Министерство образования и науки Российской Федерации
Федеральное бюджетное государственное образовательное учреждение
высшего профессионального образования
«Уральский государственный педагогический университет»
Математический факультет
Кафедра математического анализа

Теория вероятностей и математическая статистика (ТВиМС).

Часть 5. Элементы математической статистики.

Бодряков Владимир Юрьевич, д.ф.-м.н.
зав. кафедрой математического анализа МФ УрГПУ

E-mail: Bodryakov_VYu@e1.ru

Екатеринбург – 2011-2012

Литература и интернет - ресурсы

1. Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. пособие. – М.: Высшее образование, 2006. – 479 с.
2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математическая статистика: Учеб. пособие. – М.: Высшее образование, 2006. – 404 с.
3. http://e-lib.uspu.ru
4. www.exponenta.ru

Введение. Основные понятия математической статистики

- **Определение:** Основной задачей математической статистики (МС) считают создание методов сбора и обработки экспериментальных данных с целью получения достоверной информации о случайной величине, интересующей экспериментатора.
- Более подробно: *Первая задача* МС – указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или специально поставленных экспериментов.
- *Вторая задача* МС – разработать методы анализа статистических данных в зависимости от целей исследования.
- Сюда относятся: (а) Оценка неизвестной вероятности события; оценка неизвестной функции распределения, оценка параметров распределения вид которого известен; оценка зависимости с.в. от одной или нескольких случайных величин и др.;
- (б) Проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

Генеральная и выборочная совокупности. Дискретный вариационный ряд. Полигон частот. Гистограмма.

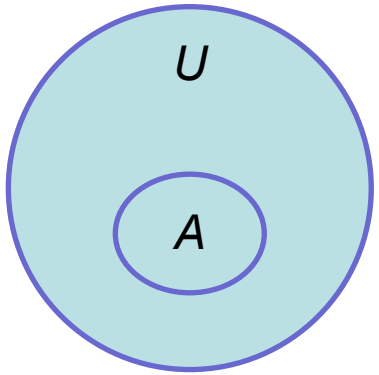
- **Определение:** Множество всех объектов, подлежащих изучению, называется *генеральной совокупностью*. На языке теории множеств аналогом генеральной совокупности является универсальное множество.

- З а м е ч а н и е: Часто сплошное изучение (по качественному или количественному признаку) всех элементов генеральной совокупности сопряжено со значительными трудностями: большое число объектов в совокупности, необходимость уничтожения объектов при проведении некоторых видов испытаний и др. В этом случае приходится ограничиваться изучением ограниченной выборочной совокупности.

- **Определение:** *Выборочной совокупностью* (*выборкой*) называется совокупность объектов, случайно отобранных из генеральной совокупности. Выборочную совокупность можно интерпретировать как подмножество A универсального множества U (рис. 1).

Продолжение …

- Рис. 1.



- **Требование:** Для того, чтобы по данным выборки можно было уверенно судить о всей генеральной совокупности, выборка должна быть *репрезентативной* (*представительной*). В силу закона больших чисел, выборка будет репрезентативной при выполнении следующих условий:
- Объем выборки достаточно велик;
- Обеспечена случайность отбора объектов совокупности;
- Обеспечена равная вероятность попадания в выборку любого объекта генеральной совокупности.

Продолжение …

- Пусть измеряется некоторый интересующий исследователя количественный или качественный признак X генеральной совокупности.
- **Определение:** Возможные значения признака $x_1, x_2, \dots, x_n$ называют вариантами. Обозначим через m_i частоту появления варианты x_i, через n – объем выборки, через $w_i = m_i/n$ - относительную частоту.
- **Определение:** Таблицу, в которой перечислены (в возрастающем порядке) все варианты признака X и соответствующие им частоты (или относительные частоты) называют *статистическим законом распределения* признака X или *дискретным статистическим рядом*.

Признак X	x_1	x_2	x_3	…	x_k	…	Σ_k
Частота m	m_1	m_2	m_3	…	m_k	…	n
Относительная частота w	w_1	w_2	w_3	…	w_k	…	1

- П р и м. При составлении статистического ряда необходимо проверять контрольные суммы частот (относительных частот).

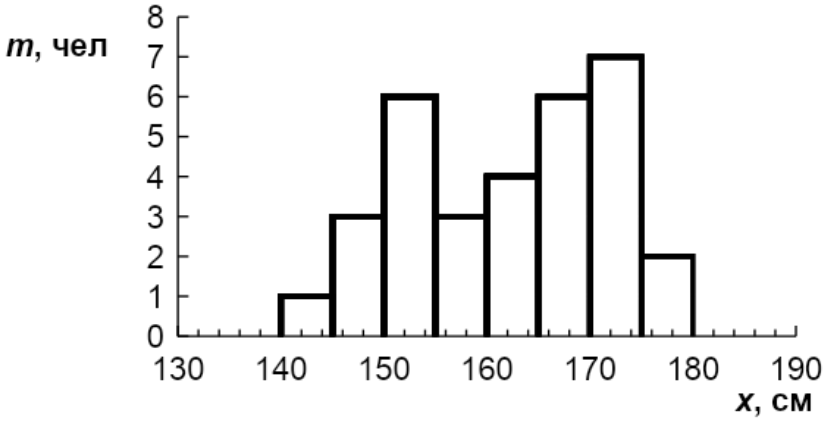
Продолжение …

- **Определение:** Интервальным называется статистический ряд, в котором значения признака отнесены к одному из непересекающихся промежутков, покрывающих в совокупности весь диапазон возможных значений признака.

- П р и м е р 1. Проведено измерение роста x (см) группы учащихся в $n = 32$ чел. «Сырые» результаты измерений таковы: 152, 160, 174, 177, 148, 149, 151, 178, 163, 170, 172, 152, 167, 171, 163, 156, 154, 168, 173, 145, 168, 155, 154, 153, 169, 144, 168, 173, 157, 172, 167, 164. Построить интервальный статистический ряд и частотную гистограмму распределения учащихся по росту.

Рост *X, см*	[140 -145)	[145 -150)	[150 -155)	[155 -160)	[160 -165)	[165 -170)	[170 -175)	[175 -180)	Σ_i
Частота *m*	1	3	6	3	4	6	7	2	32

- Рис. 2. Частотная гистограмма распределения учащихся по росту.



§1. Статистическое изучение случайной величины. Статистические оценки параметров распределения.

