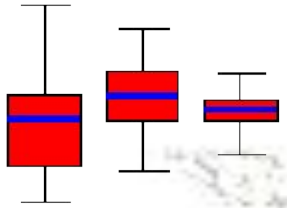
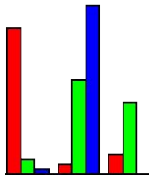
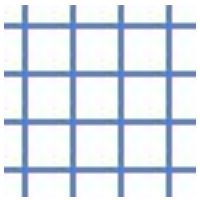
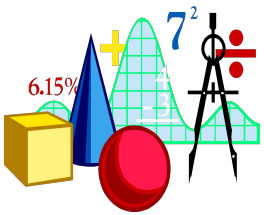




ГРАФИКИ И ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Стат. методы в психологии (Радчикова Н.П.)





Методы исследования

Наблюдение  Описательная статистика

Определение связей между переменными  Корреляционная техника

Эксперимент  Критерии различий

Методы исследования

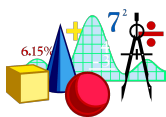
Наблюдение → **Описательная статистика**

Определение связей между переменными → Корреляционная техника

Эксперимент → Критерии различий

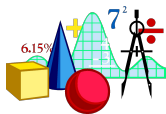
Описательная статистика

Методы и способы, используемые для «суммирования», организации и «уменьшения» большого количества наблюдений (статистических опытов).



Описательная статистика

- Частотные распределения и графики
- Меры центральной тенденции
- Меры изменчивости
- Меры положения
- Меры формы
- …



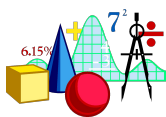
Группировка данных

Предположим, мы спрашивали студентов, насколько их провал на экзамене зависел от причин, которые они никак не могли контролировать.

Ответы даются по шкале от 1 до 7
(1 - совсем не зависел, 7 - полностью зависел)

Гипотетические данные опроса 25 студентов:

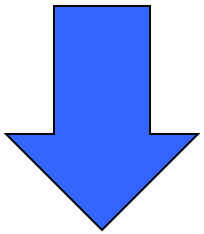
3,5,6,5,2,3,6,4,6,7,6,4,5,5,1,2,5,4,4,5,5,7,3,3,4



Группировка данных

Гипотетические данные опроса 25 студентов:

3,5,6,5,2,3,6,4,6,7,6,4,5,5,1,2,5,4,4,5,5,7,3,3,4



1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,7,7

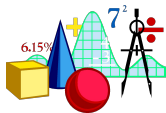
Группировка данных

1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,7,7

ответ	частота
1	1
2	2
3	4
4	5
5	7
6	4
7	2

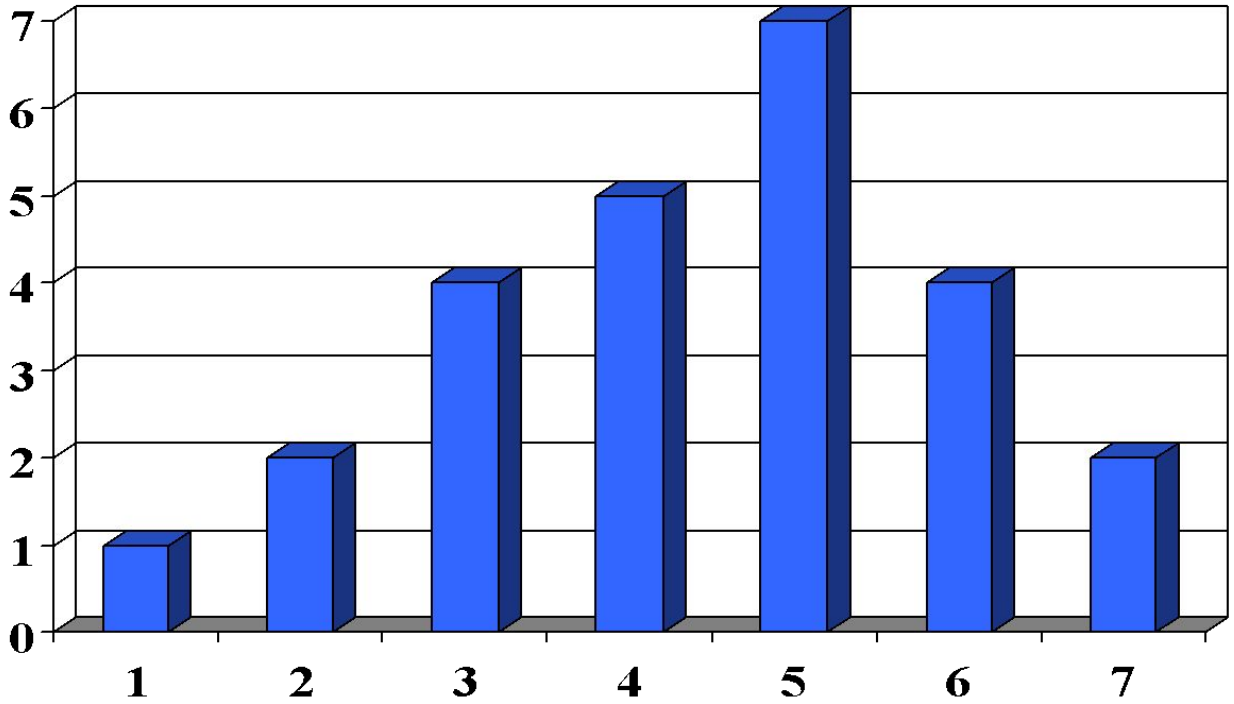
Группировка данных

ответ	частота	накопленная частота	%	накопленный процент
1	1	1	4	4
2	2	3	8	12
3	4	7	16	28
4	5	12	20	48
5	7	19	28	76
6	4	23	16	92
7	2	25	8	100



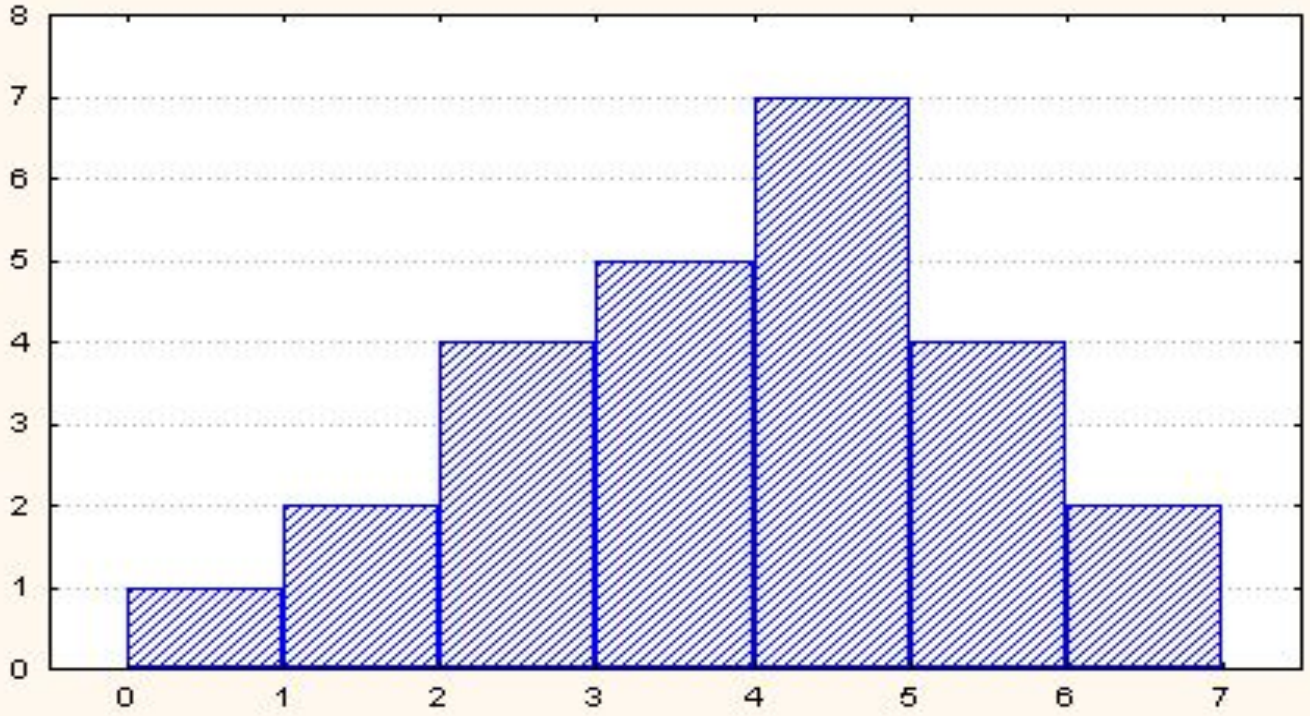
Группировка данных

Столбчатая диаграмма



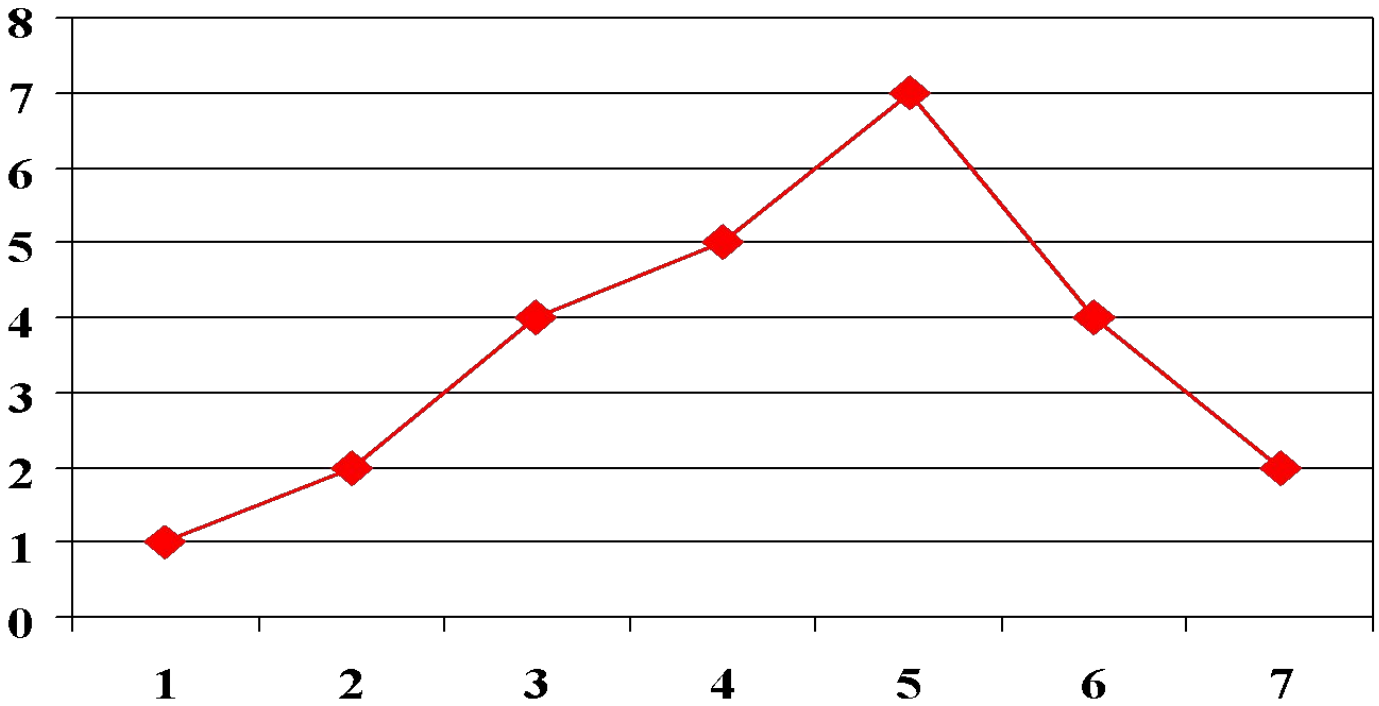
Группировка данных

Гистограмма



Группировка данных

ПОЛИГОН

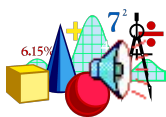


Группировка данных

Группировка данных

А если значений много?

