http://www-chemo.univer.kharkov.ua/

Учебно-методические материалы

- Рабочий план и программа курса “Хімічна інформатика та хемометрія”
- Примеры экзаменационных билетов
- Презентации

Last updated November, 2008

Случайные величины и их характеристики.

1. Типы данных. Шкалы наименований, порядковая, интервальная, отношений.
2. Непрерывные и дискретные случайные величины.Генеральная совокупность и выборка Прямые и косвенные измерения. Выборочные характеристики случайных величин – результатов прямых измерений: интервал изменения, гистограмма, мода, медиана, среднее, дисперсия и стандартное отклонение, относительное стандартное отклонение; ковариационная матрица, коэффициент корреляции. Автомасштабное преобразование.

3. Одномерные распределения случайной величины. Моменты.
4. Распределения дискретные: биомиальное, полиномиальное, Пуассона.
5. Распределения непрерывные: равномерное, Гаусса, Лапласа, хи-квадрат.
6. Центральная предельная теорема.
7. Оценивание характеристик генеральной совокупности по выборочным данным. Метод максимума правдоподобия и его применение для обоснования свойств выборочных среднего и дисперсии. Статистические веса.
8. Робастные оценки: оценки метода абсолютных модулей, бивес-оценки, М-оценки Хьюбера.
9. Правила переноса погрешностей.

Измерение - получение любых количественных характеристик материальных объектов опытным путем.

Измерения бывают ***прямыми*** (когда объект непосредственно сопоставляется с носителем единицы измерения, например, измерение длины линейкой) и ***косвенными*** (когда измеряемая величина расчитывается из других измеренных величин, например, измерение глубины с помощью эхолота)

Выборка (*выборочная совокупность*) - конечное число значений одной случайной величины

Генеральная совокупность - полное (бесконечное) множество значений. (т.е. она включает все возможные значения измеряемой величины и ничего добавить туда уже нельзя)

Под ***случайной величиной*** (СВ) понимается величина, которая в результате опыта со случайным исходом принимает то или иное значение, причем заранее, до опыта, неизвестно, какое именно.

Ω – множество возможных значений величины X.

Опыт – бросок кубика; случайные величины X – число выпавших очков; $\Omega = \{0, 1, 2, 3, 4, 5, 6\}$.

Опыт – работа ЭВМ до первого отказа; случайные величины X – время наработки на отказ; $\Omega = (0, \infty]$

Случайная величина (СВ) X называется ***дискретной***, если множество Ω – счетное, т.е. его элементы можно расположить в определенном порядке и пронумеровать.

Случайная величина *X* называется ***непрерывной*** (недискретной), если множество Ω – несчетное.

Законом распределения **случайной величины *X* называется любая функция (правило, таблица и т.п.), устанавливающая соответствие между значениями случайной величины и вероятностями их наступления и позволяющая находить вероятности всевозможных событий** $p\{a \le X < b\},\ \forall a,b$ **связанных со случайной величиной.**

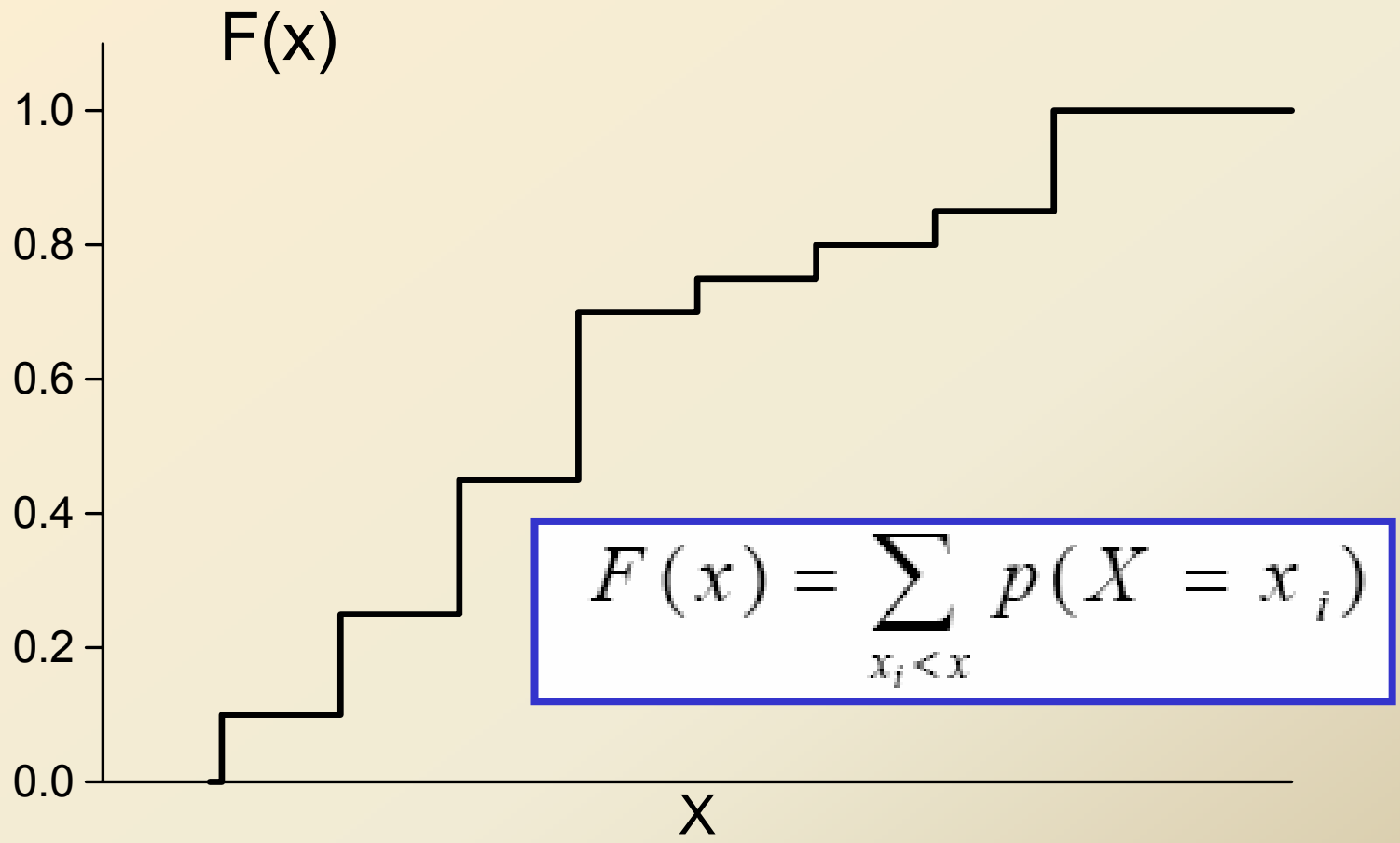
Функцией распределения $F(x)$ случайной величины X называется вероятность того, что она примет значение меньшее, чем аргумент функции x:

$$F(x) = p\{X < x\}$$

Свойства функции распределения

1. $F(-\infty) = 0$.
2. $F(+\infty) = 1$.
3. $F(x_1) \le F(x_2)$, при $x_1 < x_2$.
4. $p(x_1 \le X < x_2) = F(x_1) - F(x_2)$.

Функция распределения любой дискретной случайной величины есть разрывная ступенчатая функция



Случайная величина X называется ***непрерывной***, если ее функция распределения $F(x)$ – непрерывная и дифференцируемая функция для всех значений аргумента.

Парадокс нулевой вероятности:

Для непрерывной функции распределения $F(x)$ **вероятность любого отдельного значения случайной величины должна быть равна нулю**, т.е. не должно быть скачков ни в одной точке.

Плотность вероятности (плотность распределения)

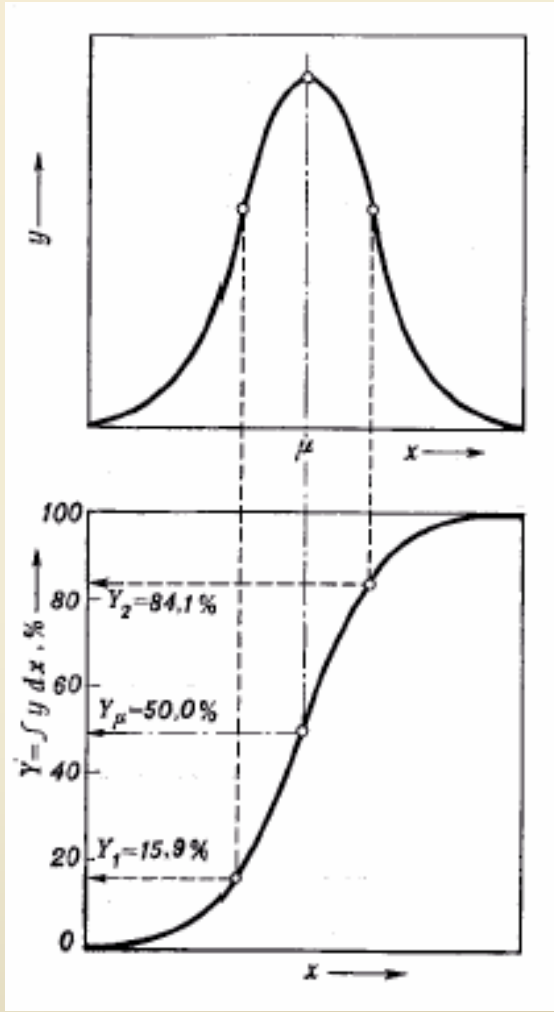
$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

$$p\{a \le X < b\} = \int_a^b f(x)dx$$

$$F(x) = p\{X < x\} = p\{-\infty < X < x\} = \int_{-\infty}^{x} f(x)dx$$

$$f(x) \ge 0$$

$$\int_{-\infty}^{\infty} f(x)dx = p(-\infty \le X < +\infty) = 1$$



Математическое ожидание

$$m_X = \mathrm{M}[X] = \begin{cases} \sum_{i=1}^{N} x_i \cdot p_i & \text{для ДСВ,} \\ \int_{-\infty}^{\infty} x \cdot f(x)dx & \text{для НСВ} \end{cases}$$

Начальный момент *k*-го порядка

$$\alpha_k(x) = \mathrm{M}[X^k] = \begin{cases} \sum_{i=1}^{N} x_i^k \cdot p_i & \text{для ДСВ,} \\ \int_{-\infty}^{\infty} x^k \cdot f(x)dx & \text{для НСВ} \end{cases}$$

При $k = 0$ $\alpha_0(x) = M[X^0] = M[1] = 1$;

$k = 1$ $\alpha_1(x) = M[X^1] = M[X] = m_X$ – математическое ожидание;

$k = 2$ $\alpha_2(x) = M[X^2]$.

Центральный момент *k*-го порядка

$$\mu_k(x) = \mathrm{M}[\overset{\circ}{X}^k] = \begin{cases} \sum_{i=1}^{N} (x_i - m_X)^k \cdot p_i & \text{для ДСВ,} \\ \int_{-\infty}^{\infty} (x - m_X)^k \cdot f(x)dx & \text{для НСВ} \end{cases}$$

При $k = 0$ $\mu_0(x) = M[\overset{\circ}{X}^0] = M[1] = 1$;

$k = 1$ $\mu_1(x) = M[\overset{\circ}{X}^1] = M[\overset{\circ}{X}] = 0$;

$k = 2$ $\mu_2(x) = M[\overset{\circ}{X}^2] = M[(X - m_X)^2] = M[X^2] - 2m_X M[X] + m_X^2 = \alpha_2 - m_x^2 = D_X$ – дисперсия.

