СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ.

"Статистика знает всё" известно, сколько какой пищи съедает в год средний гражданин республики: известно, сколько в стране охотников, балерин: станков, велосипедов, памятников, маяков и швейных машинок: Как много жизни, полной пыла, страстей и мысли, глядит на нас со статистических таблиц!..".

отрывок из романа Ильфа и Петрова "Двенадцать стульев"

Это ироничное описание даёт общее представление о статистике.

Термин "статистика" произошел от латинского слова "*статус*" (*status*), что означает "состояние и положение вещей".

Статистическая обработка данных.

Ребята, мы переходим к изучению нового раздела, связанного с вопросами обработки данных различных экспериментов и элементов теории вероятности.

Теория вероятности и математическая статистика находят свое применение практически во всех областях жизни.

Так же заметим, что, частично, мы уже изучали данный раздел раньше, так что некоторые моменты вы можете помнить.

Статистическая обработка данных.

Давайте рассмотрим какой-нибудь пример, где нам может пригодиться обработка информации.

Пусть у нас есть десять футболистов, основной состав некоторой команды.

Наши футболисты пробивают по десять пенальти и результаты каждого записываются.

После окончания у нас есть некоторый набор результатов, на первый взгляд просто набор чисел, но что можно сделать с этими числами? Какую пользу они нам могут принести?

Статистическая обработка данных.

В первую очередь надо как то сгруппировать и упорядочить нашу информацию. Группировать информацию можно различными способами, все зависит от требуемой задачи. В нашем случае мы можем сгруппировать по фамилии игрока или по номеру игрока команды.

Сгруппируем по номеру игрока.

Номер Игрока	2	3	4	5	6	7	8	9	10	11
Количество Забитых Голов	3	7	6	5	5	4	4	7	8	6

Сгруппируем по количеству забитых голов.

Количество Забитых Голов	0	1	2	3	4	5	6	7	8	9	10
Количество игроков забивших голы	0	0	0	1	2	2	2	2	1	0	0

Статистическая обработка данных.

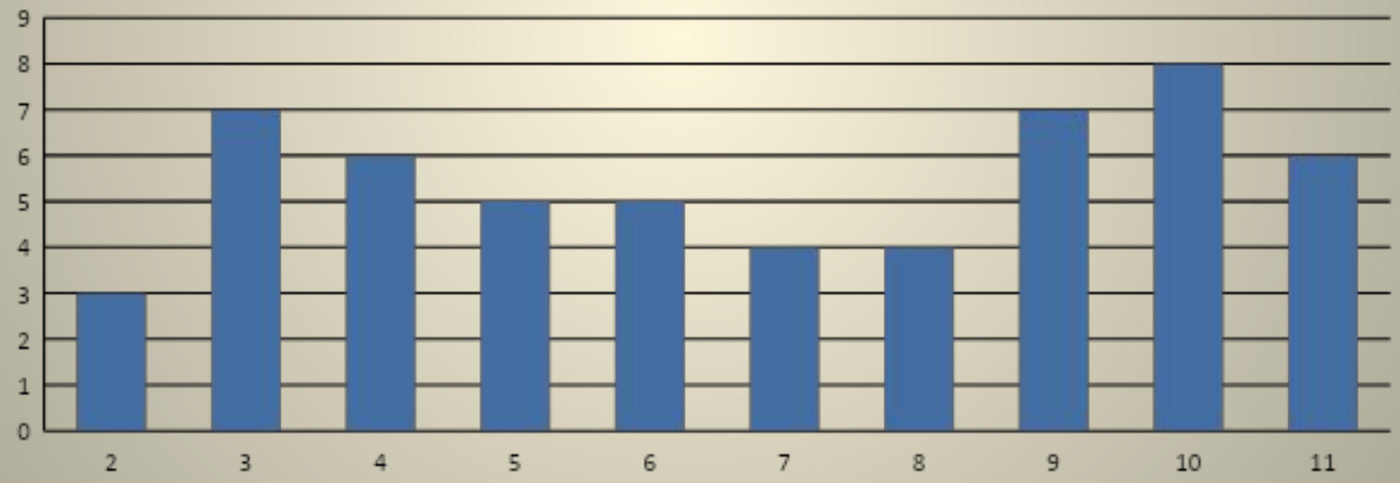
Рассмотрим, как нашу таблицу можно представить графически. В виде графиков представим первую таблицу.

