Лекция 2

Ещё про оценивание

Описательная статистика

Напоминалка

Случайная выборка $X_1, \ldots, X_n$ имеет распределение с неизвестным параметром θ.

Оценка для θ: $\hat{\theta} = f(X_1, \ldots, X_n)$

Свойства оценок

I. Несмещённость: $\mathrm{E}(\hat{\theta}) = \theta \quad \forall \theta \in \Theta$. ← множество допустимых значений

II. Состоятельность: $\lim_{n \to \infty} \mathrm{P}(|\hat{\theta}_n - \theta| \geq \epsilon) = 0 \quad \forall \epsilon > 0, \quad \forall \theta \in \Theta$.

III. Эффективность:

Несмещённая оценка $\hat{\theta}$ для параметра θ называется *эффективной*,
если для любой несмещённой оценки $\tilde{\theta}$ параметра θ по той же выборке выполняется неравенство:

$$\mathrm{D}(\hat{\theta}) \leq \mathrm{D}(\tilde{\theta}) \quad \forall \theta \in \Theta.$$

Часто используемые оценки

Оценка для математического ожидания — выборочное среднее:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Пусть $X_i \sim \text{i.i.d.}$, $\mathrm{E}(X_i) = \mu$, $\mathrm{D}(X_i) = \sigma^2$.

Тогда оценка $\bar{X}$ для μ:

- несмещённая;
- состоятельная;
- BLUE (Best Linear Unbiased Estimator);

эффективность?

Вспомним характеристики выборочного среднего:

$$\mathrm{E}(\bar{X}) = \mu, \ \mathrm{D}(\bar{X}) = \frac{\sigma^2}{n}.$$

Часто используемые оценки

Оценки для дисперсии:

- выборочная дисперсия $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2;$
- скорректированная выборочная дисперсия $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$

их свойства?

Упражнение. Докажите формулы для оценок дисперсии:

$$S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2, \qquad \hat{\sigma}^2 = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} X_i^2\right) - n(\bar{X})^2\right).$$

Эти формулы часто удобны для работы в классе, но при масштабных расчётах лучше ими не пользоваться — они вычислительно неустойчивы.

Познакомимся поближе с выборочной дисперсией

Пусть $X_i \sim \text{i.i.d.}$, $\mathrm{E}(X_i) = \mu$, $\mathrm{D}(X_i) = \sigma^2$.

Проверим выборочную дисперсию S^2 на несмещённость.

Воспользуемся формулой $S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2$.

Найдём нужные математические ожидания из формулы $\mathrm{D}(X_i) = \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2$:

Познакомимся поближе с выборочной дисперсией

Пусть $X_i \sim \text{i.i.d.}$, $\mathrm{E}(X_i) = \mu$, $\mathrm{D}(X_i) = \sigma^2$.

Проверим выборочную дисперсию S^2 на несмещённость.

Воспользуемся формулой $S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2$.

Найдём нужные математические ожидания из формулы $\mathrm{D}(X_i) = \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2$:

$$\mathrm{E}(X_i^2) = \mathrm{D}(X_i) + (\mathrm{E}(X_i))^2 = \sigma^2 + \mu^2$$

Познакомимся поближе с выборочной дисперсией

Пусть $X_i \sim \text{i.i.d.}$, $\mathrm{E}(X_i) = \mu$, $\mathrm{D}(X_i) = \sigma^2$.

Проверим выборочную дисперсию S^2 на несмещённость.

Воспользуемся формулой $S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2$.

Найдём нужные математические ожидания из формулы $\mathrm{D}(X_i) = \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2$:

$$\mathrm{E}(X_i^2) = \mathrm{D}(X_i) + (\mathrm{E}(X_i))^2 = \sigma^2 + \mu^2$$

$$\mathrm{E}((\bar{X})^2) = \mathrm{D}(\bar{X}) + (\mathrm{E}(\bar{X}))^2 = \frac{\sigma^2}{n} + \mu^2$$

Познакомимся поближе с выборочной дисперсией

Пусть $X_i \sim \text{i.i.d.}$, $\mathrm{E}(X_i) = \mu$, $\mathrm{D}(X_i) = \sigma^2$.

Проверим выборочную дисперсию S^2 на несмещённость.

Воспользуемся формулой $S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2$.

Найдём нужные математические ожидания из формулы $\mathrm{D}(X_i) = \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2$:

$$\mathrm{E}(X_i^2) = \mathrm{D}(X_i) + (\mathrm{E}(X_i))^2 = \sigma^2 + \mu^2$$

$$\mathrm{E}((\bar{X})^2) = \mathrm{D}(\bar{X}) + (\mathrm{E}(\bar{X}))^2 = \frac{\sigma^2}{n} + \mu^2$$

Наконец, математическое ожидание самой оценки:

$$\mathrm{E}(S^2) = \mathrm{E}\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathrm{E}(X_i^2)\right) - \mathrm{E}((\bar{X})^2) =$$

Познакомимся поближе с выборочной дисперсией

Пусть $X_i \sim \text{i.i.d.}$, $\mathrm{E}(X_i) = \mu$, $\mathrm{D}(X_i) = \sigma^2$.

Проверим выборочную дисперсию S^2 на несмещённость.

Воспользуемся формулой $S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2$.

Найдём нужные математические ожидания из формулы $\mathrm{D}(X_i) = \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2$:

$$\mathrm{E}(X_i^2) = \mathrm{D}(X_i) + (\mathrm{E}(X_i))^2 = \sigma^2 + \mu^2$$

$$\mathrm{E}((\bar{X})^2) = \mathrm{D}(\bar{X}) + (\mathrm{E}(\bar{X}))^2 = \frac{\sigma^2}{n} + \mu^2$$

Наконец, математическое ожидание самой оценки:

$$\mathrm{E}(S^2) = \mathrm{E}\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathrm{E}(X_i^2)\right) - \mathrm{E}((\bar{X})^2) = \frac{1}{n} n(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) =$$

Познакомимся поближе с выборочной дисперсией

Пусть $X_i \sim \text{i.i.d.}$, $\mathrm{E}(X_i) = \mu$, $\mathrm{D}(X_i) = \sigma^2$.

