 .	01111111122223333333444
20,00	5 .	55667777778888888899
27,00	6 .	000011111222333333333444444
27,00	6 .	55555566666667888888999999
24,00	7 .	000000011111122233333444
13,00	7 .	5566666788899
11,00	8 .	00001111224
2,00	8 .	67
Stem width :	10	
Each leaf:	1 case(s)	

Описание данных

Моделирование

- **Корреляционный, регрессионный, факторный анализ.**
- **Классификационные деревья, нейронные сети.**
- **Временные ряды, анализ выживаемости.**

Для графических объектов

1. **Таблицы с цифрами намного хуже диаграмм, графиков, схем.**
2. **Секторные круговые диаграммы – сопоставление частей и целого.**
3. **Столбиковые – сравнение групп.**
4. **Графики линейные – отображение динамики, но не более 5-ти линий на одном поле.**
5. **Необходимо гораздо чаще демонстрировать корреляционные поля и box&whisker plot – наглядность, полнота.**

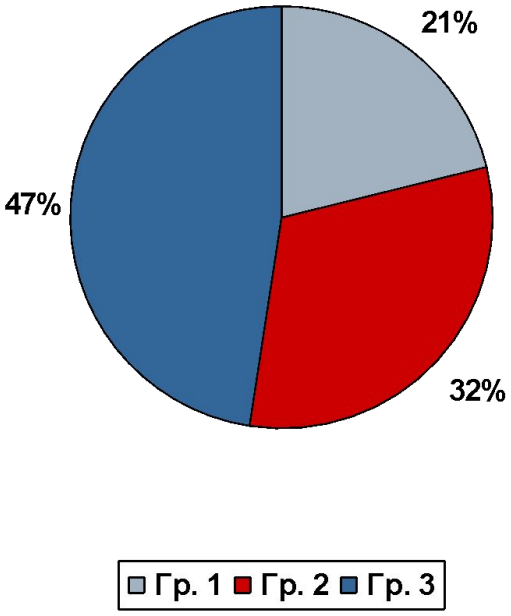
1. Таблицы с цифрами

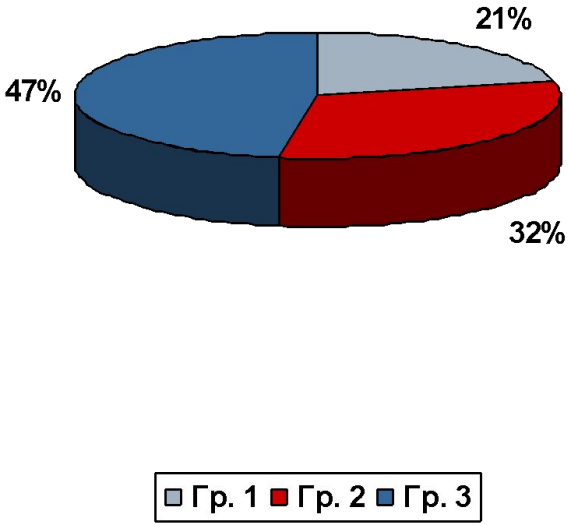
№				
1				
2				
3				

1. Таблицы с цифрами

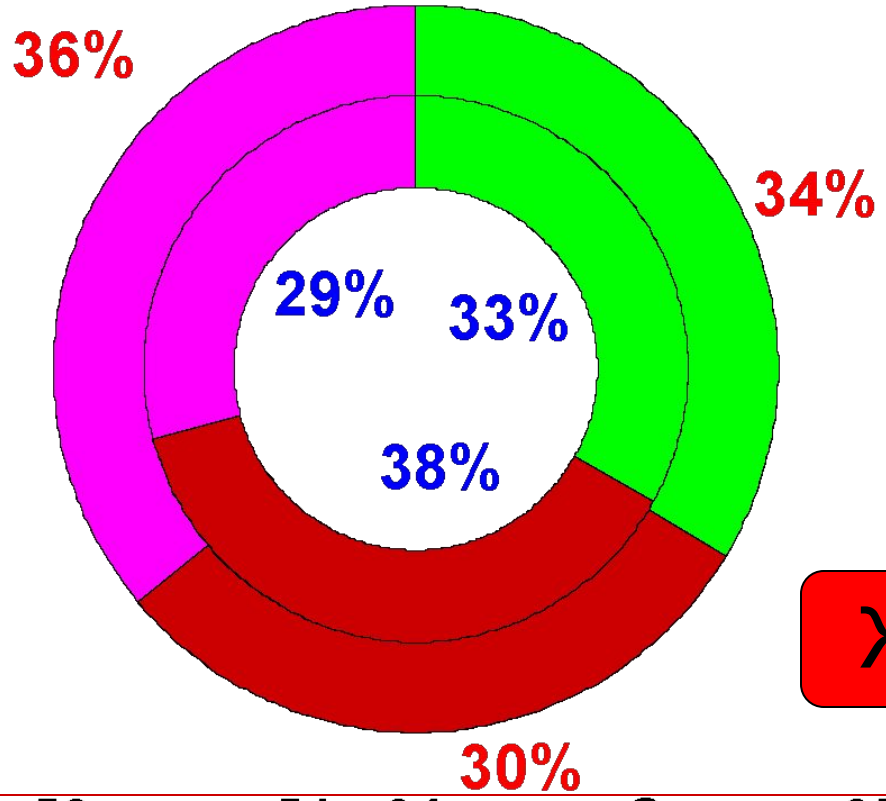
Группы	Метод лечения	**N-пациентов**
I	Фибробласты	11
II	Фибробласты через 3-е суток кератиноциты	17
III	Аналог кожи	38
IV	Многослойный пласт кератиноцитов	14
V	Группа сравнения	30
	Всего	110

Секторные круговые диаграммы – сопоставление частей и целого.





Секторные круговые диаграммы – сопоставление частей и целого.

