Лекция 1 Бакеев Мат. стат.

Выборки (экспериментальные данные)

Опр. Генеральной совокупностью наз-ся множ-во всех исходов эксперимента

Опр. Выборочной совок-ю наз-ся объём экспери-х данных из генеральной совти которую мы имели возможность наблюдать

Условно была выборочная сов-ть самолётов вернувшихся из боя. А предположение что пробоины равномерно распределены. Такой случай назвали ошибка выжившего.

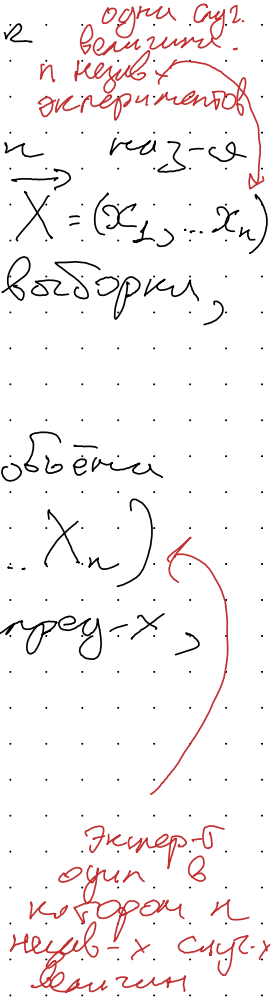
Опр. выборки наз-ся репрезентативной если её распр-е совпадает с теоретическим. (?)
Все выборки считаем репрезентатив-ми.

В каждой области существуют два способа борьбы с перерегулированием

Опр. Объёмом выборки и наз-ся кол-во экспер-х данных

Опр. А: выборкой объёма и наз-ся набор экспер-х данных $\vec{X} = (X_1, \dots X_n)$ (апостериорное определение выборки, то что после опыта)

Опр. Б: (до опыта) выборкой объёма и наз-ся набор $\vec{X} = (X_1, \dots X_n)$ независимых, одинаково распред-х случайных величин.



Выборочные характеристики

Имеется выборка $\vec{X} = (x_1, \dots x_n)$ в смысле A.

Мы можем рассматривать её, как дискретную случайную величину

X	x_1	$\dots$	x_n
p	$1/n$		$1/n$

Математическое ожидание такой дискретной случайной величины назся средне выборочным

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ясно что это приблизительная оценка математического ожидания.

Точно также мы можем вычислить дисперсию

$$D^* = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$ — выборочная дисперсия (оценка неизвестной дисперсии)

Выборочная функция распр-я

$$F^*(x) = \frac{1}{n}\sum^{n} I(X_i < x) = \frac{\text{число } X_i < X}{n}$$

Теорема: Гливенко – Кантелли

Пусть $F^*(x)$ – выбор. ф-я распр.

$F(x)$ – теор. ф-я распр-я.

n – объём выборки

$$\sup_x |F^*(x) - F(x)| \xrightarrow[n\to\infty]{P} 0$$

(сходимость по вероятности)

Начальная обработка стат. данных

1) Ранжирование выборки (упорядочиваем данные по возрастанию), в результате получаем вариационный ряд.

$(X_{(1)} > X_{(2)}, \dots, X_{(n)})$ $X_{(i)} \le X_{(j)}$ $i < j$

$X_{(i)}$ – $i^{\text{й}}$ порядковая статистика

$X_{(1)} = X_{min}$ минимальная порядковая стат-ка

$X_{(n)} = X_{max}$

Note: Если в выборке много повтор-ся данных, то бывает удобно представить в виде вариационного частотного ряда

X	X_1	...	X_n
n	n_1		n_n

где n_i это число или частота значения X_i. сколько раз оно встречалось (удобно для дискретной величины)

Note в этом случае выборочную дисперсию и другие удобно считать по формулам

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \cdot n_i$$

Дисперсия = Var = S^2

$$D^* = \frac{1}{n}\sum^{n}\left(X_i \cdot n_i - \overline{X}\right)^2$$

$V_i = \frac{n_i}{n}$ – относительная частота

их можно рассматривать как теор-е вероятности этих значений

X	X_1	...	вариационный ряд с отн-ми частотами
V	V_1	...	