Проверим выборочную дисперсию S^2 на несмещённость.

Воспользуемся формулой $S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2$.

Найдём нужные математические ожидания из формулы $\mathrm{D}(X_i) = \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2$:

$$\mathrm{E}(X_i^2) = \mathrm{D}(X_i) + (\mathrm{E}(X_i))^2 = \sigma^2 + \mu^2$$

$$\mathrm{E}((\bar{X})^2) = \mathrm{D}(\bar{X}) + (\mathrm{E}(\bar{X}))^2 = \frac{\sigma^2}{n} + \mu^2$$

Наконец, математическое ожидание самой оценки:

$$\mathrm{E}(S^2) = \mathrm{E}\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathrm{E}(X_i^2)\right) - \mathrm{E}((\bar{X})^2) = \frac{1}{n} n\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) =$$

$$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 =$$

Познакомимся поближе с выборочной дисперсией

Пусть $X_i \sim \text{i.i.d.}$, $\mathrm{E}(X_i) = \mu$, $\mathrm{D}(X_i) = \sigma^2$.

Проверим выборочную дисперсию S^2 на несмещённость.

Воспользуемся формулой $S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2$.

Найдём нужные математические ожидания из формулы $\mathrm{D}(X_i) = \mathrm{E}(X_i^2) - (\mathrm{E}(X_i))^2$:

$$\mathrm{E}(X_i^2) = \mathrm{D}(X_i) + (\mathrm{E}(X_i))^2 = \sigma^2 + \mu^2$$

$$\mathrm{E}((\bar{X})^2) = \mathrm{D}(\bar{X}) + (\mathrm{E}(\bar{X}))^2 = \frac{\sigma^2}{n} + \mu^2$$

Наконец, математическое ожидание самой оценки:

$$\mathrm{E}(S^2) = \mathrm{E}\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathrm{E}(X_i^2)\right) - \mathrm{E}((\bar{X})^2) = \frac{1}{n} n\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) =$$

$$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.$$

Смысл коррекции

Итак, $\mathrm{E}\left(S^2\right) = \frac{n-1}{n}\sigma^2.$

То есть выборочная дисперсия — смещённая оценка для генеральной дисперсии σ^2.

Убрать смещение просто.

Смысл коррекции

Итак, $\mathrm{E}\left(S^2\right) = \frac{n-1}{n}\sigma^2.$

То есть выборочная дисперсия — смещённая оценка для генеральной дисперсии σ^2.

Убрать смещение просто. Надо домножить S^2 на $\frac{n}{n-1}$. Получим:

$$\hat{\sigma}^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

$$\mathrm{E}\left(\hat{\sigma}^2\right) = \mathrm{E}\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^{2.}$$

Смысл коррекции

Итак, $\mathrm{E}\left(S^2\right) = \frac{n-1}{n}\sigma^2.$

То есть выборочная дисперсия — смещённая оценка для генеральной дисперсии σ^2.

Убрать смещение просто. Надо домножить S^2 на $\frac{n}{n-1}$. Получим:

$$\hat{\sigma}^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

$$\mathrm{E}\left(\hat{\sigma}^2\right) = \mathrm{E}\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^{2.}$$

Таким образом, скорректированная выборочная дисперсия — несмещённая оценка.

а так ли это важно?

Примечание. В англоязычной литературе термин «sample variance» используется для обозначения скорректированной оценки. Обычная выборочная дисперсия часто обходится без внимания.

В русскоязычной традиции слова «выборочная дисперсия» тоже часто относятся к несмещённой оценке.

Заметьте

Если бы нам было известно генеральное среднее μ, то мы могли бы использовать такую оценку для дисперсии:

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2.$$

Она была бы несмещённой — можете проверить.

И что?

Оценки для дисперсии

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

- смещённая;
- состоятельная.

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

- несмещённая;
- состоятельная.

Оценки для стандартного отклонения

$$S = \sqrt{S^2}$$

- смещённая;
- состоятельная.

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

- смещённая;
- состоятельная.

Статистические программы обычно выдают оценки $\hat{\sigma}$ и $\hat{\sigma}^2$.

Всё это хорошо для работы с количественными признаками.

А что делать с качественными?

Мы можем сопоставить категориям качественного признака некоторые числа и работать с таким признаком как со случайной величиной.

Вот только арифметические действия будут лишены смысла...

Выборочная доля

Пусть $X_1, \ldots, X_n$ независимы,

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

$$\mathrm{E}(X_i) = p; \quad \mathrm{D}(X_i) = p(1-p).$$

Такие признаки называют альтернативными.

p — доля «единичек» в генеральной совокупности.

Естественная оценка для p — доля «единичек» в выборке. Она совпадает с выборочным средним:

$$\hat{p} = \bar{X} = \frac{X_1 + \ldots + X_n}{n}.$$

Из ранее доказанных свойств выборочного среднего следует, что

$$\mathrm{E}(\hat{p}) = p, \quad \mathrm{D}(\hat{p}) = \frac{p(1-p)}{n}.$$

Свойства:

► несмещённая; ► состоятельная; ► эффективная.

Что если хочется оценить всё распределение целиком?

