Лекция 3.Основы математической статистики.

Лектор: Войтик В.В.

План лекции:

1. **Задачи математической статистики.**
2. **Генеральная и выборочная совокупности**
3. **Основные этапы исследования**
4. **Дискретные и интервальные ряды распределения. Числовые характеристики.**
5. **Точечные и интервальные оценки**
6. **Закономерности нормального распределения. Кривая нормального распределения и ее характеристики**
7. **Сравнение теоретических и эмпирических распределений**

Что такое математическая статистика?

Математическая статистика – это наука извлечения полезной информации из данных, полученных в результате наблюдений или экспериментов

Основные понятия математической статистики

- **Наиболее общую совокупность, подлежащих изучению объектов называют** генеральной.
- **Выборка считается** репрезентативной**, если каждый объект выборки отобран случайно из генеральной совокупности, то есть все объекты имеют одинаковую вероятность попасть в выборку.**

Основные понятия математической статистики

Объемом выборки называют число объектов этой совокупности. Таким образом, вместо большой совокупности объектов изучается совокупность объёма, значительно меньшего по количеству объектов ($n << N$).

Основные понятия математической статистики

Результаты, полученные при изучении выборки, распространяются на объекты всей генеральной совокупности. Для этого выборка должна быть репрезентативной (представительной), то есть правильно представлять генеральную совокупность. Это обеспечивается случайностью отбора.

Какие задачи нас интересуют?

- определение закона распределения случайной величины по выборочным данным;

- задача проверки правдоподобия гипотез (отличия характеристик выборки от некоторых неслучайных величин; отличия характеристик нескольких выборок; связь случайных величин из разных выборок);

- Задача нахождения неизвестных параметров распределения.

Основные этапы исследования:

- Сгруппировать исследуемый ряд по классам. Подсчитать середины интервалов и частоты попадания в интервал.
- Построить гистограмму и полигон распределения.
- Найти эмпирическую функцию распределения и построить ее график.
- Вычислить числовые (точечные) характеристики распределения.
- Проверить гипотезу о том, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону, используя критерии асимметрии и эксцесса.
- Проверить гипотезу о том, что генеральная совокупность, из которой извлечена выборка, распределена по нормальному закону, используя критерий Пирсона χ^2

Статистическое распределение выборки и его характеристики

Пусть из генеральной совокупности извлечена выборка, причем x1 наблюдалось n1 раз, x2 – n2 раз, xk – nk раз и n – объем выборки. Наблюдаемые значения xi называют вариантами, а последовательность вариант, записанных в возрастающем порядке, – вариационным рядом. Числа наблюдений называются частотами, а их отношения к объему выборки

Wi=ni/ n – относительными частотами.

Статистическим распределением выборки называют перечень вариант в порядке возрастания соответствующих им частот или относительных частот

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X<x$:

$$F^*(x) = \frac{n_x}{n}$$

где n_x – число вариант, меньших x; n – объем выборки.

Интервальная оценка (доверительный интервал) для генеральной средней

Интервальной называют оценку, которая определяется двумя числами– концами интервала.

Доверительным интервалом для параметра **Θ** называется интервал (Θ_1, Θ_2), содержащий истинное значение **Θ** с заданной вероятностью $P(\Theta_1 < \Theta < \Theta_2) = 1-\alpha$.

$\gamma = 1 - \alpha$ называется **доверительной вероятностью** (**надежностью**), а значение α – уровнем значимости.

Статистическая функция распределения случайной величины X

$$F^*(x) = P^*(X < x)$$

Рассмотрим эксперимент, который поможет понять смысл этой функции:

Дана некоторая группа людей, мы измеряем их рост и пытаемся определить закономерности распределения людей по росту.

Пример:Ряд распределения студентов по росту

148 158 149 162 170 156 186 151 161
152 171 165 174 157 172 172 177
166 157 149 159 154 164 167 173
176 147 163 185 164 161 153 168
162 184 162 169 154 167 163 166
172 158 155 165 179 165 160 159
169

Размах распределения

- Из имеющихся значений признака X выбирают наименьшее (Xmin), наибольшее (Xmax), определяют размах распределения

$$(X_{max} - X_{min})$$

186-147=39

Статистический ряд распределения

X	X_1	X_2	…	X_n
m	m_1	m_2	…	m_n
m/n	m_1/n	m_2/n	…	m_n/n

Статистический ряд распределения студентов по росту

X	140-150	150-160	160-170	170-180	180-190
m	4	14	20	9	3
m/n	4/50= 0,08	14/50= 0,28	20/50= 0,4	9/50= 0,18	3/50= 0,06
$f(x)=\frac{m}{n\cdot\Delta x}$	0,08/10 = 0,008	0,028	0,04	0,018	0,006

Гистограмма распределения студентов по росту (m, m/n, f(x))

Функция распределения вероятностей

X	<140	<150	<160	<170	<180	>180
m	0	4	18	38	47	50
m/n	0	4/50 0,08	18/50 0,36	38/50 0,76	47/50 0,94	50/50 1

График F(x)

Точечные характеристики случайной величины :выборочное среднее, дисперсия и СКО

$$\overline{X} = \frac{m_1 x_1 + m_2 x_2 + \ldots + m_n x_n}{n} = \frac{\sum_{i=1}^{n} x_i m_i}{n}$$

$$D(x) = \frac{\sum_{i=1}^{n} \left[x_i - \overline{X}\right]^2 \cdot m_i}{n}$$

$$\sigma(x) = \sqrt{D(x)}$$

Непараметрические характеристики: мода и медиана

- Me-медиана

Варианта, которая делит ряд пополам

158, 164, 172, 175, 175, 179, 186

при n- нечетном

Me=175

158, 164, 168, 172, 174, 175, 179, 186

$$Me = \frac{172 + 174}{2} = 173$$

при n- четном

Непараметрические характеристики: мода и медиана

- Мо-наиболее часто встречающаяся варианта

158, 164, 172, 175, 175, 175, 179, 186

Мо=175

158, 164, 173, 173, 175, 175, 179, 186

$$Mo = \frac{173 + 175}{2} = 174$$

бимодальные выборки- если два несмежных значения имеют одинаковые частоты

Интервальной оценкой (с надежностью γ) математического ожидания *a* нормально распределенного количественного признака *X* по выборочной средней $\bar{x}$ при известном среднем квадратическом отклонении σ служит доверительный интервал, т.е.

