Практическое занятие № 3
по курсу «Статистические методы обработки данных»

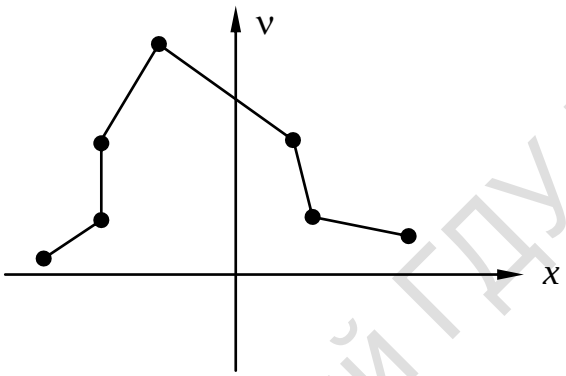
Тема: Построение гистограмм.

Цель работы: Овладеть способами построения гистограмм с помощью различных критериев выбора числа столбцов.

Краткая теория

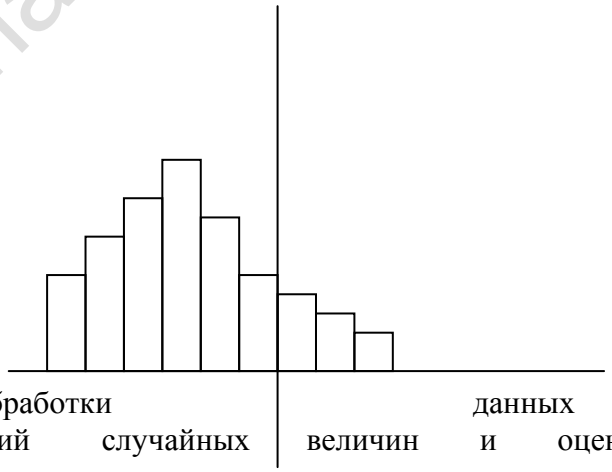
Определения гистограммы, ее назначение.

Полученные экспериментальные данные представляют, как правило, в виде таблиц. Полученные таблицы удобно представить графически. Используя частность n (отношение числа событий, попавших в данный интервал, к общему числу событий) и положение середины интервала можно построить полигон частот.



Это же распределение можно представить в виде гистограммы.

Для построения гистограммы необходимо над каждым отрезком оси 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.



Цель обработки данных заключается в выявлении вида распределений случайных величин и оценки параметров установленного распределения.

Для определения оценок математического ожидания, с.к.о., эксцесса не требуется какого-либо группирования данных.

Для определения медианы, сгибов, использования критерия согласия Колмогорова-Смирнова или для обнаружения промахов, экспериментальные данные необходимо расположить в порядке возрастания, т.е. построить вариационный ряд (упорядоченную выборку).

Для определения формы распределения, для использования критериев согласия Пирсона и др., для сопоставления гипотез о форме распределения и т.д. простого упорядочения выборки уже недостаточно, а выборка должна быть представлена в виде гистограммы, состоящей из m столбцов с определенной протяженностью d соответствующих им интервалов.

Оптимальное число интервалов для получения гистограммы экспериментальных данных. (Или как построить гистограмму). Как выбрать m и d.

Общепринято делать интервалы одинаковыми. Хотя в дальнейшем увидим, что это условие необязательно.

Условие равновеликости интервалов удобно с практической точки зрения.

Во-первых, очевидно, что существует оптимальное число интервалов группирования, когда ступенчатая огибающая гистограммы наиболее близка к плавной кривой распределения.

К примеру, при группировании данных в большое число меньших интервалов, некоторые из них окажутся пустыми или малозначительными. Гистограмма будет отличаться от плавной кривой распределения вследствие изрезанности многими всплесками и провалами.

Т.е. 1-е требование:

Размер интервала (ячейки) должен быть достаточно широким для обеспечения хороших статистических свойств будущей гистограммы (достаточно большая пуассоновская статистика, минимальные корреляции (связи) с соседними ячейками).

При слишком малом числе m интервалов, гистограмма отличается от действительной кривой распределения вследствие слишком крупной ступенчатости. Из-за чего будут потеряны характерные особенности. Например, если взять m=1, т.е. d равно размаху экспериментальных данных, то любое распределение сводится к равномерному, а если m=3, то любое колоколообразное распределение сведется к треугольному.

В примере для обработки линейчатых спектров большая ячейка может привести к потере спектральной линии

Т.о. 2-е требование:

Размер ячейки должен быть достаточно узким для того, чтобы прорисовывалась «тонкая структура» исследуемой величины.

Как видим, требования являются противоречивыми.

Укрупнение интервалов группирования является методом «фильтрации различных случайных выбросов и провалов», но слишком протяженные интервалы сглаживают особенности искомого закона распределения.

Таким образом, задача выбора оптимального числа интервалов при построении гистограммы – это задача оптимальной фильтрации, а оптимальным числом m интервалов является максимальное возможное сглаживание случайных флуктуаций данных, которое сочетается с минимальным искажением от сглаживания самой кривой искомого распределения.

Рекомендации по выбору m.

***I группа: эвристические критерии (без доказательства*).**

1) Формула Старджеса

$$m = \log_2 n + 1 \approx 3.3 \lg n + 1 \qquad (1)$$

2) Формула Брукса и Каррузера

$$m = 5\lg n \qquad (2)$$

3) $$m = \sqrt{n} \qquad (3)$$

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

II группа: с использованием χ^2 **.**

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

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

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

$$K = 4\sqrt[5]{2}\left(\frac{n}{Z_\alpha}\right)^{0.4} \qquad (4)$$

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

$$Z_\alpha = (2\pi)^{-1} \int_{-\infty}^{z_\alpha} e^{-\frac{z^2}{2}} dz = 1 - \alpha$$

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

$$K \approx 1.9\, n^{0.4} \qquad (5)$$

III группа.

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.

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 æ.

Примером является соотношение, полученное в работе Алексеевой:

$$m = \frac{4}{\kappa}\lg\frac{n}{10} \qquad (6.6)$$

Трудность использования III группы состоит в том, что число интервалов часто приходиться выбирать прежде, чем будут найдены оценки κ , $\overline{x}$ 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 ε =3). Для этих граничных точек имеем:

$$m_{min}=0.55n^{0.4} \quad \text{и} \quad m_{max}=1.25n^{0.4} \quad (6.9)$$

Искомое m сложно 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.

«Практические» рекомендации:

1) Для практического определения числа интервалов воспользоваться формулой (6.9), 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.

Задание

ЗАДАЧА 1. Провести сортировку заданных трех выборок при помощи функций MS-Excel.

ЗАДАЧА 2. Построить полигон частот для данных выборок. Число интервалов m=10. Построить соответствующие гистограммы.

ЗАДАЧА 3. Построить гистограммы, используя эвристические критерии и критерии с учетом формы распределения. Сравнить результаты и сделать выводы.

ЗАДАЧА 4. Подготовить отчет.

Составители: Андреев В.В. (2014),
Бабич К.С. (2016)