Оценка функции распределения

(Теоретическая) функция распределения:

$$F(x) = P(X \le x)$$

Выборочная (эмпирическая) функция распределения:

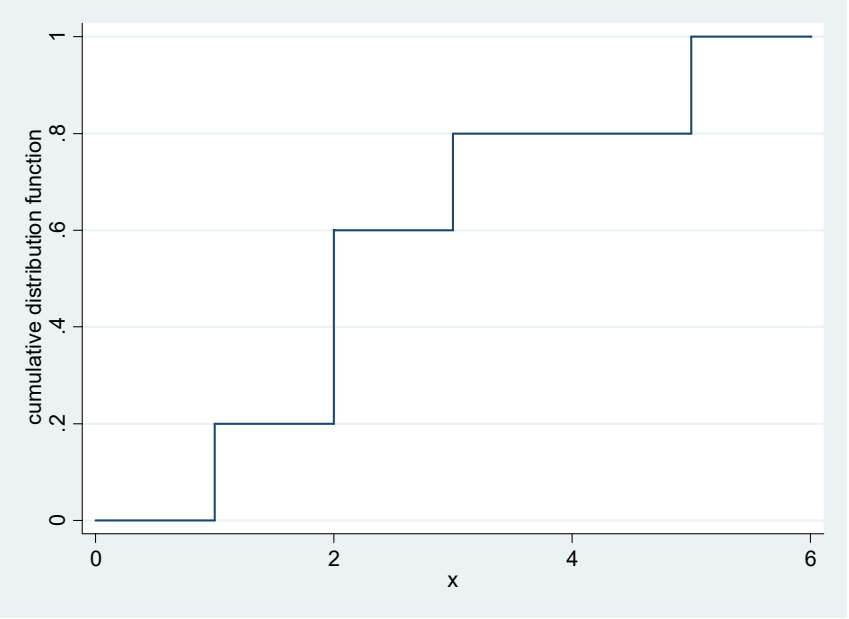
$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I_{X_i \le x}}{n} = \text{доля наблюдений, не превышающих } x$$

Какие свойства у этой оценки?

Что вообще такое — оценка для функции?

Пример. График выборочной функции распределения для выборки

1 3 2 2 5



Оценка функции квантилей

(Теоретическая) функция квантилей:

$$Q(p) = min\{x : F(x) \geq p\}$$

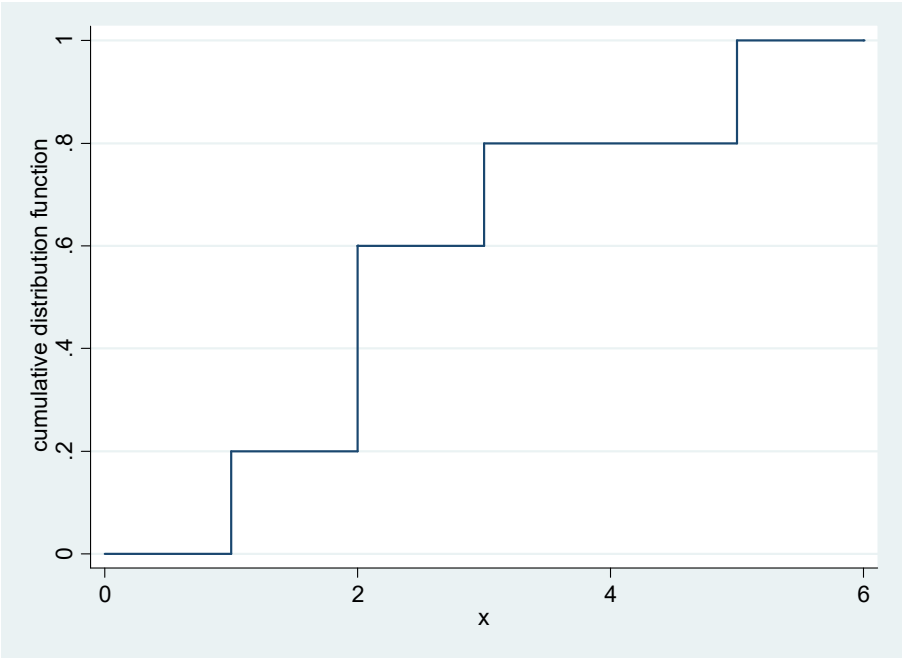
Выборочная (эмпирическая) функция квантилей:

$$\hat{Q}(p) = min\{x : \hat{F}(x) \geq p\}$$

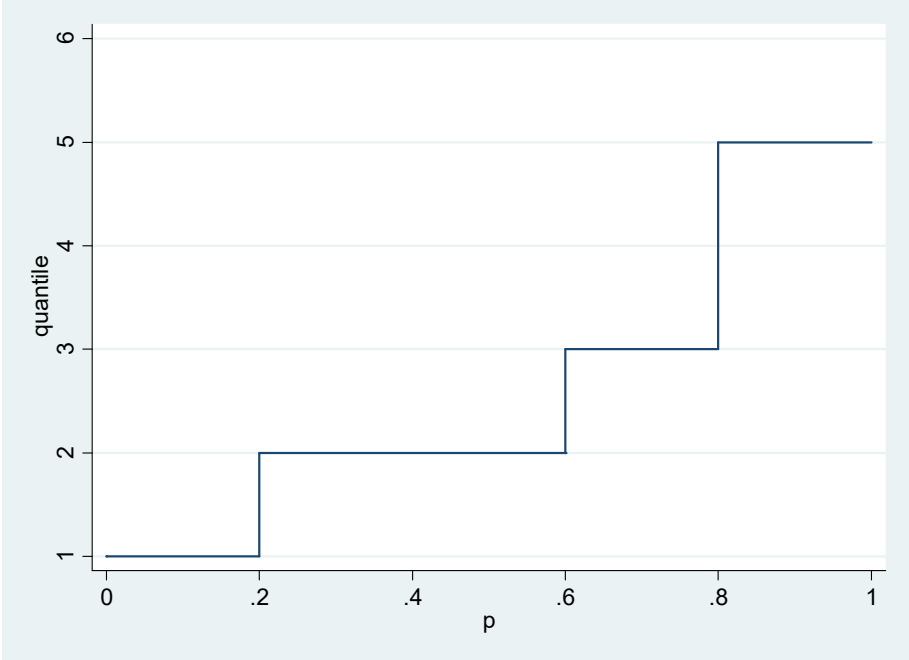
а вдруг не так?