Дисперсия

$$D_x = D[X] = \mu_2(x) = \alpha_2(\mathrm{x}) - m_X^2 = \begin{cases} \sum_{i=1}^{N}(x_i - m_X)^2 p_i = \sum_{i=1}^{N} x_i^2 p_i - m_X^2 \text{ для ДСВ,} \\ \int_{-\infty}^{\infty}(x - m_X)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - m_X^2 \text{ для НСВ} \end{cases}$$

Докажите!

Среднее квадратическое отклонение

Мода

Медиана

Квантиль χ_p случайной величины X

$$p\{X < \chi_p\} = F(\chi_p) = p$$

Вопрос:

медиана – это квантиль χ…?

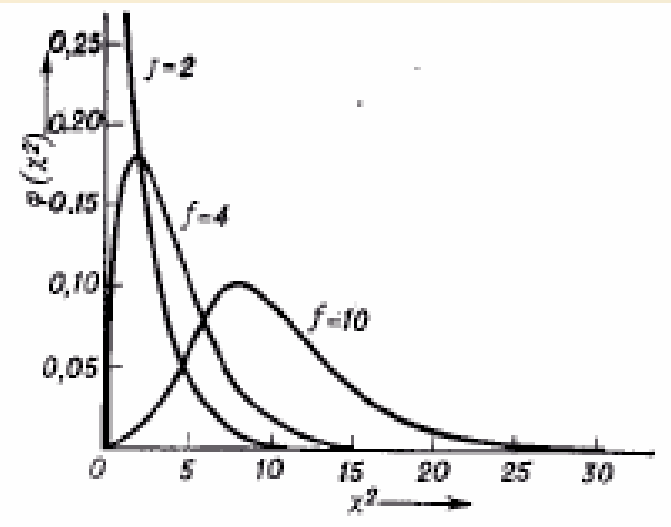
Асимметрия

$$\tilde{A} = \mu_3 / \mu_2^{3/2}$$

Эксцесс

$$\gamma_2 = \mu_3 / \mu_2^2 - 3$$

Запишите полные формулы для НСВ и ДСВ!



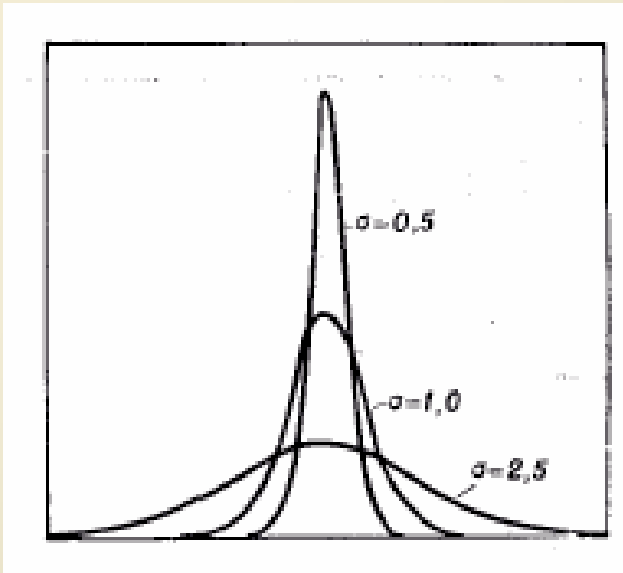


СХЕМА БЕРНУЛЛИ

Индикатор случайного события *A* **–** это дискретная случайная величина X, которая равна 1 при осуществлении события *A* и 0 при осуществлении события не-A

x_i	0	1
p_i	q	p

$$q + p = 1$$
$$m_x = p$$
$$D_x = q\,p$$

Проверьте дома!

Биномиальное распределение

X: 0, 1, …, n

$$p(X=i)=p_i=\frac{n!}{i!(n-i)!}p^i q^{n-i}$$

Откуда берется?

$$m_x = np,\ D_x = npq$$

Постройте функции для

1) n = 5, p = 0.1
2) n = 5, p = 0.3
3) n = 5, p = 0.5

Полиномиальное распределение

Подбрасывание k-гранной кости. Вероятность получить грань i при одном бросании = p_i.

Совершается n независимых бросаний.

Вероятность получить n_1 раз грань 1, n_2 – грань 2, …, n_k – грань k

$$p(X_1 = n_1, X_2 = n_2, \ldots X_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

$$\sum_{i=1}^{k} p_i = 1, \sum_{i=1}^{k} n_i = n$$

$$m_{X_i} = np_i$$

$$D_{X_i} = np_i(1 - p_i)$$

Распределение Пуассона

X: 0, 1, …, ∝

n → ∝

p → 0

Радиоактивный распад

$$\lim np = a$$

$$m_X = a, D_X = a$$

$$p(X = i) = p_i = \frac{a^i}{i!} e^{-a}$$

Задача о кексах

160 кексов, 300 изюминок в 10 кг теста.

Какова вероятность, что в одном наугад выбранном кексе изюминок не окажется?

Сравните результаты, полученные при решении задачи на основе свойств биномиального и полиномиального распределений.