$\bar{x}-\frac{t\sigma}{\sqrt{n}}<a<\bar{x}+\frac{t\sigma}{\sqrt{n}}$, где *n* – объем выборки; *t* – значение аргумента функции Лапласа Ф(*t*), при котором Ф(t) =$\gamma/2$.

Доверительные вероятности и доверительные интервалы

- **Вероятности 0,95 и 0,99 (95% и 99%) – доверительные вероятности**
- **$\Delta x = \pm \sigma t$ – доверительный интервал** Доверительным **называется интервал, в который попадает случайная величина с заданной вероятностью**

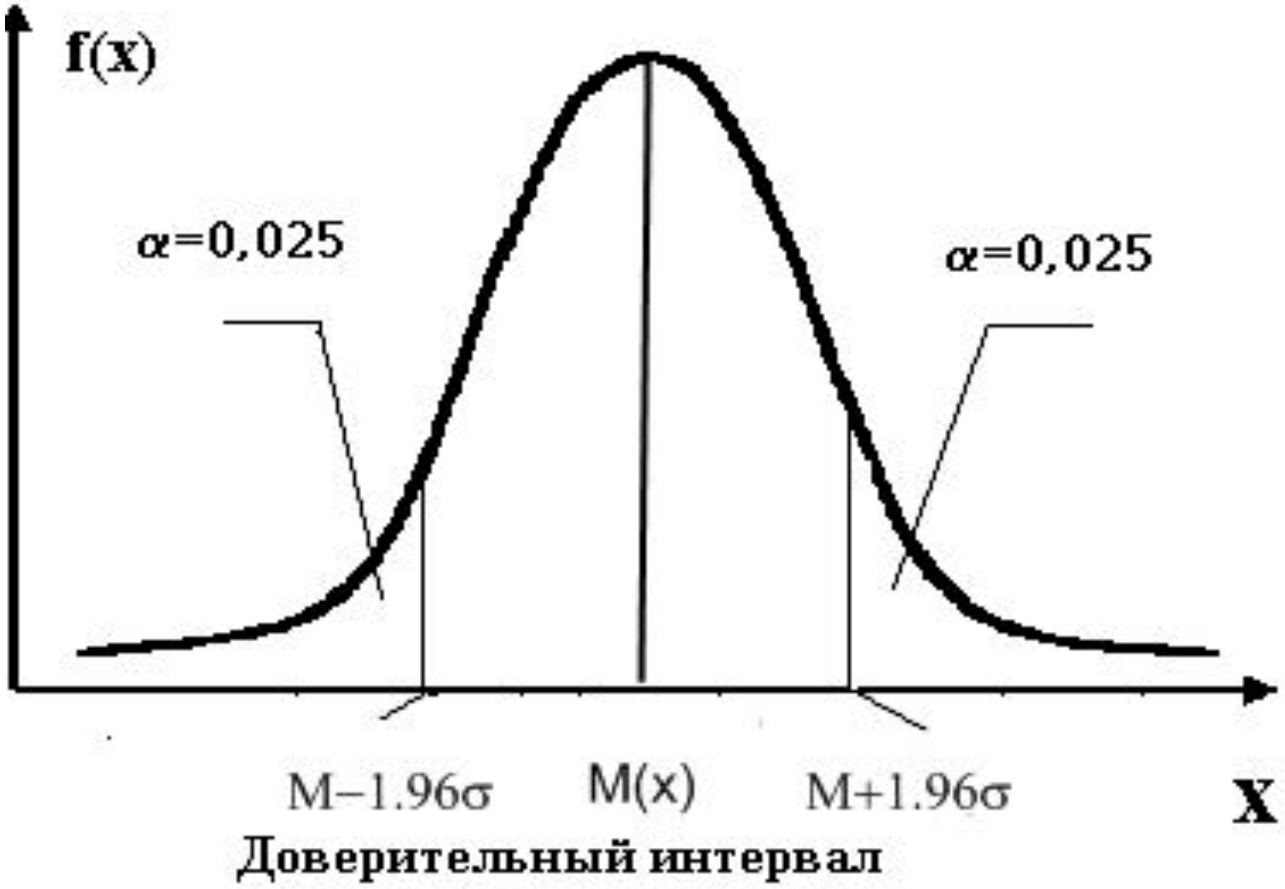
Вероятности	Интервалы
0,95	±1,96σ
0,99	±2,58σ
0,999	±3,03σ

Уровни значимости

- Определенным значениям доверительных вероятностей соответствуют так называемые уровни значимости (α).
- Уровень значимости обозначает вероятность выхода случайной величины за пределы доверительного интервала. Если доверительную вероятность обозначить – P, а уровень значимости – α, то α=1 – P.

Доверительные вероятности	Уровни значимости
0,95	0,05
0,99	0,01
0,999	0,001

95% доверительный интервал



Задача:

- **Найти доверительный интервал для роста студентов с вероятностью $p=0,95$ ($\alpha=0,05$); $M(x)=170$ см, $\sigma=5$ см**

$$\Delta x=1,96\cdot 5\approx 10 \text{ см}$$

Следовательно, рост студентов находится в интервале: $170-10<x<170+10$

$$160 \text{ см}<x<180 \text{ см}$$

Нормальный закон распределения случайных величин

Нормальное распределение **возникает тогда, когда на изменение случайной величины действует множество различных независимых факторов, каждый из которых в отдельности не имеет преобладающего значения.**

Главная особенность - **это предельный закон, к которому при определенных условиях стремятся другие законы**