Пример. График выборочных функций распределения и квантилей для выборки

1 3 2 2 5



функция распределения



функция квантилей

Оценивание функции плотности. Гистограмма.

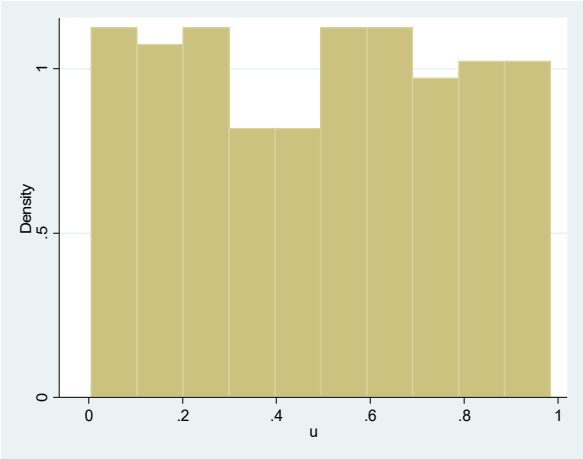
Область значений признака разбивается на интервалы.

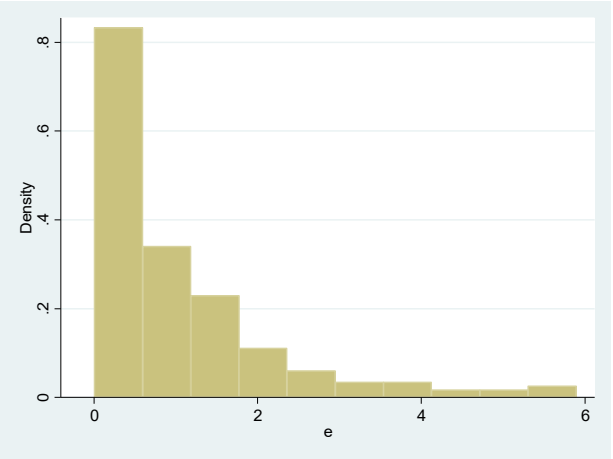
Оценка функции плотности в интервале [l; r) рассчитывается так:

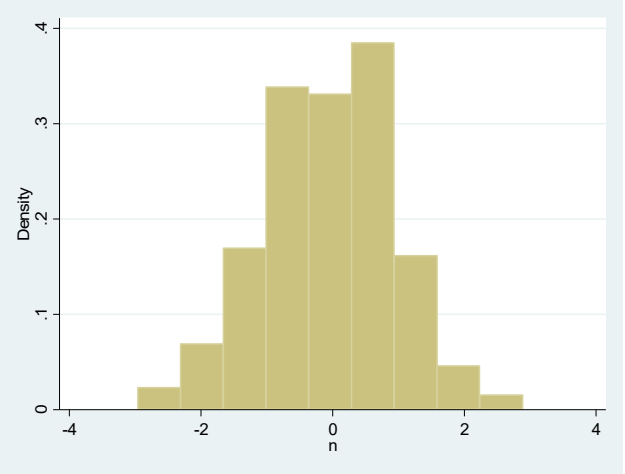
$$\hat{f}(x) = \frac{\textit{доля наблюдений в интервале } [l\,;r) \textit{ среди всех наблюдений}}{r-l}, \quad l \le x < r.$$

График оценённой функции плотности называют гистограммой.

Пример. Гистограммы, полученные по выборкам из равномерного, показательного и нормального распределений:







Есть альтернатива: ядерная оценка плотности (kernel density estimator).

Описательная статистика

Выборочные характеристики (среднее, дисперсия и т. п.) могут рассматриваться не только как оценки, но и просто как способы описания данных. При этом не важно, действительно ли данные получены в результате наблюдения за случайной выборкой, стоит ли какая-либо вероятностная модель за этими данными.

Пример: оценки за контрольную работу 2019 г. в разбивке по полу.

пол	число наблюдений	среднее	станд. откл.	мин.	макс.
юноши	95	22.52	8.17	0	36
девушки	20	25.75	7.39	12	35

Впрочем, мне больше нравится описание через квартили (five-point summary):

пол	число наблюдений	мин.	25%	50%	75%	макс.
юноши	95	0	17	22	29	36
девушки	20	12	22	29	32	35

Какие выводы можно сделать из этих таблиц?

Описательная статистика качественных признаков

Среднее и дисперсия — негодные характеристики (арифметические операции над данными нечисловой природы).

Альтернативный признак можно описать долей одной из категорий.
Доля девушек среди писавших контрольную: 17.4%.

А если категорий больше двух?

Можем рассчитать выборочные доли для каждой из категорий.
При желании — отобразить их на графике. Например, в виде круговой диаграммы или гистограммы.
При построении гистограммы не нужно делать поправку на длину интервала группировки — её можно считать равной единице.

Для описания порядковых признаков можно использовать и квантили, ведь они опираются на отношение порядка, но не на арифметические операции.
Квантили могут быть полезны при большом числе категорий.

К сожалению, меры разброса, основанные на квантилях, всё же теряют смысл, ведь они опираются на вычитание или деление.

Есть в теории вероятностей понятие энтропии, которое можно применять для описания разброса качественных (даже номинальных) величин, но мы обойдём его.

Применимость методов описательной статистики к разным классам признаков

Способ представления	*количественные*	*порядковые*	*номинальные*
Функция распределения	**+**	**+***	**-**
Функция квантилей	**+**	**+***	**-**
Гистограмма, таблица частот	**+**	**+**	**+**

* Описывая распределение *порядкового* признака, помните:
- расстояние по горизонтальной оси на графике функции распределения лишено смысла — оно полностью условно;
- на графике функции квантилей лишено смысла расстояние по вертикальной оси.

