Нормальный закон распределения и его применение

Нормальное распределение (распределение Гаусса) характеризуется тем, что крайние значения признака в нем встречаются достаточно редко, а значения, близкие к средней величине – достаточно часто.

Это распределение следует закону, открытому тремя учеными в разное время: Муавром в 1733 г. в Англии, Гауссом в 1809 г. в Германии и Лапласом в 1812 г. во Франции.

Нормальным такое распределение называется потому, что оно очень часто встречалось в естественно-научных исследованиях и *казалось "нормой" всякого массового случайного проявления признаков.*

Оно применимо только для метрических данных!

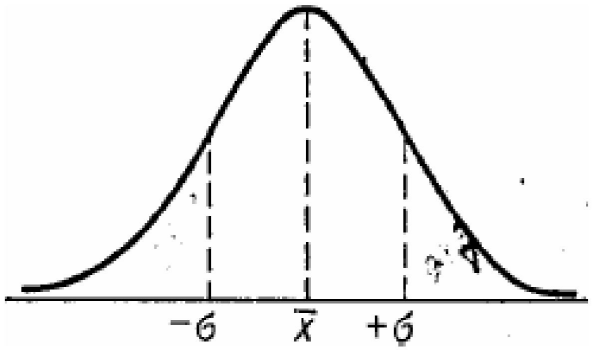
Это распределение описывается формулой:

$$f_{отн} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - M)^2}{2\sigma^2}}.$$

где f отн. – относительные частоты появления каждого конкретного значения случайной величины xi. Предполагается, что переменная xi, может принимать бесконечно большие и бесконечно малые значения, количество измерений бесконечно.