Говорят, что X имеет нормальное (гауссовское) распределение с параметрами μ и σ, где $\mu \in R$, $\sigma>0$, если X имеет следующую плотность распределения:

$$f(x)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

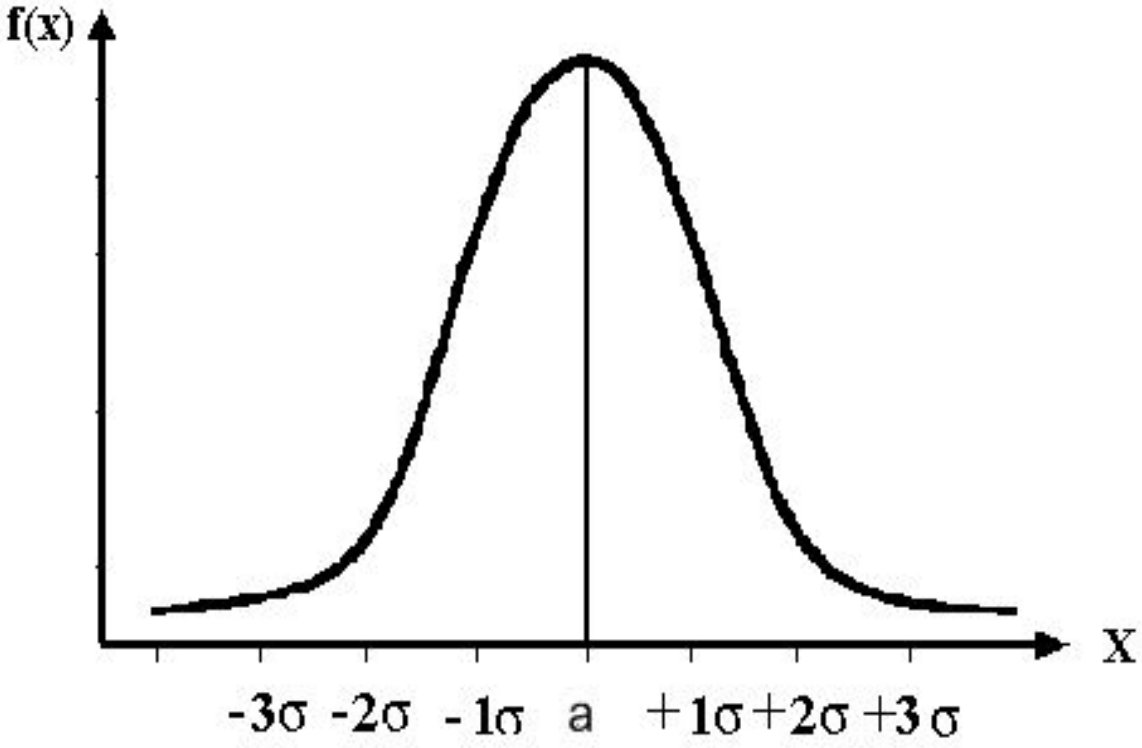
дифференциальная функция распределения

Функция распределения вероятностей

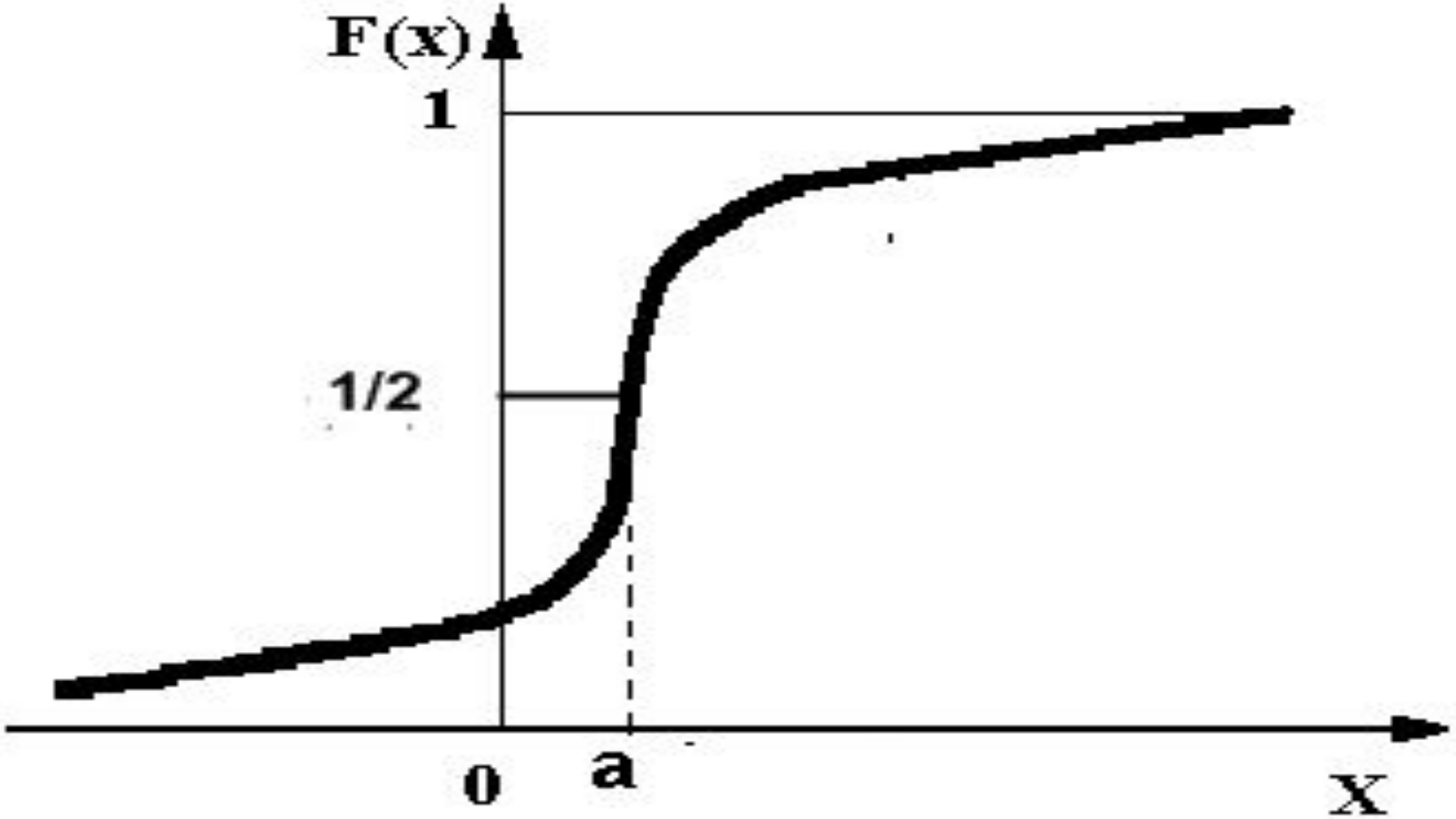
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

