

Курс: Статистические Методы Обработки Данных

Лекция 3. Гистограммы и Операции над ними

Специальность: 1-53 01 02 – Автоматизированные системы обработки информации

УО «ГГУ им. Ф. Скорины»

Преподаватель: Бабич К.С, ст. преподаватель, 2016

Раздел 2 – Гистограммы и Операции над ними

5. Определение гистограммы, её назначение

Полученные экспериментальные данные представляют, как правило, в виде таблиц.

А полученные таблицы удобно представить графически.

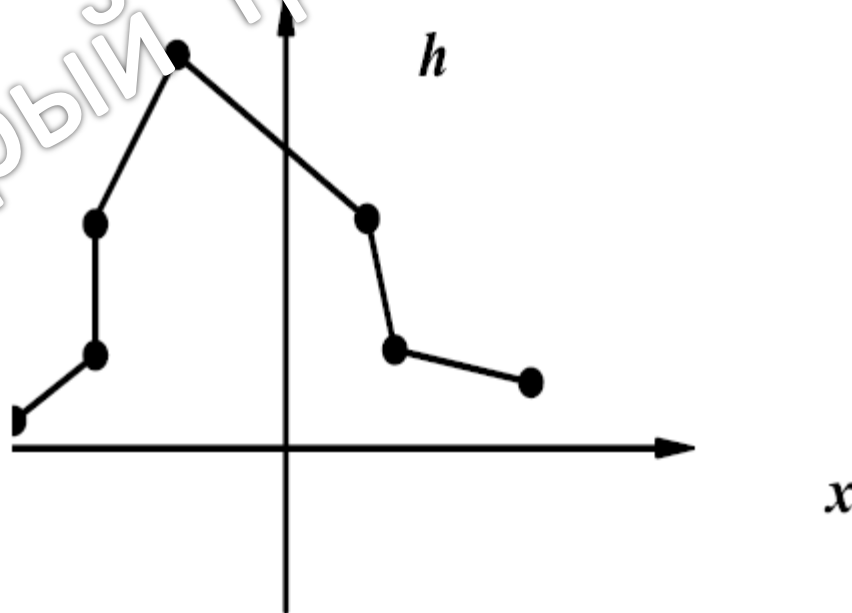
Раздел 2 – Гистограммы и Операции над ними

Полигон частот

Можно построить т.н. **полигон частот** используя частность

$$h = \frac{\text{число событий попадающих в данный интервал}}{\text{общее число событий}}$$

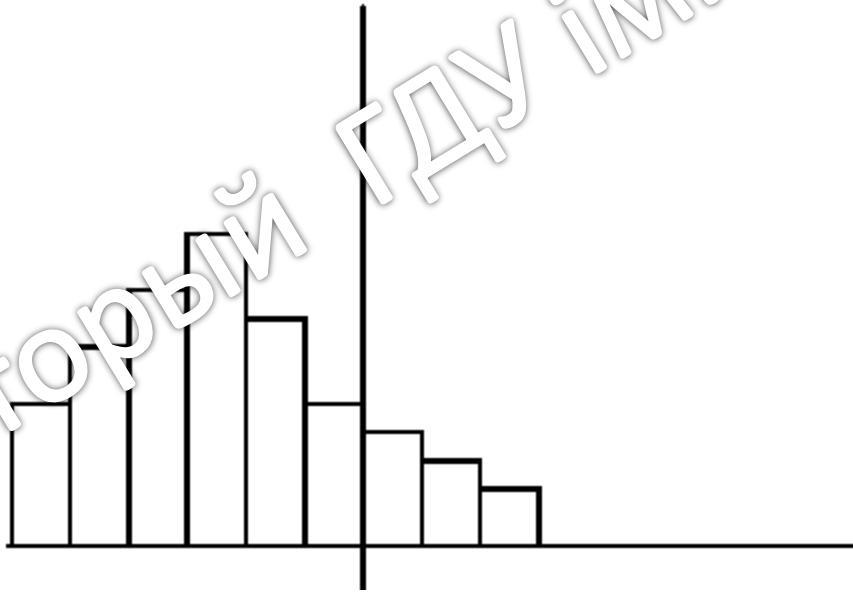
и положение середины интервала



Раздел 2 – Гистограммы и Операции над ними

Гистограмма

Это же распределение можно представить в виде гистограммы. Для построения гистограммы необходимо над каждым отрезком Δx абсцисс, соответствующим интервалу значений измеряемой величины, построить прямоугольник, площадь которого пропорциональна плотности или частоте попадания в этот интервал. Обычно выбирают интервалы одинаковой ширины, поэтому высота прямоугольников различна.

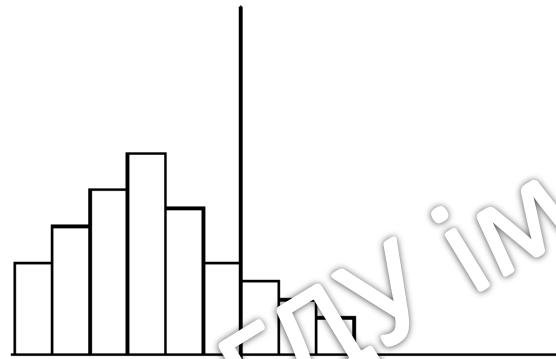


Прим.: Если интервалы одинаковой ширины, то высота прямоугольников прямо пропорциональна частоте попадания.

Раздел 2 – Гистограммы и Операции над ними

Гистограмма

Смысл Гистограммы -- плотность распределения исследуемой случайной величины.



для построения Гистограммы:

1. построить вариационный ряд;
2. разбить ряд на m столбиков с некоторой определенной длиной d ;
3. Построить ступенчатый график.

Итак: Назначение Гистограммы -- использовать её для составления гипотез о форме распределения.