- Пусть имеется генеральная совокупность объема N и X – изучаемый признак распределения. Для изучения этого признака генеральной совокупности произведена репрезентативная выборка объема n с выборочным частотным распределением:

X	x_1	x_2	…	x_k	Σ_i
m	m_1	m_2	…	m_k	n

- **Определение:** *Взвешенным выборочным средним* (*выборочной средней*) распределения называют сумму произведений всех ее возможных значений на соответствующие относительные частоты:
  - $\overline{x_в} = \sum_{i=1}^{k} x_i w_i = \frac{1}{n} \sum_{i=1}^{k} x_i m_i.$
- **Определение:** *Выборочной дисперсией* распределения называют сумму произведений относительных частот и квадратов отклонений:
  - $D_в = \frac{1}{n} \sum_{i=1}^{k} m_i \cdot (x_i - \overline{x_в})^2.$
- **Определение:** Выборочное СКО равно квадратному корню из $D_в$:
  - $\sigma_в = \sqrt{D_в}.$

§1. Продолжение …

- **Определение:** *Генеральной средней* $\overline{x_г}$ называют среднее арифметическое значений признака X по генеральной совокупности.
- Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объема N различны, то
  - $\overline{x_г} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}(x_1 + x_2 + \ldots + x_N).$
- Если же значения $x_1, x_2, \ldots, x_k$ признака X имеют, соответственно, частоты $N_1, N_2, \ldots, N_k$, причем $N_1 + N_2 + \ldots + N_k = N$, то
  - $\overline{x_г} = \frac{1}{N}\sum_{i=1}^{k} N_i x_i = \frac{1}{N}(N_1 x_1 + N_2 x_2 + \ldots + N_k x_k).$
- **Определение:** *Генеральной дисперсией* $D_г$ называют среднее из квадратов отклонений значений признака от их среднего значения $\overline{x_г}$:
- Если все значения $x_1, x_2, \ldots, x_N$ признака X различны, то
  - $D_г = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x_г})^2$
- Если же значения $x_1, x_2, \ldots, x_k$ признака X имеют, соответственно, частоты $N_1, N_2, \ldots, N_k$, причем $N_1 + N_2 + \ldots + N_k = N$, то
  - $D_г = \frac{1}{N}\sum_{i=1}^{k} N_i \cdot (x_i - \overline{x_г})^2$.
- **Определение:** Генеральное СКО равно квадратному корню из $D_г$:
  - $\sigma_г = \sqrt{D_г}.$

§1. Продолжение …

- **Философия статистического анализа случайной величины:**
- Пусть $\Theta_г$ некоторая «генеральная» числовая характеристика генеральной совокупности. Например, $\Theta_г = \{M(X_г); D(X_г); \ldots\}$. Пусть $\Theta_в$ некоторая числовая характеристика генеральная совокупности. Например, $\Theta_в = \{\overline{x}_в; D_в; \ldots\}$.
- Выборочная оценка $\Theta_в$ является случайной величиной, поскольку сам процесс выборки носит случайный характер. Поэтому в общем случае $\Theta_в \neq \Theta_г$.
- Можно ли по величине выборочной характеристики $\Theta_в$ можно судить о величине интересующей исследователя генеральной характеристики $\Theta_г$? С какой достоверностью можно по величине $\Theta_в$ судить о величине $\Theta_г$?
- Существуют два вида оценок параметров (числовых характеристик) изучаемого признака Θ генеральной совокупности по данным выборки: точечные и интервальные оценки.
- **Определение:** В результате точечной оценки получается конкретное числовое значение оцениваемого параметра; интервальная оценка дает диапазон в котором с определенной вероятностью лежит оцениваемое значение статистической случайной величины и, следовательно, позволяет судить о точности оценки.

§1. Продолжение …. Точечные оценки.

- Пусть найдено выборочное значение $\Theta_в$ интересующей исследователя числовой характеристики генеральной совокупности. Если можно утверждать, что $\Theta_в = \Theta_г$, то говорят, что оценка $\Theta_в$ является несмещенной, состоятельной и эффективной оценкой числовой характеристики генеральной совокупности $\Theta_г$.
- **Определение:** Выборочная оценка $\Theta_в$ называется *несмещенной*, если ее математическое ожидание равно $\Theta_г$, т.е. $M(\Theta_в) = \Theta_г$. Если $M(\Theta_в) \neq \Theta_г$, то оценка будет *смещенной*.
- **Определение:** Выборочная оценка $\Theta_В$ называется *эффективной*, если при данном объеме выборки n из всех возможных оценок она имеет наименьшую дисперсию.
- **Определение:** Выборочная оценка $\Theta_в$ называется *состоятельной*, если $\Theta_в$ сходится по вероятности к $\Theta_г$ при $n \to \infty$, т.е. если:
  - $\lim\limits_{n\to\infty} P(|\Theta_в - \Theta_г| < \varepsilon) = 1.$
- Можно показать, что $\Theta_в = \overline{x}_г$ является несмещенной и состоятельной оценкой генеральной средней $\overline{x}_г$. Если, к тому же, признак в генеральной совокупности распределен по нормальному закону, то эта оценка является и эффективной.

§1. Продолжение …. Точечные оценки.

- **Теорема 1.** Об оценке генеральной средней по выборочной средней.
- Пусть из генеральной совокупности в результате наблюдений над количественным признаком X извлечена повторная выборка объемом n с различными значениями признака $x_1, x_2, \ldots, x_n$. Пусть определена выборочная средняя: $\overline{x_в} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Тогда
  - $M(\overline{x_в}) = \overline{x_г}$.
- Док-во: Будем рассматривать выборочную среднюю $\overline{x_В}$ как случайную величину и $x_1, x_2, \ldots, x_n$ как независимые одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$. Поскольку эти величины одинаково распределены, они имеют одинаковые числовые характеристики, в частности, математические ожидания, которые мы обозначим через a. В силу свойств математического ожидания имеем:
  - $M(\overline{x_в}) = M\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n}M(X_1 + X_2 + \ldots + X_n) = \frac{na}{n} = a = \overline{x_г}$,
- т.е. выборочная средняя есть *несмещенная* оценка генеральной средней, ч.т.д.
- З а м е ч а н и е: можно показать, что выборочная средняя является и *состоятельной* оценкой генеральной средней.

§1. Продолжение …. Точечные оценки.

- Из доказанной теоремы 1 следует, что при увеличении объема выборки n выборочная средняя стремится по вероятности к генеральной средней. В результате, если по нескольким выборкам достаточно большого объема из одной и той же генеральной совокупности будут найдены выборочные средние, то они будут приближенно равны между собой (тем точнее, чем больше n). В этом и состоит **свойство устойчивости выборочных средних**.
- З а м е ч а н и е. До сих пор выборка предполагалась повторной. Однако полученные выводы справедливы и для бесповторной выборки, если ее объем значительно меньше объема генеральной совокупности. Это положение широко используется на практике.
- **Теорема 2. Формула дисперсии.** Дисперсию D (как выборочную, так и генеральную) можно вычислить как
  - $D = \overline{x^2} - (\bar{x})^2$.
- Док-во: *СРС.*

§1. Продолжение …. Точечные оценки.