На обычной координатной плоскости, по оси абсцисс отложим номер игрока, а по оси ординат - количество забитых голов.

Полученная кривая называется **полигоном частот.**

Статистическая обработка данных.

Теперь давайте построим **гистограмму**: она позволяет так же наглядно наблюдать за значениями нашего ряда распределений. Мы строим прямоугольники с “центром” в значениях нашего ряда. Получаются такие прыгающие столбики.



Статистическая обработка данных.

Нам осталось построить еще один тип диаграммы – **круговую диаграмму.** Представим, что наш круг занимает все - 100% забитых голов (55 голов), тогда игрок с номером два займет 3/55 площади круга, игроки с номерами 5 и 6 займут 1/11 часть круга, так как 5/55=1/11.

Давайте построим для всех игроков круговую диаграмму.

Статистическая обработка данных.

Ну вот, мы с вами научились немного обрабатывать данные.

Давайте напишем небольшой алгоритм первичной обработки данных:

1) Упорядочить и сгруппировать данные.

2) Составить таблицу распределения данных.

3) Графическое представление данных. В зависимости от задачи построить один из графиков распределения: **Полигон частот (относительных частот), Гистограмму или Круговую диаграмму.**

Статистическая обработка данных.

Но на этом обработка информации не заканчивается, для нашего ряда распределения можно найти многие числовые характеристики. Давайте рассмотрим их.

Первая числовая характеристика это ***объем выборки*** (***N***), в нашем случае он равен десяти, так как мы рассматривали десять футболистов, т.е. $N = 10$.

Размах измерения – разница между наибольшим и наименьшим значениями выборки.

Больше всего голов забил игрок под номером 10 – 8 голов. Меньше всего, игрок под номером 2 – 3 гола. Тогда размах нашего измерения: $R = 8 - 3 = 5$.

Самое популярное или наиболее часто встречаемое значение называется ***модой*** ***(M_0)*** *выборки*.

В нашем примере $M_0 = 10$ – игрок забивший наибольшее количество голов. В реальности тренер команды мог назначить этого игрока штатным пенальтистом.

Статистическая обработка данных.

Среднее значение выборки ($X_{ср}$). Суммируя все результаты и поделив на объем выборки можно получить среднее значение.

$$X_{ср} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

В нашем примере для подсчета среднего значения удобнее использовать данные второй построенной таблицы.

$$X_{ср} = \frac{0 \cdot 0 + 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 1 + 4 \cdot 2 + 5 \cdot 2 + 6 \cdot 2 + 7 \cdot 2 + 8 \cdot 1 + 9 \cdot 0 + 10 \cdot 0}{10} = 5{,}5$$

Округлив до … али по шесть голов. Тренер команды мог бы запомнить данное значение, и через некоторое время провести еще раз такой эксперимент и проверить растут ли показатели команды или нет.

Статистическая обработка данных.

Варианта измерения – каждое число встретившиеся в результате измерения. В нашем случае для первой таблицы – количество забитых голов, для второй – количество игроков забивших гол.

Медиана измерения (M_e) – средняя варианта встречающаяся в выборке. Она делит нашу выборку пополам.

Для второй выборки $M_e = 5$, так как это значение делит наш ряд ровно пополам.

Если число вариант четно, как в первой выборке, то берутся два средних значения и делятся пополам: $M_e = (6+7)/2=6{,}5$.