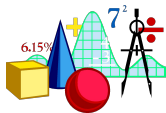
40, 48, 11, 16, 52, 64, 21, 33, 39, 69, 45, 8,35, 22, 57, 74, 13, 25, 47, 27, 38, 43, 15, 33, 66, 52, 47, 37, 0, 24, 43, 61, 35, 29, 52, 40, ….



Группировка данных

Частотная таблица получается большой:

балл	f	балл	f	балл	f
0	1	**8**	2	**15**	3
1	0	**9**	0	**16**	1
2	0	**10**	0	**17**	4
3	1	**11**	0	**18**	5
5	0	**12**	1	**19**	2
6	1	**13**	2	**…**	
7	1	**14**	0	**74**	1



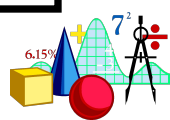
Группировка данных

Тогда стоит сгруппировать значения переменной в интервалы

4. Следующий интервал начинается с числа, которое следует за наибольшим значением предыдущего интервала

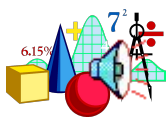
7+i-1=7+7-1=13

Второй интервал будет от 7 до 13

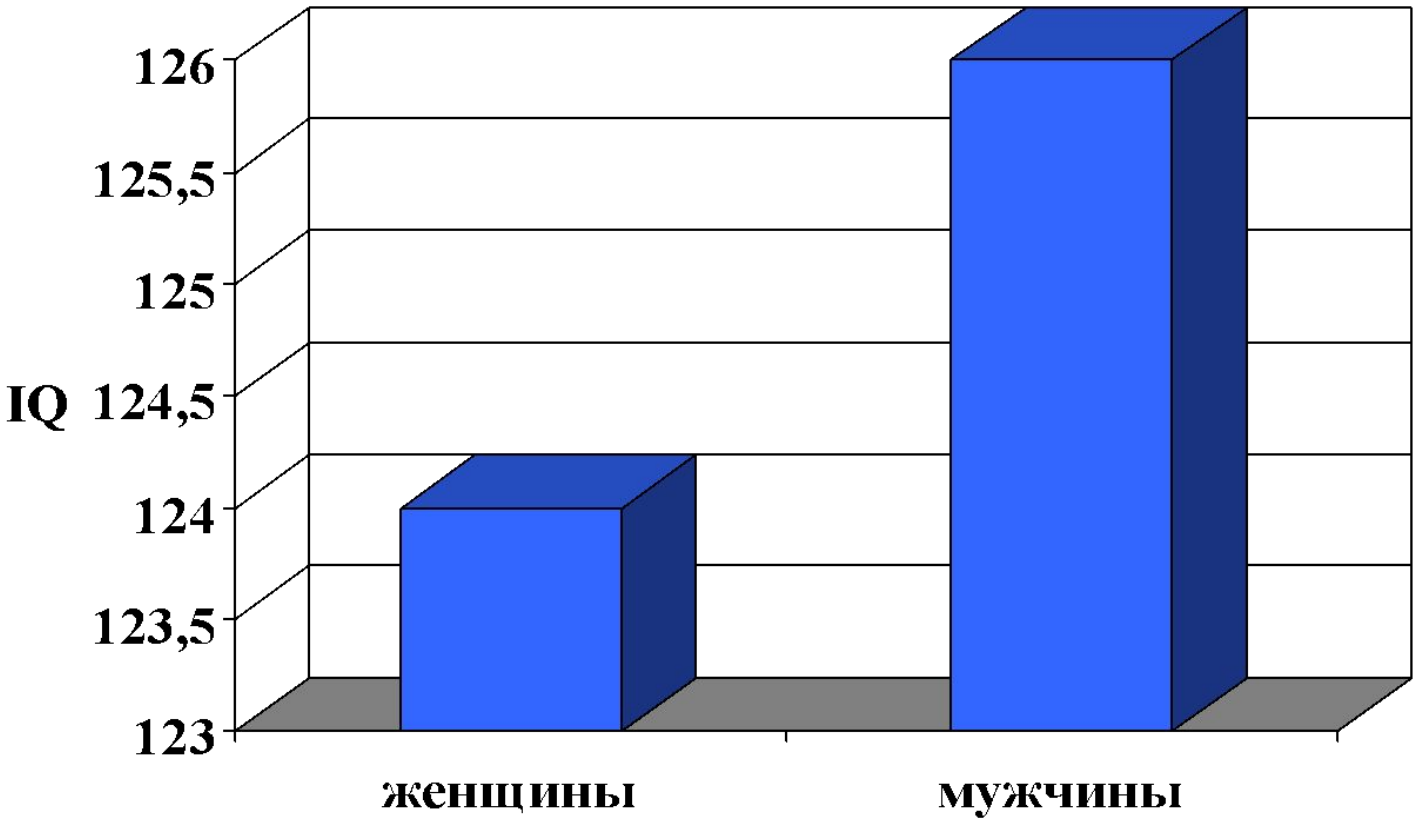


Группировка данных

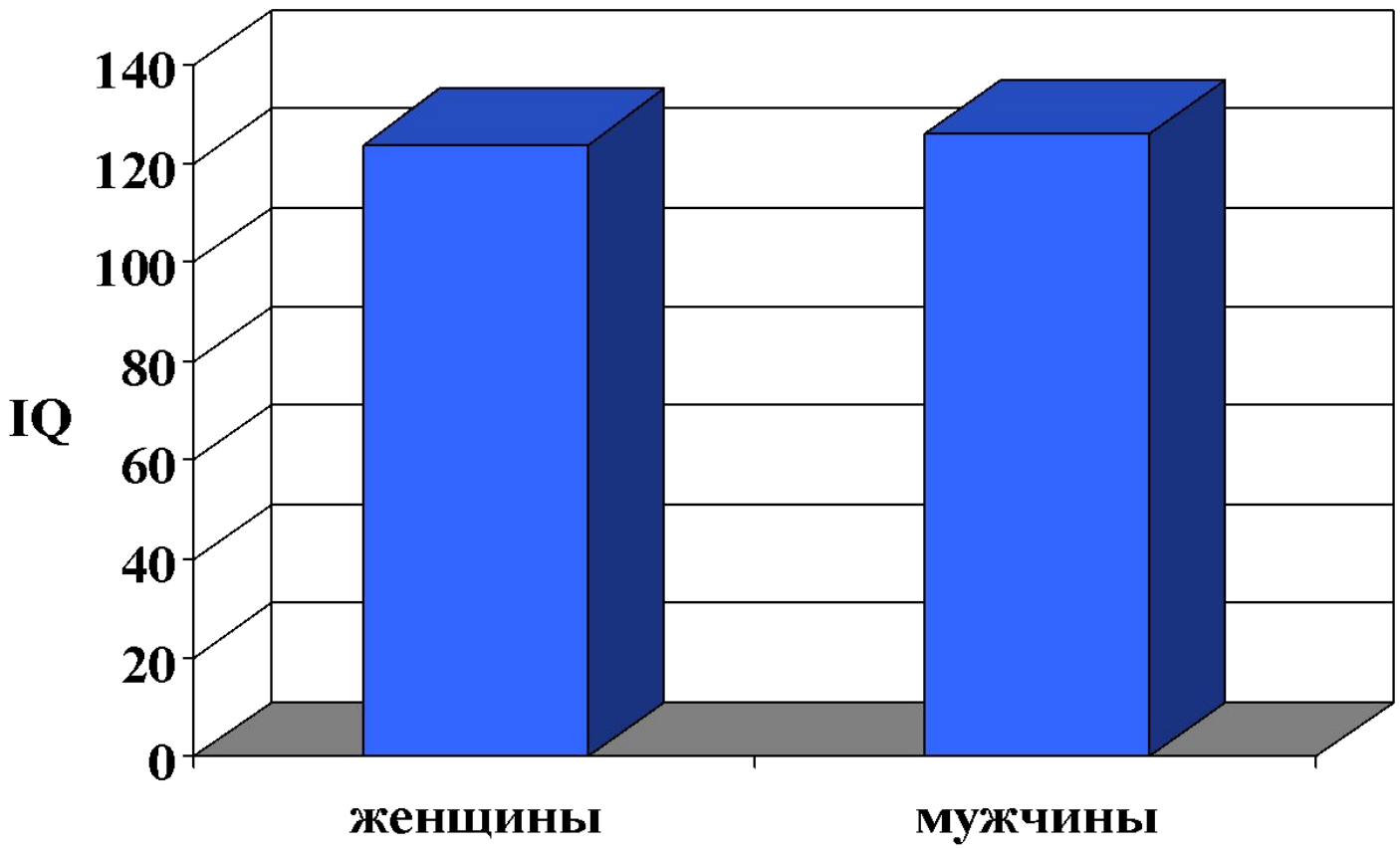
возраст	f	возраст	f
0-6	2	**50-56**	14
7-13	4	**57-63**	4
14-20	5	**64-70**	5
21-27	7	**71-77**	3
28-35	10		
36-42	13		
43-49	17		