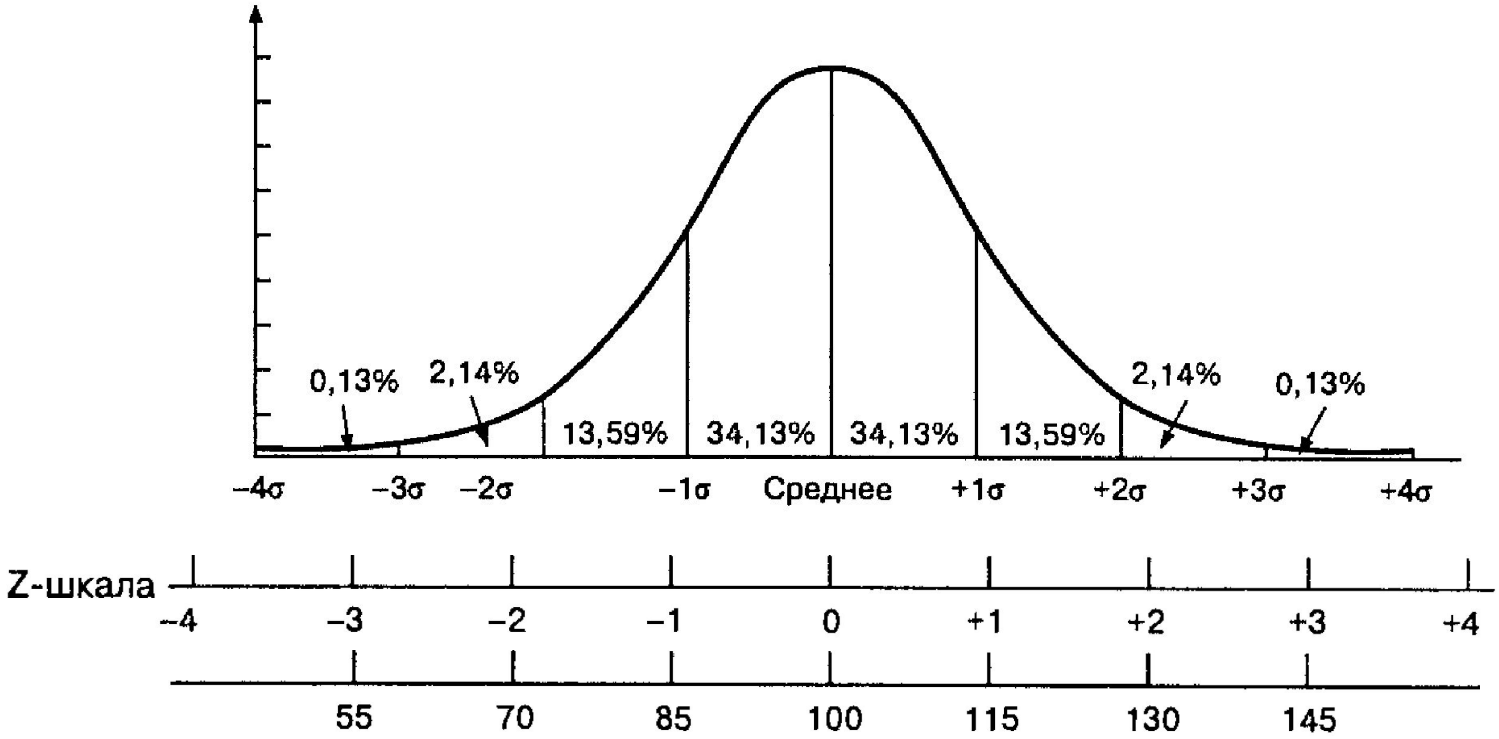
Нормальное распределение

График нормального распределения представляет собой колоколообразную кривую (симметричен относительно среднего арифметического значения).



Характерное свойство нормального распределения состоит в том, что 68,26% из всех его наблюдений всегда лежат в диапазоне «плюс - минус» одно стандартное отклонение от среднего арифметического (какова бы ни была величина стандартного отклонения). 95,44% - в пределах двух стандартных отклонений и 99,72% - в пределах трех стандартных отклонений.