интегральная функция распределения

Кривая нормального распределения (Гаусса)



Функция распределения вероятностей



ЗАКОНОМЕРНОСТИ РАСПРЕДЕЛЕНИЯ:

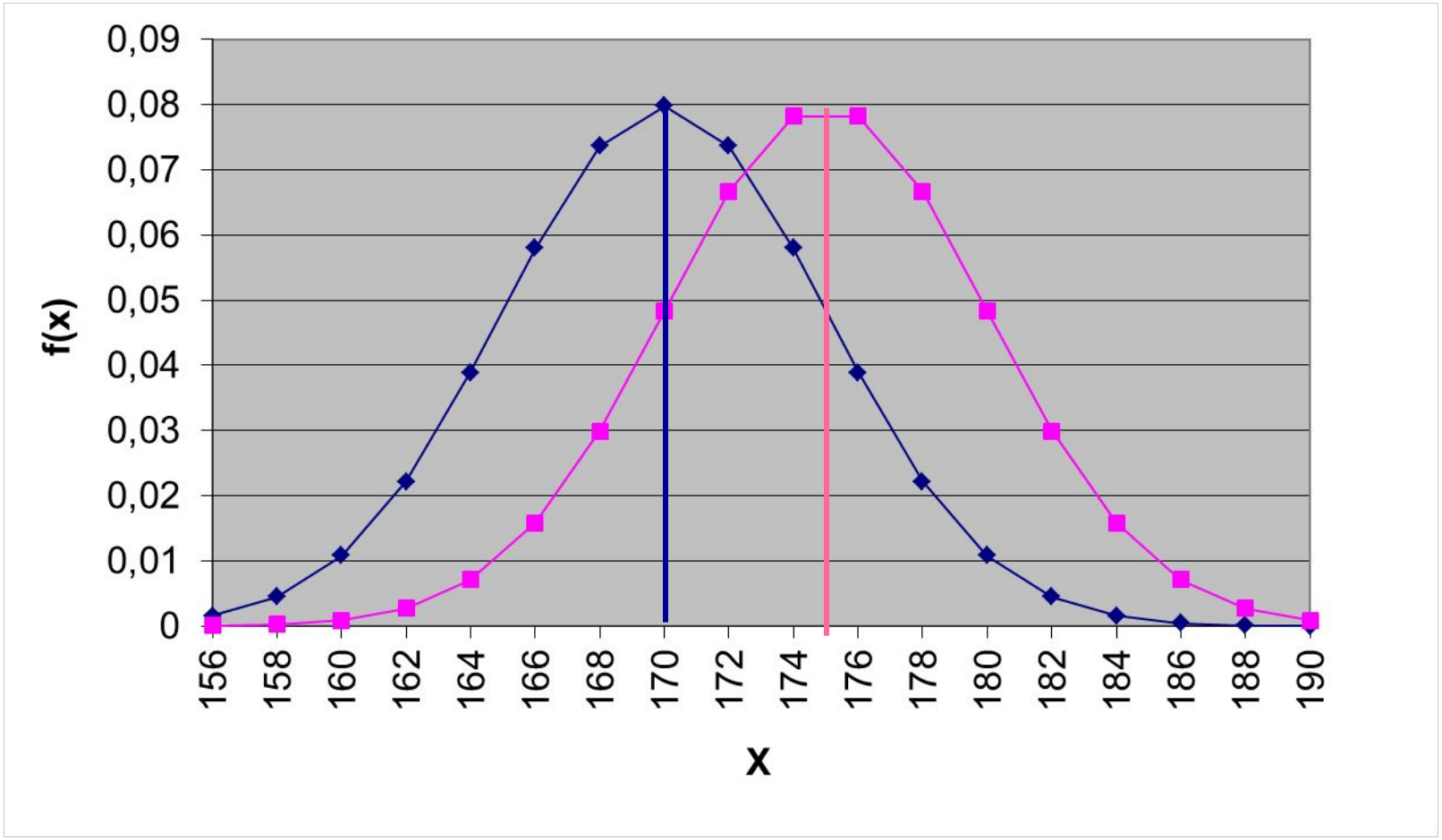
- Параметр μ характеризует математическое ожидание (среднее арифметическое) случайной величины, являясь центром распределения и наиболее вероятным значением. Изменение математического ожидания не влияет на форму кривой, а только вызывает ее смещение вдоль оси x.

Пример:

Рост в группе П101-M(x)=170 см, σ=5 см

П102-M(x)=175 см, σ=5 см

Пример:


0,09
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0
f(x)
156
158
160
162
164
166
168
170
172
174
176
178
180
182
184
186
188
190
x