Применимость методов описательной статистики к разным классам признаков (2)

Характеристика	*количественные*	*порядковые*	*номинальные*
среднее (мат. ожидание)	+	-	-
дисперсия, станд. отклонение, коэффициент вариации	+	-	-
медиана и прочие квантили	+	+	-
квартильный размах, децильный коэффициент	+	-	-
мода	+	+	+
частота попадания в категории или интервалы значений	+	+	+

Пример описания данных

Данные выборочного обследования 200 респондентов:
Зарплата, тыс. руб.
Богатство — ответ респондента на следующий вопрос: «Представьте себе лестницу из 9 ступеней, где на нижней, первой ступени, стоят нищие, а на высшей, девятой — богатые. На какой ступени находитесь сегодня вы лично?»
Семейное положение — один из четырёх вариантов: «не женат и никогда не был», «состоит в браке», «разведён/разведена», «вдовец/вдова».

первые 10 наблюдений:

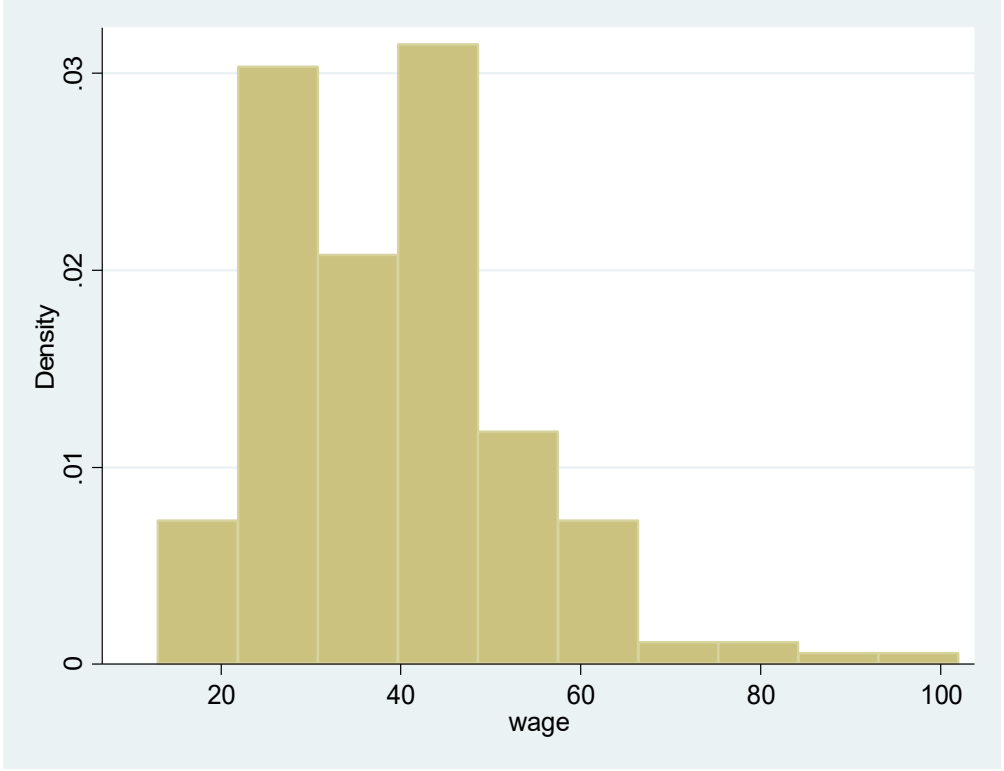
№	Зарплата	Богатство	Семейное положение
1	42	7	в браке
2	47	6	в браке
3	46	9	вдовец/вдова
4	58	9	в браке
5	27	.	разведён/разведена
6	29	5	не женат/не замужем
7	57	4	разведён/разведена
8	58	5	не женат/не замужем
9	51	4	в браке
10	28	5	не женат/не замужем

Зарплата

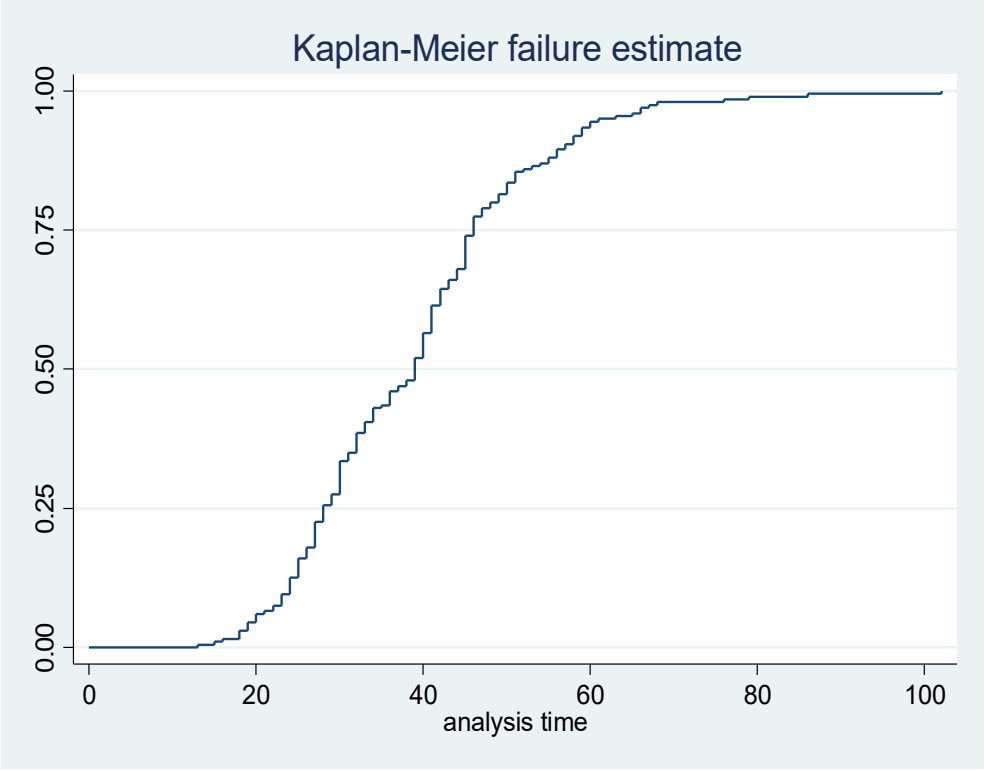
200 наблюдений,
Наименьшее значение: 13
Наибольшее значение: 102
Среднее: 38.9
Стандартное отклонение: 13.7
Квантили:

25%	-	28
50%	**-**	**39**
75%	-	46

Гистограмма:



Выборочная функция распределения:



Субъективное ощущение богатства

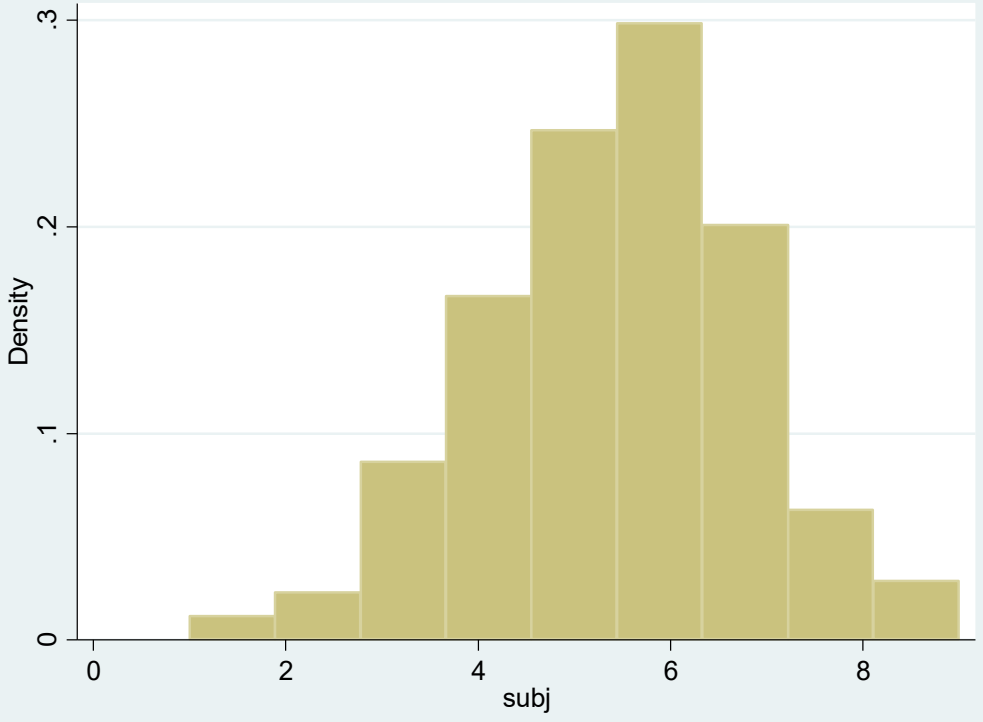
196 наблюдений
(4 пропуска),
Наименьшее значение: 1
Наибольшее значение: 9
Квантили:

25%	-	4
50%	**-**	**6**
75%	-	7

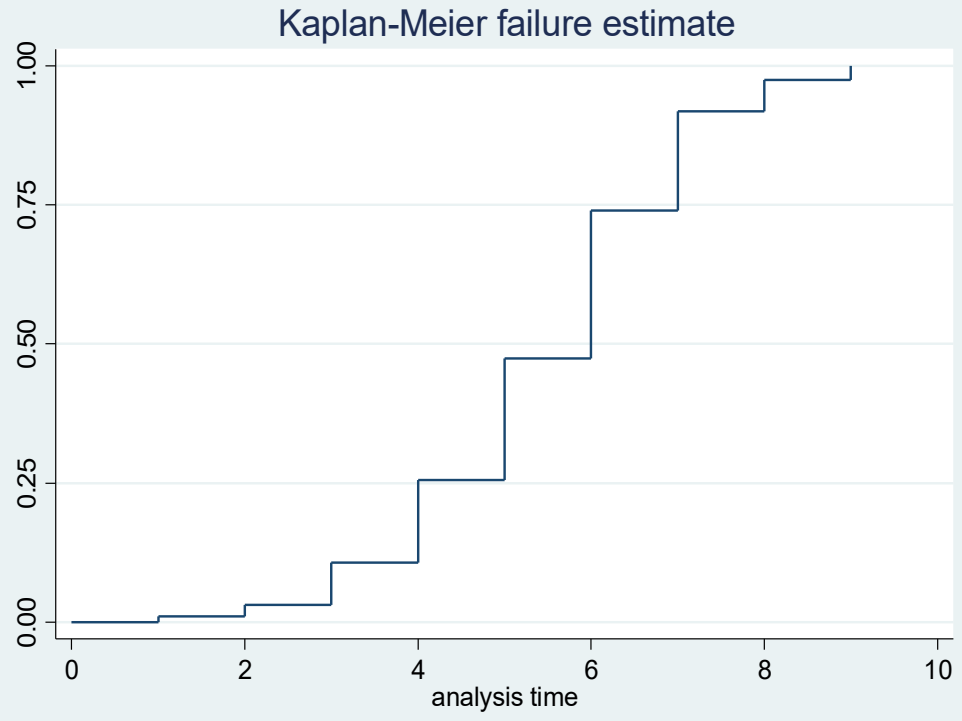
Мода: 6

subjwealth	Freq.	Percent	Cum.
1	2	1.02	1.02
2	4	2.04	3.06
3	15	7.65	10.71
4	29	14.80	25.51
5	43	21.94	47.45
6	52	26.53	73.98
7	35	17.86	91.84
8	11	5.61	97.45
9	5	2.55	100.00
Total	196	100.00	