Нормальное распределение



$M \pm \sigma$ соответствует **≈68%** (точно — 68,26%) площади;

$M \pm 2\sigma$ соответствует **≈95%** (точно — 95,44%) площади;

$M \pm 3\sigma$ соответствует **≈100%** (точно — 99,72%) площади.

90% всех случаев располагается в диапазоне значений $M \pm 1{,}64\sigma$;

95% всех случаев располагается в диапазоне значений $M \pm 1{,}96\sigma$;

99% всех случаев располагается в диапазоне значений $M \pm 2{,}58\sigma$.

Проверка нормальности распределения

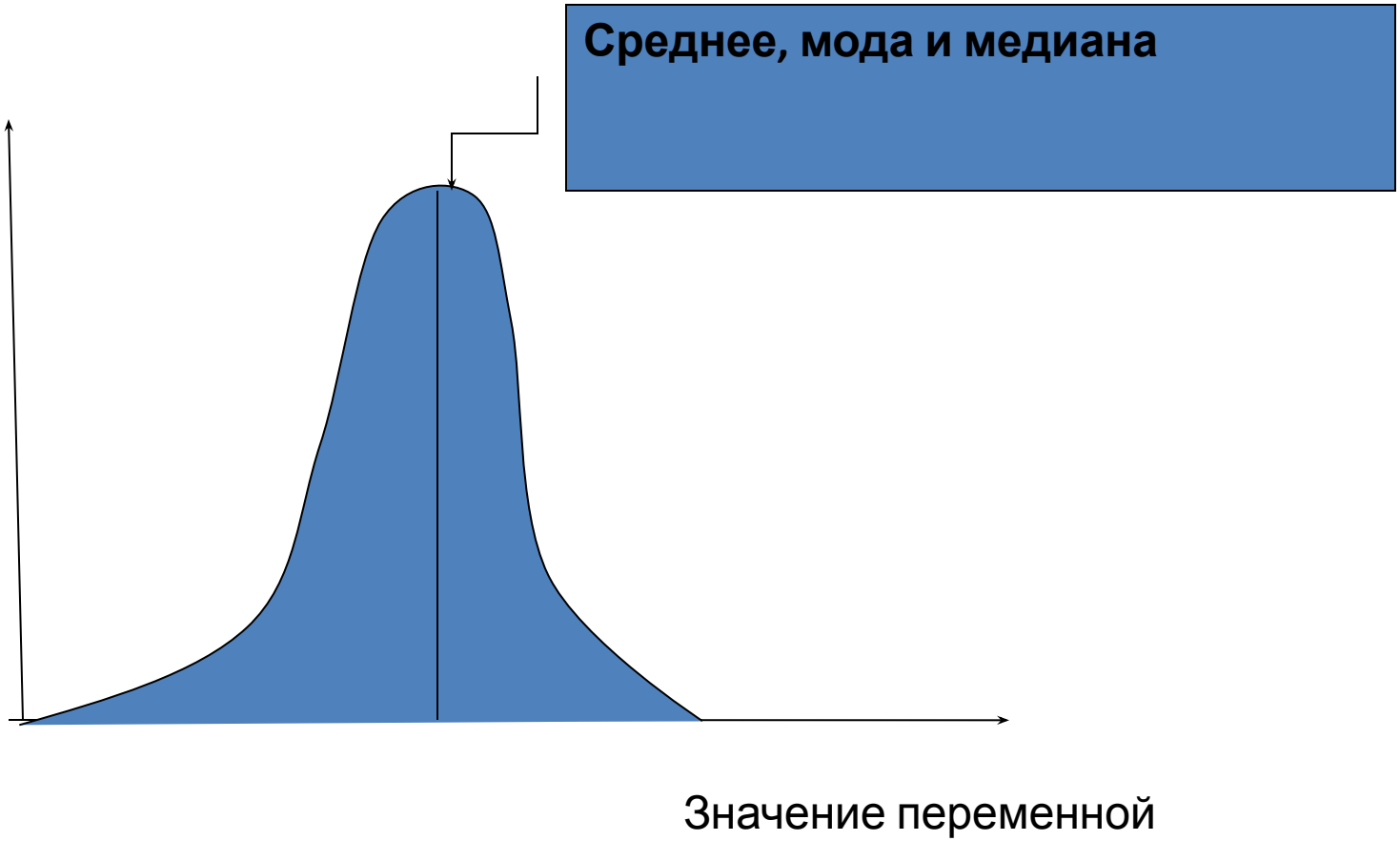
1. Среднее арифметическое, мода и медиана равны.