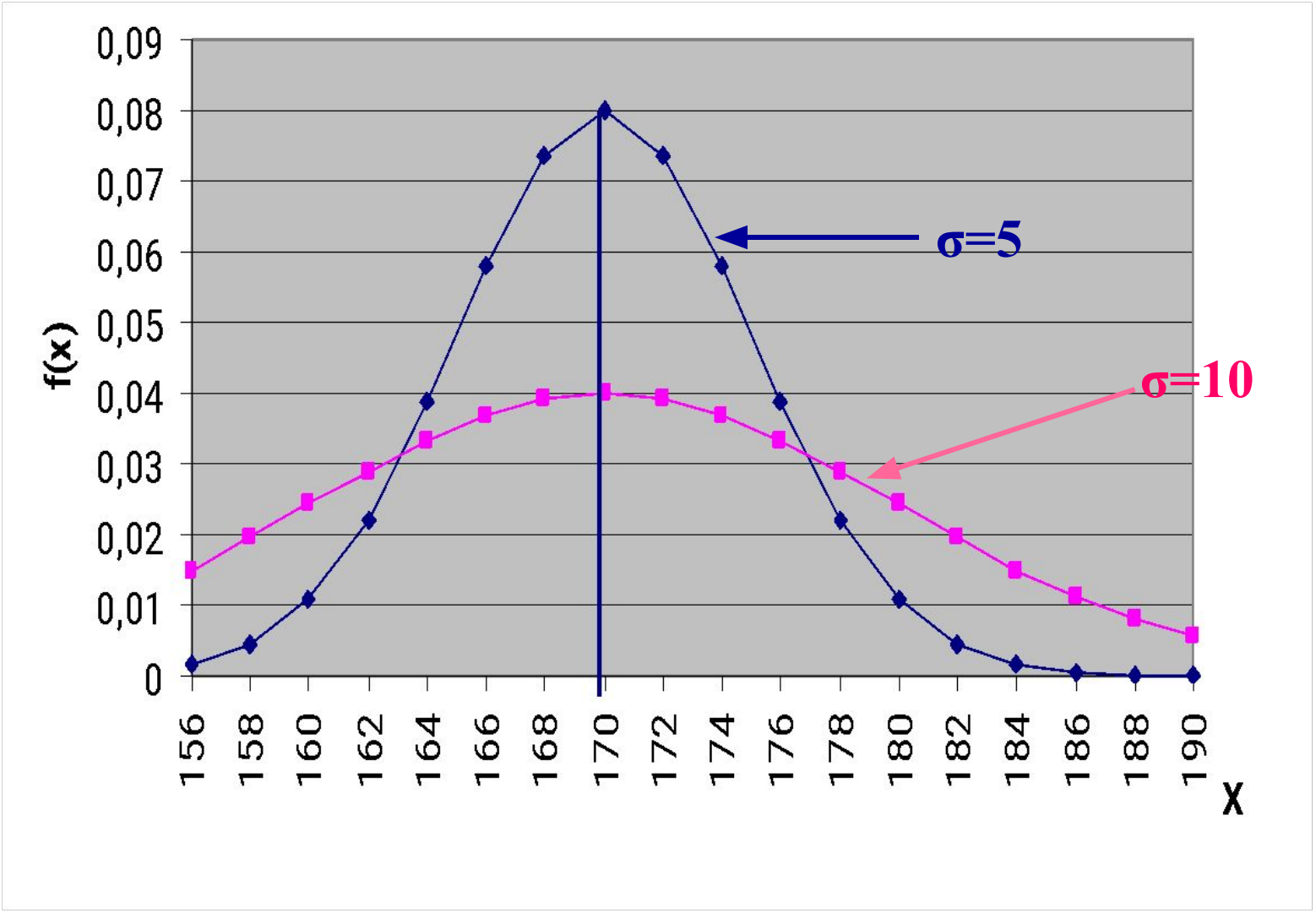
ЗАКОНОМЕРНОСТИ РАСПРЕДЕЛЕНИЯ:

- Параметр σ характеризует изменчивость случайной величины (меру растянутости кривой вдоль оси x): чем больше σ, тем больше кривая растянута.

Пример:
Рост в группе Л101-M(x)=170 см, σ=5 см

Л132-M(x)=170 см, σ=10 см

Пример:


0,09
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0
f(x)
156
158
160
162
164
166
168
170
172
174
176
178
180
182
184
186
188
190
X
σ=5
σ=10

ЗАКОНОМЕРНОСТИ РАСПРЕДЕЛЕНИЯ:

- График нормальной кривой симметричен относительно прямой $x=\mu$ (одинаковые по абсолютной величине отрицательные и положительные отклонения случайной величины от центра равновероятны).

По мере увеличения разности $(x-\mu)$ значение $f(x)$ убывает. Это значит, что большие отклонения менее вероятны, чем малые.

При $(x-\mu) \to \infty$ значение $f(x)$ стремится к нулю, но никогда его не достигает.

ЗАКОНОМЕРНОСТИ РАСПРЕДЕЛЕНИЯ:

- **По мере увеличения разности** $(x-\mu)$ **значение** $f(x)$ **убывает. Это значит, что большие отклонения менее вероятны, чем малые. При** $(x-\mu)$ **значение** $f(x)$ **стремится к нулю, но никогда его не достиг**

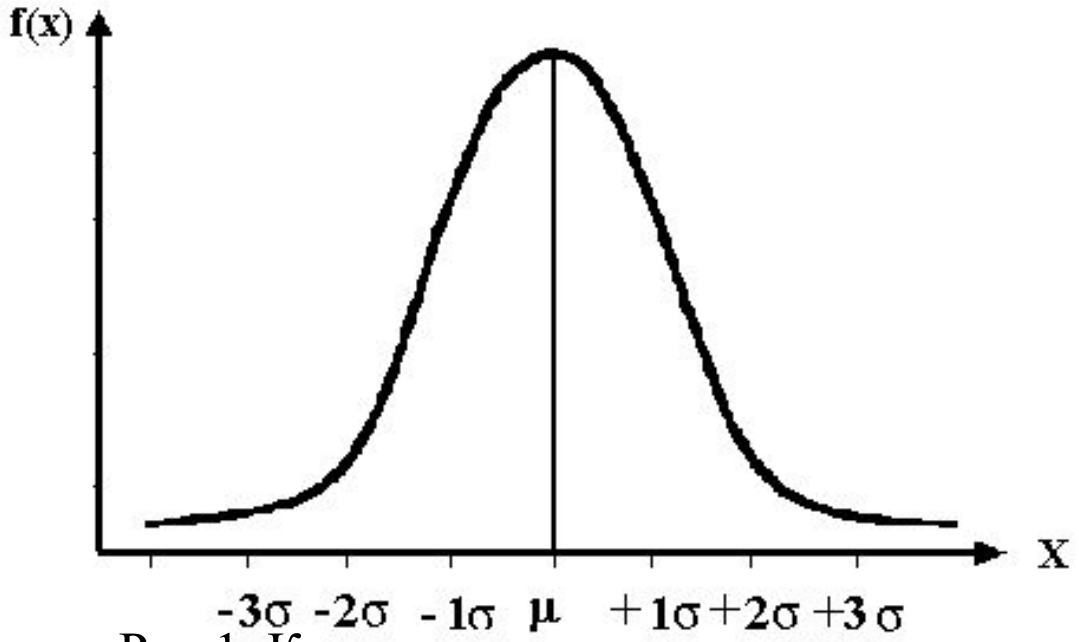


Рис.1. Кривая нормального распределения

Функция нормального закона

$$f(x)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

функция плотности распределения вероятностей

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

функция распределения вероятностей

$$t=\frac{x-\mu}{\sigma}\Rightarrow \qquad F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