Гистограмма:



Выборочная функция распределения:



Семейное положение

Категория	Число наблюдений	Доля
не женат/не замужем	56	28,0%
в браке	71	35,5%
разведён/разведена	37	18,5%
вдовец/вдова	36	18,0%
Итого	*200*	*100,0%*

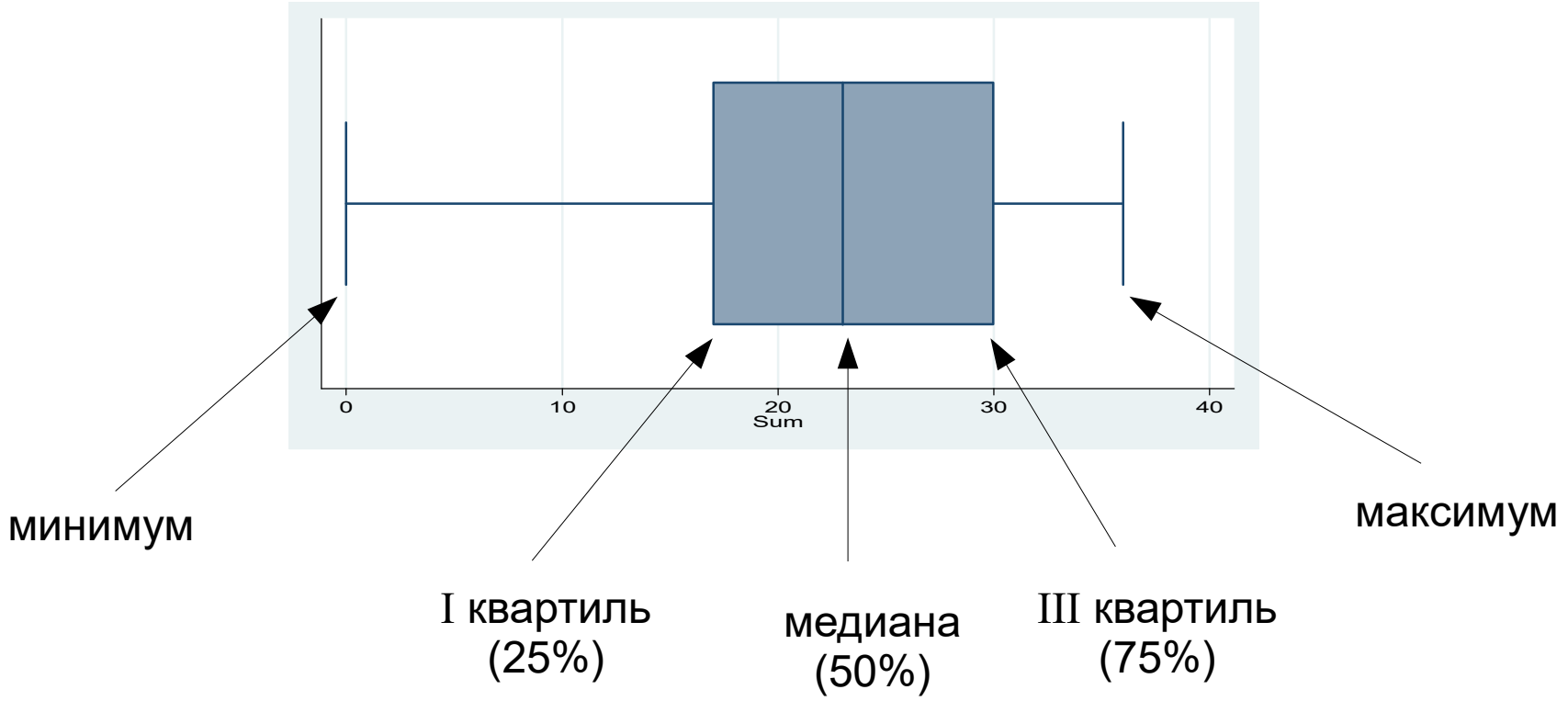
мода

Box-and-whiskers plot

Как это перевести?
«Ящик-с-усами», «ларь-да-усы», «ящичковая диаграмма»?

Этот график — наглядное представление распределения признака по квартилям.

Вот усатый ящик для оценок за контрольную (юноши и девушки вместе):



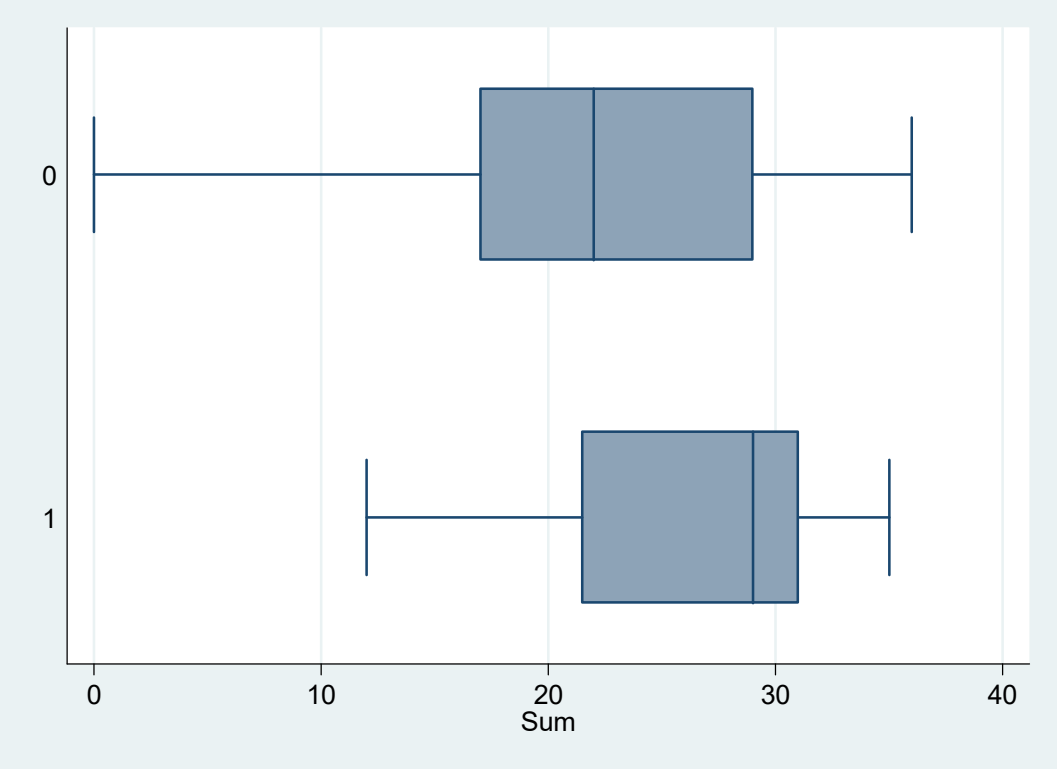
На самом деле, усы — это не всегда минимум и максимум. Об этом потом.