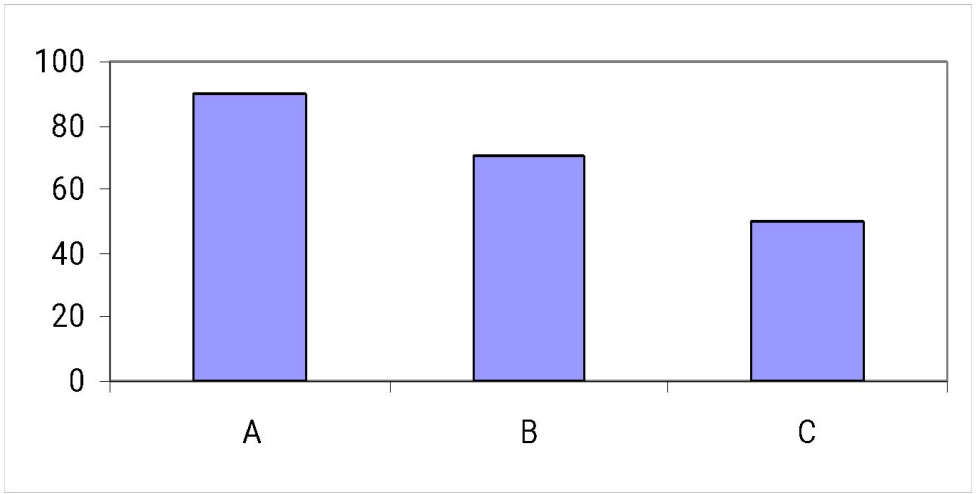
Использование графиков

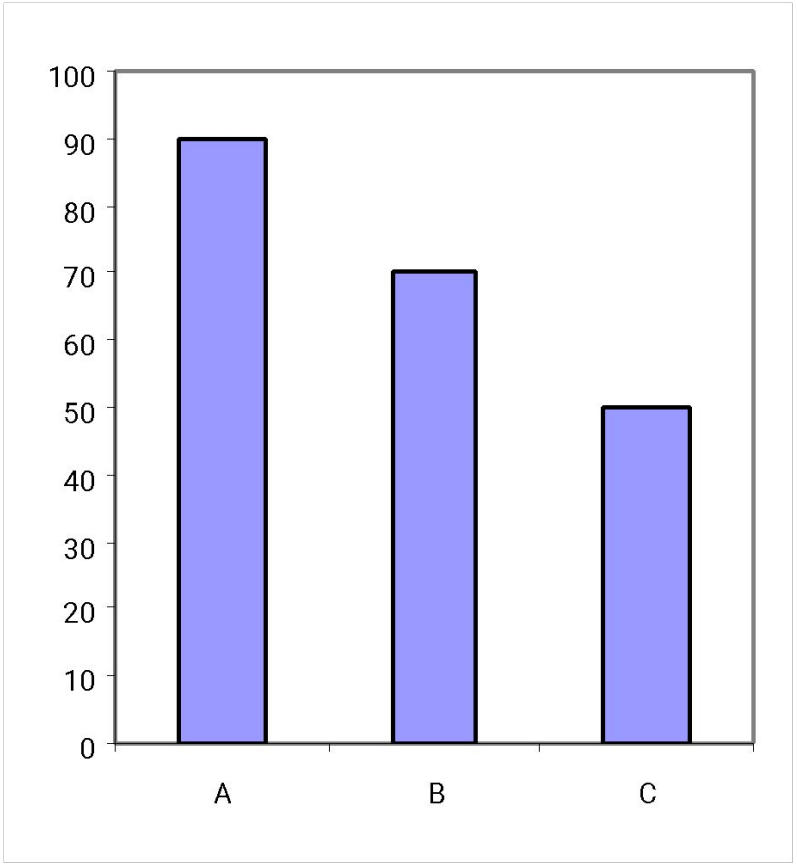


Использование графиков



Использование графиков

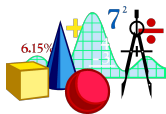




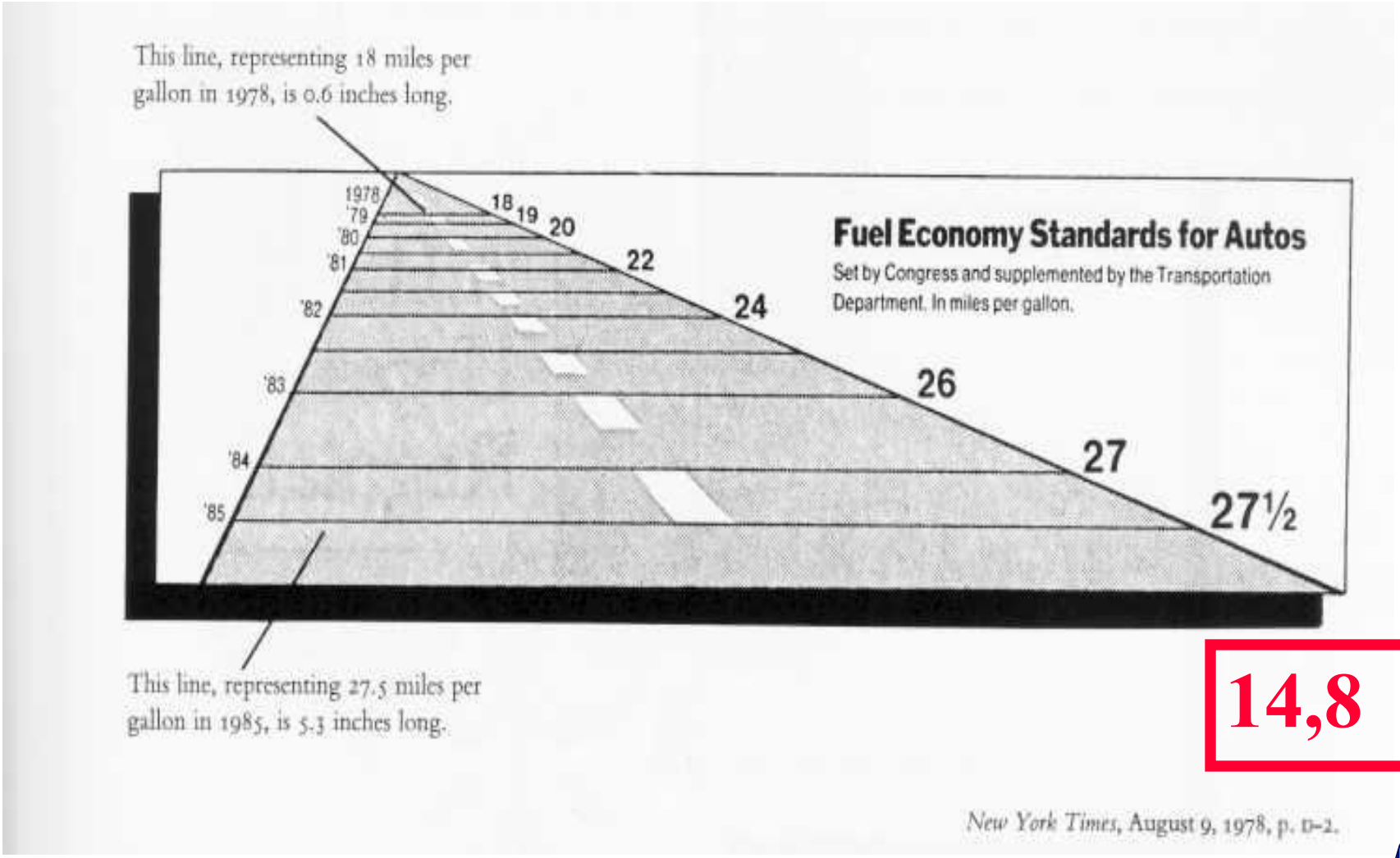
Использование графиков

Lie factor – отношение разницы в размере элементов графика к разнице величин, которые они представляют

Наиболее информативные («честные») графики имеют Lie factor =1



Использование графиков

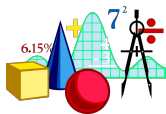


New York Times, August 9, 1978, p. D-2.

14,8

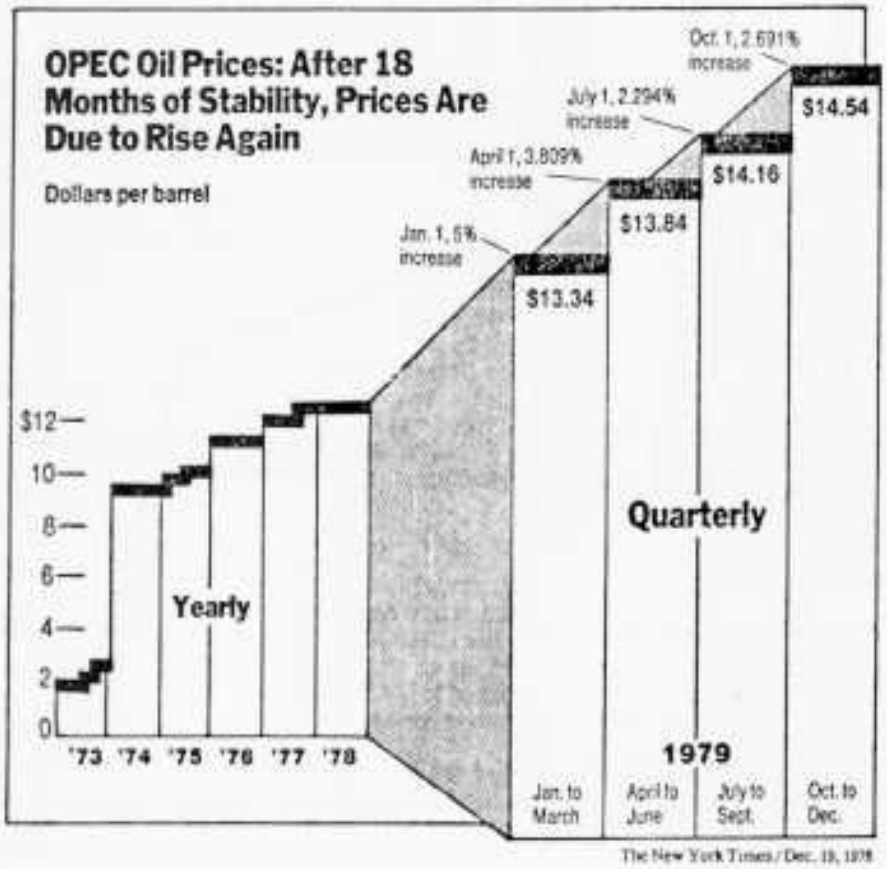
Использование графиков

Следует избегать соединения изменений в оформлении графика с изменениями в данных



Использование графиков

Design variation corrupts this display:

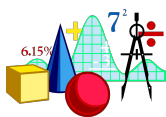


New York Times, December 19, 1978, p. D-7.

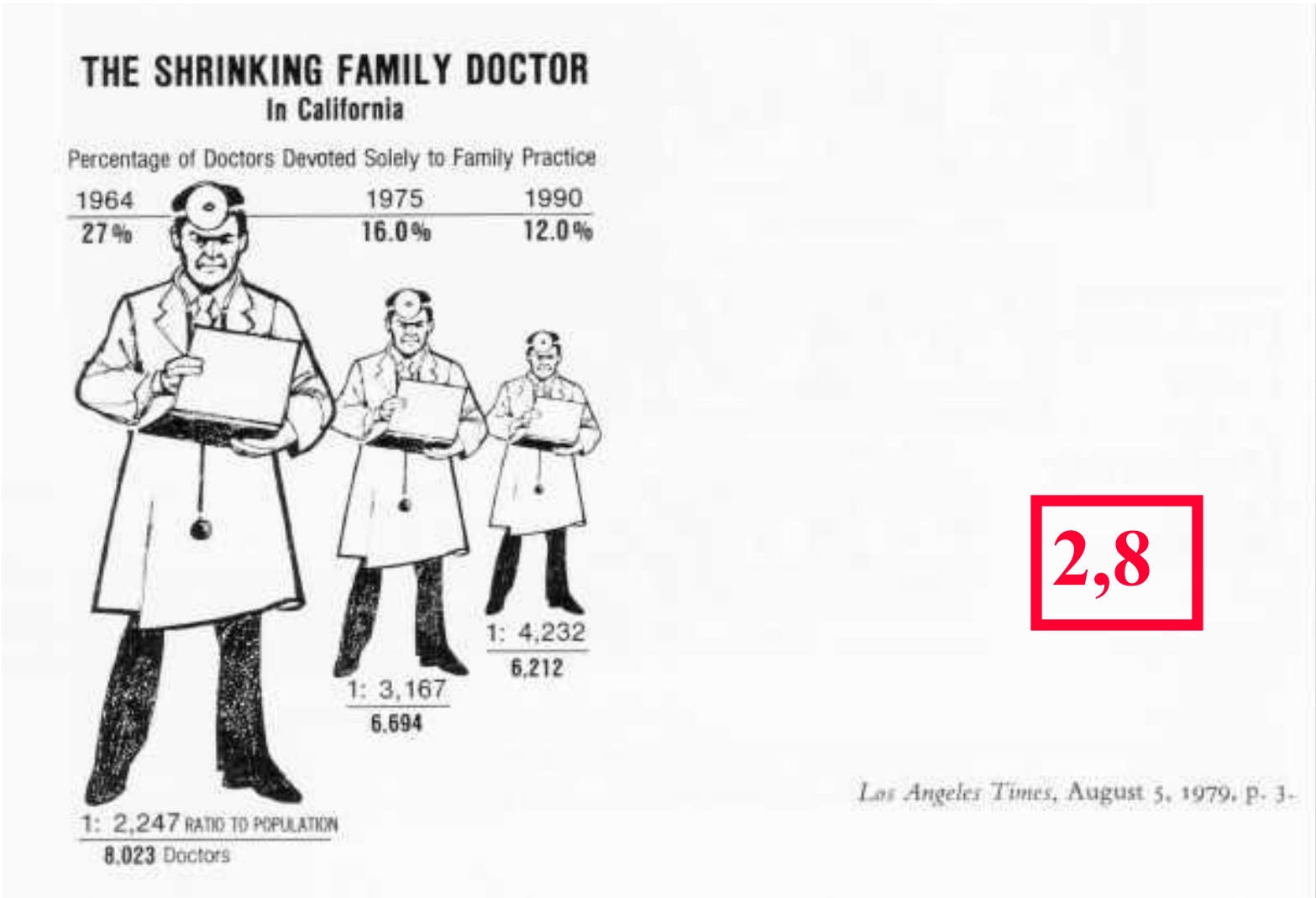
Использование графиков

Еще одна проблема – многомерные изменения, т.е. изменения сразу по нескольким размерностям, например, по высоте и ширине.

Если масштабирование ведется сразу по двум измерениям, площадь изменяется пропорционально квадрату изменений!