Вероятность попадания значения случайной величины в интервал от a до b:

$$P(a\langle x\langle b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

причем $\Phi(-t) = 1 - \Phi(t)$

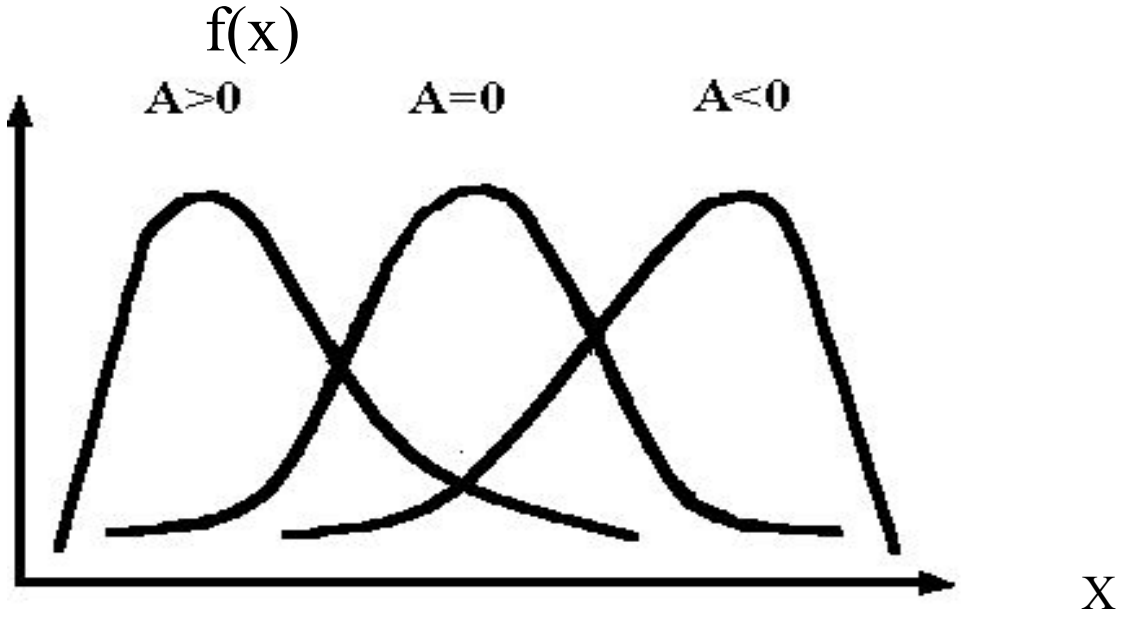
Характеристики кривой:

- Коэффициент асимметрии
- Показатель эксцесса

КОЭФФИЦИЕНТ **АСИММЕТРИИ**

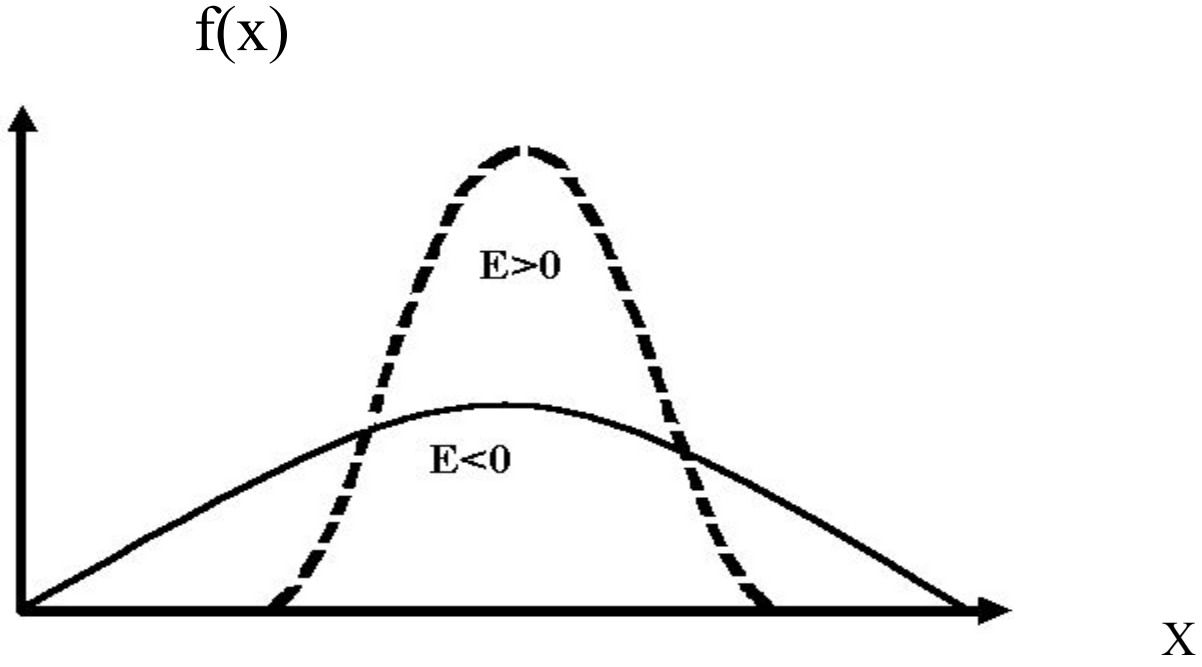
$$A = \frac{M(x - M(x))^3}{\sigma^3}$$

A>0 - правоасимметричные,
A<0 - левоасимметричные



ПОКАЗАТЕЛЬ **ЭКСЦЕССА**

$$E = \frac{M(x - M(x))^4}{\sigma^4} - 3$$



Для нормального распределения показатели А=0 и Е=0

Задача:

- **Записать функции нормального закона для распределения студентов по росту:**

M(X)=170 см; σ=5 см

$$f(x)=\frac{1}{5\sqrt{2\pi}}e^{-\frac{(x-170)^2}{2\cdot 5^2}}$$

$$F(x)=\frac{1}{5\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(x-170)^2}{2\cdot 5^2}}dx$$

Нормальное распределение с параметрами M(x)=0 и σ=1 называется стандартным $N_{0,1}$ (нормированным нормальным распределением)

Функция плотности распределения вероятностей

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

Функция распределения вероятностей

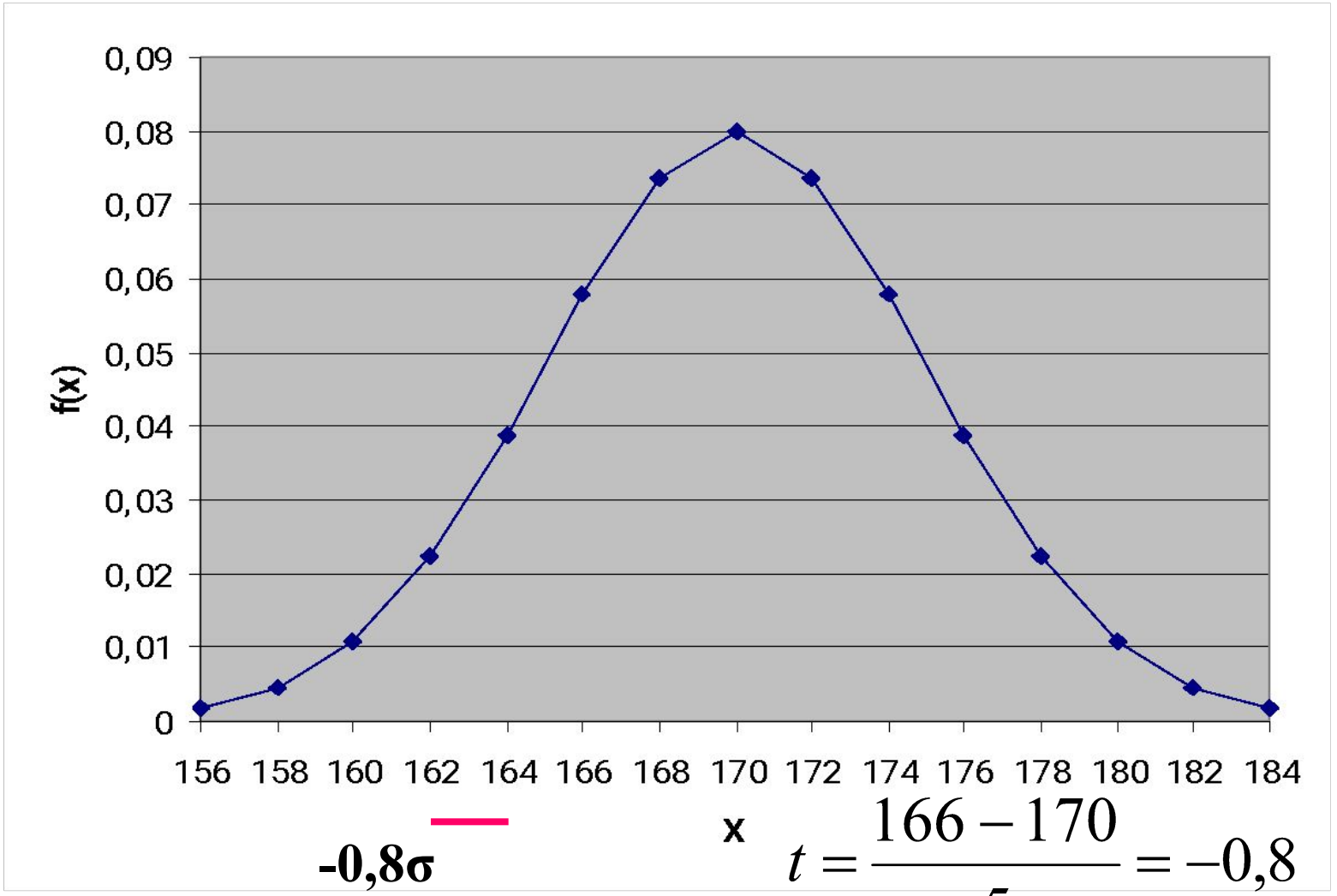
$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Нормированное отклонение:

Нормированным отклонением называется отклонение случайной величины x,от её математического ожидания, выраженное в единицах σ

$$t = \frac{x - M(x)}{\sigma}$$

Найти нормированное отклонение для x=166 см, если M(x)=170 см, σ=5 см.



-0,8σ

$$t = \frac{166 - 170}{5} = -0{,}8$$

Вероятность попадания значения случайной величины в интервал от -∞ до x:

$$t = \frac{x - \mu}{\sigma} \quad \Rightarrow \quad F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Функция F(x) не выражается через элементарные функции, но для нее составлены таблицы, которые называются таблицами нормального интеграла вероятности

Вероятность попадания значения случайной величины в интервал от a до b:

$$P(a < x < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$= \Phi(t_2) - \Phi(t_1)$$

причем $\Phi(-t) = 1 - \Phi(t)$

Задача:

- **Найти вероятность попадания случайной величины в интервал от 155 см до 160 см если M(x)=a=170 см, σ=5 см.**

$$P(155 < x < 160) = \Phi\left(\frac{160-170}{5}\right) - \Phi\left(\frac{155-170}{5}\right) =$$

$$\Phi(-2)-\Phi(-3)=(1-\Phi(2))-(1-\Phi(3))=(1-0,9772)-(1-0,9986)=$$

$$0,0228-0,0014=0,0214 \quad (2,14\%)$$

Интервальные оценки

$$t = \frac{x - \mu}{\sigma}$$ ← нормированное отклонение

x –

μ=σt

± 1σ – 68,3%;