2. Нормальность распределения результативного признака можно проверить путем расчета показателей асимметрии и эксцесса и сопоставления их с критическими значениями (формулы Н.А. Плохинского и Е.И. Пустыльника).

3. Нормальным распределением может быть только распределение с числом наблюдений не менее 30 (при наличии и других условий соответствий).

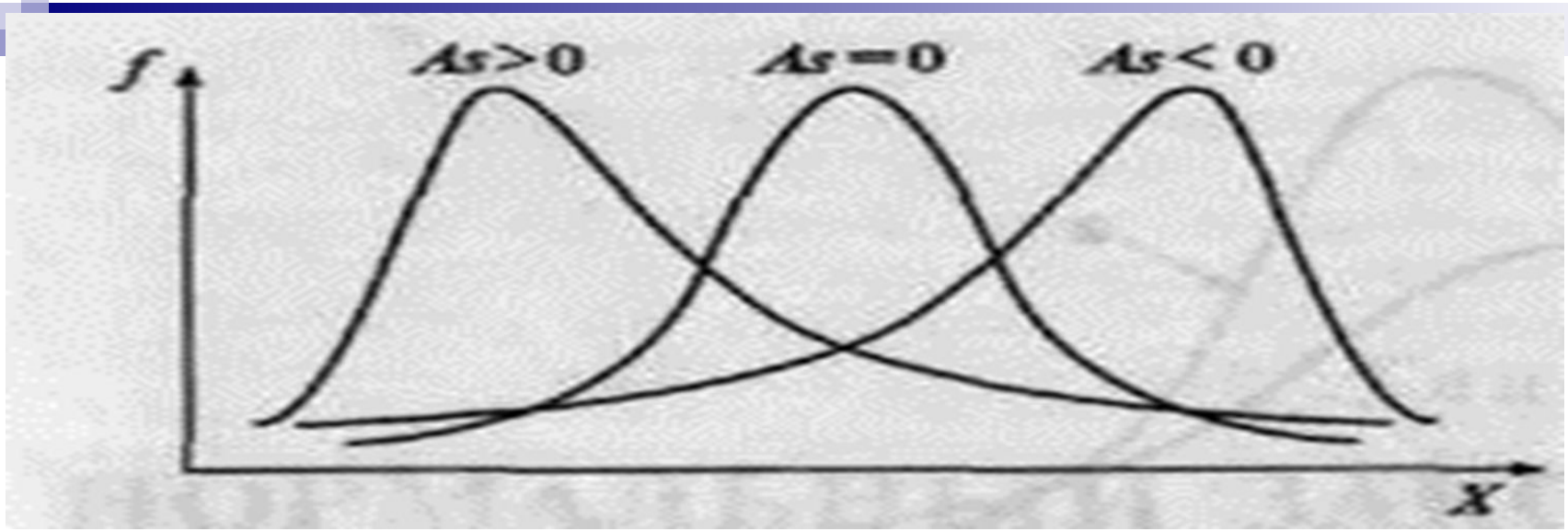
Нормальное распределение

- Частота



Меры распределения

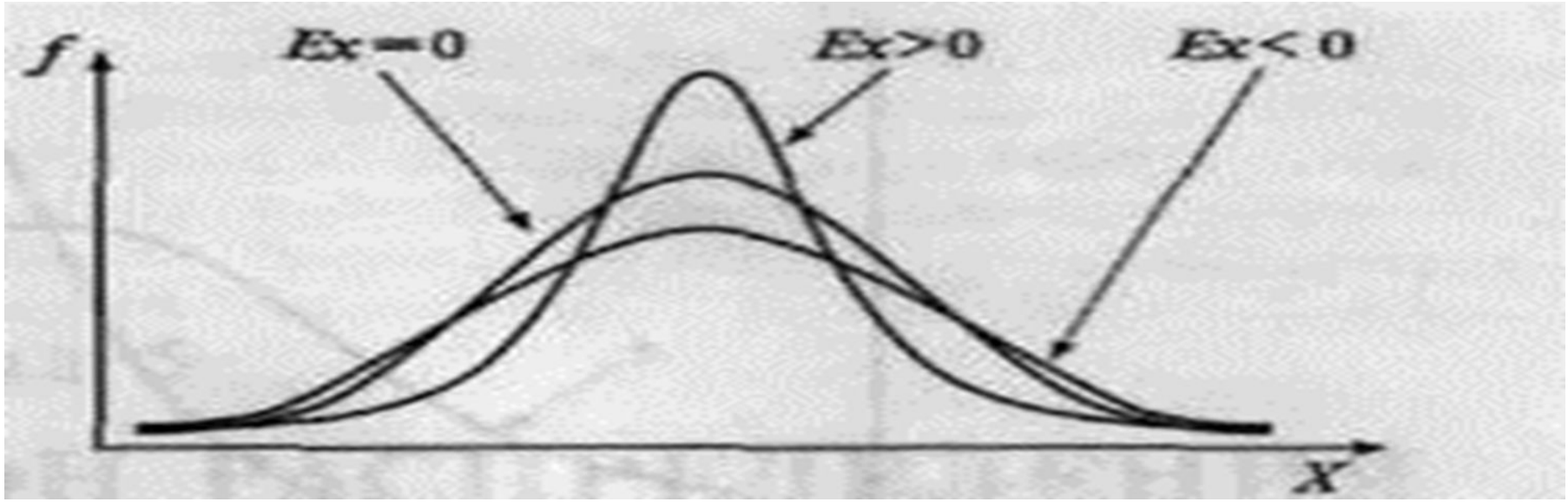
- Асимметрия
- Эксцесс



Если чаще встречаются **значения меньше среднего**, то говорят о **левосторонней**, или **положительной асимметрии** *(As > 0)*.

Если же чаще встречаются **значения больше среднего**, то **асимметрия – правосторонняя**, или **отрицательная** *(As<0)*.

Чем больше отклонение от нуля, тем больше асимметрия.