- **Теорема 3.** Оценка генеральной дисперсии по исправленной выборочной дисперсии.
- Пусть из генеральной совокупности в результате наблюдений над количественным признаком X извлечена повторная выборка объемом n со значениями признака $x_1, x_2, \ldots, x_k$ и соответствующими частотами $m_1, m_2, \ldots, m_k$. Пусть определена выборочная дисперсия: $D_в = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot (x_i - \bar{x}_в)^2$. Тогда
  - $M(D_в) = \frac{n-1}{n} D_г.$
- Док-во: Без доказательства.
- З а м е ч а н и е: Легко «исправить» выборочную дисперсию $D_в$ так, чтобы ее математическое ожидание $M(D_в)$ было равно генеральной дисперсии $D_г$. В силу теоремы 3 имеем:
- **Определение:** Исправленной выборочной дисперсией s^2 называется величина:
  - $s^2 = \frac{n-1}{n} D_в = \frac{1}{n-1}\sum_{i=1}^{k} m_i \cdot (x_i - \bar{x}_в)^2$.
- **Определение:** Исправленным выборочным среднеквадратическим отклонением (СКО) s называется корень квадратный из исправленного СКО s^2.
- З а м е ч а н и е: Исправленное СКО s является смещенной оценкой $\sigma_г$.

§2. Точность оценки, доверительная вероятность (надежность). Доверительный интервал.

- **Определение:** *Точечной* называют оценку $\Theta_в$ статистической случайной величины X, которая определяется одним числом.
- Все рассмотренные выше оценки – точечные. Например, точечной оценкой генерального СКО $\sigma_г$ является исправленное СКО s. Однако точечная оценка не позволяет судить о точности этой оценки. По этой причине предпочтительнее пользоваться интервальными оценками.
- **Определение:** *Интервальной* называют оценку с.в. X, которая определяется двумя числами – концами интервала, заключающего в себе точечную оценку этой с.в.
- Все рассмотренные выше оценки – точечные. Например, точечной оценкой генерального СКО $\sigma_г$ является исправленное СКО s. Однако точечная оценка не позволяет судить о точности этой оценки. По этой причине предпочтительнее пользоваться интервальными оценками.
- Пусть найденная по данным выборки величина $\Theta_в$ служит *случайной* оценкой неизвестного параметра $\Theta_г$ генеральной совокупности. Число $\delta > 0$ служит *мерой точности* случайной оценки, если $|\Theta_г - \Theta_в| < \delta$. Однако случайный характер выборочной оценки $\Theta_в$ не позволяет утверждать, что неравенство $|\Theta_г - \Theta_в| < \delta$ выполняется всегда; можно лишь говорить о *вероятности* γ осуществления этого неравенства.

§2. Точность оценки … Продолжение.

- **Определение:** *Надежностью* (*доверительной вероятностью*) оценки $\Theta_г$ по $\Theta_в$ называют вероятность γ, с которой осуществляется неравенство $|\Theta_г - \Theta_в| < \delta$.
- Как правило, надежность оценки задается наперед в виде числа, близкого к единице. Стандартными значениями в этом случае являются доверительные вероятности 0,95; 0,99; 0,995.
- Пусть вероятность того, что $|\Theta_г - \Theta_в| < \delta$, равна γ. Иными словами,
  - $P(|\Theta_г - \Theta_в| < \delta) = P(-\delta < \Theta_г - \Theta_в < \delta) = P(\Theta_в - \delta < \Theta_г < \Theta_в + \delta) = \gamma$.
- Это соотношение следует понимать так: вероятность того, что интервал $(\Theta_в - \delta; \Theta_в + \delta)$ заключает в себе (покрывает) оцениваемый неизвестный параметр $\Theta_г$, равна γ.
- **Определение:** *Доверительным* называют интервал $(\Theta_в - \delta; \Theta_в + \delta)$, который покрывает оцениваемый неизвестный параметр $\Theta_г$ генеральной совокупности с заданной надежностью γ.
- З а м е ч а н и е: В силу случайного характера оценки $\Theta_г$ по $\Theta_в$, концы интервала $(\Theta_в - \delta; \Theta_в + \delta)$ сами являются случайными числами (их называют *доверительными границами*).
- Метод доверительных интервалов разработал американский статистик Ю. Нейман, развивая идеи английского статистика Р. Фишера.

§2. Точность оценки … Доверительные интервалы для оценки математического ожидания нормального распределения при известном σ.

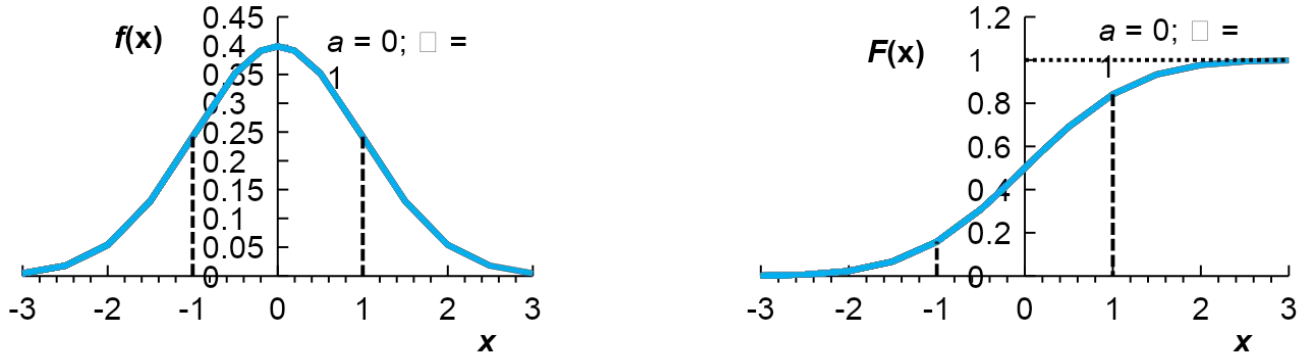
- **Постановка задачи:** Пусть количественный признак X генеральной совокупности распределен нормально, причем СКО σ этого распределения известно. Оценим неизвестное математическое ожидание a по выборочной средней $\bar{x}_в \equiv \bar{x}$ с оценкой доверительных интервалов, покрывающих параметр a, с надежностью γ.
- Решение: Будем рассматривать выборочную среднюю $\bar{x}$ как случайную величину $\bar{X}$. Такой подход оправдан, ибо $\bar{x}$ меняется от выборки к выборке. Будем считать выборочные значения признака x_1, x_2, …, x_n частными значениями одинаково распределенных случайных величин X_1, X_2, …, X_n, имеющих одинаковые математическое ожидание a и СКО σ.
- Можно показать, что если с.в. X распределена нормально, то выборочная средняя $\bar{X}$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения $\bar{X}$ таковы:
  - $M(\bar{X}) = a;\ \sigma(\bar{X}) = \sigma/\sqrt{n}.$
- Потребуем, чтобы выполнялось соотношение
  - $P(|\bar{X} - a| < \delta) = \gamma,$
- где γ – заданная надежность.