Равномерное распределение

$$f(x)=\begin{cases}0, x<a,\\ \dfrac{1}{b-a}, a\le x\le b\\ 0, x>b.\end{cases}$$

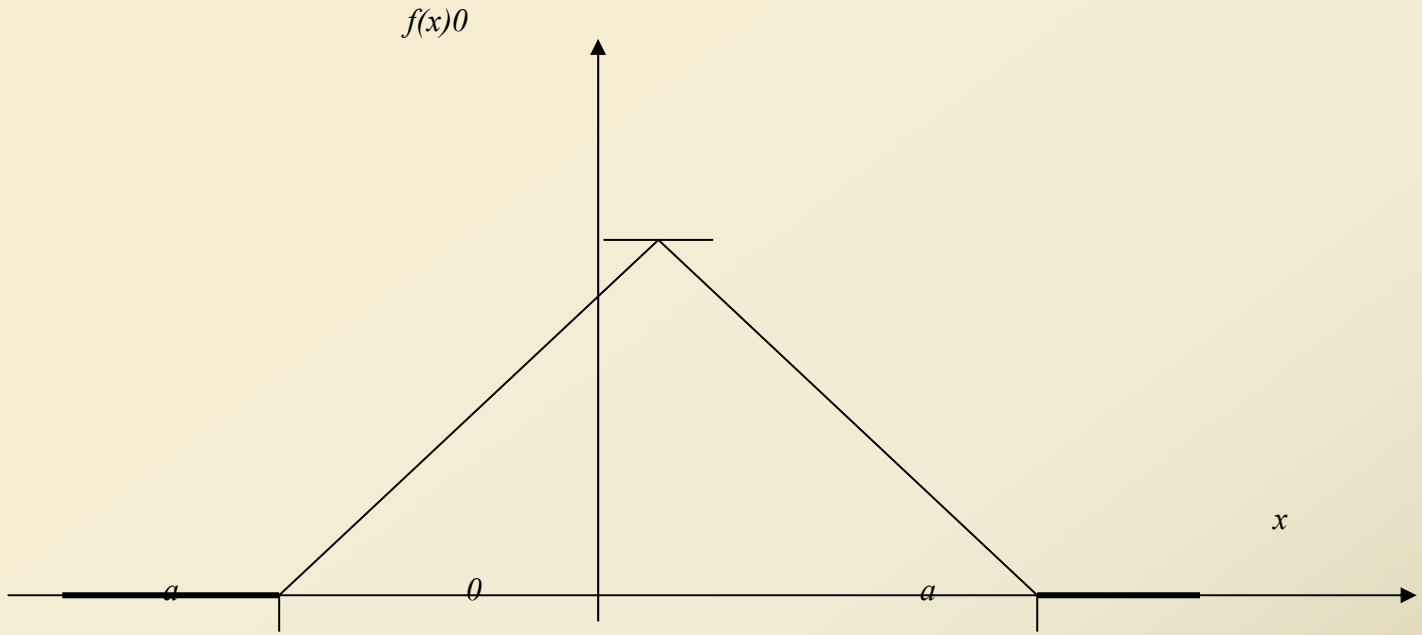
$$F(x)=\begin{cases}0, x<a,\\ \dfrac{x-a}{b-a}, a\le x\le b\\ 1, x>b.\end{cases}$$

Постройте графики плотности и функции равномерного распределения

$m_x = ?$

$$D_x = \frac{(b-a)^2}{12}$$

Случайная величина x распределена по закону Симпсона (по "закону равнобедренного треугольника")



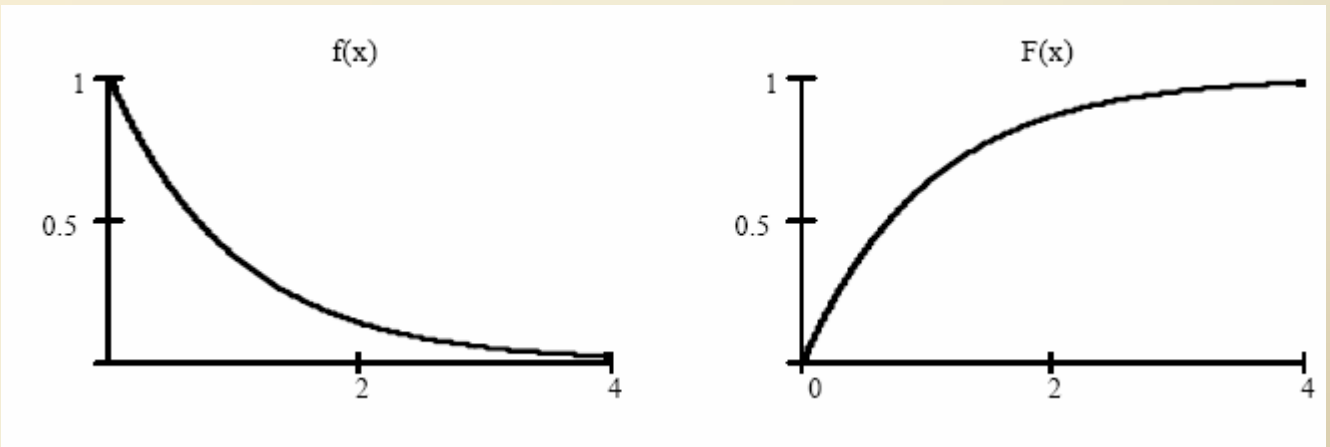
Напишите выражение плотности вероятности. Найдите функцию распределения и постройте ее график

Экспоненциальное распределение

Условия возникновения. Случайная величина $T > 0$ – интервал времени между двумя соседними событиями в пуассоновском потоке случайных событий, причем параметр распределения $\lambda > 0$ – интенсивность потока.