Раздел 2 – Гистограммы и Операции над ними

6. Оптимальное число интервалов для получения гистограммы

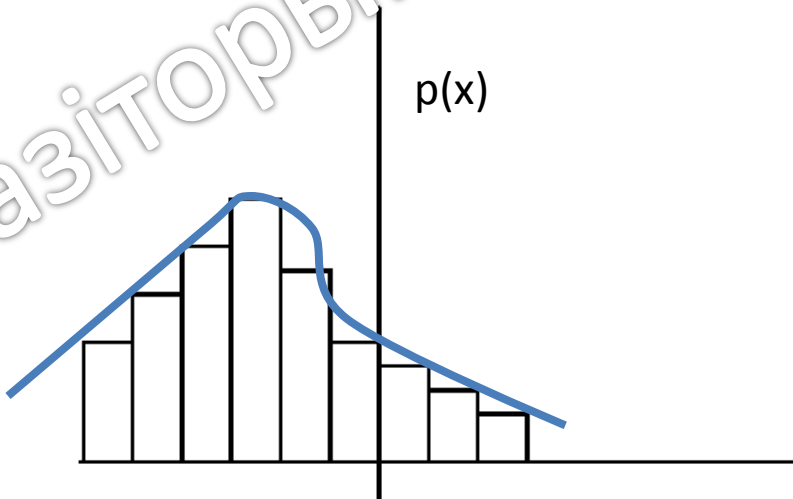
(Как выбрать m (кол-во столбцов) и d (ширины))

1) Общепринято делать интервалы одинаковой длины. (но это не обязательно)

Каким выбирать m ?

- Это число д.б. таким чтобы выполнялся **Критерий** –

Гистограмма д.б. максимально близка к идеальному значению $p(x)$



Раздел 2 – Гистограммы и Операции над ними

6. Оптимальное число интервалов для получения гистограммы

(Как выбрать m (кол-во столбцов) и d (ширины))

Пр.1

Пусть m - мало.

В этом случае любое распределение сведется к равномерному.



Раздел 2 – Гистограммы и Операции над ними

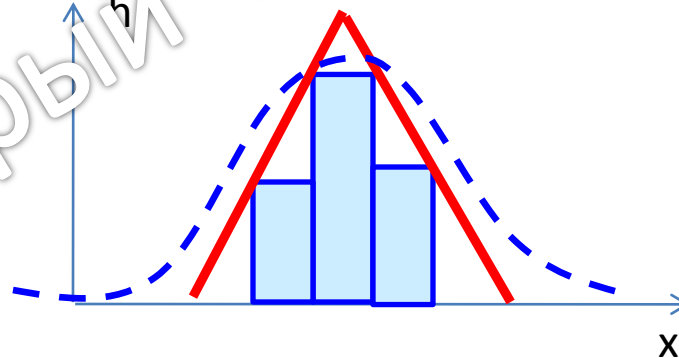
6. Оптимальное число интервалов для получения гистограммы

(Как выбрать m (кол-во столбцов) и d (ширины))

Пр.2

Пусть $m = 3$.

В этом случае любое колокообразное распределение сведется к **треугольному**.



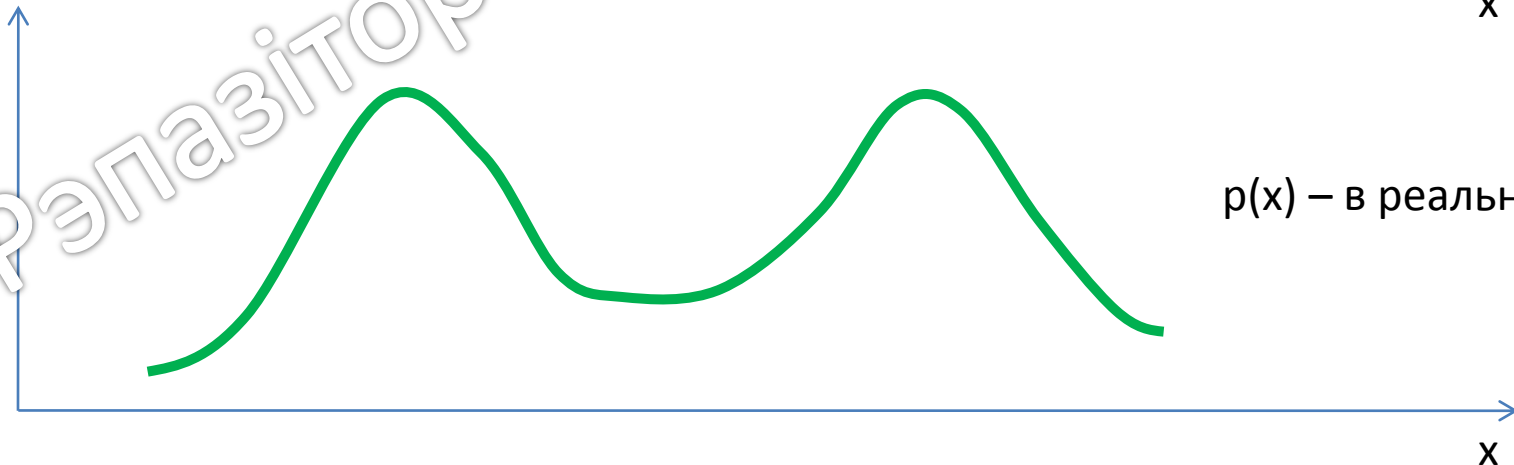
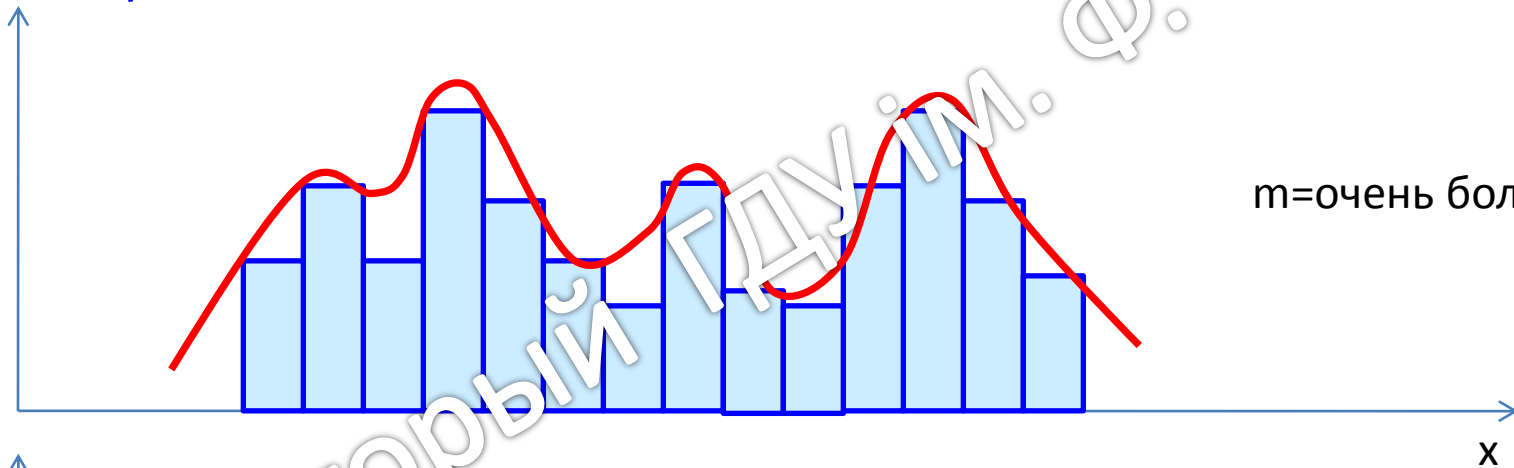
Раздел 2 – Гистограммы и Операции над ними