Кратность или абсолютная частота варианты – то сколько раз встречается конкретная варианта.

Для второй таблица кратность 0 равна 0, кратность 4 равна двум, кратность 8 равна единице.

Статистическая обработка данных.

При составлении таблицы, не всегда получается, что варианты расположены через равные промежутки.

Варианта измерения может принимать *фактически любые значения и положительные и отрицательные.*

Кратность варианты всегда больше нуля, если кратность равна нулю то фактически в нашем эксперимента данное значение не встретилось, поэтому вторую таблицу распределения целесообразней записать в таком виде:

Количество Забитых Голов	3	4	5	6	7	8
Количество игроков забивших голы	1	2	2	2	2	1

Статистическая обработка данных.

Частота варианты– числовая характеристика, показывающая часть или долю которую составляет варианта от всей выборки, которая рав

$$\text{Частота варианты} = \frac{\text{Кратность варианты}}{\text{Обьем выборки}}$$

$$\text{Частота варианты \%} = \frac{\text{Кратность варианты}}{\text{Обьем выборки}} \cdot 100\%$$

Перепишем нашу вторую таблицу с учетом частот и объема выборки:

Количество Забитых Голов	3	4	5	6	7	8	Объем
Количество игроков забивших голы	1	2	2	2	2	1	10
Частота	0,1	0,2	0,2	0,2	0,2	0,1	
Частота, %	10%	20%	20%	20%	20%	10%	

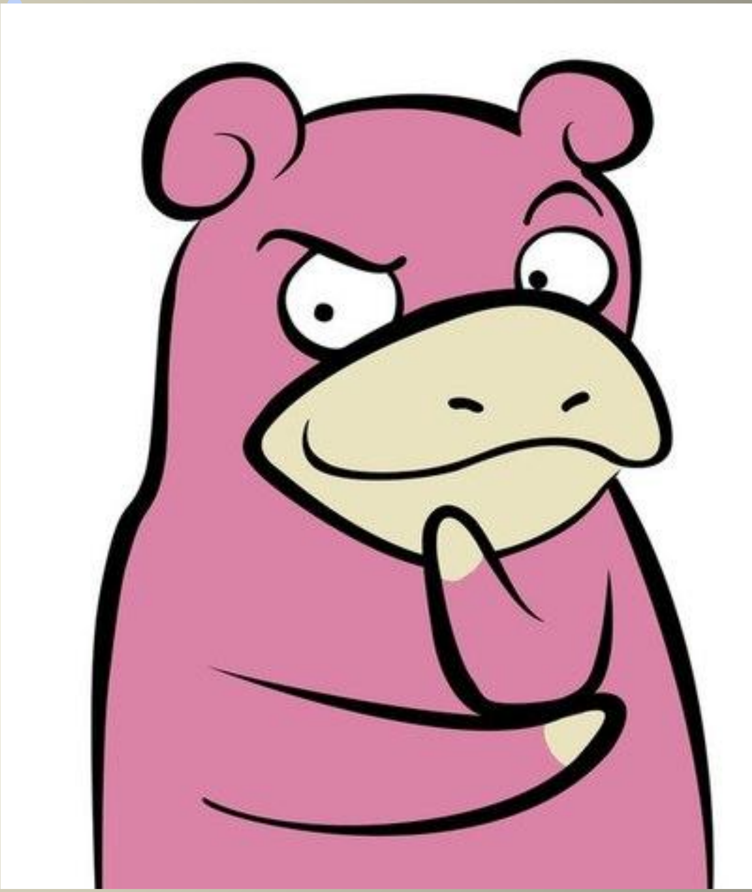
Сумма всех частот всегда равна 1, а сумма частот в процентах всегда равна 100%.

Статистическая обработка данных.

Вернемся к среднему значению, данная числовая характеристика часто является очень полезной.

Но не во всех задачах имеет смысл ее вычислять.