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

$$F(t) = \begin{cases} 1 - e^{-\lambda t}, & t \geq 0 \\ 0, & t < 0. \end{cases}$$



Нормальное распределение

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}$$

$$F(x)=0.5+\Phi\left(\frac{x-m}{\sigma}\right)$$

Функция Лапласа

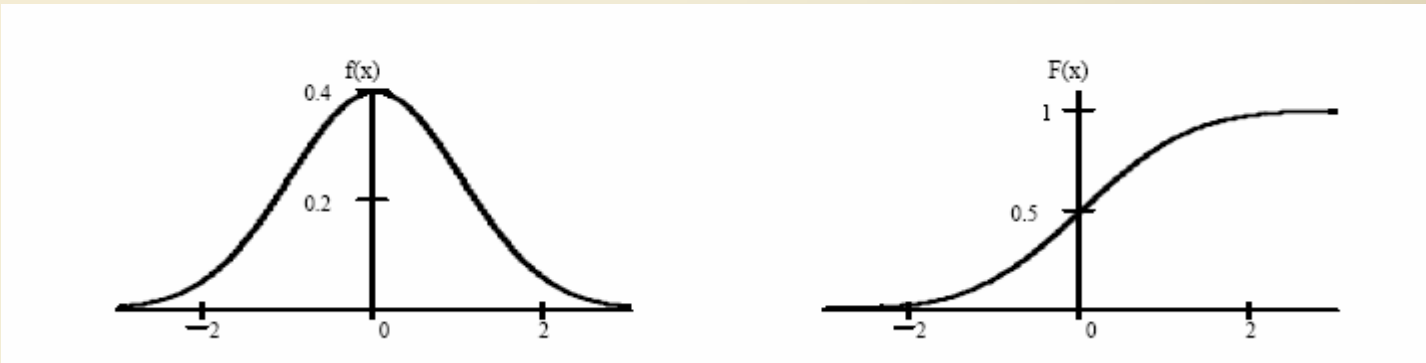
Узнаете?

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$$

$$\Phi(-x)=-\Phi(x),\ \Phi(0)=0,\ \Phi(\infty)=0{,}5$$

m = 0

σ = 1



Моменты: $m_x = m, D_x = \sigma^2$

Чему равна асимметрия?
А эксцесс равен 0!

Центральная предельная теорема Чебышева (очень грубо)**:**
Если случайная величина подвержена воздействию бесконечного числа бесконечно малых случайных факторов, то она имеет нормальное распределение.

Формулировка Линдеберга-Леви

Центральная предельная теорема (для одинаково распределенных слагаемых). Пусть $X_1, X_2, \ldots, X_n, \ldots$ – независимые одинаково распределенные случайные величины с математическими ожиданиями $M(X_i) = m$ и дисперсиями $D(X_i) = \sigma^2$, $i = 1, 2, \ldots, n, \ldots$ Тогда для любого действительного числа x существует предел

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}} < x\right) = \Phi(x),$$

где $\Phi(x)$ – функция стандартного нормального распределения.

«Каждый уверен в справедливости нормального закона:

экспериментаторы – потому, что они думают, что это математическая теорема;

математики – потому, что они думают, что это экспериментальный факт»

Приписывается Анри Пуанкаре

Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. 1991 Т.57. No.7 С.64-66.

В большинстве случаев распределения существенно отличаются от нормальных, в других нормальные распределения могут, видимо, рассматриваться как некоторая аппроксимация, но никогда нет полного совпадения.

Проф. П. В. Новицкий: распределение погрешностей электромеханических приборов, электронных приборов для измерения температур, цифровых приборов с ручным уравновешиванием.

46 из 47 распределений значимо **отличались** от нормального.

Лаборатория прикладной математики Тартуского университета: 2500 выборок из архива реальных статистических данных.

В 92% гипотезу нормальности пришлось **отвергнуть**.

Мудров, Кушко:

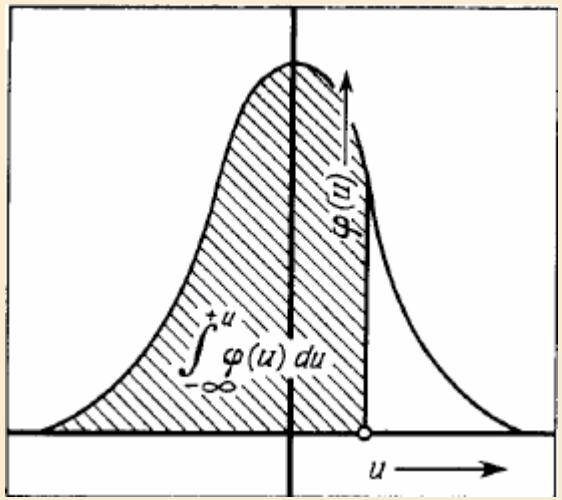
Если суммарная погрешность измерения ε формируется как сумма большого числа независимых ошибок δ_i, но дисперсии последних непостоянны, а колеблются вокруг некоторых средних значений, распределение ε **подчиняется закону Лапласа** с функцией плотности распределения

$$f(x)=\frac{1}{2\lambda}\exp\left\{-\frac{1}{\lambda}|x-m|\right\}$$

Моменты $m_X = m \quad D_X = 2\lambda^2$

$$\tilde{A}=0 \quad \gamma_2=3$$

Интеграл Гаусса, или как пользоваться таблицами, когда компьютера с Excel'ем рядом нет