6. Оптимальное число интервалов для получения гистограммы

(Как выбрать m (кол-во столбцов) и d (ширины))

Пр.3

Пусть $m = \infty$.



Раздел 2 – Гистограммы и Операции над ними

6. Оптимальное число интервалов для получения гистограммы

(Как выбрать m (кол-во столбцов) и d (ширины))

Но оптимальное m существует!

Будем определять его их двух требований:

- 1) Число m д.б. таким чтобы d была достаточно большой для обеспечения хороших статистических свойств.
- 2) Размер ячейки d д.б. достаточно узким чтобы мы могли видеть тонкую структуру исследуемой величины.

Раздел 2 – Гистограммы и Операции над ними

6. Оптимальное число интервалов для получения гистограммы

(Как выбрать m (кол-во столбцов) и d (ширины))

Возникает задача выбора m -- задача оптимальной фильтрации.

m -- д.б. такое, чтобы было максимально возможное сглаживание флуктуаций при минимальном искажении самой кривой искомого распределения.

Раздел 2 – Гистограммы и Операции над ними

6. Оптимальное число интервалов для получения гистограммы

(Как выбрать m (кол-во столбцов) и d (ширины))

Как этого добиться?

-- Ввести количественные критерии

Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

I-я группа: Эвристические критерии (без доказательства)

1) Формула Старджеса

$$m = \log_2 n + 1 = 3.3 \lg n + 1.$$

2) Формула Брукса и Каррузера

$$m = 5 \lg n.$$

3) $m = \sqrt{n}$.

Эти три формулы являются часто встречающимися в литературе по математической статистике.

Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

I-я группа: Эвристические критерии (без доказательства)

1) Формула Старджеса

$$m = \log_2 n + 1 = 3.3 \lg n + 1.$$

2) Формула Брукса и Каррузера

$$m = 5 \lg n.$$

3) $m = \sqrt{n}$.

Эти три формулы являются часто встречающимися в литературе по математической статистике.

Чтобы учесть форму распределения вводят критерии использующие критерий χ^2 .

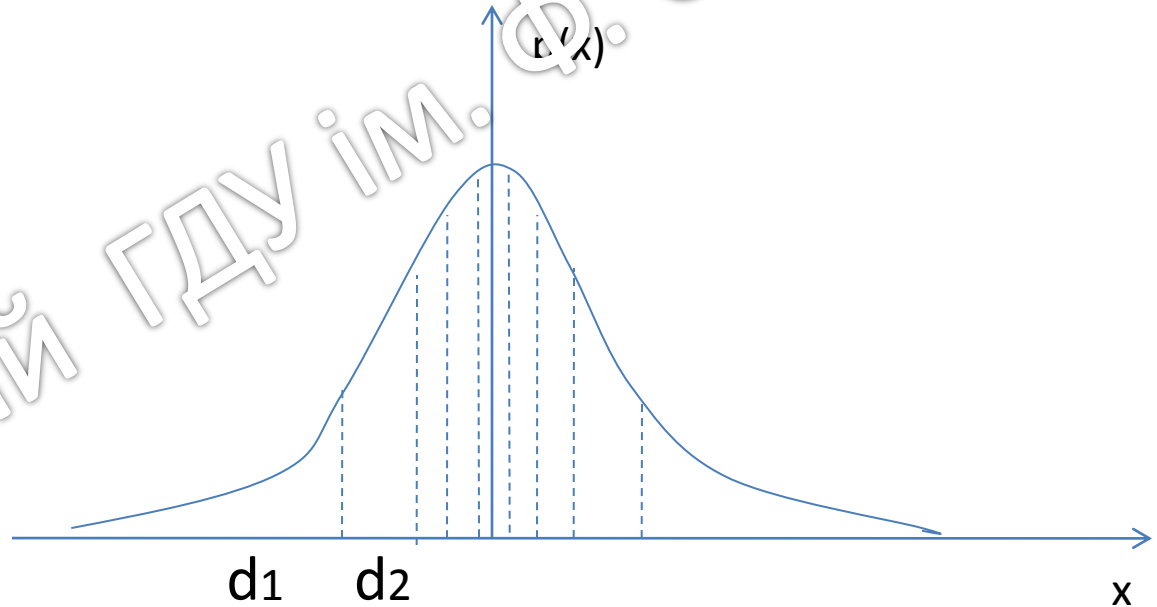
Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

II-я группа: Использующие критерий χ^2

В ней используется рассмотрение интервалов не с равной длиной, а с равной вероятностью в соответствии с принимаемой моделью (т.е. предположением о законе распределения).

$$p_i = \text{const} = \frac{\text{число } x_i}{n}$$



Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

II-я группа: Использующие критерий χ^2

Число интервалов с равной вероятностью, которые мы обозначим как K , отличаются от числа m с равной длиной d (в несколько раз).