Использование графиков

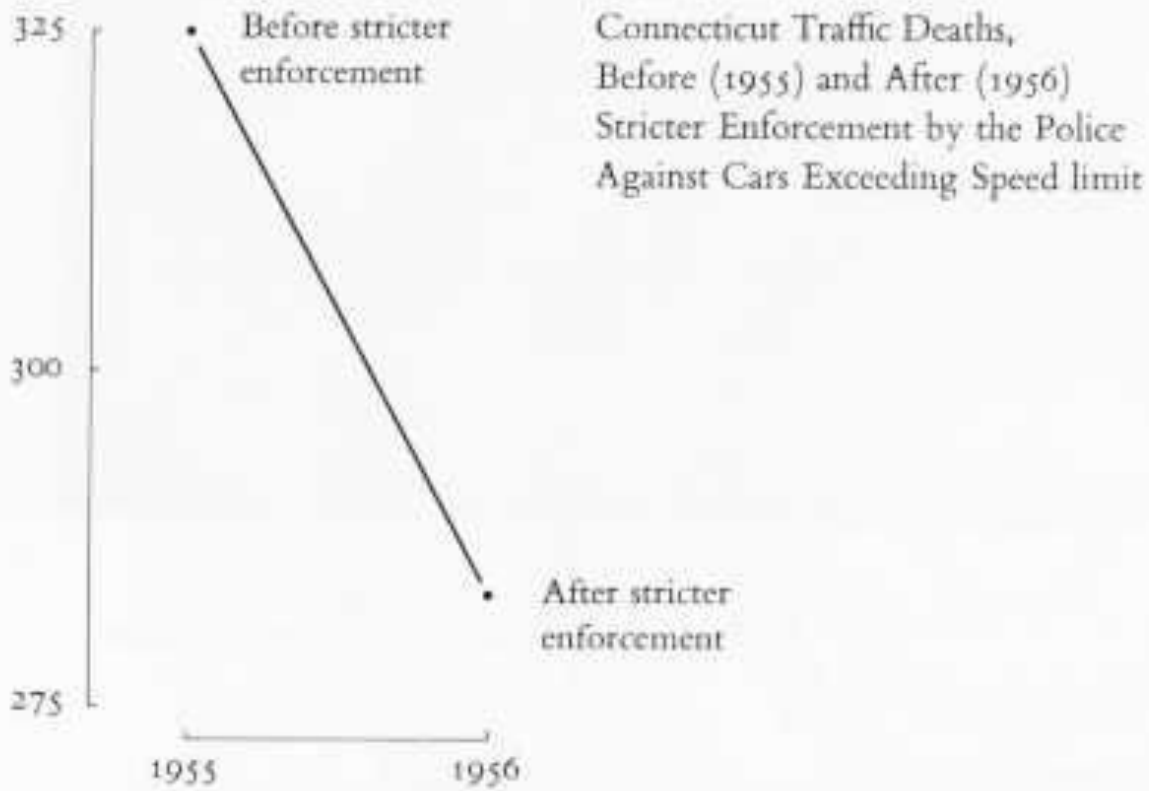


2,8

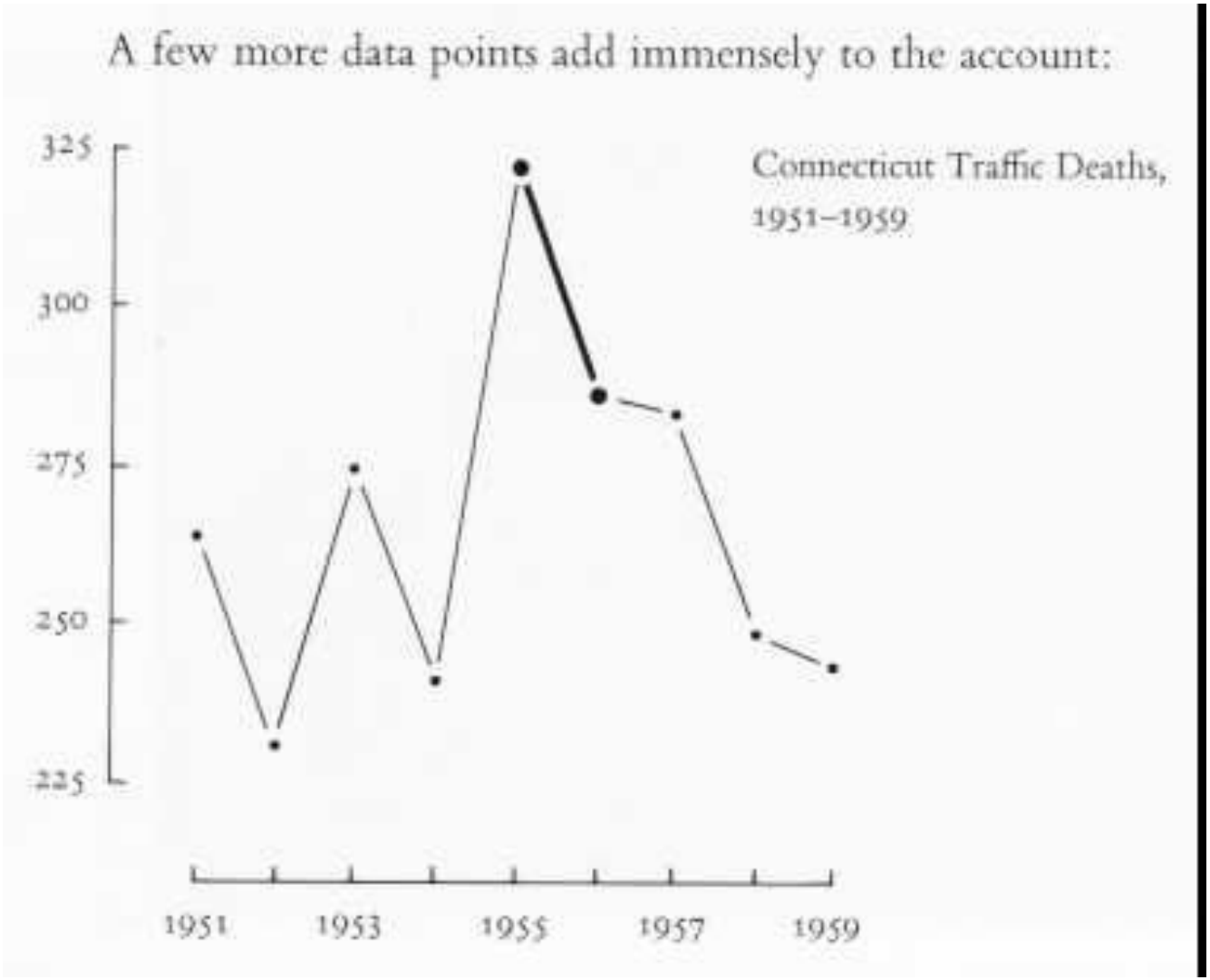
Использование графиков

Graphics must not quote data out of context.

Nearly all the important questions are left unanswered by this display:



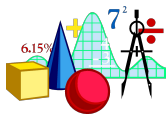
Использование графиков



Основные понятия

Выборочной совокупностью или просто ***выборкой*** называют совокупность случайно отобранных объектов.

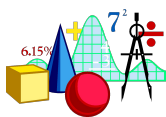
Генеральной совокупностью называют совокупность объектов, из которых производится выборка.



Основные понятия

Параметры – это меры описания, полученные при сплошном описании (описании генеральной совокупности).

Статистики (или оценки параметров) – это те же меры, но полученные при выборочном наблюдении (т.е. параметры описывают генеральную совокупность, а статистики – ее выборку).



Генеральная и выборочная совокупности

Генеральная совокупность → **Параметр**

Выборка → **Статистика**

Выборки

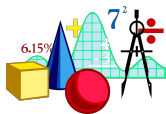
Выборки бывают разные!

Классификация Л.Мюллера и К. Шусслера

По критерию методов отбора выборки бывают

1) Не случайные

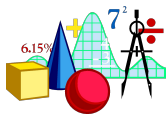
2) Случайные (вероятностные, пробабилистские)



Выборки

Классификация Л.Мюллера и К. Шусслера

1) **Не случайные** – не имеют теоретико-вероятностного обоснования и, следовательно, не соответствуют критерию репрезентативности, т.е. статистики не могут выступать оценками генеральной совокупности



Выборки

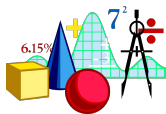
Классификация Л.Мюллера и К. Шусслера

1) Не случайные

1.1) Бессистемная выборка

1.2) Доступная выборка

1.3) Целенаправленная выборка



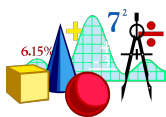
Выборки

Классификация Л.Мюллера и К. Шусслера

1.1) Бессистемная выборка

Отбор любых случайно встретившихся прохожих, согласившихся принять участие в исследовании.

Может использоваться только для самого первого ознакомления с проблемной ситуацией



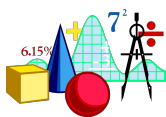
Выборки

Классификация Л.Мюллера и К. Шусслера

1.2) Доступная выборка

Формируется из числа лиц, которые по субъективным и объективным факторам могут быть включены в число респондентов, т.е. доступны физически.

Используется для накопления данных о латентных или аномальных явлениях



Выборки

Классификация Л.Мюллера и К. Шусслера

1.3) Целенаправленная выборка

Преднамеренный отбор определенной категории респондентов, которые по оценке исследователя в наибольшей степени информированы по проблеме или заинтересованы в ее изучении

Используется в экспертных опросах, лабораторных исследованиях и социальных экспериментах

Выборки

Классификация Л.Мюллера и К. Шусслера

2) Случайные

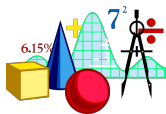
2.1) Простая случайная

2.2) Серийная

2.3) Систематическая (интервальная)

2.4) Стратифицированная

2.5) Комбинированная

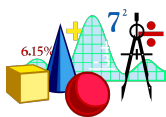


Выборки

Классификация Л.Мюллера и К. Шусслера

2.1) Простая случайная – формируется путем случайного отбора единиц наблюдения из однородной генеральной совокупности (жребий, таблицы случайных чисел, компьютерное моделирование)

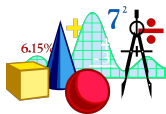
.



Выборки

Классификация Л.Мюллера и К. Шусслера

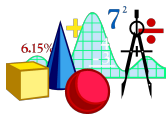
2.2) Серийная – единицами отбора являются статистические серии (таксоны, гнезда) – территориальные общности, коллективы, семьи и т.д. Серии выбираются по методике простой случайной выборки



Выборки

Классификация Л.Мюллера и К. Шусслера

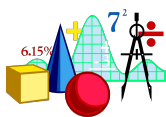
2.3) Систематическая (интервальная) – отбор единиц производится через один и тот же интервал, при этом начало отсчета определяется случайным образом



Выборки

Классификация Л.Мюллера и К. Шусслера

2.4) Стратифицированная выборка на основе предварительного выделения в генеральной совокупности однородных частей, типических групп (страт). В каждой страте производится случайный отбор единиц наблюдения, как правило, пропорционально их доле в генеральной совокупности.

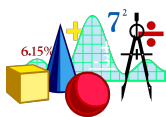


Выборки

Классификация Л.Мюллера и К. Шусслера

2.5) Комбинированная – выборка, в которой используются различные способы отбора.

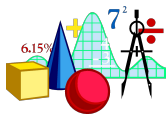
Например: Гнездовая выборка – по два предприятия из типичных групп (сильных, средних и слабых). Далее отбор респондентов осуществляется интервальным методом.



И это все?

Меры центральной тенденции

- Среднее арифметическое (M или $\overline{x}$)
- Медиана M_e или срединное значение
- Мода M_d (наиболее вероятное значение)



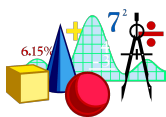
Меры центральной тенденции

1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,7,7

Среднее арифметическое

$$M=(x_1+\ldots+x_N)/N$$

$$M=(1+2+2+3+3+\ldots.+6+7+7)/25=4,4$$



Меры центральной тенденции

1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,7,7

$M_e = 5$

Меры центральной тенденции

1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,7

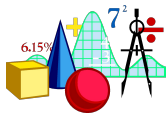
$$M_e=(4+5)/2=4,5$$

Меры центральной тенденции

1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,7,7

Мода

$M_d = 5$

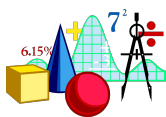


Доверительный интервал

Доверительный интервал

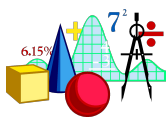
(***95% confidence limits of mean***)

для среднего представляет интервал значений вокруг оценки, где с данным уровнем доверия находится «истинное» (неизвестное) среднее генеральной совокупности.



