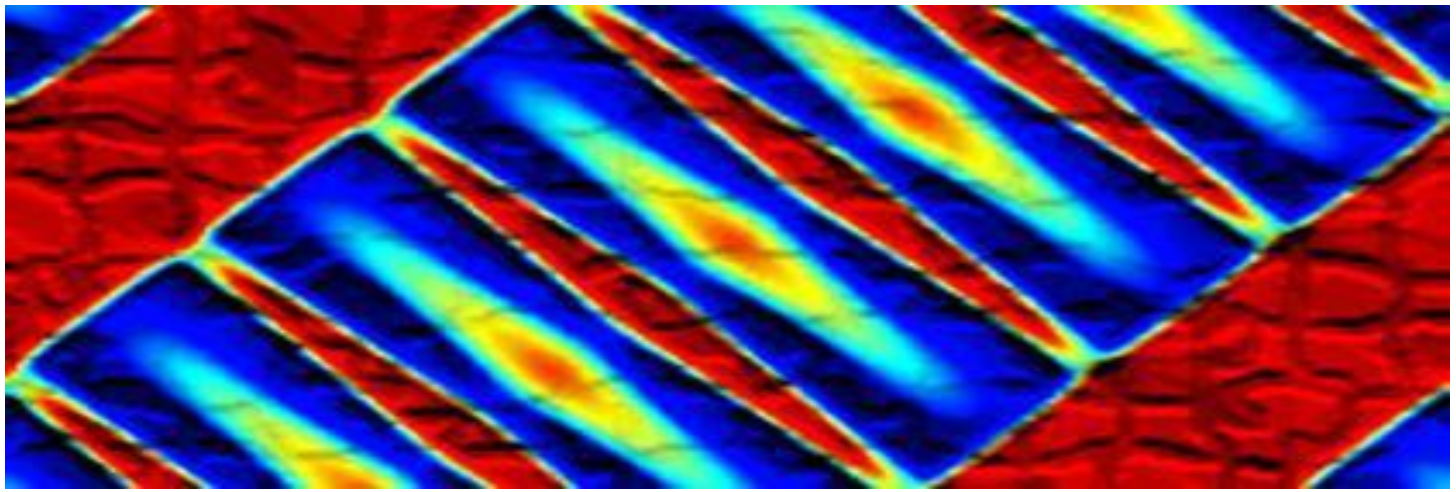
Информационные технологии в биологических исследованиях

Раздел: «Информационные технологии и математическая обработка результатов биологического эксперимента»

Лекция 2: «Первичный анализ и обработка данных»



Базовые понятия и операции первичной обработки экспериментальных данных

1. Распределения, их виды и характеристики
2. Оценка сильно отклоняющихся значений
3. Основные параметры совокупности – средняя, арифметическая, ошибка средней, достоверность
4. Мера варьирования величин – среднеквадратичное отклонение, коэффициент вариации
5. Оценка репрезентативности выборки
6. Некоторые конкретные примеры

Базовые понятия и операции первичной обработки экспериментальных данных

- В биологических исследованиях основной интерес представляют сведения, относящиеся не к индивидуальному объекту, а к целой группе или некоторому статистическому среднему объекту.

- Необходимость использования статистических методов в биологических исследованиях связана с тем, что свойства биологических объектов варьируют в пределах популяции, а физиологические и другие параметры одной особи испытывают флуктуации во времени.

Статистическая совокупность – это и объекты исследования и полученные данные

Объекты каждого исследования (растения, животные, микроорганизмы, урожаи с опытных делянок или вегетационных сосудов, образцы плодов, семян и пр.) образуют общую, или генеральную, совокупность.

Термин совокупность относят и к полученным в опыте или путем наблюдений числам, характеризующим с какой-либо одной количественной стороны объекты, входящие в данную генеральную совокупность.
В статистическую совокупность следует включать лишь числа, относящиеся к качественно однородным **признакам** (свойствам) объекта исследования.

Признаки (их количественная мера, варианта) варьируют случайным образом по причине естественной изменчивости и ошибок измерений

Основное – естественная изменчивость, вызванная биологическими причинами

Характер самого наблюдаемого явления, особенности причин, вызывающих колебания данного признака определяют особенности колебаний данных.