Г. Манн и А. Вальд установили, что при $n \rightarrow \infty$ оптимальное число K равновероятных интервалов задается соотношением

$$K \sim 1.5 \left(\frac{n}{Z_\alpha} \right)^{0.4},$$

Z_α – квантиль нормального распределения, соответствующий вероятности $P = 1 - \alpha$, α – принятый уровень значимости.

$$Z_\alpha = (2\pi)^{-1/2} \int_{-\infty}^{Z_\alpha} e^{-\frac{Z^2}{2}} dZ = 1 - \alpha.$$

На практике часто берут $\alpha = 0.1$, тогда

$$K \simeq 1.9 n^{0.4}$$

Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

II-я группа: Использующие критерий χ^2

Поскольку для K интервалы получаются не равной длины, то это приводит к ряду неудобств при построении гистограмм, но зато при этом мы неявно закладываем при использовании χ^2 выбор K в зависимости от формы распределения.

Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

III-я группа

III группа рекомендаций "устраняет" недостаток II группы, возвращаясь к интервалам m с равной длиной d , но при этом и учитывает, в отличие от I группы, форму распределения (форма характеризуется эксцессом ε и контрэксцессом κ).

Примером является соотношение, полученное в работе Алексеевой:

$$m = \frac{4}{\kappa} \lg \frac{n}{10}.$$

На практике эту формулу в зависимости от n при $\kappa = \text{const}$ удобнее аппроксимировать выражением

$$m = \frac{\varepsilon + 1.5}{6} n^{0.4}$$

или

$$m = \frac{1}{3} \sqrt[5]{\varepsilon^4 n^2} = \frac{1}{3} \sqrt[5]{\frac{n^2}{\kappa^8}}.$$

Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

III-я группа

$$m = \frac{\varepsilon + 1.5}{6} n^{0.4}$$

$$m = \frac{1}{3} \sqrt[5]{\varepsilon^4 n^2} = \frac{1}{3} \sqrt[5]{\frac{n^2}{\kappa^8}}.$$

Предпоследняя формула более точна, а последняя более удобна в аналитических вычислениях, из нее четко видно, что зависимость от формы (κ^8) существенно сильнее от объема выборки (n^2).

Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

III-я группа

Трудность использования III группы состоит в том, что часто интервалов часто приходится выбирать прежде, чем будут найдены оценки K , $\bar{x}$, и т.д.

Раздел 2 – Гистограммы и Операции над ними

Рекомендации по выбору m

III-я группа

Трудность использования III группы состоит в том, что число интервалов часто приходится выбирать прежде, чем будут найдены оценки K , $\bar{x}$, и т.д.

Эту трудность обходят следующим образом: наиболее часто встречаются распределения с ε от 1.8 до 6 (от равномерного до Лапласа, включая нормальное $\varepsilon=3$). Для этих граничных точек имеем

$$m_{min} = 0.55 n^{0.4}$$

и

$$m_{max} = 1.25 n^{0.4}$$

Искомое m можно выбрать близким к этому интервалу, при этом m лучше выбрать нечетным, т.к. при четном m для островершинных распределений в центре гистограммы оказывается два столбца равных по высоте и середина распределения принудительно утолщается.

Раздел 2 – Гистограммы и Операции над ними

Практические Рекомендации

1) Для практического определения числа интервалов воспользоваться формулами для m_{max} и m_{min} , выбрав при этом m нечетным.

2) Так как крайние точки могут располагаться несимметрично, то ширина d столбца гистограммы определяется по отклонению от центра ΔX_m наиболее удаленной точки.

$$d = \frac{2\Delta X_m}{m}$$

3) При этом полученное значение d необходимо округлять в большую сторону, чтобы крайняя точка не оказалась за пределами крайнего столбца.

4) Величину d при этом удобно выбирать так, чтобы она делилась на 2 так, чтобы потом центральный столбец можно было бы поделить пополам для уточнения центра распределения.

Раздел 2 – Гистограммы и Операции над ними

7. Промахи и Методы исключения промахов

Одним из условий правомерности статистической выборки является требование ее однородности, т.е. принадлежности всех ее членов к одной и той же генеральной совокупности.

Однако на практике это требование очень часто нарушается. И, если скажем, при обработке вручную еще можно вспомнить как (при каких условиях) были получены "подозрительные" данные, то при автоматической обработке данных необходимы методы исключения "чужих" для данной выборки результатов.

Отсчеты, резко отклоняющиеся по своим значениям от большинства других отсчетов принято называть промахами и исключать их из выборки.