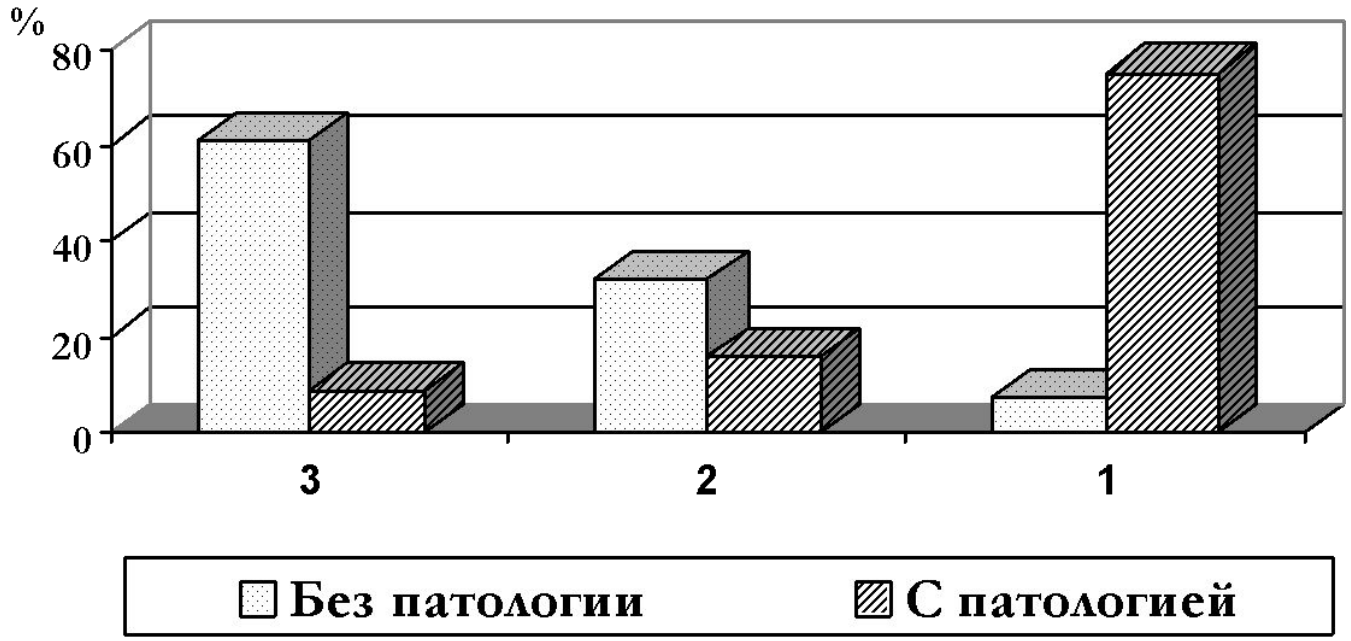


М – 45 чел

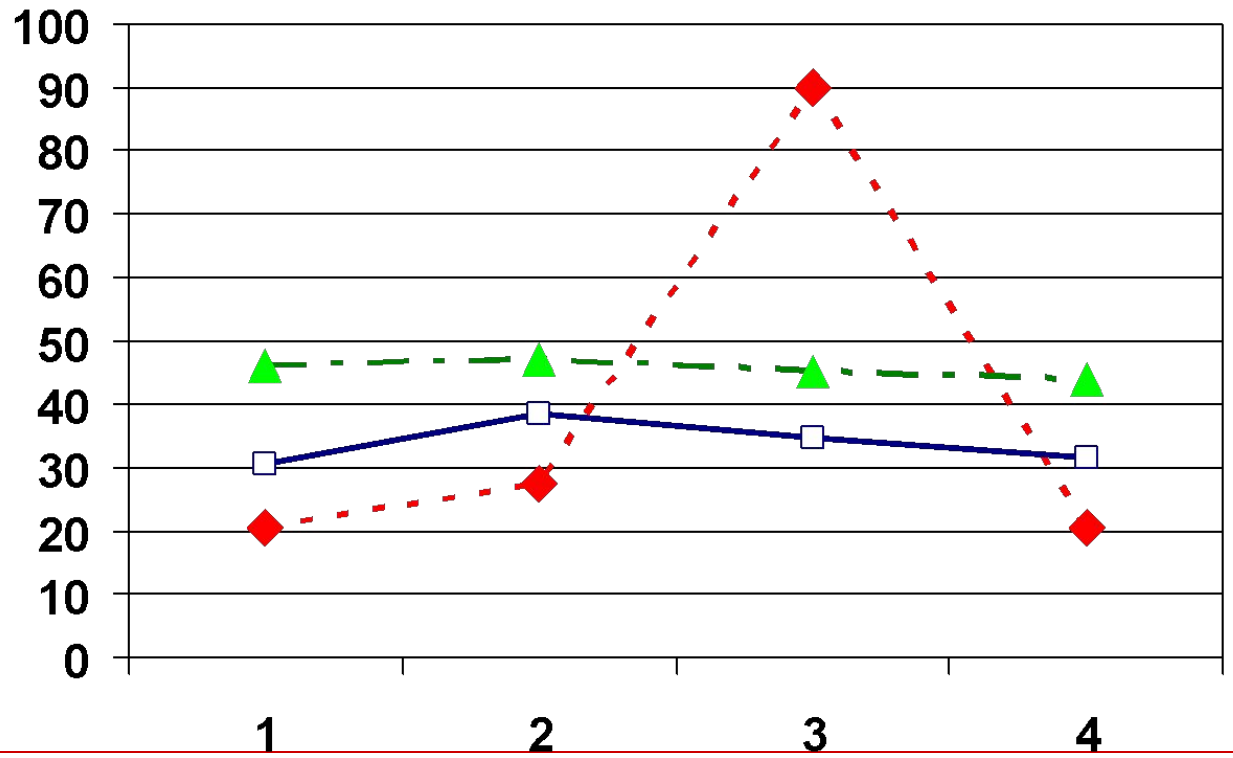
Ж – 65 чел

Секторные круговые диаграммы – сопоставление частей и целого.

Столбиковые – сравнение групп.



4. Графики линейные – отображение динамики,



box&whisker plot – наглядность, полнота.

box&whisker plot – наглядность, полнота.

box&whisker plot – наглядность, полнота.

box&whisker plot – наглядность, полнота.

box&whisker plot – наглядность, полнота.

Классификационное дерево

Кривые выживаемости

Статистические системы

I. BMDP, SAS

II. Statistica for Windows, SPSS, Stadia

III. Stata, Statgraphics, EPI, MEDcalc

Благодарю за внимание!