§2. Точность оценки … Продолжение.

- Как известно, для нормального распределения
  - $P(|X - a| < \delta) = 2\Phi(\delta/\sigma)$,
- где $\sigma = \sigma(X)$ и $\Phi(z)$ – табулированная функция Лапласа.
- Заменяя X на $\bar{X}$ и $\sigma(\bar{X}) = \sigma/\sqrt{n}$, получим
  - $P(|\bar{X} - a| < \delta) = 2\Phi(\delta\sqrt{n}/\sigma) = 2\Phi(t)$,
- где параметр $t = \delta\sqrt{n}/\sigma$. Выражая δ через параметр t,
  - $\delta = t\sigma/\sqrt{n}$,
- имеем окончательно
  - $P(|\bar{X} - a| < t\sigma/\sqrt{n}) = 2\Phi(t)$.
- В исходных обозначениях задачи получаем рабочую формулу:
  - $P(\bar{x} - t\sigma/\sqrt{n} < a < \bar{x} + t\sigma/\sqrt{n}) = 2\Phi(t) = \gamma.$ (*)
- Задача решена.
- И н т е р п р е т а ц и я полученного соотношения (*) такова: с надежностью γ можно утверждать, что доверительный интервал
  - $(\bar{x} - t\sigma/\sqrt{n} < a < \bar{x} + t\sigma/\sqrt{n})$
- покрывает неизвестный параметр a; точность оценки $\delta = t\sigma/\sqrt{n}$. Параметр t, определяющий точность оценки δ, находится из равенства
  - $\Phi(t) = \gamma/2$.

§2. Точность оценки … Продолжение.

- Альтернативное решение: Пусть количественный признак X генеральной совокупности распределен нормально, причем $M(X) = a$, $\sigma(X) = \sigma$. Тогда, как отмечено выше, выборочная средняя имеет нормальное распределение с параметрами $M(\bar{X}) = a$; $\sigma(\bar{X}) = \sigma/\sqrt{n}$.
- Если из этой генеральной совокупности извлекать выборки объема n и по ним находить выборочные средние, то можно доказать, что случайная величина
- $$Z = \frac{\bar{X} - a}{\sigma/\sqrt{n}},$$
- где $\bar{X}$ – выборочная средняя; a – генеральная средняя; σ – известное СКО по генеральной совокупности; n – объем выборки, также имеет нормальное распределение как линейная функция нормального аргумента $\bar{X}$, причем $M(Z) = 0$, $\sigma(Z) = 1$ (рис. 3).



- Рис. 3. Плотность и интегральная функция нормального распределения.

§2. Точность оценки … Продолжение.

- П р и м е р 2. С.в. Z принимает значения в промежутке $(-1; +1)$, т.е. в промежутке $(-\sigma; +\sigma)$ с вероятностью
- $P\left(\left|\frac{\bar{X}-a}{\sigma/\sqrt{n}}\right| < t_\gamma\right) = \int_{-t_\gamma}^{t_\gamma} f(1;0;t)dt = 2\int_0^{t_\gamma} f(1;0;t)dt = 2\Phi(t_\gamma{=}1) = 0{,}68 = \gamma.$
- Иными словами, доверительная вероятность того, что
  - $a \in (\bar{x} - 1/\sqrt{n}\,;\, \bar{x} + 1/\sqrt{n}) = 2\Phi(t) = \gamma$
- равна
  - $P(\bar{x} - 1/\sqrt{n} < a < \bar{x} + 1/\sqrt{n}) = 2\Phi(1) = 0{,}683.$
- Аналогично получаем, что с.в. Z принимает значения в промежутке $(-2; +2)$, т.е. в промежутке $(-2\sigma; +2\sigma)$ с вероятностью
  - $P(|Z| < 2) = 2\Phi(2) = 0{,}955;$
- Аналогично, с.в. Z принимает значения в промежутке $(-3; +3)$, т.е. в промежутке $(-3\sigma; +3\sigma)$ с вероятностью
  - $P(|Z| < 3) = 2\Phi(3) = 0{,}997.$

§2. Точность оценки … Продолжение.

- З а м е ч а н и е 1. Оценку $|\bar{x} - a| < t\sigma/\sqrt{n}$ называют классической. Из формулы $\delta = t\sigma/\sqrt{n}$, определяющей точность классической оценки, можно сделать следующие выводы:
- 1) при возрастании объема выборки n число δ убывает и, следовательно, точность оценки $\bar{x}$ возрастает ~ $1/\sqrt{n}$;
  - $P(|\bar{X} - a| < t\sigma/\sqrt{n}) = 2\Phi(t)$;
- 2) увеличение надежности оценки γ в силу соотношения $\gamma = 2\Phi(t)$ влечет возрастание t и, как следствие δ (функция $\Phi(t)$ возрастающая). Иными словами, увеличение надежности оценки влечет за собой уменьшение ее точности.
- З а м е ч а н и е 2. Если доверительная вероятность γ и точность δ заданы, то объем n требуемой выборки дается оценкой $n = (t\sigma/\delta)^2$.
- П р и м е р 3. С.в. X нормально распределена с известным СКО $\sigma = 3$. Найти доверительный интервал для оценки математического ожидания a по выборочной средней $\bar{x}$, если объем выборки $n = 36$; заданная надежность оценки $\gamma = 0{,}95$.
- Решение: Найдем t из соотношения $2\Phi(t) = \gamma = 0{,}95$, откуда $t = 1{,}96$ (см. таблицу функции Лапласа $\Phi(z)$). Точность оценки $\delta = t\sigma/\sqrt{n} = 1{,}96 \cdot 3/\sqrt{36} = 0{,}98$. Доверительный интервал: $(\bar{x} - 0{,}98 < a < \bar{x} + 0{,}98)$. Т.е., с надежностью $\gamma = 0{,}95$ можно утверждать, что генеральное среднее a накрывается интервалом $(\bar{x} - 0{,}98; \bar{x} + 0{,}98)$. Задача решена.