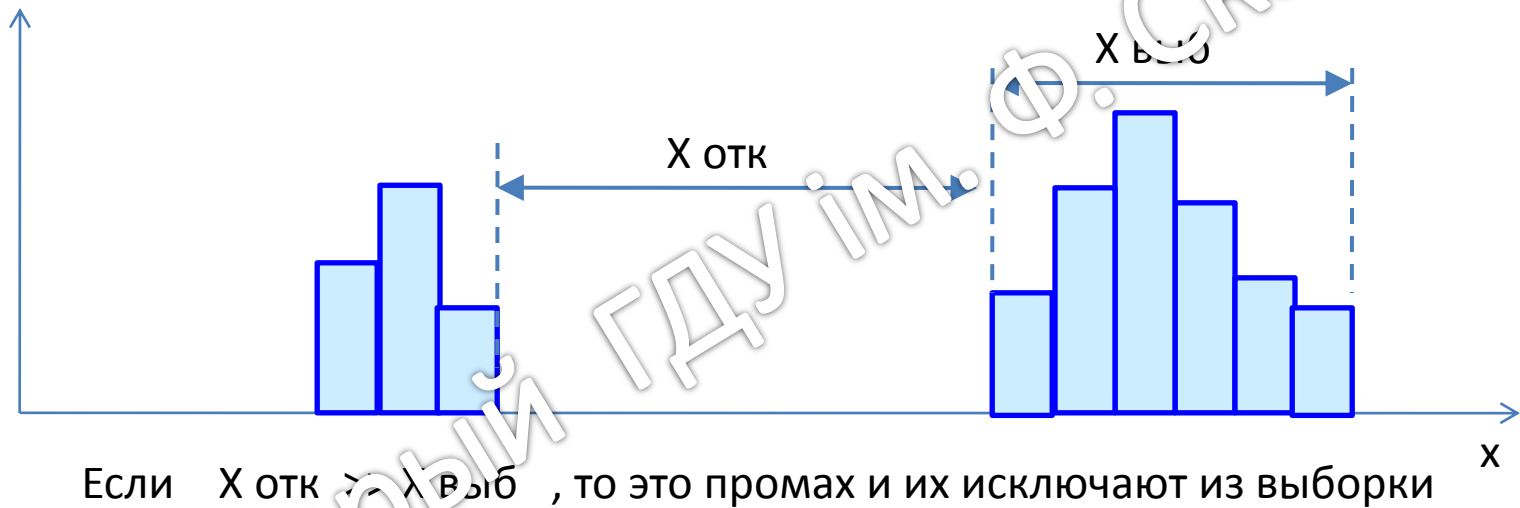
Но особую неприятность доставляют отсчеты, которые и не входят в компактную группу отсчетов, но и не удалены от нее на значительное расстояние. Такой отсчет называют предполагаемым промахом.

В экспериментальной практике исследователи просто отбрасывали крайние, "слишком удаленные от центра наблюдения". Это получило название цензурирования выборки.

Однако для принятия решения необходимы какие-либо формальные критерии.

Раздел 2 – Гистограммы и Операции над ними

7. Промехи и Методы исключения промехов

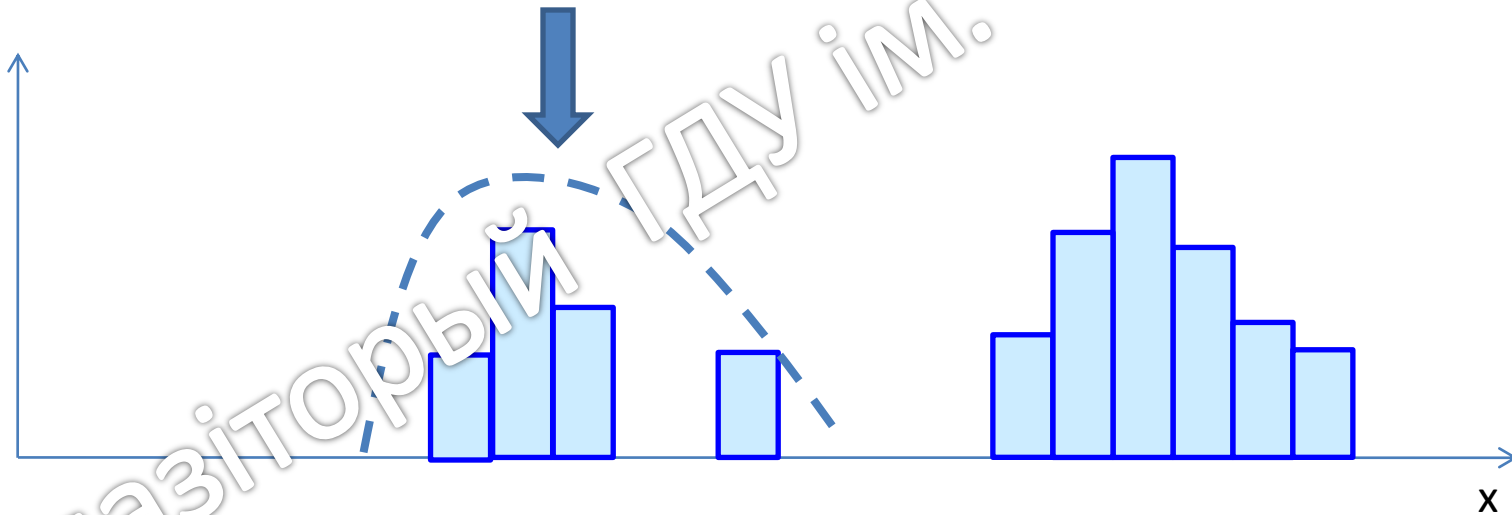


Раздел 2 – Гистограммы и Операции над ними

7. Промехи и Методы исключения промехов

Но часто бывает, что $X_{\text{отк}} \sim$ или $< X_{\text{выб}}$

Предполагаемые промехи



Процедура исключения промехов носит название **цензурирование** выборки.

Раздел 2 – Гистограммы и Операции над ними

Правило «3σ»

Простейший метод заключается в использовании "правила 3σ", когда по выборке с удаленными отсчетами (предполагаемыми промахами) вычисляется оценка σ и граница $|X_b| = 3\sigma$, а все $|x_i| \geq 3\sigma$.