Вычисления можно проводить как угодно точно, но результат вычисления не может быть точнее тех данных, на которых оно основано

Распределения

Чаще всего в природе наблюдается закономерность: большие по величине колебания данных встречаются значительно реже, чем меньшие по величине

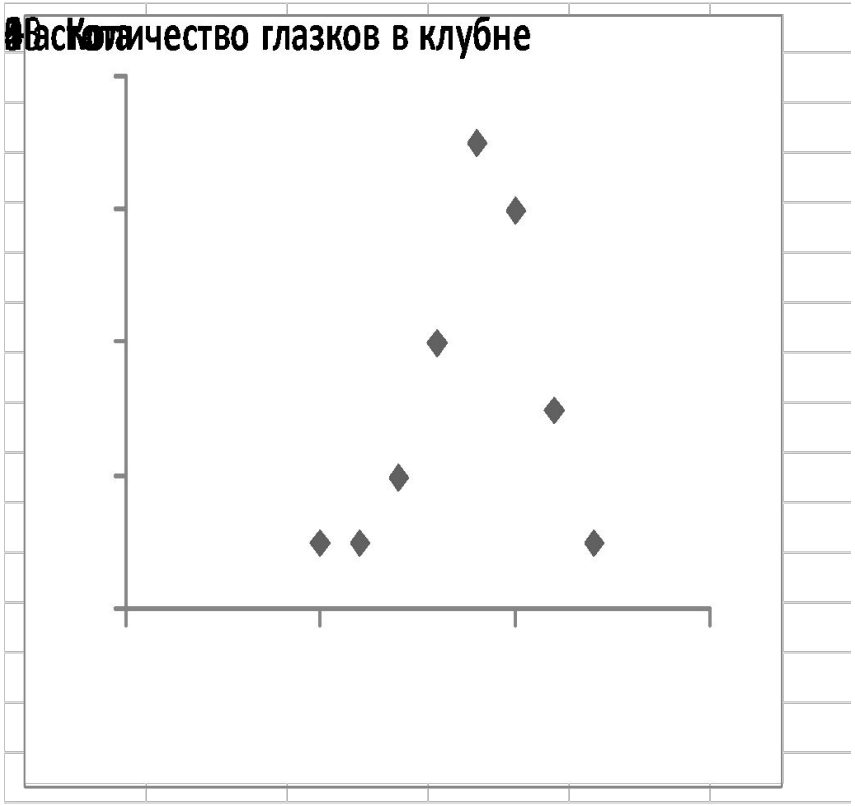
Большинство членов статистической совокупности оказывваются среднего или близкого к нему размера. Чем дальше они отстоят от среденего уровня, тем реже стречаются.

Существует связь между числовыми значениями варьирующих признаков и частотой их встречаемоссти в данной совокупности - это и есть распределение

Пример распределения

Вариационный ряд

В случае, если глубина выборки, т. е. количество чисел, полученных в результате измерений, невелико, можно составить вариационный ряд