§2. Точность оценки … Доверительные интервалы для оценки математического ожидания нормального распределения при неизвестном σ.

- **Постановка задачи:** Пусть количественный признак X генеральной совокупности распределен нормально, причем СКО σ этого распределения неизвестно. Оценим неизвестное математическое ожидание a по выборочной средней $\overline{x_в} \equiv \bar{x}$ с оценкой доверительных интервалов, покрывающих параметр a, с надежностью γ.
- Решение: Вследствие того, что величина СКО σ генеральной совокупности теперь неизвестна, нельзя напрямую пользоваться результатами решения предыдущей задачи. Однако и в этом случае можно применять ту же идеологию оценки доверительного интервала.
- Построим по данным выборки с.в. T с возможными значениями t:

$$T = \frac{\bar{X} - a}{S/\sqrt{n}},$$

- где $\bar{X}$ – выборочная средняя; a – генеральная средняя; S – исправленное выборочное СКО; n – объем выборки.
- Можно показать, что с.в. T имеет t-распределение Стьюдента с $k = n - 1$ степенями свободы. Свойства распределения Стьюдента хорошо изучены (см. Приложение 1). Плотность $S(t; n)$ распределения Стьюдента подобна нормальному распределению, и зависит только от двух параметров: действительного t и натурального n.

§2. Точность оценки … Продолжение..

- С учетом четности распределения Стьюдента, вероятность осуществления неравенства
  - $\left|\frac{\bar{X}-a}{S/\sqrt{n}}\right| < t_\gamma$
- определяется из соотношения:
  - $P\left(\left|\frac{\bar{X}-a}{S/\sqrt{n}}\right| < t_\gamma\right) = \int_{-t_\gamma}^{t_\gamma} S(t;n)dt = 2\int_0^{t_\gamma} S(t;n)dt = \gamma.$
- Выписанный интеграл табулирован и позволяет по заданным величине доверительной вероятности γ и объему выборки n найти величину параметра t_γ, определяющего границы доверительного интервала.
- Заменим неравенство в круглых скобках, «раскрыв модуль»:
  - $P(\bar{X} - t_\gamma S/\sqrt{n} < a < \bar{X} + t_\gamma S/\sqrt{n}) = \gamma.$
- Итак, пользуясь распределением Стьюдента, установлен доверительный интервал $(\bar{x} - t_\gamma s/\sqrt{n} < a < \bar{x} + t_\gamma s/\sqrt{n})$, покрывающий неизвестное генеральное среднее a с надежностью γ.
- З а м е ч а н и е. При большом объеме выборки (практически уже при $n > 30$) распределение Стьюдента стремится к нормальному и может быть им заменено. Однако при малом объеме выборки применение нормального распределения вместо распределения Стьюдента ведет к существенным ошибкам.

§2. Точность оценки … Продолжение…

- П р и м е р 4. Количественный признак X генеральной совокупности распределен нормально. По выборке объема $n = 16$ найдены выборочная средняя $\bar{x} = 20{,}2$ и исправленное СКО $s = 0{,}8$. Оценить неизвестное математическое ожидание a с надежностью $\gamma = 0{,}95$.
- Решение: Пользуясь таблицей значений интегрального распределения Стьюдента, по заданным объему выборки $n = 16$ и доверительной вероятности $\gamma = 0{,}95$ находим величину $t_\gamma = 2{,}13$. Остается выписать доверительные границы:
  - $\bar{x} - t_\gamma s/\sqrt{n} = 20{,}8 - 2{,}13 \cdot 0{,}8/\sqrt{16} = 19{,}77;$
  - $\bar{x} + t_\gamma s/\sqrt{n} = 20{,}8 + 2{,}13 \cdot 0{,}8/\sqrt{16} = 20{,}63.$
- Ответ: С надежностью $\gamma = 0{,}95$ математическое ожидание с.в. X, распределенной нормально с неизвестным СКО, заключено в доверительном интервале $19{,}77 < a < 20{,}63$.
- З а м е ч а н и е: Полученный результат можно интерпретировать и следующим образом, употребительным при представлении результатов обработки экспериментальных статистических данных: оцениваемое значение = среднее ± статистическая погрешность.
- Итак, с надежностью $\gamma = 0{,}95$ математическое ожидание с.в. X, распределенной нормально с неизвестным СКО равно
  - $a = \bar{x} \pm t_\gamma s/\sqrt{n} = 20{,}80 \pm 0{,}43.$

§2. Точность оценки … Оценка истинного значения измеряемой величины

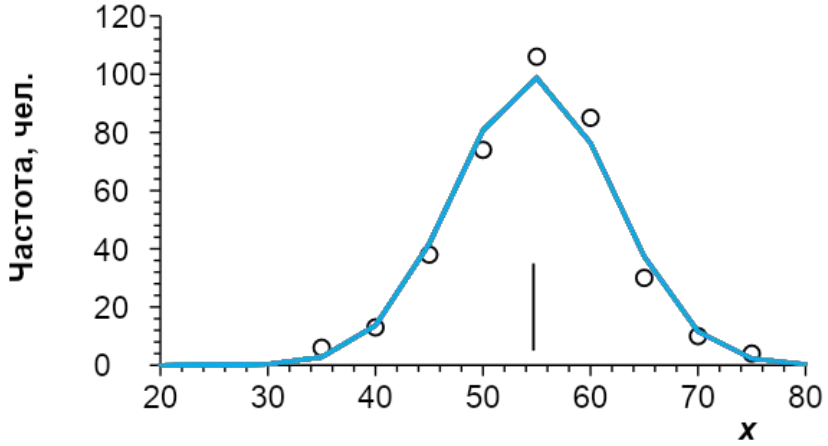
- **Постановка задачи.** Пусть производится n независимых равноточных измерений некоторой физической величины X, истинное значение которой неизвестно. Будем рассматривать результаты отдельных измерений как случайные величины $X_1, X_2, \ldots, X_n$. Эти величины:
- независимы (ибо измерения независимы);
- имеют одно и то же математическое ожидание a (истинное значение измеряемой величины – ибо величины $X_1, X_2, \ldots, X_n$ взяты из одной генеральной совокупности);
- имеют одинаковые дисперсии σ^2 (по той же причине, что и a);
- В силу сказанного, можно непосредственно применить результаты решения предшествующей задачи.