"Правило 3σ" обосновано на неравенстве Чебышева:

$$P\{|x_i - \bar{x}| \geq a\} \leq \frac{\sigma^2}{a^2}, \quad (a > 0)$$

Если все $x_i \geq 0$, то

$$P(x \geq a) \leq \frac{\bar{x}}{a}, \quad (a > 0)$$

Если x имеет одномодальное распределение (непрерывное), то справедлива более сильная оценка

$$P_s\{|x_i - \bar{x}| \geq a\} \leq \frac{4}{9} \frac{1 + S^2}{\left(\frac{a}{\sigma} - |S|\right)^2},$$

где S – гирсоновская мера асимметрии (для распределений симметричных относительно моды $S = 0$).

Раздел 2 – Гистограммы и Операции над ними

Правило «3σ»

Если x имеет одномодальное распределение (непрерывное), то справедлива более сильная оценка

$$P_s\{|x_i - \bar{x}| \geq a\} \leq \frac{4}{9} \frac{1 + S^2}{\left(\frac{a}{\sigma} - |S|\right)^2},$$

где S – пирсоновская мера асимметрии (для распределений симметричных относительно моды $S = 0$).

$$S = \frac{\bar{x} - \xi}{\sigma},$$

где ξ – максимальная вероятность

$$S = \frac{\gamma_1 (\gamma_2 + 6)}{2 (5\gamma_2 - 6\gamma_1^2 + 6)},$$

где

$$\gamma_1 = \frac{m_3}{m_2^{3/2}}, \quad \gamma_2 = \frac{m_4}{m_2^2} - 3.$$

и определяются по формуле:

$$m_j = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^j.$$

Раздел 2 – Гистограммы и Операции над ними

Правило «3σ»

Если положить $a = 3\sigma$, то имеем

$$P_s\{|x_i - \bar{x}| \geq 3\sigma\} \leq \frac{1}{9} \approx 0.11$$

или для одномодальных симметричных:

$$P\{|x_i - \bar{x}| \geq 3\sigma\} \leq \frac{4}{9} \cdot \frac{1}{9} = \frac{4}{81} \approx 0,05$$

т.е. для произвольного распределения вероятность, что $x_i \geq \bar{x}$ на 3σ составляет 11%, а для одномодальных симметричных 5%.

Однако правило "3σ" только для нормального распределения.

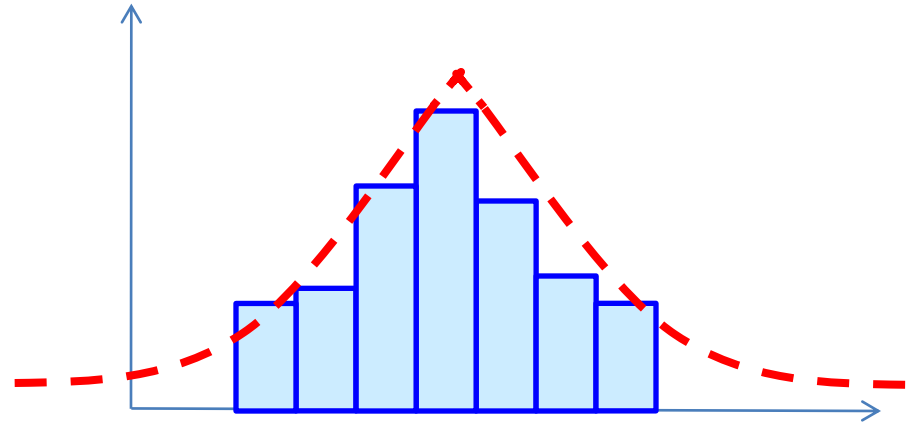
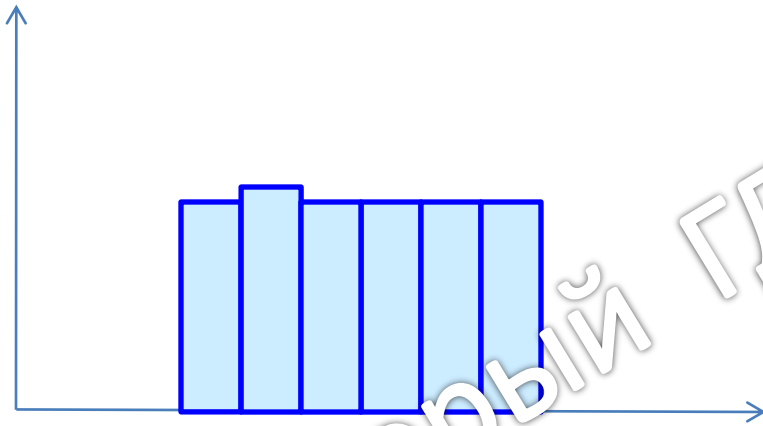
Действительно, при $n = 100$ появление $|x_i| \geq 3\sigma$ можно считать промахом, то для равномерного промахом можно считать уже $|x_i| = 1.8\sigma$, а для распределения Лапласа $|x_i| = 3\sigma$ есть отсчет, принадлежащий данной выборке.

Раздел 2 – Гистограммы и Операции над ними

Правило «3 σ »

Однако правило "3 σ " хорошо для нормального распределения.

Действительно, при $n = 100$ появление $|x_i| \geq 3\sigma$ можно считать промахом, то для равномерного промахом можно считать уже $|x_i| = 1.8\sigma$, а для распределения Лапласа $|x_i| = 3\sigma$ есть отсчет, принадлежащий данной выборке.



Раздел 2 – Гистограммы и Операции над ними

Правило «3 σ »

На этом примере видно, что граница цензурирования зависит не только от объема выборки n , но также и от формы распределения.

Зависимость от n полуколичественно можно оценить из условия: границы цензурирования должны отсекают в среднем менее одной точки, тогда назначение границ с уровнем значимости $g = 1 - P$, где $P = \frac{n}{n+1}$ дает зависимость от n .

В заключение отметим, что все оценки μ , σ и т.д. должны пересчитываться после цензурирования выборки.