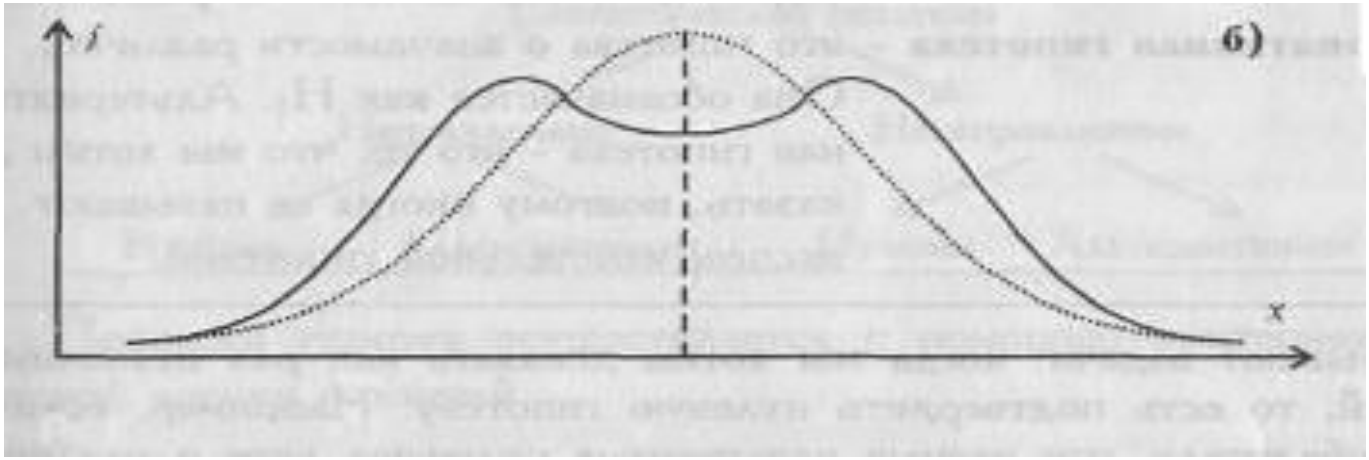
Островершинное распределение характеризуется положительным эксцессом *(Ex* > 0)

Плосковершинное - отрицательным (*Ex* < 0)

«**Средневершинное» (нормальное)** распределение имеет нулевой эксцесс *(Ex* = 0).

Эксцесс (E, Ex)

Если в распределении преобладают крайние значения, причем одновременно и более низкие, и более высокие, то такое распределение характеризуется отрицательным эксцессом и в центре распределения может образоваться впадина, превращающая его в **двувершинное**:



Асимметрия, эксцесс

Формула показателя асимметрии следующая:

$$As = \frac{\sum_i (x_i - M)^3}{n\sigma^3}.$$

Показатель эксцесса определяется по формуле:

$$Ex = \frac{\sum_i (x_i - M)^4}{n\sigma^4} - 3.$$

Проверка нормальности распределения

Рассмотрим применение **метода Е.И. Пустыльника.**

Действовать будем *по следующему алгоритму:*

1) рассчитаем критические значения показателей асимметрии и эксцесса по формулам Е.И. Пустыльника и сопоставим с ними эмпирические значения;

2) если эмпирические значения показателей окажутся ниже критических, сделаем вывод о том, что распределение признака не отличается от нормального.

Формулы для определения критических значений асимметрии и эксцесса (формулы Е.И. Пустыльника):

$$A_{кр} = 3 \cdot \sqrt{\frac{6 \cdot (n-1)}{(n+1) \cdot (n+3)}}$$

$$E_{кр} = 5 \cdot \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n+1)^2 \cdot (n+3) \cdot (n+5)}}$$

где n - количество наблюдений.

Для обработки данных понадобятся такие последовательные шаги: вычисление *Mx, σ, A, E* и подсчет *n.*

|A|<Акр } → **распределение совпадает с нормальным**
|E|<Екр}

Проверка нормальности распределения (пример)

№	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
1	11	0,94	0,884	0,831	0,781
2	13	2,94	8,644	25,412	74,712
3	12	1,94	3,764	7,301	14,165
4	9	−1,06	1,124	−1,191	1,262
5	10	−0,06	0,004	−0,000	0,000
6	11	0,94	0,884	0,831	0,781
7	8	−2,06	4,244	−8,742	18,009
8	10	−0,06	0,004	−0,000	0,000
9	15	4,94	24,404	120,554	595,536
10	14	3,94	15,524	61,163	240,982
11	8	−2,06	4,244	−8,742	18,009
12	7	−3,06	9,364	−28,653	87,677
13	10	−0,06	0,004	−0,000	0,000
14	10	−0,06	0,004	−0,000	0,000
15	5	−5,06	25,604	−129,554	655,544
16	8	−2,06	4,244	−8,742	18,009
Суммы	161		102,944	30,468	1725,467

Проверка нормальности распределения (пример)

Для расчетов в таблице, необходимо значение среднего арифметического, которое вычисляется по формуле:

$$\bar{x} = \frac{\sum x_i}{n}$$

где x_i - каждое наблюдаемое значение признака;

n - количество наблюдений.

В данном случае:

$$\bar{x} = \frac{161}{16} = 10{,}06$$

Стандартное отклонение (сигма) вычисляется по формуле:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

где x_i - каждое наблюдаемое значение признака;

$\bar{x}$ - среднее значение (среднее арифметическое);

n - количество наблюдений.