Доверительный интервал

Если среднее выборки равно 23, а нижняя и верхняя границы доверительного интервала с уровнем p=.95 равны 19 и 27 соответственно, то можно заключить, что с вероятностью 95% интервал с границами 19 и 27 накрывает среднее генеральной совокупности.



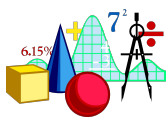
Подумай,

Примени

Найдите среднее, моду и медиану для следующих данных

10, 8, 6, 0, 8, 3, 2, 5, 8, 0

среднее=5,
медиана=5,5,
мода=8

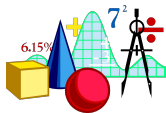


Подумай,

Примени

Среди мужчин, приговоренных к пожизненному заключению, только 10 % подвергаются повторному наказанию.

Среди тех, кого осудили на срок до 6 месяцев, повторно судимых (и опять приговоренных) 60 %. Следовательно, более длительное тюремное заключение более эффективно



Подумай,

Примени

Смертность американских солдат во время войны в Персидском заливе была 9 человек на 1000. В это же время смертность гражданских лиц, например в Нью-Йорке была 16 человек на 1000. Следовательно, во время войны действующая армия – самое безопасное место.

Подумай,

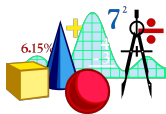
Примени

Это поистине очевидно. Среди 57 млн. жителей Великобритании около 5 000 имеют одну ногу. Следовательно, среднее количество ног будет ((5000*1)+(56995000*2))/ 57000000= 1.999123

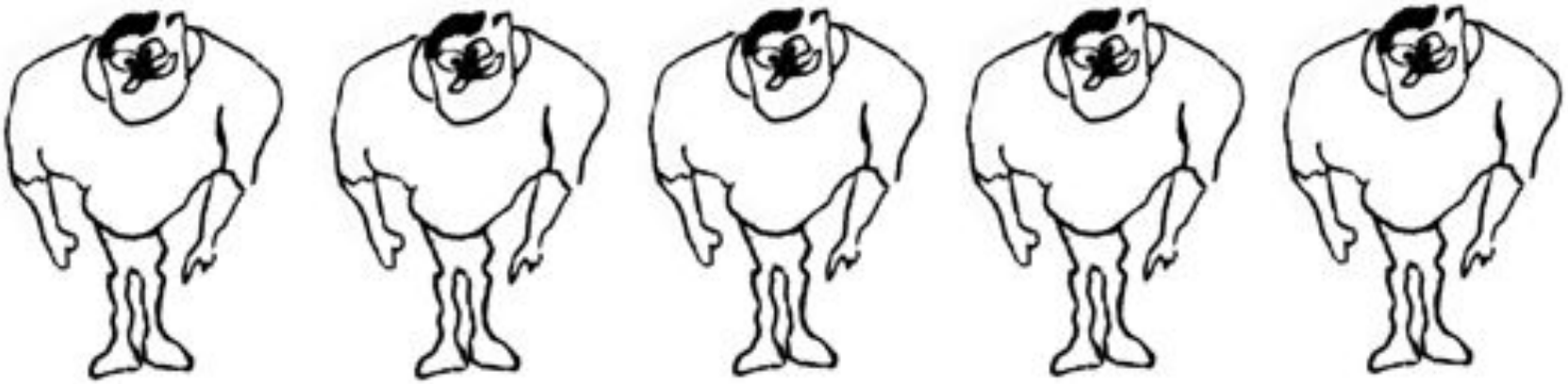
Так как большинство имеют две ноги…

Меры изменчивости

- Размах
- Дисперсия
- Стандартное (среднеквадратичное) отклонение
- Стандартная ошибка

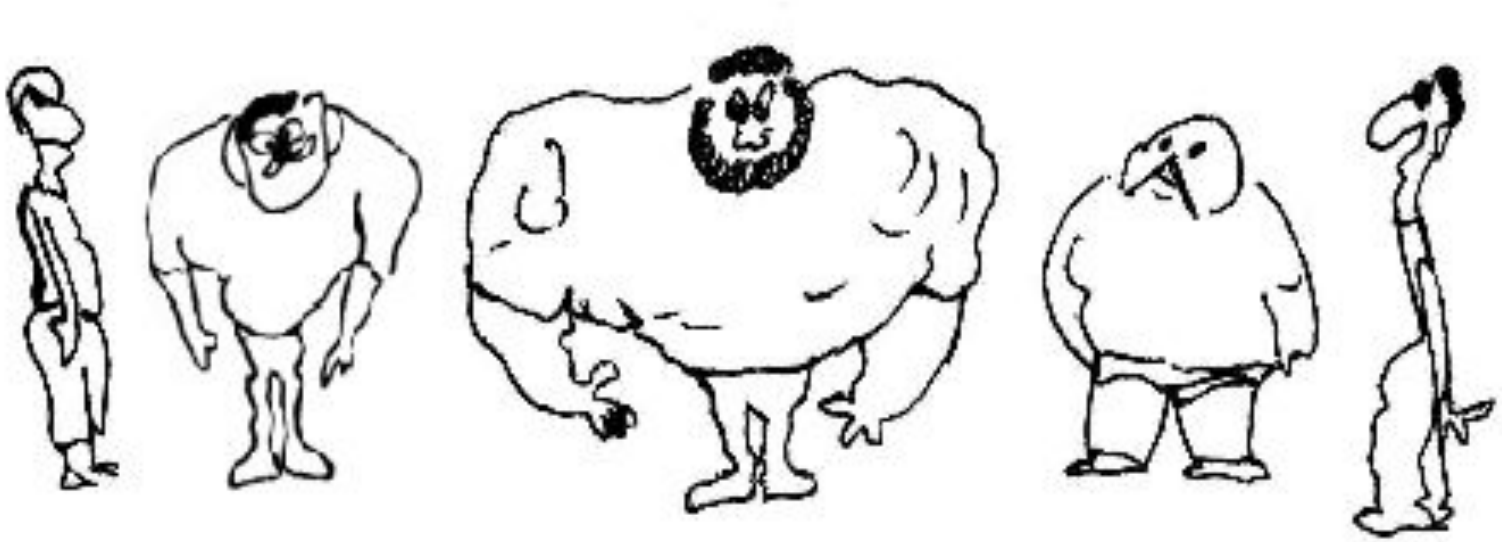


Меры изменчивости



Средний вес команды = 95 кг

Меры изменчивости



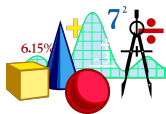
Средний вес команды тоже = 95 кг

Меры изменчивости

- Размах $R = X_{max} - X_{min}$

1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,7

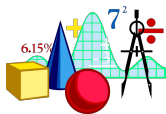
$$R = X_{max} – X_{min} = 7-1=6$$



Меры изменчивости

- **Дисперсия**

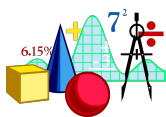
$$S^2 = \frac{\sum \left(X_i - \overline{X} \right)^2}{N - 1}$$



Меры изменчивости

Пример. **Вычислить дисперсию для следующей выборки:**
5, 6, 3, 8, 5, 9

Вычисляем среднее арифметическое: = (5+6+3+8+5+9)/6=6



Меры изменчивости

№	$x_i - x$	$(x_i - x)^2$
6	9-6=3	9
$\sum$		**24**

Подставляем в формулу:

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{N - 1} = 24/(6-1) = 4,8$$

Меры изменчивости

- **Другая формула для дисперсии:**

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}$$

Меры изменчивости

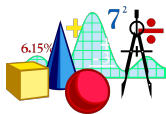
- **Стандартное отклонение**

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}}$$

Меры изменчивости

Стандартная ошибка среднего значения - это стандартное отклонение, деленное на квадратный корень из объема выборки.

$$SE(\overline{X}) = \frac{s}{\sqrt{N-1}}$$



Ме… …вости

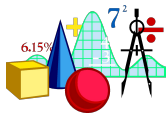
Гляньте-ка! СЕКС!
И прямо тут, в формуле!

…ачения -
…енное на
…а выборки.

$$SE(\overline{X}) = \frac{s}{\sqrt{N-1}}$$

Меры изменчивости

В диапазоне удвоенной стандартной ошибки по обе стороны от среднего значения с вероятностью примерно 95% находится среднее значение генеральной совокупности.



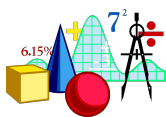
Подумай,

Примени

Найдите размах и дисперсию для следующих данных

10, 8, 6, 0, 8, 3, 2, 5, 8, 0

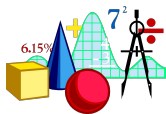
размах=10,
дисперсия=12,8889



Меры положения

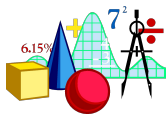
Квантили - структурные характеристики вариационного ряда, отсекающие в пределах ряда определенную часть его членов.

К ним относятся ***квартили, децили и перцентили (центили).***



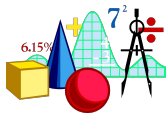
Меры положения

Квантиль – это точка на числовой оси, на которой откладываются результаты наблюдений. Эта точка делит всю совокупность наблюдений на части (группы) с определенными пропорциями между ними.



Процентили

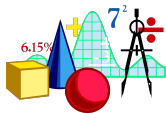
Перцентили (центили, процентили) отделяют от совокупности по 0,01 части (делят совокупность на 100 равных частей), их 99.



Процентили

В 1985 году примерно 24,7 миллионов людей в Соединенных Штатах были в возрасте 65 лет и старше

Таня набрала 41 балл по тесту по математике в этом году

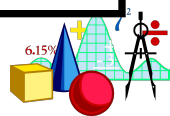


Процентили

В 1985 году примерно 24,7 миллионов людей в Соединенных Штатах были в возрасте 65 лет и старше

89% населения США находится в возрасте не старше 65 лет

89 – это и есть процентиль для 65-летних

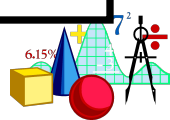


Процентили

Процентиль какого-либо значения, таким образом, представляет собой процент случаев, которые имеют то же самое или меньшее значение

Сказать «**возрасту 65 лет соответствует 89 процентиль**» - это сказать, что

«**89% населения США находится в возрасте 65 лет и меньше**»

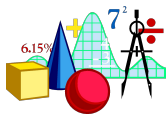


Процентили

Таня набрала 41 балл по тесту по математике в этом году, и это соответствует 62 процентилю.

62% белорусских абитуриентов сдали так же, как Таня или еще хуже,

и только 38% были лучше ее.



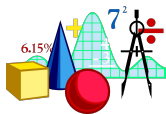
Процентили

Можно определить прямо по графику накопленных процентов

Процентили

Какой процентиль соответствует ответу 4?

Какой процент студентов считает, что результат провала на экзамене скорее зависел от них, чем от причин, которые они не могли контролировать?



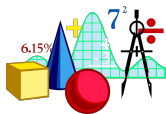
Процентили

Какой процентиль соответствует ответу 4?