2) Иногда бывает полезно отбросить некоторое количество самых малых и больших значений выборки (5, 5%). Причины
1) улучшить репрезентативность
2) исключить выбросы (потом обсудим)

Применять по ситуации. Если распр-е равномерное или нормальное, то не следует отбрасывать.

3) Часто если объём данных большой и распределение равномер-е вариационный ряд удобнее интервальным рядом. Данное разбивается на k интервалов
$[a_i, a_{i+1})$ - интервал
n_i - частота данного интервала
Note:
число интервалов выбирается по формуле Стёрджеса
$k \approx 1 + \log_2 n$

$k \approx \sqrt[3]{n}$

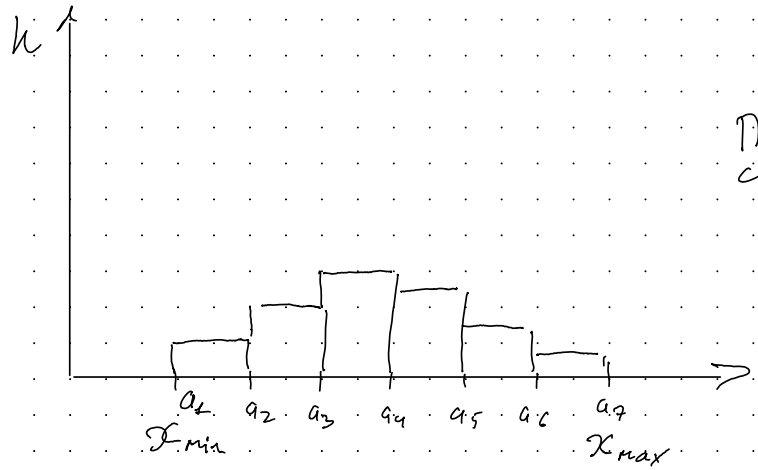
а) Можно разбивать на интервалы отрезки $[X_{min}, X_{max}]$ или сегменты $[X_{min} - \frac{l}{2}, X_{max} + \frac{l}{2}]$ где l длина интервала

б) Интервалы берутся либо одинаковой длины $l \approx (X_{max} - X_{min})/k$ или равнонаполненные, в каждом интервале приблизительно равное число элементов.

Число интервалов может зависеть от разных интервалов

Гистограмма

Строится ступенчатая фигура из прямоугольников основания которых равны интервалам длины h а высоты прямоугольников $h_i = \frac{n_i}{n \cdot h}$

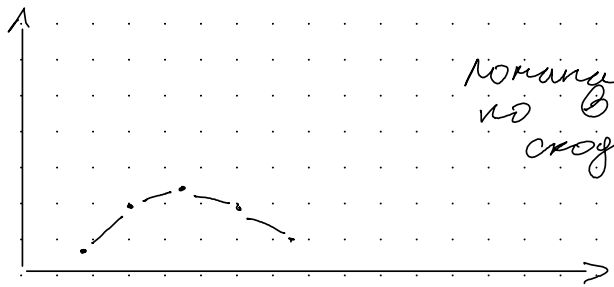


При $n \to \infty$ гистограмма сходится по вероятности к плотности распред-я

Полигон

На числовой прямой отм-ся точки
где c_i k_i $c_i = \frac{q_i + q_{i+1}}{2}$ - сер-а i-го интервала

$k_i = \frac{n_i}{n}$



Ломаная которая по вер-м будет сходиться к плотности распр-я вер-ти

Иногда с помощью полигона изображается выборочная функция распред-я.

Можно специализироваться в экспертных системах и анализе данных.

Нужно учиться, в курсе только база

$$\frac{6+3+9+7+5+9}{6} = \frac{10+10+14}{6} = \frac{34}{6} = \frac{17}{3} =$$

$$= 5,67$$

$$\frac{1}{5}\left(0,333^2 + 2,667 + 1,667 + 1,333 + 0,667 + 3,333\right) \approx 4,667$$