Площадь F под **нормированной** кривой Гаусса

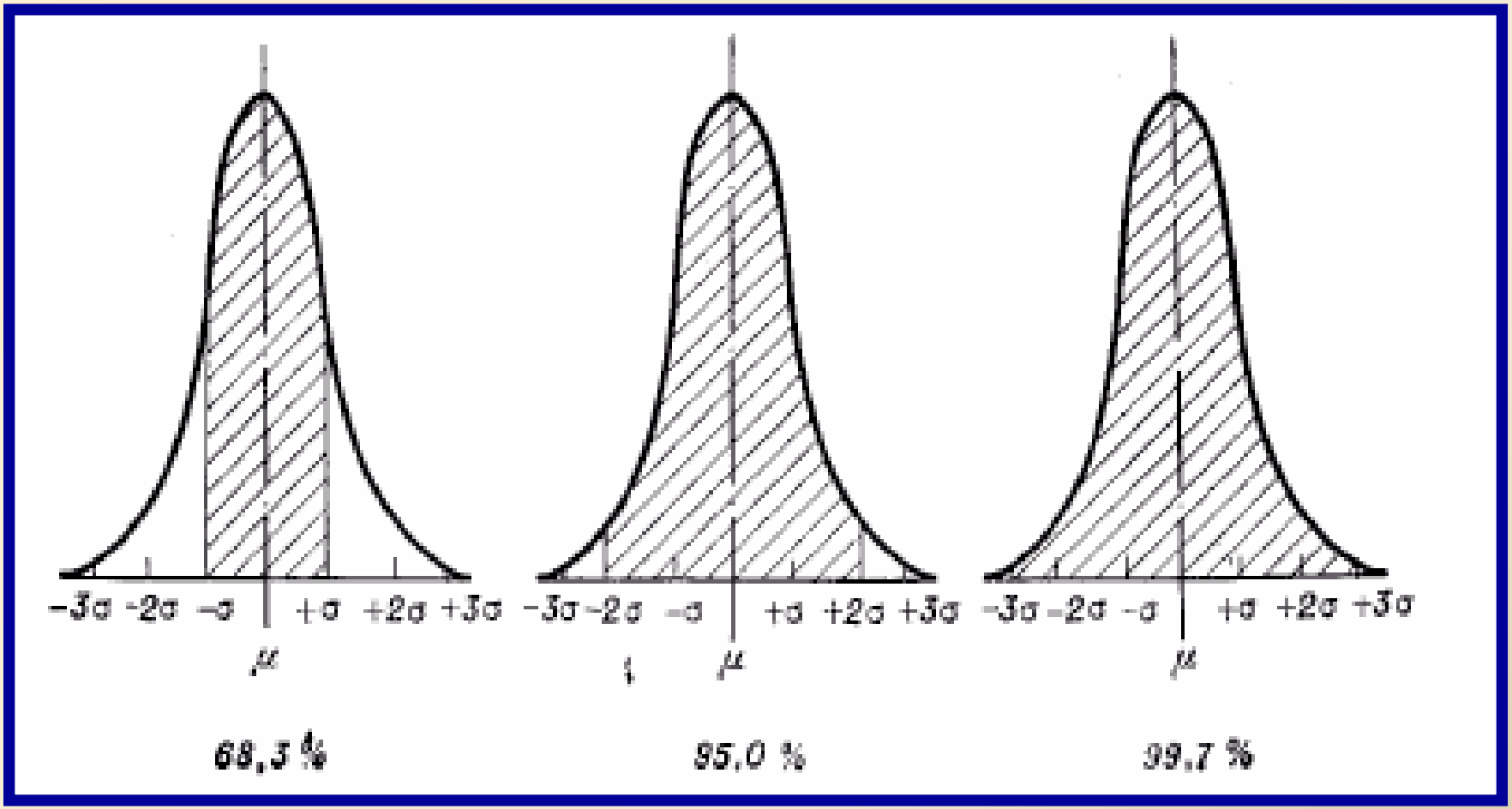
F(0.5) =

F(-0.1) = ?

F(?) = 0.3

P(-0.1 < x < 0.3) = ?

u	0,00	0,01	0,02
0,0	0,500000	0,503989	0,507978
0,1	0,539828	0,543795	0,547758
0,2	0,579260	0,583166	0,587064
0,3	0,617911	0,621720	0,625616
0,4	0,655422	0,659097	0,662757
0,5	0,691462	0,694974	0,698468
0,6	0,725747	0,729069	0,732371
0,7	0,758036	0,761148	0,764238
0,8	0,788145	0,791030	0,793892
0,9	0,815940	0,818589	0,821214
1,0	0,841345	0,843752	0,846136
1,1	0,864334	0,866500	0,868643
1,2	0,884930	0,886861	0,888768
1,3	0,903200	0,904902	0,906582
1,4	0,919243	0,920730	0,922196
1,5	0,933193	0,934478	0,935744
1,6	0,945201	0,946301	0,947384
1,7	0,955434	0,956367	0,957284
1,8	0,964070	0,964852	0,965620
1,9	0,971283	0,972933	0,971571

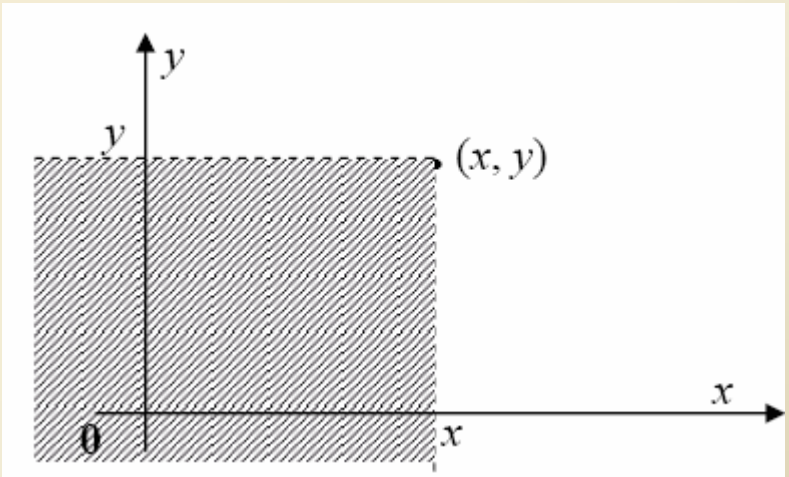

−3σ −2σ −σ +σ +2σ +3σ
μ
68,3 %
−3σ −2σ −σ +σ +2σ +3σ
μ
95,0 %
−3σ −2σ −σ +σ +2σ +3σ
μ
99,7 %

Многомерные распределения (на примере нормального)

Двухмерная случайная величина (*X*, *Y*) – совокупность двух одномерных случайных величин, которые принимают значения в результате проведения одного и того же опыта.