Процентили

Можно определить по формуле

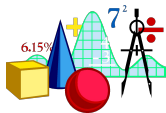
Процентиль=(накопленная частота/N)*100



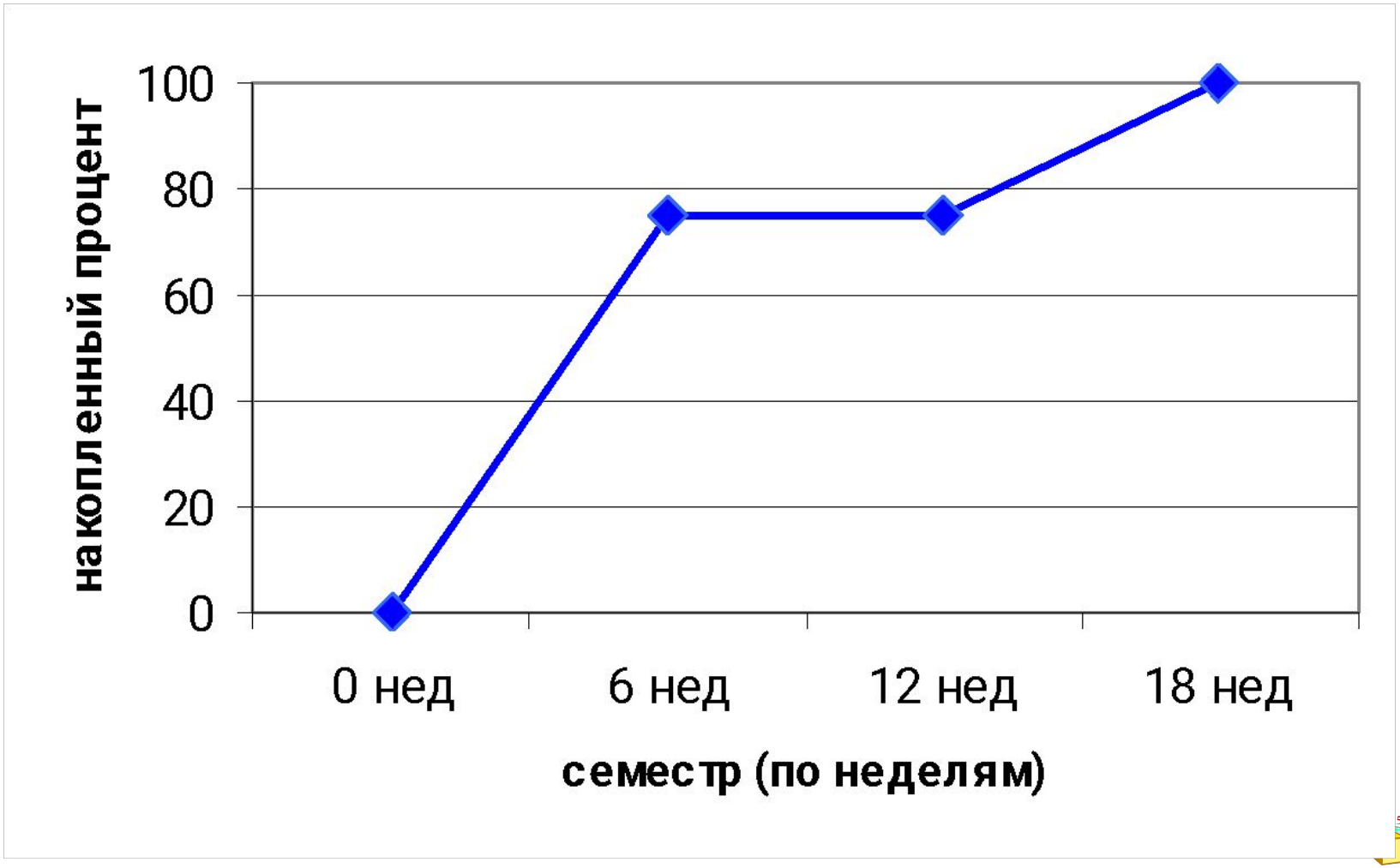
Процентили

Правда ли, что сессионная пара – необычайно стрессовая ситуация для студента, которая приводит даже к самоубийствам?

Seiden, R.H. (1966) “Campus Tragedy: A Story of Students Suicide” Journal of Abnormal Psychology, 71, 389-399



Процентили

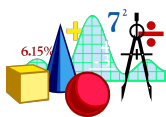

накопленный процент
100
80
60
40
20
0
0 нед
6 нед
12 нед
18 нед
семестр (по неделям)

Процентили

Процентиль всегда выражает положение значения по отношению к какой-либо выборке:

Таня набрала такое количество баллов по тесту по математике, которое соответствует 93 процентилю.

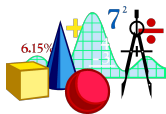
1) Она сдавала математику с 8-классниками обычной школы
2) Она сдавала математику с 11-классниками математической школы



Меры положения

Квартили - значения, которые делят две половины выборки (разбитые медианой) еще раз пополам.

Таким образом, медиана и квартили делят диапазон значений переменной на четыре равные части.

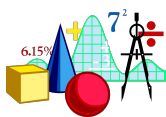


Меры положения

Верхний квартиль ($\mathbf{Q_3}$) делит пополам верхнюю часть выборки (значения переменной больше медианы).

Нижний квартиль ($\mathbf{Q_1}$) делит пополам нижнюю часть выборки (значения переменной меньше медианы).

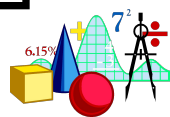
Внутриквартильный (квартильный) размах $= \mathbf{Q_3 - Q_1}$



Меры положения

Нижний квартиль часто обозначают символом 25%, это означает, что 25% значений переменной меньше нижнего квартиля.

Верхний квартиль часто обозначают символом 75%, это означает, что 75% значений переменной меньше верхнего квартиля.

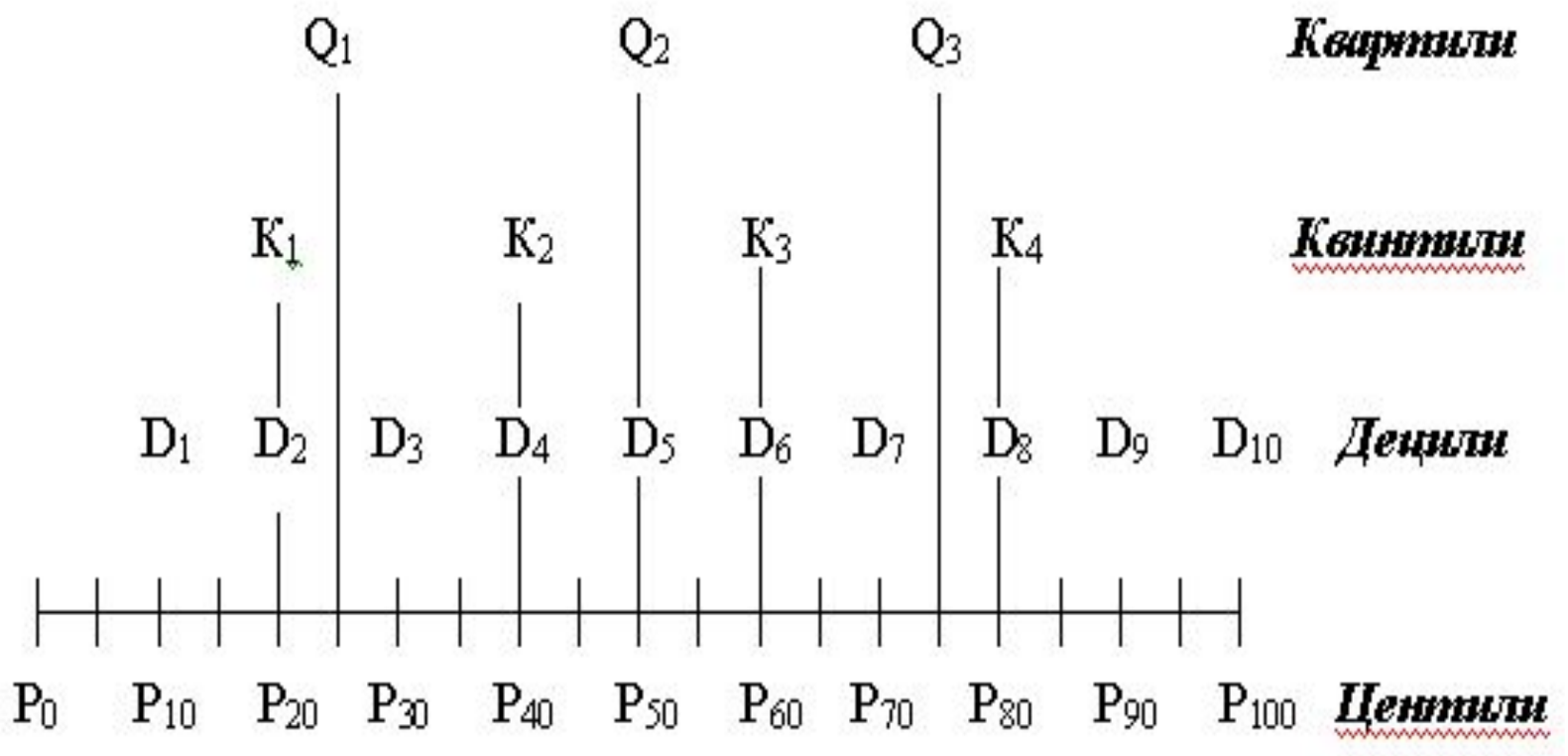


Меры положения

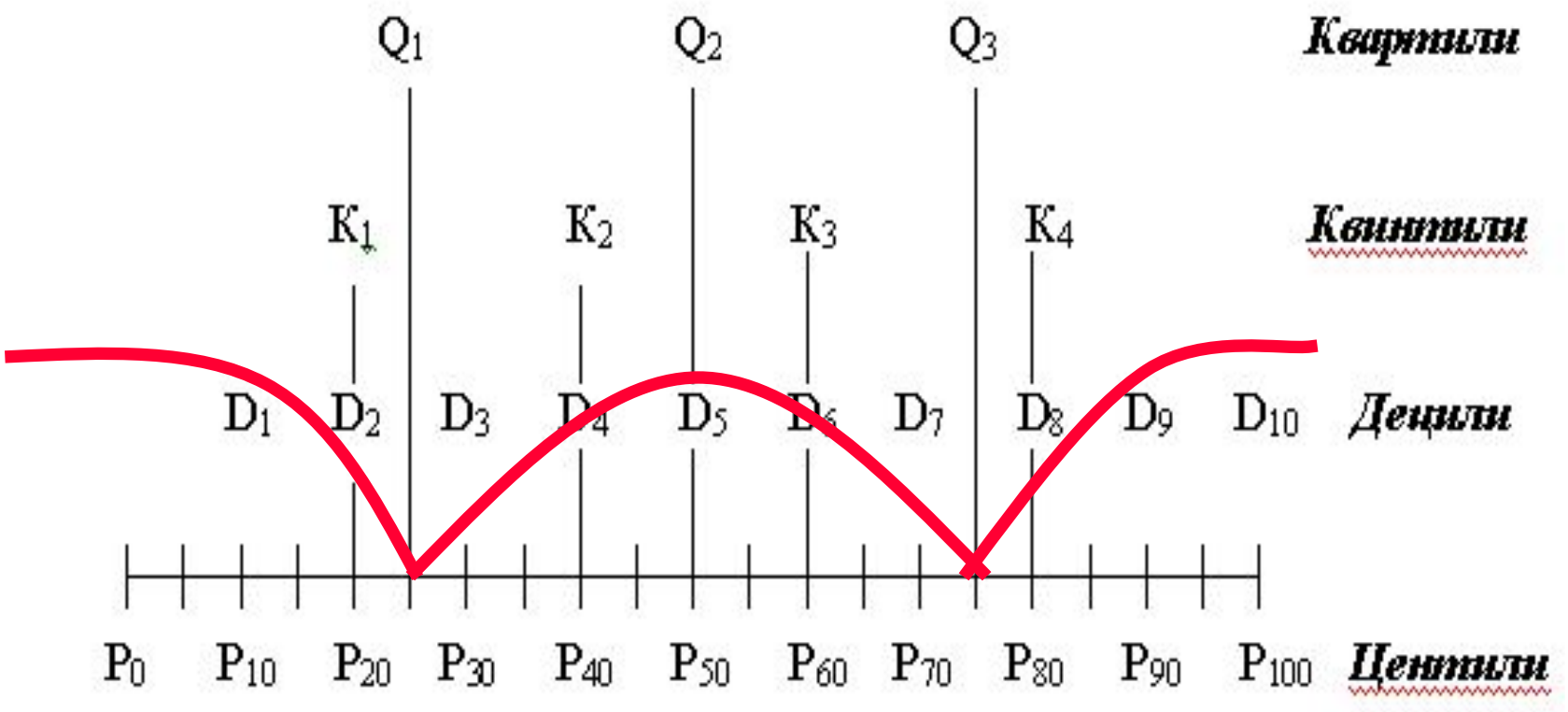
Квинтили делят значения наблюдений на 5 частей, их 4 ($К_1$, $К_2$, $К_3$, $К_4$).

Децили делят совокупность на 10 частей, их 9 (D_1, …, D_9).