Раздел 2 – Гистограммы и Операции над ними

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Рэпазіторый ГДУ ім. Ф. Скарныны

Раздел 2 – Гистограммы и Операции над ними

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.1 Наиболее распространенным является критерий согласия Пирсона (χ^2 - критерий).

Суть χ^2 - критерия состоит в вычислении величины:

$$\chi^2_{\text{ýê}} = n \sum_{j=1}^m \frac{(\nu_j - p_j)^2}{p_j}$$

где m – число интервалов,

ν_j - частота попадания экспериментальных данных в j -й интервал,

p_j - вероятности в том же j -м столбце, рассчитанном по заданной модели.

Если бы наша выбранная модель во всех центрах столбцов имела

$$\nu_j = p_j \Rightarrow \chi^2 = 0$$

Раздел 2 – Гистограммы и Операции над ними

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.1 Наиболее распространенным является критерий согласия Пирсона (χ^2 - критерий).

Суть χ^2 - критерия состоит в вычислении величины:

$$\chi^2_{\text{ýê}} = n \sum_{j=1}^m \frac{(\nu_j - p_j)^2}{p_j}$$

где m – число интервалов,

ν_j - частота попадания экспериментальных данных в j -й интервал,

p_j - вероятности в том же j -м столбце, рассчитанном по заданной модели.

Если бы наша выбранная модель во всех центрах столбцов

$$\nu_j = p_j \Rightarrow \chi^2 = 0$$

т.о. хи-квадрат – мера суммарного отклонения между модельным и экспериментальным распределением.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Величина χ^2 получила своё обозначение не случайно, т.к. она подчиняется распределению χ^2 с $k = m - 1 - r$ степенями свободы.

где r – число параметров модели, определяемых по экспериментальным данным, необходимых для совмещения модели и гистограммы.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Величина χ^2 получила своё обозначение не случайно, т.к. она подчиняется распределению χ^2 с $k = m - 1 - r$ степенями свободы.

где r – число параметров модели, определяемых по экспериментальным данным, необходимых для совмещения модели и гистограммы.

Критерий χ^2 :

$$\chi^2 \leq \chi_{p,k}^2$$

$p = 1 - \alpha$ – доверительная вероятность,
или вероятность принятия модели.

т.е. гипотеза о распределении при данном уровне значимости
не противоречит модели

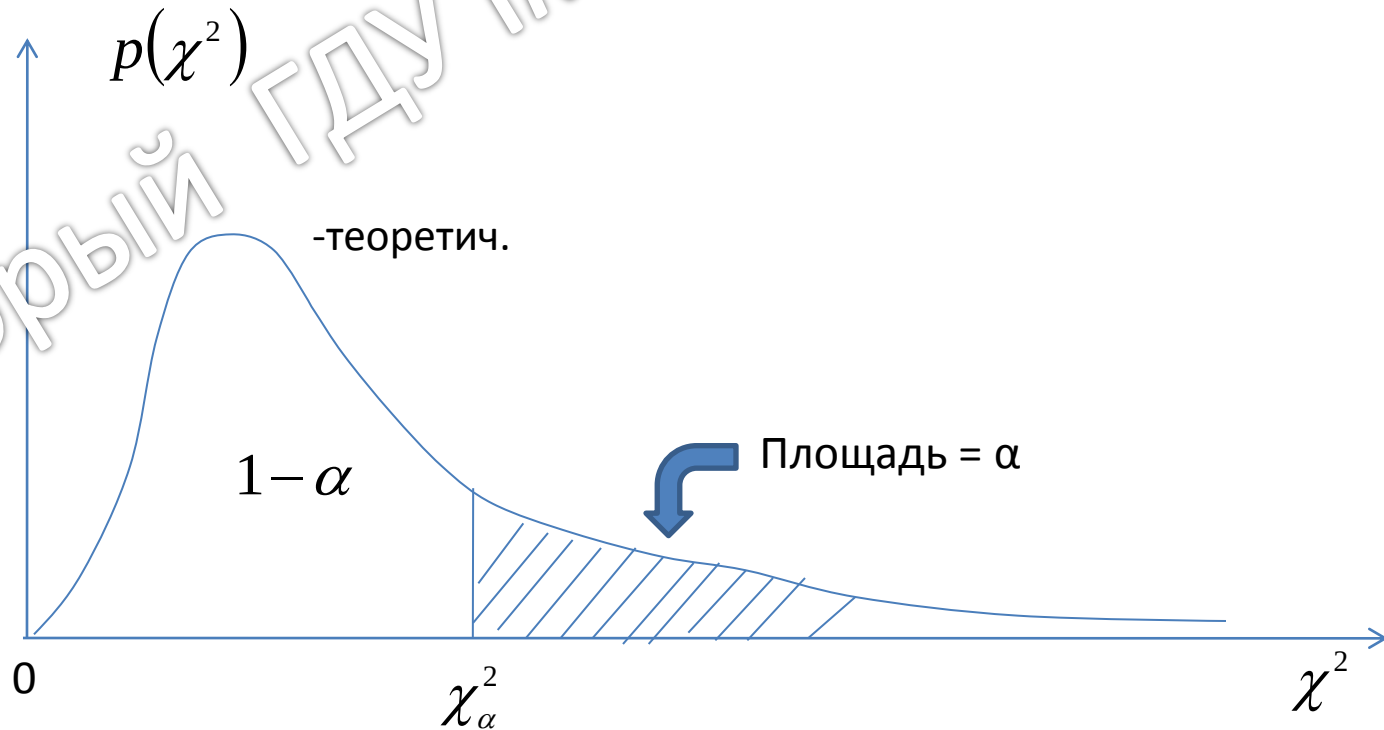
$$\chi^2 > \chi_{\alpha,k}^2$$

α – уровень значимости,
или вероятность отклонения модели

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

Как правило, критерий хи-квадрат дает лишь отрицательный ответ, однако на практике используется и для принятия положительного решения.