± 2σ – 95,5%;

± 3σ – 99,7% всех вариант

Закон 3σ: в пределах 3σ находится 99,7% всех вариант

Сравнительная характеристика

Характеристики	Совокупность	
	Генеральная	Выборочная
Математическое ожидание	μ	$\overline{x}$
Среднее квадратическое отклонение	σ	s
Средняя квадратическая ошибка (стандартная ошибка)	$s_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$	$s_{\overline{x}} = \frac{s}{\sqrt{n}}$

$$\mu = \overline{X} \pm t s_{\overline{x}}$$ ← значение генеральной средней с доверительным интервалом

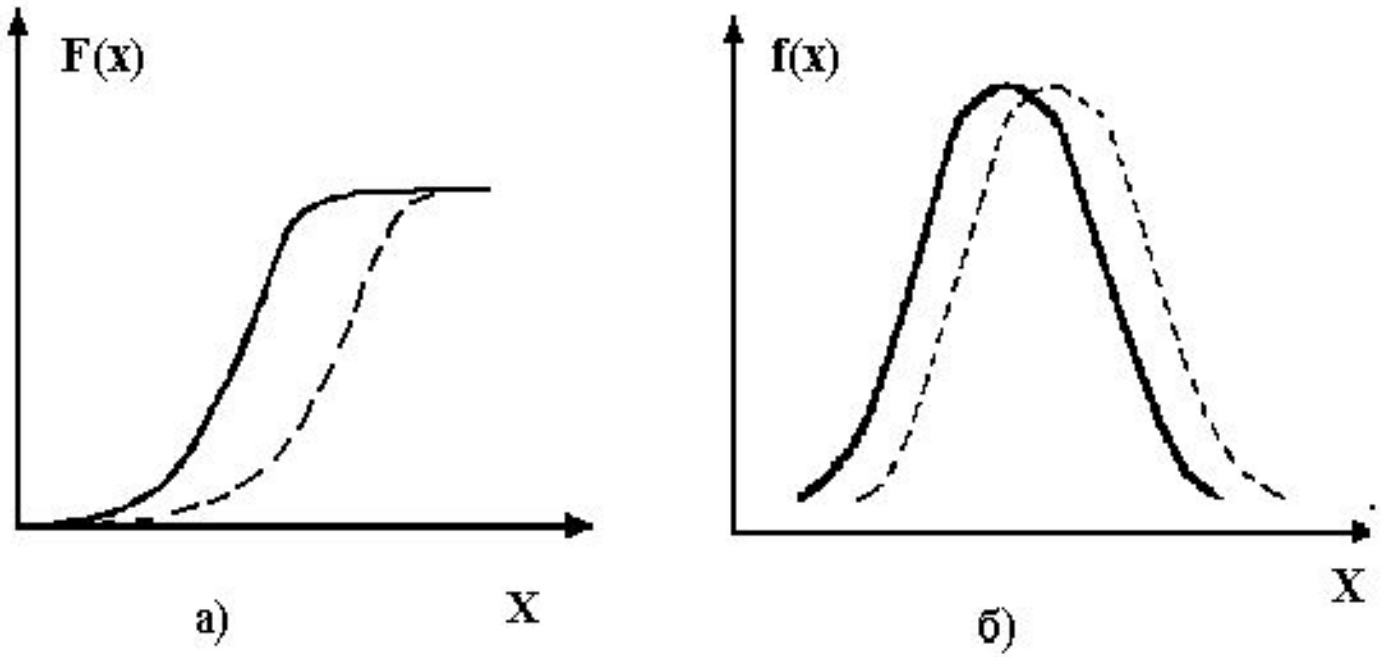
Сравнение теоретических и эмпирических распределений

- Нулевая гипотеза. **Согласно этой гипотезе первоначально принимается, что между эмпирическим и теоретическим распределением признака в генеральной совокупности достоверного различия нет.**

Средние квадратические ошибки s_A (асимметрии) и s_E (эксцесса)

$$s_A = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}} \qquad s_E = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}$$

Для достаточно большой выборки (n>30), если показатели асимметрии (A) и эксцесса (E) в два и более раза превышают показатели их средних квадратических ошибок, гипотезу о нормальности распределения нужно отвергнуть.



Сравнение теоретических и экспериментальных распределений по:

а) критерию Колмогорова – Смирнова,

б) критерию Пирсона.

Пунктирная линия – эмпирическое распределение, сплошная – теоретическое распределение.

Критерий Пирсона

$$\chi^2_{\text{эмп.}} = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i}$$

где m_i – экспериментальные частоты попадания значения случайной величины в интервал,

np_i **– теоретические частоты.**

- Число степеней свободы **– это общее число величин, по которым вычисляются соответствующие статистические показатели, минус число тех условий, которые связывают эти величины, то есть уменьшают возможности вариации между ними. Число степеней свободы определяется по следующей** формуле:

  $df=k-r-1$, **где k – число интервалов, r – число параметров предполагаемого распределения. Для нашего случая** $r=2$**, следовательно,** $df=k-3$.
- **По заданному уровню значимости** (α) **и числу степеней свободы** df, **находим критическое значение** $\chi^2_{кр}(\alpha,df)$.

- Если $\chi^2_{эмп} < \chi^2_{кр}$ **гипотеза о согласии эмпирического и теоретического распределения** подтверждается.

Заключение

Нами рассмотрены:

- Основные параметры нормального распределения;
- Понятие доверительной вероятности и доверительного интервала;
- Нулевая гипотеза и ее применение для сравнения теоретического и практического распределений.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА:

- **Основная литература:**
- Павлушков И.В. Основы высшей математики и математической статистики. М., ГЭОТАР-Медиа, 2005, с.251-269.
- Ремизов А.Н., Максина А.Г. Сборник задач по медицинской и биологической физике. М., Дрофа, 2001.