- П р и м е р 5. Баллы X за ЕГЭ по математике группы абитуриентов объемом n = 366 чел. сгруппированы с 5-балльным шагом. Предполагая, что распределение нормально, оценить математическое ожидание с.в. X и доверительный интервал для него с надежностью γ = 0,95. Сравнить визуально эмпирическое и расчетное распределения.

X	35	40	45	50	55	60	65	70	75	Σm_i
m	6	13	38	74	106	85	30	10	4	366

§2. Точность оценки … Продолжение

- Решение: В качестве оценки математического ожидания a с.в. X по известным статистическим формулам вычислим выборочное среднее значение $\bar{x} = 54{,}70$; исправленную выборочную дисперсию $D = 54{,}57$ и исправленное СКО $s = 7{,}39$.
- Пользуясь таблицей значений интегрального распределения Стьюдента, по заданным объему выборки $n = 366$ и доверительной вероятности $\gamma = 0{,}95$ находим величину $t_\gamma = 1{,}960$. Остается выписать доверительные границы для математического ожидания:
  - $a = \bar{x} \pm t_\gamma s/\sqrt{n} = 54{,}70 \pm 1{,}960 \cdot 7{,}39/\sqrt{366} = 54{,}70 \pm 0{,}76.$
- Ответ: С надежностью $\gamma = 0{,}95$ М.О. $a = \bar{x} \pm t_\gamma s/\sqrt{n} = 54{,}70 \pm 0{,}76$ (рис.4).



- Рис. 4. Эмпирическое (кружки) и теоретическое частотное распределение с.в. X.

Приложение 1. Распределение Стьюдента

- **Определение:** t - распределением Стьюдента (рис. 5) называется распределение вероятностей, заданное плотностью распределения:

$$S(t; n) = B_n \cdot \left[1 + \frac{t^2}{n-1}\right]^{-n/2},$$

- где t – действительный аргумент распределения; n – натуральный параметр распределения (объем выборки). Нормирующие множители:

$$B_n = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)\cdot\Gamma((n-1)/2)}};$$

- В свою очередь, $\Gamma(z)$ – гамма – функция Эйлера. По определению

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t}.$$

- Основные свойства гамма – функции определяются равенствами:

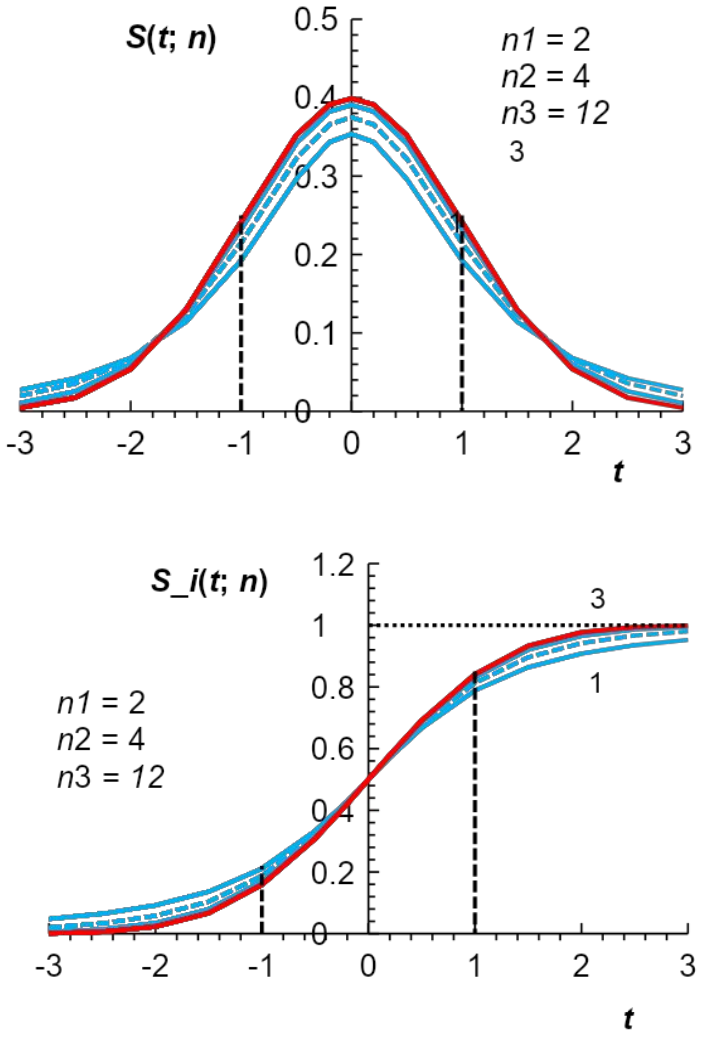
$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}; \quad \Gamma(1) = 1; \quad \Gamma(n+1) = n!; \quad \Gamma(z+1) = z\,\Gamma(z).$$

- **Определение:** Интегральным t - распределением Стьюдента (рис. 4) называется распределение вероятностей, заданное интегралом от плотности t - распределения:

$$S_{int}(t; n) = \int_{-t}^{t} S(t; n)dt = 2\int_0^t S(t; n)dt.$$

- Свойства интегральной функции t - распределения Стьюдента:
- 1) Функция $S_{int}(t; n)$ четна; 2) $0 \le S_{int}(t; n) \le 1$;
- 3) При $n \to \infty$ t - распределение переходит в нормальное $N(a=1; \sigma=1; x)$.

Приложение 1. Распределение Стьюдента. Продолжение



- Рис. 5. Плотность и интегральная функция t – распределения Стьюдента (синие линии) для нескольких значений n. Для сравнения показано (красная линия) нормальное распределение $N(a = 1; \sigma = 1; x)$.

Приложение 2. William Sealy Gosset

Уильям Сили Госсет

William Sealy Gosset

Стьюдент в 1908 г.

Дата рождения:	13 июня 1876
Место рождения:	Кентербери, Кент, Англия
Дата смерти:	16 октября 1937 (61 год)
Место смерти:	Беконсфильд (*англ.*), Бакингемшир, Англия
Научная сфера:	Математическая статистика

Уи́льям Си́ли Го́ссет (13 июня 1876 г. — 16 октября 1937 г.) — известный учёный-статистик, более известный под своим псевдонимом ***Стьюдент*** и за свои работы по исследованию т.н. Распределения Стьюдента.

Приложение 2. William Sealy Gosset. Продолжение

- Родился в Кентербери, у Агнес Сили Видал и полковника Фредерика Госсета. Госсет посещал колледж Винчестер (англ.), а затем прослушал курсы химии и математики в Новом колледже Оксфорда. По окончании университета в 1899 году он поступил на работу на пивоваренный завод Arthur Guinness Son & Co в Дублине.