Меры положения



Меры положения

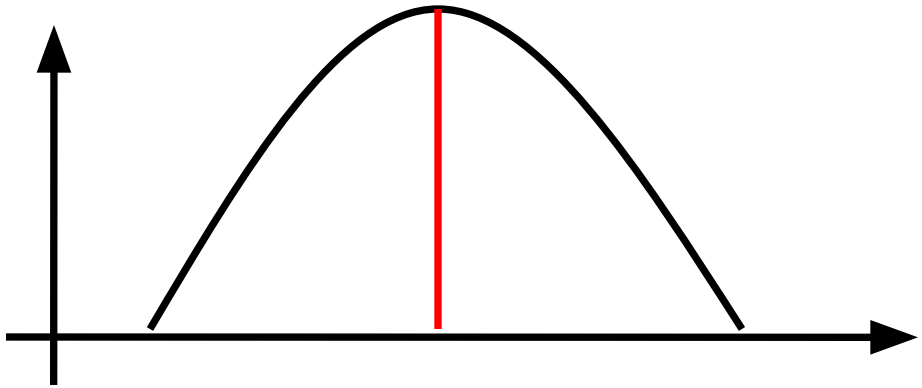


Меры формы

- **Асимметрия** является мерой несимметричности распределения. Если этот коэффициент значительно отличается от 0, распределение является асимметричным

$$A = \frac{\sum (x - \overline{x})^3}{N s^3}$$

Меры формы

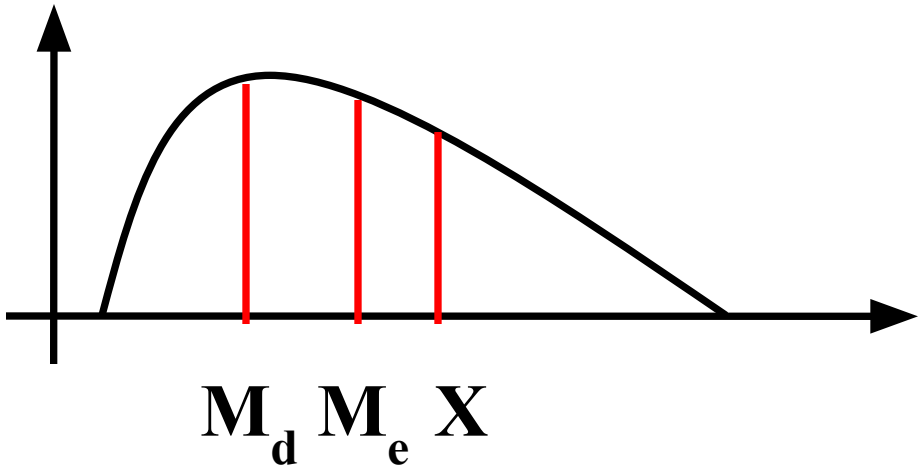


$$\overline{X}=M_e=M_d$$

Симметричное распределение (A=0)

Когда распределение симметрично, среднее, мода и медиана совпадают

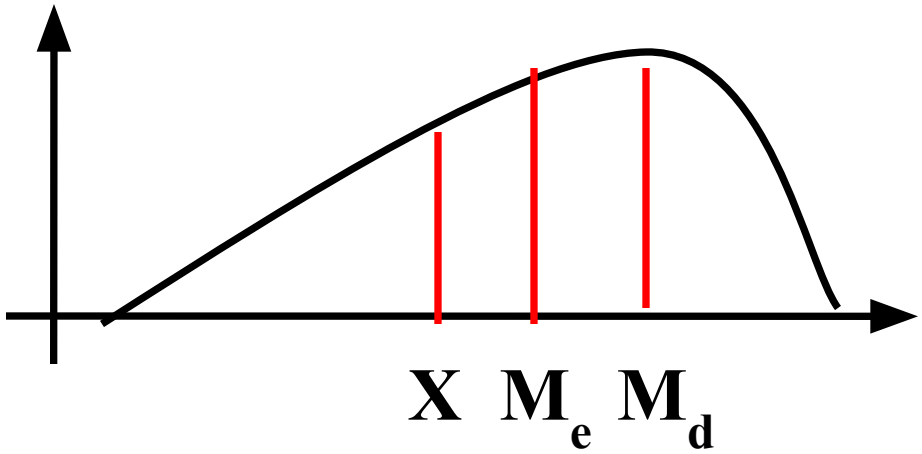
Меры формы



Левостороннее, положительное распределение

Если среднее больше медианы, то распределение называется левосторонним или положительно асимметричным (по знаку числовой характеристики A>0).

Меры формы



Отрицательное, правостороннее распределение

Если среднее меньше медианы, то распределение называется правосторонним или отрицательно асимметричным (A<0).

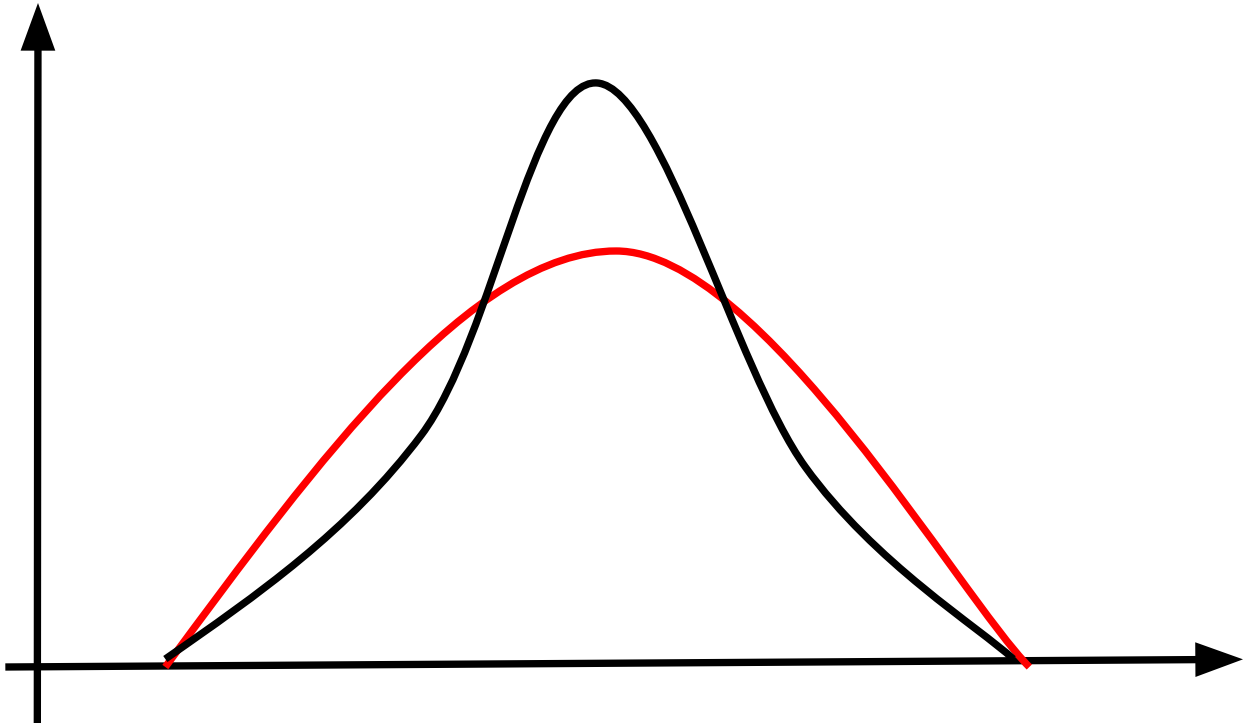
Меры формы

- **Эксцесс** измеряет остроту пика распределения

$$\mathbf{E}= \frac{\sum (x-\overline{x})^4}{N s^4} - 3$$

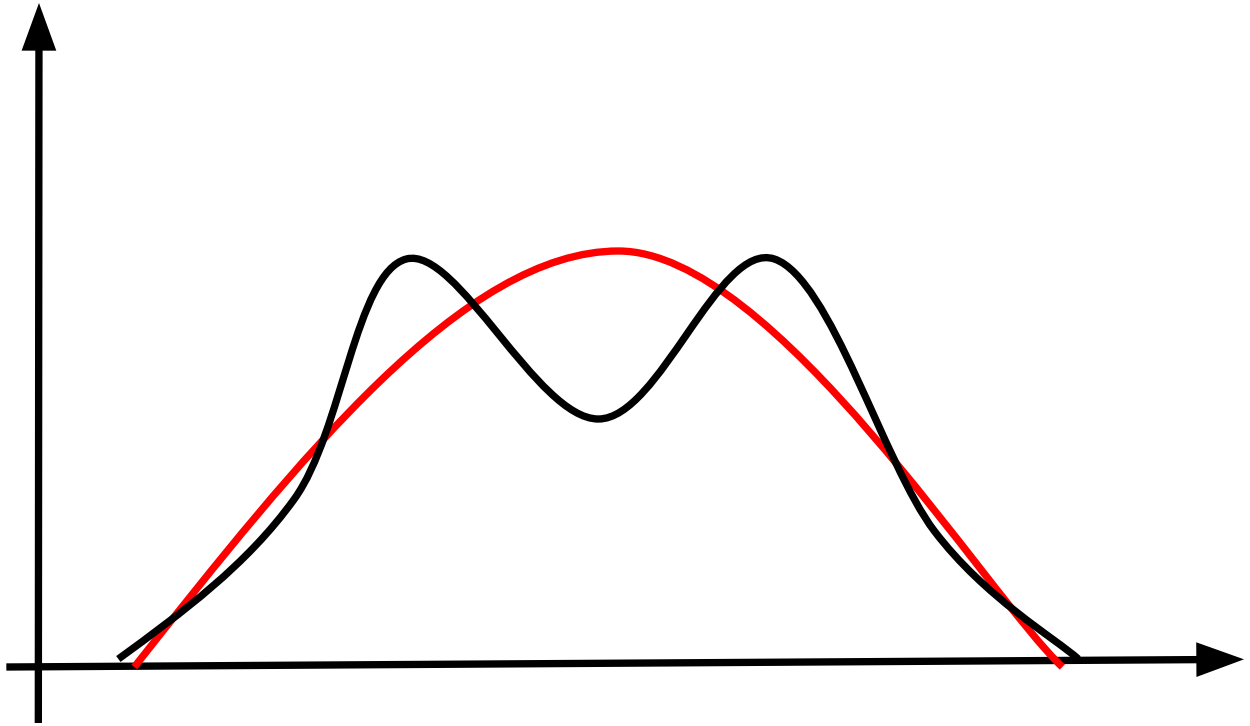
Меры формы

Положительный эксцесс



Меры формы

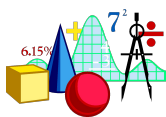
Отрицательный эксцесс



Нормальное распределение

- Нормальное распределение: $f(x)=(1/\sigma\sqrt{2\pi})\exp\{(x-m)^2/2\sigma^2\}$
  - среднее значение m
  - дисперсия σ^2
  - асимметрия A = 0
  - эксцесс E = 3