Распределение χ^2 :



8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

В этом критерии используется максимальное значение модуля разности между экспериментальным и теоретическим распределением интегральных функций распределений.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

Математически это означает, что находят:

$$D = \max | F_{\text{экспер}}(x) - F_{\text{теор}}(x) | / n$$

где $F(x)$ – интегральная функция распределения.

и сравнивается с граничным значением вероятности

$$P \geq 2 \exp(-2nD^2) = 2 \exp(-2\Delta^2 / n)$$

$$\Delta = | F_{\text{экспер}} - F_{\text{теор}} |$$

Здесь P – искомая вероятность того, что искомая выборка может соответствовать выбранной модели.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

Можно построить и доверительный интервал следующего типа:

$$P_{\text{дов}} [F_{\text{экспер}} - d_{\alpha} \leq F_{\text{теор}} \leq F_{\text{экспер}} + d_{\alpha}] = 1 - \alpha$$

где d – ширина полосы в которой лежит $F_{\text{экспериментальная}}$.

где критические значения

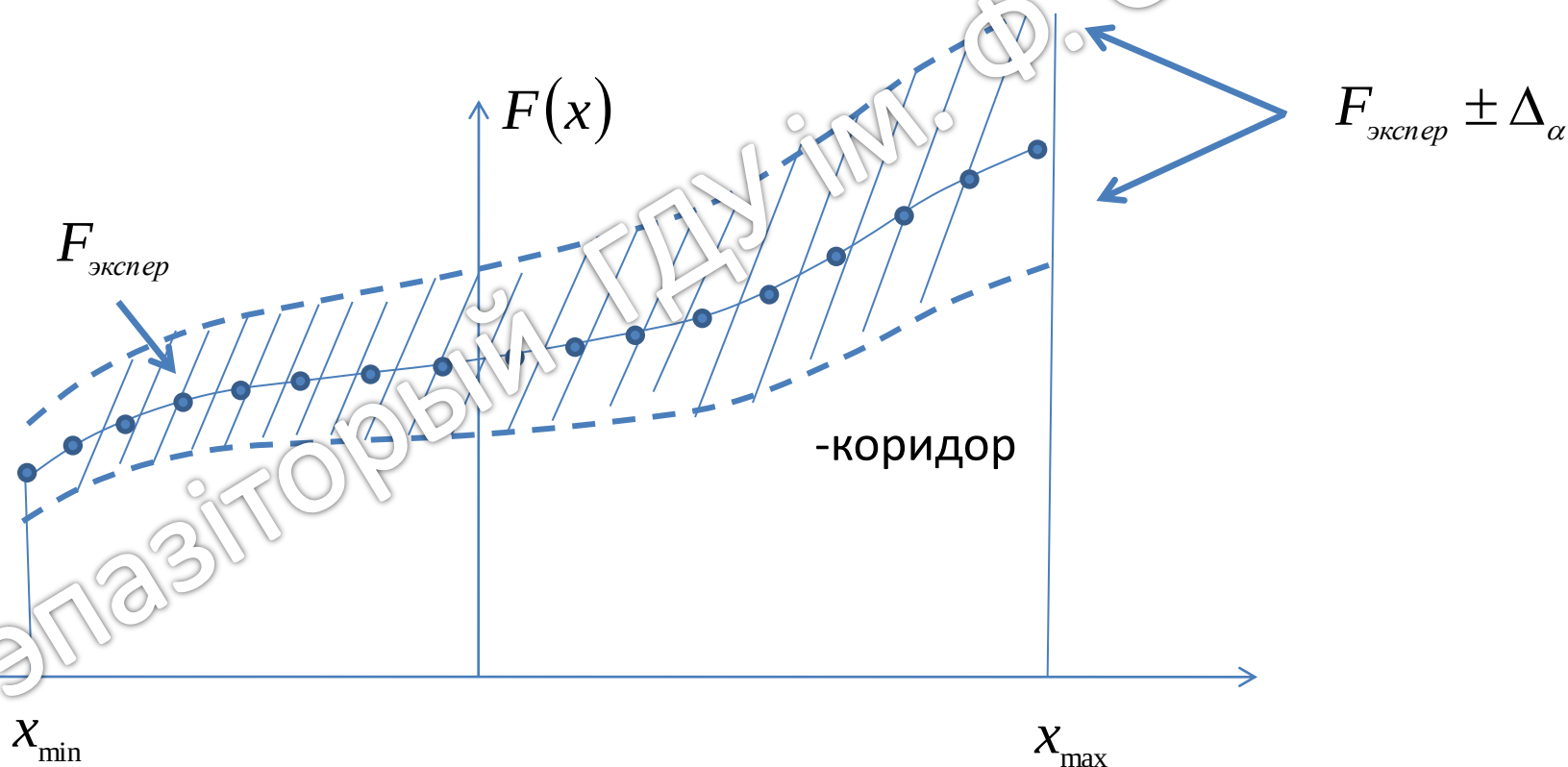
$$d_{\alpha} \approx \sqrt{-0.5 \ln \left(\frac{\alpha}{2} \right) / n} \quad (n > 35)$$

Здесь P – искомая вероятность того, что искомая выборка может соответствовать полученной модели, α – уровень значимости.

8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

Доверительный интервал представляет собой полосу с шириной $\pm d_\alpha$ около выбранного нами (измеренного) $F_{\text{экспер}}$ и с вероятностью $1-\alpha$ истинная функция $F_{\text{теор}}(x)$ лежит внутри полосы.



8. Идентификация формы распределения экспериментальных данных с помощью критериев согласия.

8.2 Критерий Колмогорова-Смирнова

Преимуществом перед χ^2 -критерием состоит в том, что не требуется группирование данных в интервалы.

Если же есть гистограммы. То они д.б. построены одинаковых границах числе интервалов группирования.