Двухмерная функция распределения двухмерной случайной величины

$$F(x,y) = p\left(\{X < x\} \cdot \{Y < y\}\right)$$



Двухмерная плотность распределения

$$f(x,y)=\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\frac{p(\{x\le X<x+\Delta x\}\,\mathrm{I}\;\{y\le Y<y+\Delta y\}}{\Delta x\Delta y}=\frac{\partial^2 F(x,y)}{\partial x\partial y}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy=1$$

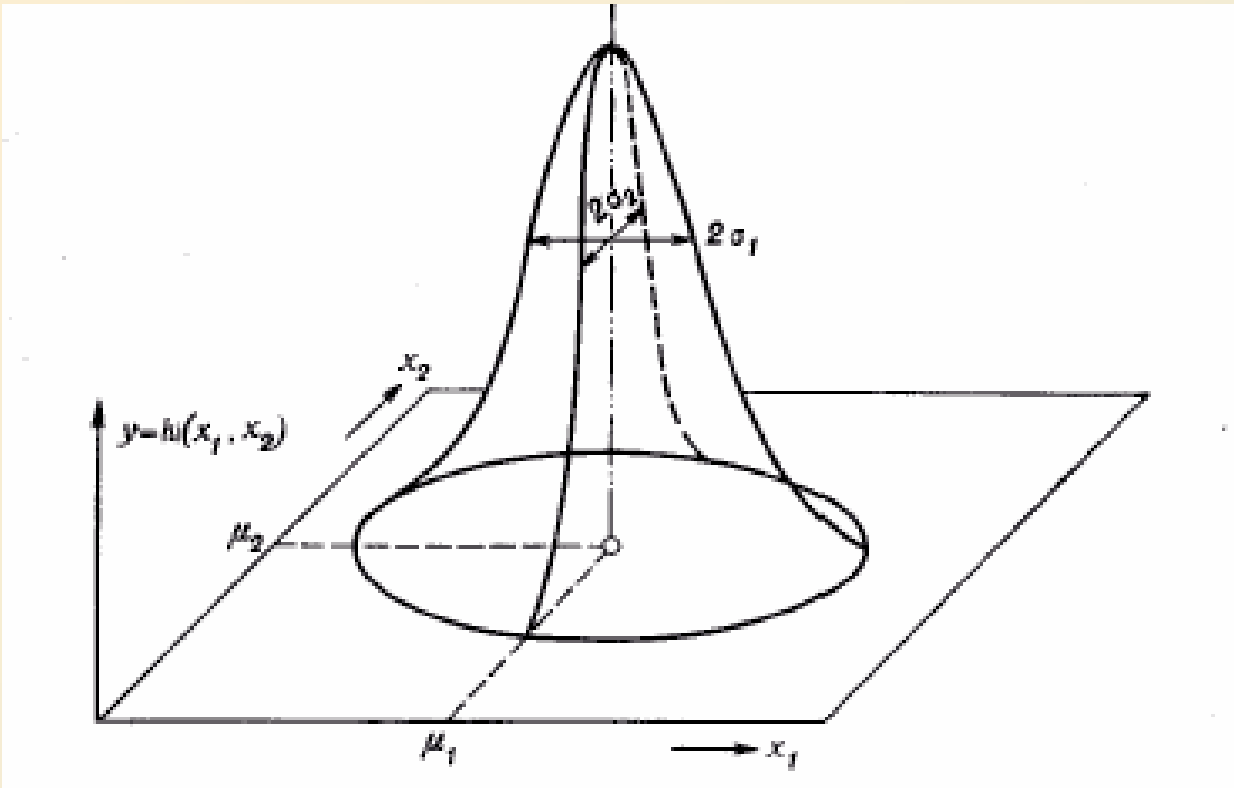
$$f(x,y)\ge 0$$

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)\,dx\,dy$$

$$f_X(x)=\int_{-\infty}^{\infty} f(x,y)\,dy$$

$$f_Y(y)=\int_{-\infty}^{\infty} f(x,y)\,dx$$

Плотность двумерного стандартного нормального распределение



Вместо дисперсии – ковариационная матрица!

Вспомним о ковариации и коэффициенте корреляции!

Смешанный центральный момент порядка *k* + *s*

$$\mu_{k,s}(x,y)=\mathrm{M}[(X-m_X)^k(Y-m_Y)^s]=\begin{cases}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i-m_x)^k(y_j-m_y)^s p_{i,j} & \text{для ДСВ,}\\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-m_x)^k(y-m_y)^s f(x,y)dxdy & \text{для НСВ,}\end{cases}$$