- «Гиннесс» был передовым предприятием пищевой промышленности, и Госсет мог применить свои знания в области статистики как при варке пива, так и на полях — для выведения самого урожайного сорта ячменя. Госсет приобретал эти знания путём изучения, методом проб и ошибок, проведя два года (1906—1907 гг.) в биометрической лаборатории Карла Пирсона. Госсет и Пирсон были в хороших отношениях, и Пирсон помогал Госсету в математической части его исследований. Так, Пирсон был причастен к публикациям 1908 года (принёсших славу Стьюденту), но придавал мало значения этому открытию. Исследования были обращены к нуждам пивоваренной компании и проводились на малом количестве наблюдений. Биометристы же обычно имели дело с сотнями наблюдений и не испытывали необходимости в развитии методов, основанных на малом их количестве.

- Ранее другой исследователь, работавший на «Гиннесс», опубликовал в своих материалах сведения, составлявшие коммерческую тайну этой пивоваренной компании. Чтобы предотвратить дальнейшее раскрытие конфиденциальной информации, «Гиннесс» запретил своим работникам публикацию любых материалов, независимо от содержавшейся в них информации. Это означало, что Госсет не мог опубликовать свои работы под своим именем. Поэтому он избрал себе псевдоним Стьюдент, чтобы скрыть себя от работодателя. Поэтому его самое важное открытие получило название Распределение Стьюдента, иначе бы оно могло называться теперь распределением Госсета.

Приложение 2. William Sealy Gosset. Продолжение

- Госсет практически все свои работы, включая работу «Вероятная ошибка среднего» (англ. *The probable error of a mean*) опубликовал в журнале Пирсона «Биометрика» под псевдонимом Стьюдент. Первым, кто понял значение работ Госсета по оценке параметров малой выборки, был биолог Рональд Фишер. Госсет написал ему: «Я посылаю вам копию таблиц Стьюдента, поскольку вы, похоже, единственный человек, который когда-либо станет пользоваться ими!» Фишер считал, что Госсет совершил «логическую революцию». По иронии судьбы, t - статистика, благодаря которой знаменит Госсет, была фактически изобретением Фишера. Госсет считал статистику для $z = t/\sqrt{(n-1)}$. Фишер предложил вычислять статистику для t, потому что такое представление укладывалось в его теорию степеней свободы. Фишер также применил распределение Стьюдента в регрессионном анализе.

- Стьюдентизированные остатки также названы в честь Стьюдента, хотя их предложили другие учёные. Подобно проблемам, которые привели к распределению Стьюдента, в их основе лежит та же идея — исправление (adjusting) выборочного стандартного отклонения. Интерес Госсета к выращиванию ячменя привёл его к мысли, что опыт надо планировать с той целью, чтобы не просто повысить среднюю урожайность, но чтобы вывести такие сорта ячменя, чья урожайность была бы устойчива к колебаниям состава почвы или климата. Этот принцип встречается только позднее у Фишера и затем в 1950-х в работе Гэнъити Тагути.

- В 1935 году он покинул Дублин, чтобы занять должность главного пивовара, ответственного за научную сторону производственного процесса, в новой пивоварне Гиннесса в Парк Ройял (англ.), в северо-западной части Лондона. Он скончался от сердечного приступа в городе Беконсфильд (англ.) в Англии.

- Госсет был другом Пирсона и Фишера и был достаточно скромным человеком. Известен случай, когда он оборвал речь своего почитателя словами «Фишер всё равно бы сумел открыть всё это сам».

§3. Элементы корреляционного анализа. Функциональная, статистическая и корреляционная зависимости

- Во многих задачах требуется установить наличие и оценить степень зависимости изучаемой величины Y от одной или нескольких других случайных величин X (или $X_1, X_2, \ldots, X_n$). Прежде всего, рассмотрим зависимость с.в. Y от одной случайной или неслучайной величины X.
- Две с.в. Y и X могут быть: а) связаны функционально; б) связаны статистической, в частности, корреляционной зависимостью; в) независимы.
- **Определение:** *Статистической* называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. Если, при этом, изменение одной из величин влечет изменение среднего значения другой, то такую статистическую зависимость называют *корреляционной*.
- П р и м е р 6. Пусть с.в. Y – урожай зерна; с.в. X – количество внесенных удобрений. С одинаковых по площади участков снимают различный урожай, т.е. Y не является функцией X, ибо на величину урожая, помимо количества удобрений, влияет множество случайных факторов (осадки, температура воздуха, качество почвы и др.). Вместе с тем, как показывает опыт, средний урожай определенно зависит от количества внесенных удобрений. Иными словами с.в. Y и X связаны *корреляционной* зависимостью.

§3. Элементы корреляционного анализа. Отыскание параметров выборочного уравнения прямой линии среднеквадратической регрессии.

- Пусть изучается система количественных признаков (X; Y). В результате n независимых опытов получены n пар чисел $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$.
- Найдем по данным наблюдений выборочное уравнение прямой линии среднеквадратичной регрессии. Уравнение будем искать в виде регрессии Y на X:
  - $y = kx + b$,
- где угловой коэффициент k называют *выборочным коэффициентом регрессии* Y на X и обычно обозначают ρ_{yx}.
- **Определение:** Выборочным уравнением линейной регрессии Y на X называют уравнение вида
  - $Y = \rho_{yx} x + b$,
- наилучшим образом (в определенном смысле) описывающее n пар чисел $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$.
- **Определение:** Назовем отклонением разность $Y_i - y_i$, $i = 1, 2, \ldots, n$. Здесь Y_i – расчетная по уравнению регрессии ордината, соответствующая абсциссе x_i, y_i – экспериментальная i-ая ордината.

§3. Элементы корреляционного анализа. Продолжение...

- Подберем параметры ρ_{yx} и b так, чтобы сумма квадратов отклонений $(Y_i - y_i)^2$ была минимальной:
  - $F(\rho_{yx}; b) = \sum_{i=1}^{n}(Y_i - y_i)^2 = \sum_{i=1}^{n}(\rho_{yx}x_i + b - y_i)^2 \to \min.$
- Такой подход называется методом наименьших квадратов (МНК).
- Для минимизации функции $F(\rho_{yx}; b)$ приравняем нулю частные производные по параметрам (для краткости обозначено $\rho_{yx} = \rho$):
  - $\frac{\partial F}{\partial \rho} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i)\cdot x_i = 0;$
  - $\frac{\partial F}{\partial b} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i) = 0.$
- Выполнив элементарные преобразования (*СРС*), получим систему линейных уравнений относительно параметров ρ_{yx} и b, откуда
  - $\rho_{yx} = \frac{n\sum xy - \sum x\cdot\sum y}{n\sum x^2 - (\sum x)^2};$
  - $b = \frac{\sum x^2 \sum y - \sum x\cdot\sum xy}{n\sum x^2 - (\sum x)^2}.$
- Аналогично можно найти выборочное уравнение прямой линейной регрессии x на y:
  - $x = \rho_{xy}y + d,$
- где ρ_{xy} – выборочный коэффициент регрессии X на Y.