В данном случае:

$$\sigma = \sqrt{\frac{102{,}944}{16-1}} = \sqrt{6{,}893} = 2{,}62$$

Проверка нормальности распределения (пример)

Подставляя в формулы для расчета А и Е полученные значения n, σ и соответствующие значения из таблицы, получаем:

$$A = \frac{+30{,}468}{16 \cdot 2{,}62^3} = +0{,}106$$

$$E = \frac{1725{,}467}{16 \cdot 2{,}62^4} - 3 = -0{,}711$$

Теперь рассчитаем критические значения для показателей A и Е по формулам Е.И. Пустыльника:

$$A_{кр} = 3 \cdot \sqrt{\frac{6 \cdot (n-1)}{(n+1) \cdot (n+3)}}$$

$$E_{кр} = 5 \cdot \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n+1)^2 \cdot (n+3) \cdot (n+5)}}$$

где n - количество наблюдений.

Проверка нормальности распределения (пример)

В данном случае:

$$A_{кр} = 3 \cdot \sqrt{\frac{6 \cdot (16-1)}{(16+1) \cdot (16+3)}} = 3 \cdot \sqrt{\frac{90}{323}} = 1{,}58$$

$$E_{кр} = 5 \cdot \sqrt{\frac{24 \cdot 16 \cdot (16-2) \cdot (16-3)}{(16+1)^2 \cdot (16+3) \cdot (16+5)}} = 5 \cdot \sqrt{\frac{69888}{115311}} = 3{,}89$$

$A_{эмп}$=0,106

$A_{эмп} \angle A_{кр}$

$E_{эмп}$=-0,711

$E_{эмп} \angle E_{кр}$

Так как эмпирические значения А и Е меньше критических значений, то можно сделать следующий вывод: распределение результативного признака в данном примере не отличается от нормального распределения.

Процедура стандартизации

Приведение распределения к стандартной форме. Любое множество значений показателя со средним значением *Mx* и стандартным показателем σ можно преобразовать в другое множество, среднее значение которого равно 0, а стандартное отклонение - равно1.

Необходимость в таком преобразовании возникает когда требуется сопоставить значения показателей, имеющих разную размеренность, т.е. измеренных по шкалам с различными единицами измерения (баллы, секунды, см и т. д.).

Такое преобразование называется **стандартизация или нормирование** и позволяет получить стандартизированные или нормированные значения исходных данных.

Стандартизация (нормирование) осуществляется по формуле:

$$Zi=\frac{Xi\ -Mx}{\sigma}$$

где Zi – стандартная тестовая оценка i-го испытуемого,

Xi – нормальная оценка i-го испытуемого.

Смысл этой процедуры состоит в том, что на шкале интервалов **вводится новая единица измерения**, равная значению средне квадратичного отклонения σ и исходное значение показателя xi и его отклонения от среднего $xi\ -Mx$ начинают измеряться в единицах этого средне квадратичного отклонения.

Чтобы избежать дробных и отрицательных значений z используют линейное преобразование значений показателя:

$$Xi\ -Mx$$

Из наиболее известных шкал, образованных путем указанной процедуры стандартизации исходных значений показателя:

а) **шкала Гилфорда**, построенная им для оценки интеллекта:

$$IQ = 100 + 15 * \frac{Xi - Mx}{\sigma}$$

б) **шкала Векслера**, которая была построена для этих же целей:

$$IQ = 10 + 3 * \frac{Xi - Mx}{\sigma}$$

в) шкала общего значения Мак-Колла (**шкала Т-баллов**):

$$T = 50 + 10 * \frac{Xi - Mx}{\sigma}$$

Эти преобразования корректны только для статистически

В проведенном школьном обследовании по следующим методикам (логического мышления, воображения, объема памяти, общительность) ученик получил следующие результаты (см. таблицу).

Рассчитайте Т-баллы данного ученика и постройте его индивидуально-психологический профиль.

Методика измерения:	**Индивидуальные показатели ученика**	**Среднее значение по группе**	**σ по группе**
Воображение	13	10,2	2,3
Логическое мышление	92	103	12,4
Объем памяти	6	5,4	1,3
Общительность	6	7,7	1,6

Статистическая норма

Принято считать, что в пределах $Mx \pm 2\sigma$ располагаются значения, относящиеся к статистической норме, то есть те значения, которые включены в так называемый 95%-ный доверительный интервал. Знание Mx и σ можно использовать для выведения статистической нормы.