Два ящика и четыре уса

Графики box-and-whiskers особенно удобны при сравнении выборок:

юноши:

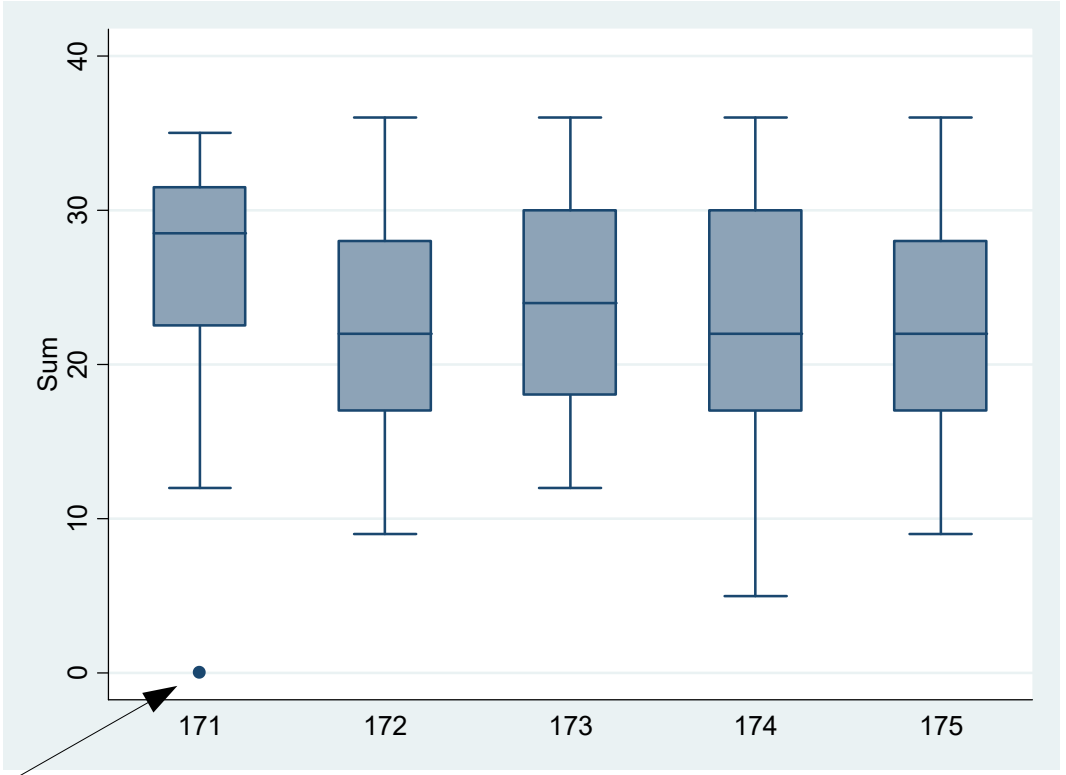
девушки:



Вообще-то, слово «выборка» здесь неуместно — не было никакого отбора.

Пять ящиков, десять усов и одна точка

А вот разбивка оценок по группам:



Кто это?

Часто при построении ящиков стат. программы отделяют «выбросы» (outliers) — наблюдения, далеко отстоящие от основной массы. Вот один из возможных алгоритмов отделения выбросов.

Пусть $I\hat{Q}R(X) = \hat{Q}_X(0.75) - \hat{Q}_X(0.25)$ — межквартильный размах признака в выборке.

Тогда основная часть — те наблюдения, которые находятся в границах

$$\{\hat{Q}_X(0.25) - 1.5\, I\hat{Q}R(X);\ \hat{Q}_X(0.75) + 1.5\, I\hat{Q}R(X)\}.$$

Усы — это минимум и максимум основной части. Отдельные точки — выбросы.

Правило это берётся с потолка, относиться к нему серьёзно не стоит.

Уточнение в конце

Существует три типа данных:

- *пространственные (перекрёстные?) данные* (cross-section data): данные обследования различных объектов приблизительно в один момент времени;

примеры : результаты опросов, выборочных обследований качества продукции...

- *временные ряды* (time series): данные об одном объекте в разные моменты времени;

примеры : динамика числа посетителей сайта, объёмы продаж...

- *панельные данные* (panel or longitudinal data): результаты продолжительного наблюдения за группой объектов (несколько наблюдений за объектом в разные периоды времени).

примеры: РМЭЗ, NLSY, World Bank и т.п.

То, что было в сегодняшней лекции (кроме определений из мат. статистики) относится, в основном, к анализу пространственных данных. Для описания временных рядов и панельных данных, как и для оценивания параметров соответствующих им вероятностных моделей, разработаны свои подходы.

Следующая лекция

Выборка из нормальной генеральной совокупности

Распределение среднего, дисперсии и доли в выборке

График «квантиль-квантиль»