§3. Элементы корреляционного анализа. Продолжение...

- П р и м е р 7. Найти выборочное уравнение прямой линии регрессии Y на X по данным $n = 5$ наблюдений:

x	1,00	1,50	3,00	4,50	5,00
y	1,25	1,40	1,50	1,75	2,25

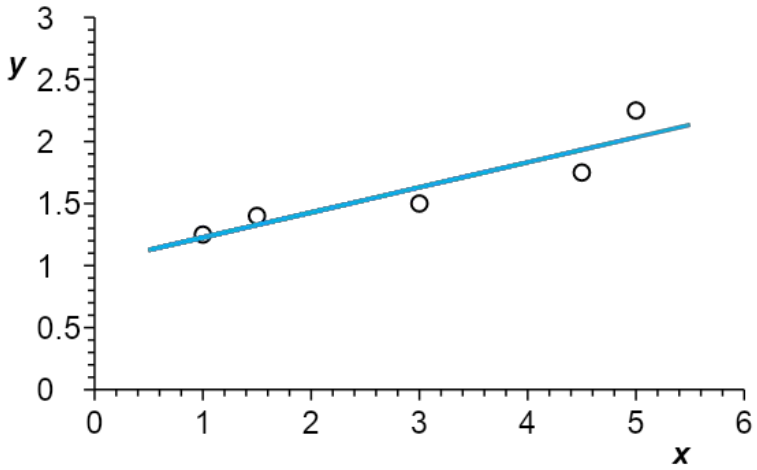
- Решение: Составим расчетную таблицу и вычислим ρ_{yx} и b:

i				
1	1,00	1,25	1,00	1,250
2	1,50	1,40	2,25	2,100
3	3,00	1,50	9,00	4,500
4	4,50	1,75	20,25	7,875
5	5,00	2,25	25,00	11,250
Σ_i	15,00	8,15	57,50	26,975

- $\rho_{yx} = (5 \cdot 26{,}975 - 15 \cdot 8{,}15)/(5 \cdot 57{,}5 - 15^2) = 0{,}202;$
- $b = (57{,}5 \cdot 8{,}15 - 15 \cdot 26{,}975)/(5 \cdot 57{,}5 - 15^2) = 1{,}024.$

§3. Элементы корреляционного анализа. Продолжение...

- Напишем искомое уравнение регрессии и построим ее в сопоставлении с эмпирическими данными (см. рис. 6):
  - $Y = 0{,}202 \cdot x + 1{,}024.$



- Рис. 6. Линейная регрессия: символы – эмпирические данные; линия – расчет. Уравнение линейной регрессии: $Y = 0{,}202 \cdot x + 1{,}024$.

- Ответ: Выборочное уравнение линейной регрессии получено методом МНК; уравнение имеет вид $Y = 0{,}202 \cdot x + 1{,}024$ и находится в разумном согласии с эмпирическими данными.

§3. Элементы корреляционного анализа. Продолжение...

- Обобщим полученный результат, заметив, что:
- $\Sigma x = n \cdot \bar{x}$; $\Sigma y = n \cdot \bar{y}$; $\Sigma x^2 = n \cdot \overline{x^2}$; $\Sigma xy = n \cdot \overline{xy}$.
- Теперь система уравнений для коэффициентов ρ_{yx} и b примет вид:
  - $n \cdot \overline{x^2} \cdot \rho_{yx} + n \cdot \bar{x} \cdot b = n \cdot \overline{xy}$;
  - $n \cdot \bar{x} \cdot \rho_{yx} + n \cdot b = n \cdot \bar{y}$.
- Исключив b из системы, получим для ρ_{yx} представление:
  - $\rho_{yx} = \frac{n \cdot \overline{xy} - n \cdot \bar{x} \cdot \bar{y}}{n(\overline{x^2} - \bar{x}^2)} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2}$.
- Здесь использовано обозначение: $\sigma_x^2 = n \cdot (\overline{x^2} - (\bar{x})^2)$ для выборочной дисперсии случайной величины X.
- В корреляционном анализе, как правило, пользуются не выборочным коэффициентом регрессии ρ_{yx}, а более показательной величиной - *выборочным коэффициентом линейной корреляции* $r_в$, отражающем тесноту корреляционной связи между с.в. Y и X.
- **Определение:** Выборочным коэффициентом линейной корреляции между с.в. Y и X называют с.в.:
  - $r_в = \rho_{yx} \frac{\sigma_x}{\sigma_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}$.

§3. Элементы корреляционного анализа. Продолжение...

- З а м е ч а н и е 1: Выборочный коэффициент линейной корреляции является выборочной оценкой генерального коэффициента корреляции:

$$r_{\text{г}} = \frac{M(XY) - M(X) \cdot M(Y)}{\sigma_x \sigma_y}.$$

- З а м е ч а н и е 2: Коэффициент линейной корреляции $-1 \le r_{\text{в}} \le +1$.
- З а м е ч а н и е 3: Если с.в. X и Y независимы, то коэффициент корреляции $r = 0$; если с.в. X и Y связаны линейной функциональной зависимостью, то коэффициент корреляции $r = \pm 1$.
- З а м е ч а н и е 4: Если генеральная совокупность имеет нормальное распределение, то для оценки генерального коэффициента линейной корреляции $r_{\text{г}}$, можно пользоваться выборочным коэффициентом линейной корреляции $r_{\text{в}}$ (объем выборки $n \ge 50$):

$$r_{\text{в}} - 3\frac{1 - r_{\text{в}}^2}{\sqrt{n}} \le r_{\text{г}} \le r_{\text{в}} + 3\frac{1 - r_{\text{в}}^2}{\sqrt{n}}.$$

- З а м е ч а н и е 5: Важный с точки зрения корреляционного анализа вопрос о значимости коэффициента корреляции $r_{\text{г}}$ решается средствами теории статистической проверки статистических гипотез.

Спасибо за внимание!
Данный раздел закончен.

Ваши вопросы, замечания, предложения …