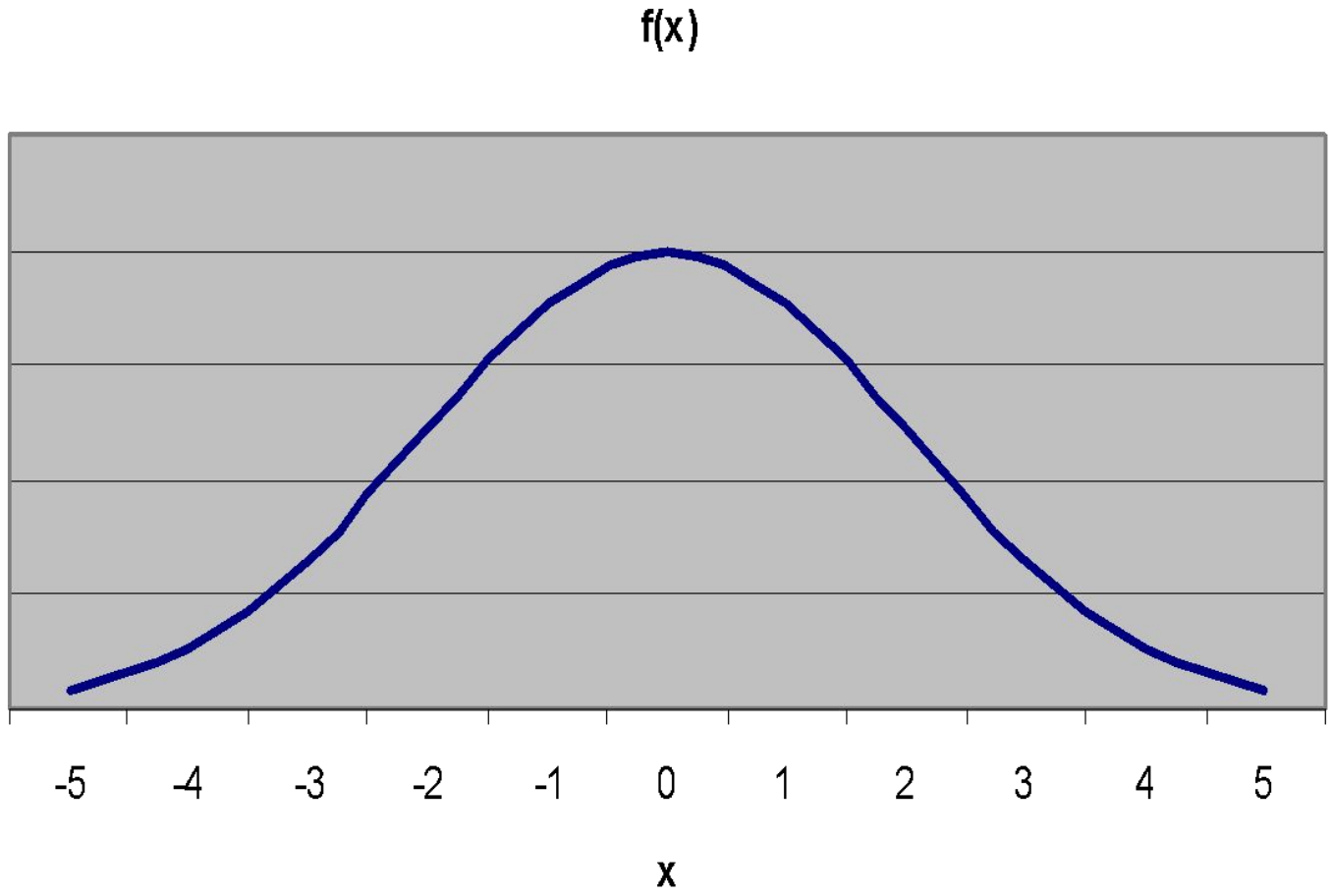
  Стандартное нормальное распределение имеет нулевое среднее и единичную дисперсию



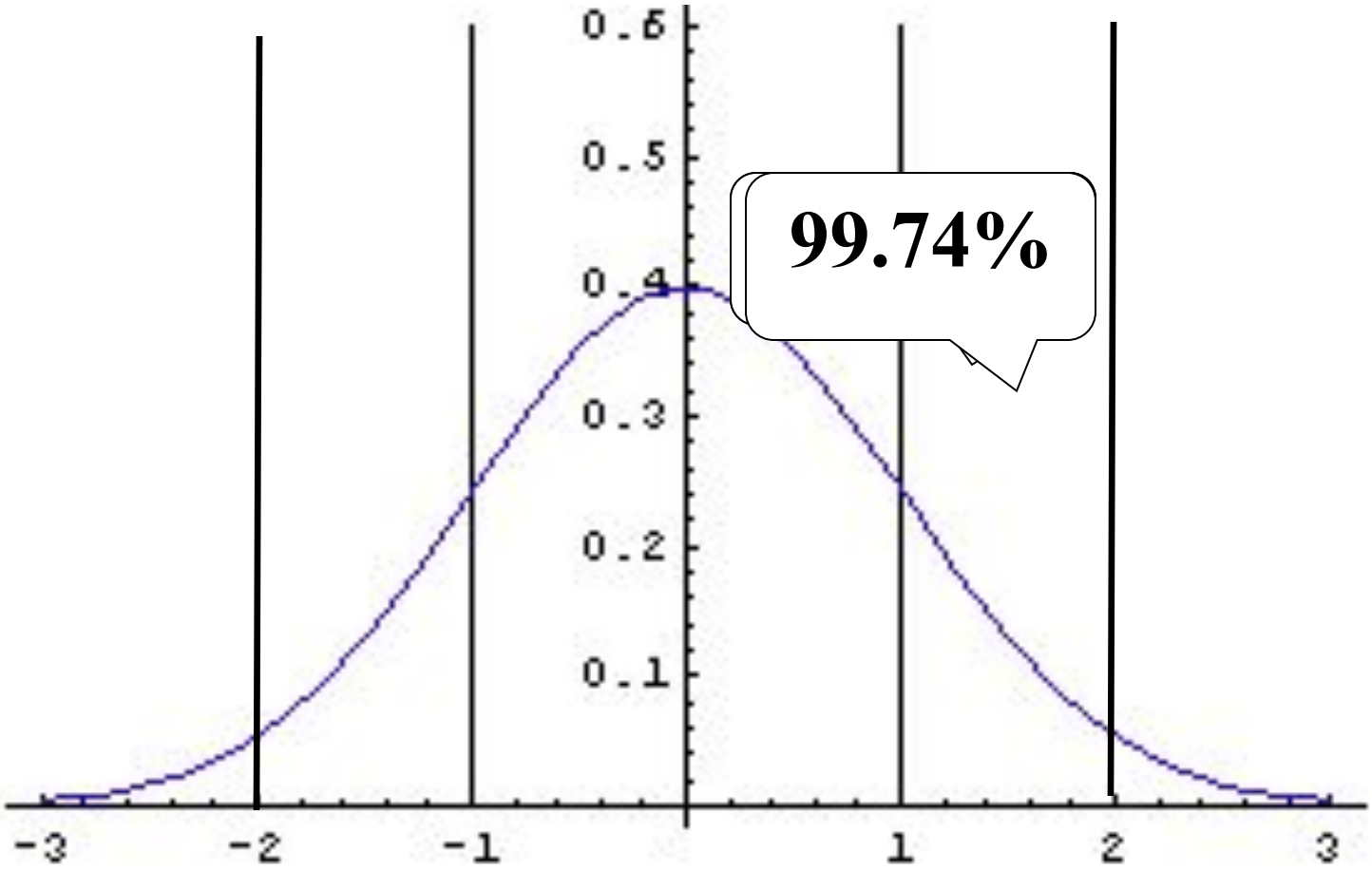
Нормальное распределение

Форма, которую надо запомнить!

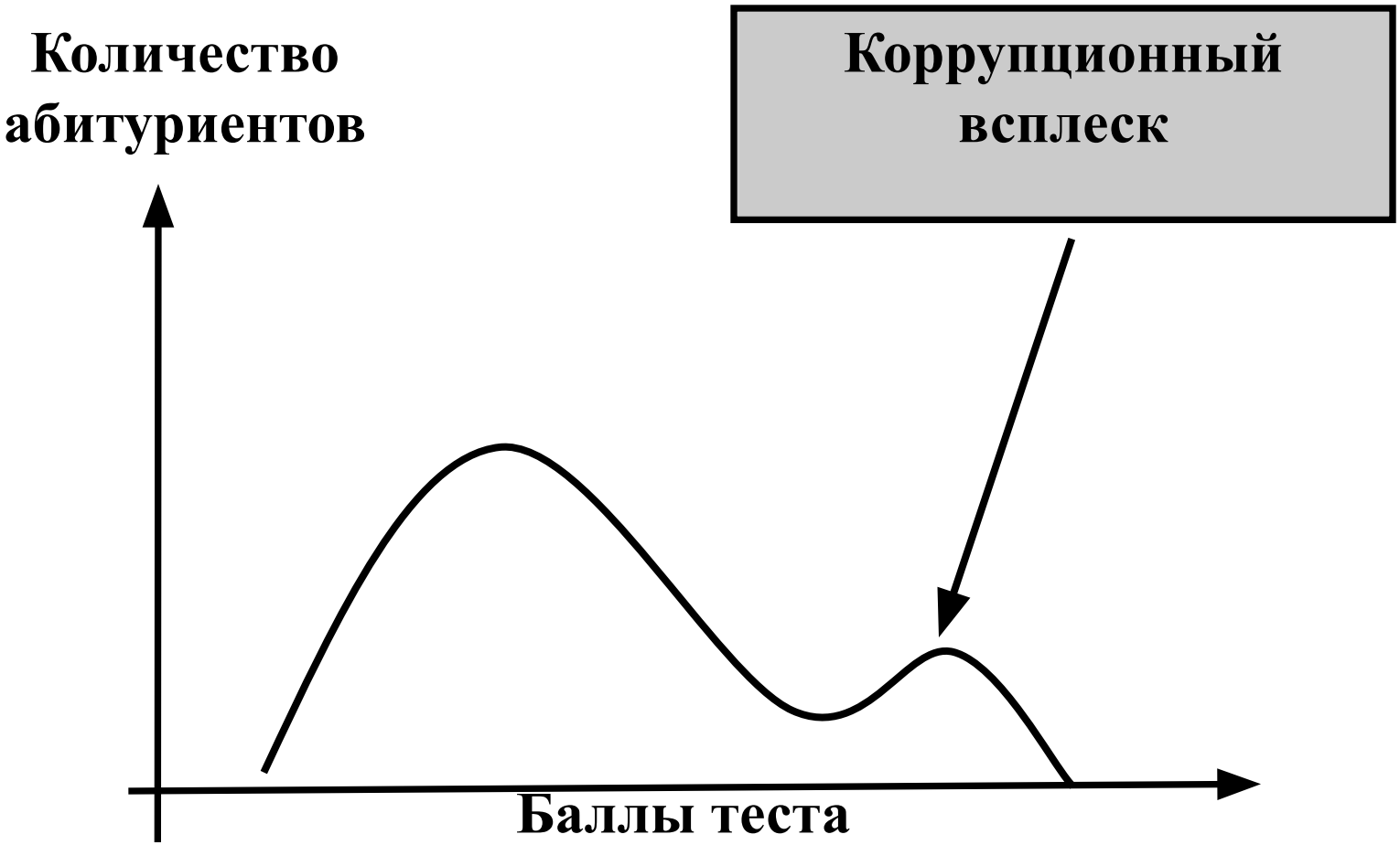
Нормальное распределение



Нормальное распределение



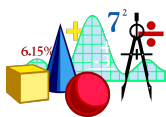
Меры формы


Количество абитуриентов
Коррупционный всплеск
Баллы теста

Нормальное распределение

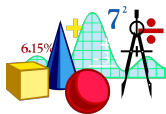
Нормальная кривая человеческих достижений:

2 года – не писать в штаны
10 лет – иметь много друзей и много тусоваться
20 лет – иметь сексуальные отношения
30 лет – много зарабатывать и иметь крутую тачку
50 лет – много зарабатывать и иметь крутую тачку
60 лет – иметь сексуальные отношения
70 лет – иметь много друзей и много тусоваться
78 лет – не писать в штаны



Какую меру выбрать?

Шкала	Мера
Интервальная или отношений	Среднее Стандартное отклонение
Порядка	Медиана Внутриквартильный размах
Наименований	Мода



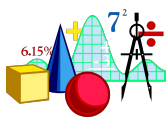
Какую меру выбрать?

Медиана используется когда

1) распределение асимметрично

2) есть опасность перекоса из-за экстремальных значений. Медиана не чувствительна к экстремальным значениям, в то время как среднее очень чувствительно.

3)медиану можно вычислять для данных шкалы порядка и выше.



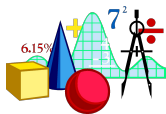
Что мы должны знать?

1) Как строить частотные таблицы и графики
2) Меры центральной тенденции
3) Меры изменчивости

2) Меры положения

3) Меры формы

4) Свойства нормального распределения



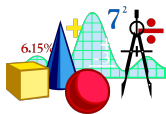
Полезная литература:

К следующей лекции прочитать:

- **Clay Helberg: Pitfalls of Data Analysis (or How to Avoid Lies and Damned Lies)**

- **Barnett A. How Numbers can trick you// Technology Review, October 1994 (на русском)**

(есть в эл.виде в папке «Дополнительная литература»)



ФУХ! ВСЕ!