В нашем примере эта числовая характеристика показывала, сколько в среднем забивает команда. Со временем можно делать выводы об эффективности или неэффективности методов тренировки. Если среднее значение забитых голов растет, то видимо и тренировка эффективна, если не растет, а даже падает то видимо, методы тренировки неэффективны.



Статистическая обработка данных.

Одной из наиболее распространённых характеристик выборки значений случайной величины, чьё распределение по вероятностям известно, является ***математическое ожидание (М).***

Пусть распределение по вероятностям *P* значений некоторой случайной величины *X* задано таблицей:

X	X_1	X_2	X_3	…	X_{n-1}	X_n
P	P_1	P_2	P_3	…	P_{n-1}	P_n

тогда математическое ожидание вычисляется по формуле:

$$M = X_1P_1 + X_2P_2 + X_3P_3 + \ldots + X_{n-1}P_{n-1} + X_nP_n,$$

Статистическая обработка данных.

Рассмотрим на примере вычисление математического ожидания:

X	*1*	*2*	*3*	*4*	*5*	*6*
P	*0,1*	*0,25*	*0,3*	*0,2*	*0,1*	*0,05*

Применим формулу вычисления математического ожидания

$$\boldsymbol{M = X_1P_1 + X_2P_2 + X_3P_3 + \ldots + X_{n-1}P_{n-1} + X_nP_n,}$$

$M = 1\cdot 0{,}1 + 2\cdot 0{,}25 + 3\cdot 0{,}3 + 4\cdot 0{,}2 + 5\cdot 0{,}1 + 6\cdot 0{,}05 = 3{,}1$

Ответ: М = 3,1

Статистическая обработка данных.

Еще одна важная числовая характеристика – **Дисперсия**, или **разброс значений** вокруг среднего значения. Чем меньше дисперсия, тем плотнее результаты эксперимента сосредоточены около своего среднего значения.

Подсчет дисперсии довольно таки трудоемкая операция, опишем алгоритм поиска дисперсии.

Пусть нам даны данные измерений: $x_1, x_2, \ldots, x_n$. Найдём

1. среднее значение X_{cp}

$$X_{\text{ср}} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

2. отклонение данных от среднего: $x_1 - X_{cp}$, $x_2 - X_{cp}$, …, $x_n - X_{cp}$;

3. квадраты отклонений найденных на предыдущем шаге: $(x_i - X_{cp})^2$, $(i = 1,\ldots, n)$;

4. среднее значе… …***ия D***

$$D = \frac{(x_1 - X_{\text{ср}})^2 + (x_2 - X_{\text{ср}})^2 + \cdots + (x_n - X_{\text{ср}})^2}{n}$$

$$\delta = \sqrt{D}$$

- среднее квадратическое отклонение

Статистическая обработка данных.

Давайте вычислим дисперсию для нашего примера:

1. Вспомним, среднее значение у нас равнялось 5.5
2. Вычислим каждое отклонение и квадрат отклонения

Номер Игрока	2	3	4	5	6	7	8	9	10	11
Количество Забитых Голов	3	7	6	5	5	4	4	7	8	6
Отклонение от среднего	-2,5	1,5	0,5	-0,5	-0,5	-1,5	-1,5	1,5	2,5	0,5
Квадрат отклонения	6,25	2,25	0,25	0,25	0,25	2,25	2,25	2,25	6,25	0,25

3. Вычислим дисперсию

$$D = \frac{6{,}25 + 2{,}25 + 0{,}25 + 0{,}25 + 0{,}25 + 2{,}25 + 2{,}25 + 2{,}25 + 6{,}25 + 0{,}25}{10} = 2{,}25$$

Статистическая обработка данных.

Методы математической статистики позволяют обрабатывать практически любые данные, главное подходить к обработке данных обдуманно и исходя из здравого смысла.

СПАСИБО ЗА ВНИМАНИЕ !