Обязательные для этой процедуры условия: *соответствие распределения нормальному и* $n \geq 30$.

Например, необходимо определить границы нормы для российской выборки у переведенного недавно с английского языка теста. После перевода и адаптации мы проводим исследование на оптантах, чьим родным языком является русский.

По окончании обработки результатов получаем: $n = 80$, $Mx = 30$, $\sigma = 5{,}9$.

Границы статистической нормы для теста лежат в диапазоне $Mx \pm 2\ \sigma$, то есть $30 \pm 11{,}8$. Таким образом, верхняя граница нормы = 18, нижняя = 42.

Схема деления выборки на подгруппы

- Деление выборки **на три подгруппы.**

Первая центральная подгруппа образуется из испытуемых, имеющих значение показателя в пределах $Mx \pm \sigma$. Во вторую подгруппу выделяются испытуемые со значениями показателя, превышающего $Mx + \sigma$. Третью группу образуют испытуемые, у которых значение показателя ниже $Mx - \sigma$.

Значения показателей центральной подгруппы испытуемых рассматривают в качестве **нормы**; второй и третье подгрупп – соответственно, **выше и ниже нормы**.

Другой распространённой шкалой являются **стены**, для которых *Mx* = **5,5 и σ = 2** (стен – от англ. sten, сокр. standart ten – стандартная десятка).

Для перевода в стены можно использовать формулу стандартизации, но чаще всего используют более формальную процедуру: находят среднее (*Mx*), стандартное отклонение (σ), от среднего в обе стороны отсчитывают по пять интервалов по σ/2 (половине σ). Получившиеся 10 интервалов и являются стенами.

Таким образом, **первый стен** получен при *Mx* - 2 ·σ , а **10 стен** – при *Mx* + 2 ·σ. К среднему диапазону принято относить стандартные оценки **от 4 до 7 стенов**. Говорить о значимых отклонениях, выходящих за границы средней нормы, можно при получении стандартных оценок до 3

Выбор типа шкалы зависит от исходных данных.

Если сырой балл принимает значения от 0 до 100 и мы стандартизируем его в стены, то явно теряем слишком много информации, т.к. внутри одного стандартного интервала может находиться достаточно много сырых баллов. Это неприемлемо.

Поэтому, при большом диапазоне сырых баллов используются Т-баллы. В тестах интеллекта традиционно используется IQ, если интервал значений сырых баллов невелик, то можно использовать стены.

Нормализация исходных данных

Процедура приведения распределения к нормальному виду носит название **нормализация**, а преобразованные исходные данные называются **нормализованными.**

Нормализованные значения могут быть найдены с помощью таблиц, в которых приводится процент случаев (процентили) разных отклонений в единицах σ от среднего значения для нормальной кривой.

Алгоритм: сначала определяется процент испытуемых в исследуемой выборке с тем же или более высоким исходным значением показателя (вычисляются соответствующие кумуляты распределения - распределение признака в вариационном ряду по накопленным частотам). Затем этот процент отыскивается в таблице нормального распределения частот и по нему находится соответствующее значение нормализованного стандартного показателя. Далее распределению этих нормализованных значений путем соответствующего линейного преобразования можно придать любую удобную для последующего анализа форму.

Примеры

1. Процедура нормализации исходного распределения испытуемых по возрасту.

Возраст испытуемого (лет)	17	18	19	20	21	22	23
Кол-во испытуемых данного возраста	2	15	14	6	6	5	2
Доля испытуемых данного возраста (%)	4	30	28	12	12	10	4
Кумулята распределения (%)	4	34	62	74	86	96	100

С помощью таблицы соответствия процентилей и нормированных значений z для нормального распределения по значениям кумулят находим соответствующие нормализованные значения возраста:

Нормализованные значения Z	-1,75	-0,41	0,31	0,64	1,08	1,75	3,09

Преобразуем полученные значения z в более удобные значения Т-баллов: $T = 50 + 10 * z$ (при нормальном распределении Mx=0, σ =1).

Значения Т -баллов	32,5	45,9	53,1	56,4	60,8	67,5	80,9