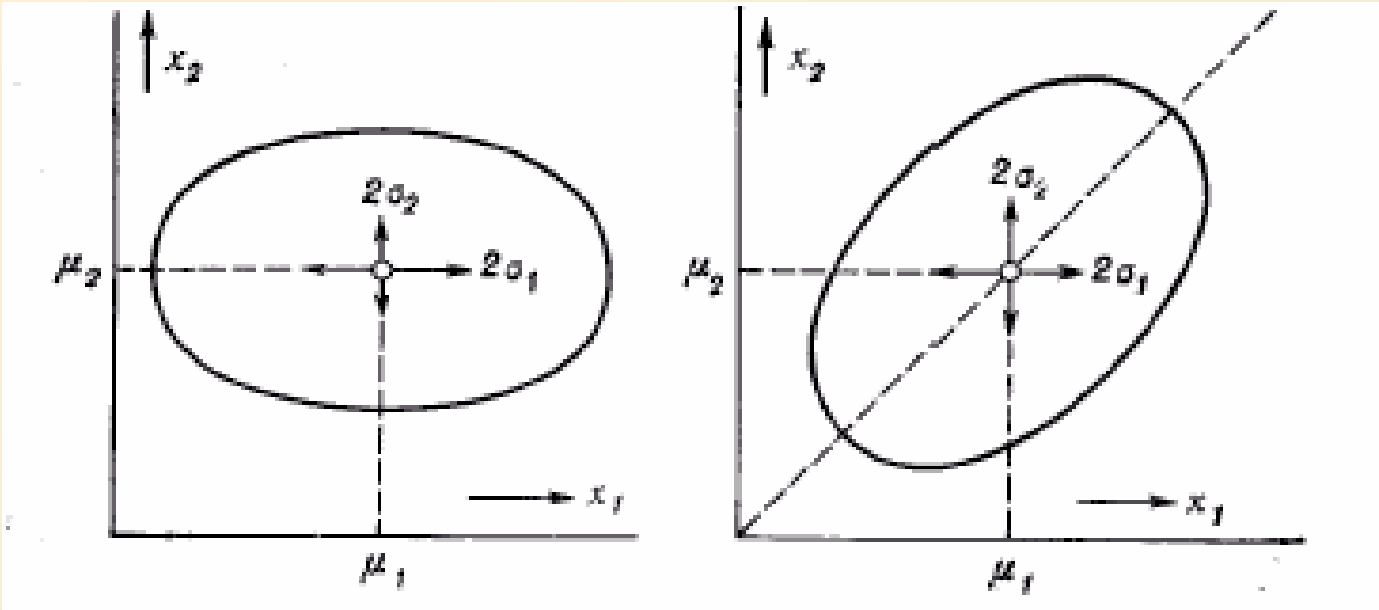
Ковариация случайных величин X, Y:

$$K_{XY}=\mu_{1,1}(x,y)=\alpha_{1,1}(x,y)-m_X m_Y$$

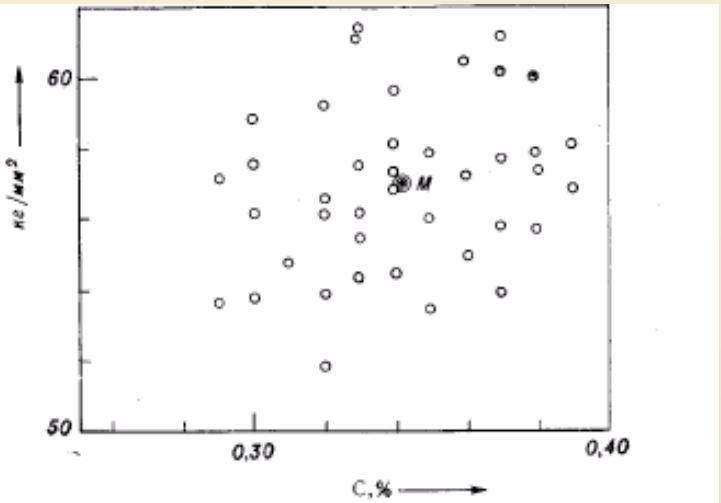
$K_{XY} = K_{YX}$ для независимых X, Y $K_{XY} = 0$

Коэффициент корреляции XY

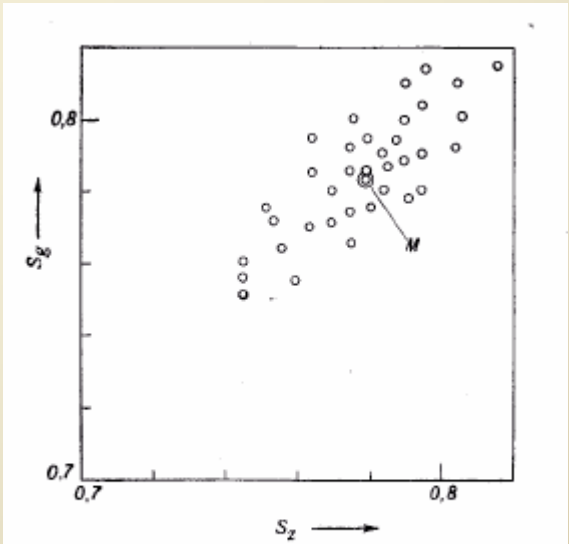
$$R_{XY}=\frac{K_{XY}}{\sqrt{D_X D_Y}}=\frac{K_{XY}}{\sigma_X\sigma_Y}$$



R = 0 **R > 0**



Р и с. 2.8. Двумерное распределение частот, характеризующее качество стали марки GS 50.



Р и с. 2.9. Двумерное распределение частот почернений, получаемых при количественном эмиссионно-спектральном анализе,

$$L(\theta \mid X^{(n)}) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

$$\mathcal{L}(\theta \mid X^{(n)}) = \sum_{i=1}^{n} \ln f(X_i \mid \theta)$$

$$\frac{\partial \mathcal{L}(\theta \mid X^{(n)})}{\partial \theta_i} = 0, \; i = 1, \ldots, k$$

$$\mathcal{L}(\theta \mid X^{(n)}) = \ln f_n(X^{(n)} \mid \theta) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_1^n (X_k - \mu)^2.$$

$$f_n(x^{(n)} \mid \theta) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_1^n (x_k - \mu)^2 \right\}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{2\sigma^2} \sum_1^n (X_k - \mu) = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_1^n (X_k - \mu)^2 = 0.$$