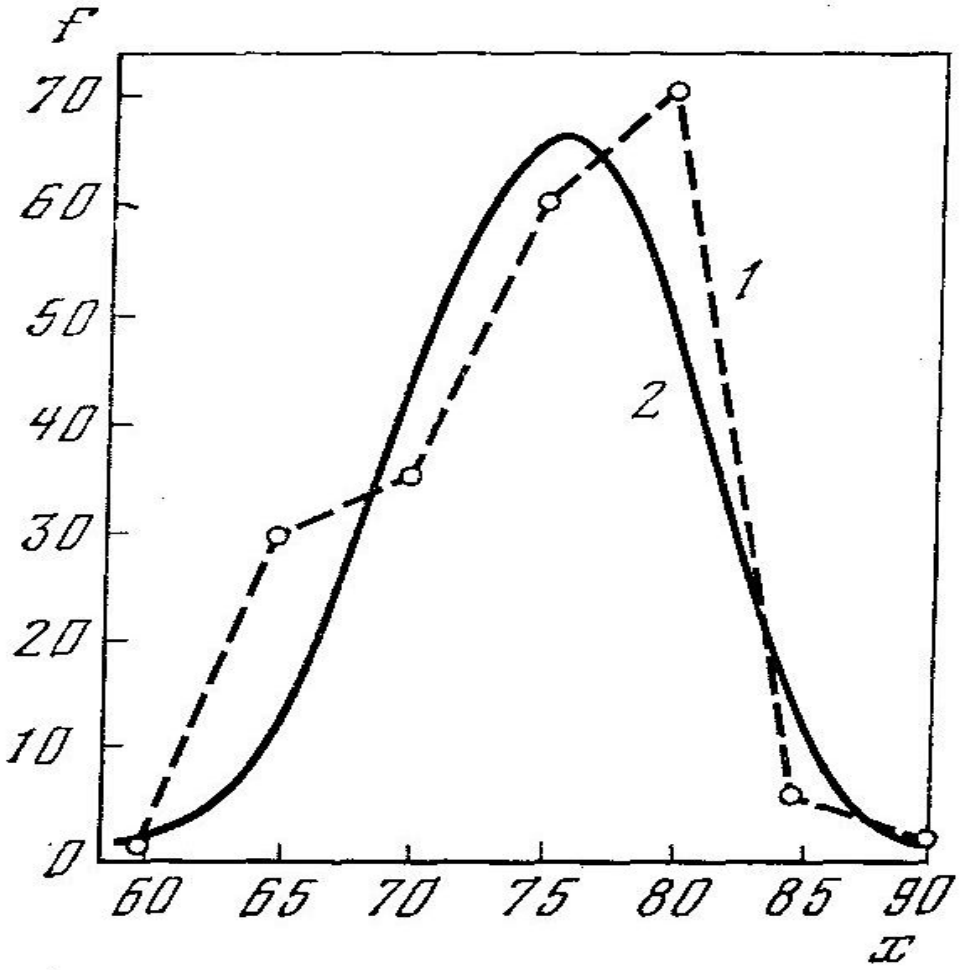


Например, подсчет количества глазков в 25 клубнях картофеля.
Всего:
6, 9, 5, 7. 10, 8 9 10 8 11, 9. 12, 9, 8, 10, 11, 9, 10, 8, 10, 7, 9, 11, 9, 10.

Варианты, x	6	9	5	7	10	8	11	12
Число вариант, f	1	7	1	2	6	4	3	1

Непрерывное распределение

где f' частоты
нормальной кривой;
x — варианты
(середины классов)
ряда

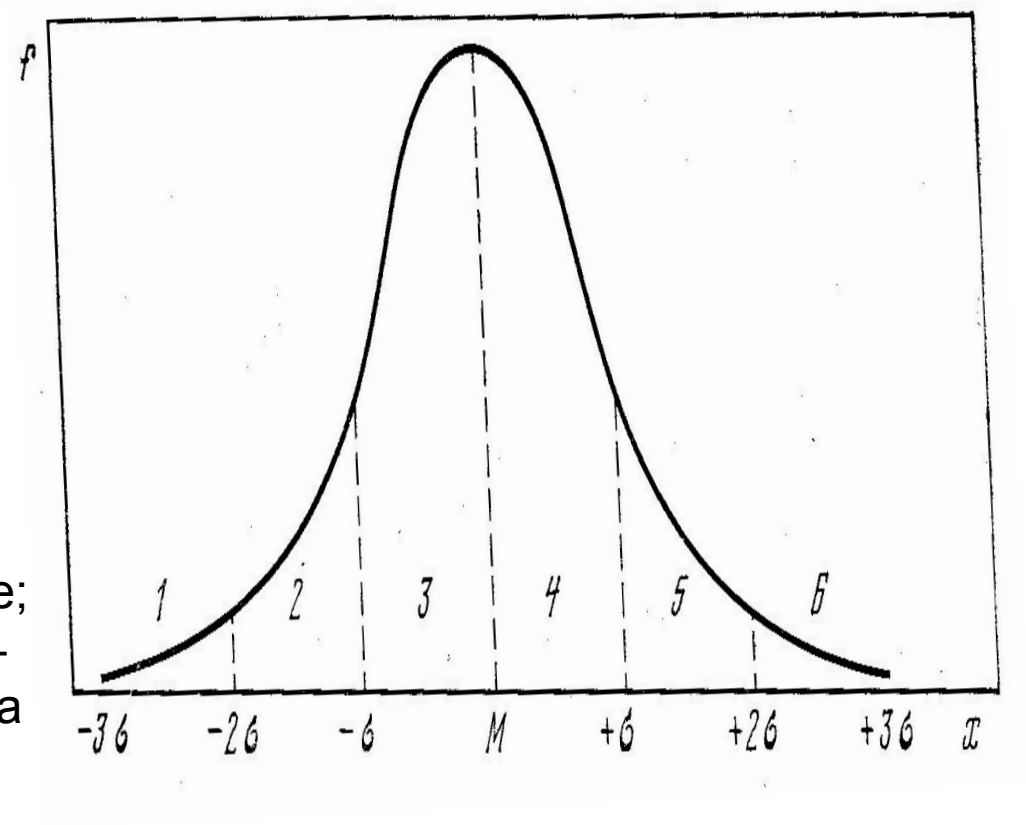


Нормальное распределение

Распределение – это соотношение между значениями случайной величины и частотой их встречаемости. Большое число случайных величин, распространенных в природе, может быть описано с помощью закона нормального распределения, который задается уравнением:

$$f' = \frac{Nc}{\sigma\sqrt{2\pi}} \cdot e^{-0,5t^2},$$

где f' — теоретические частоты нормальной кривой; *N* — объем выборки; *c* — классовый интервал; *σ* — среднее квадратическое отклонение; e — основание натуральных логарифмов; *t* =*(x—M)/ σ* — нормированное отклонение; *M* — средняя арифметическая; *x* — варианты (середины классов) ряда



Характеристики нормального распределения

Основные параметры нормального распределения – среднее арифметическое (**M**) и среднеквадратическое отклонение – сигма (**σ**)

На расстоянии $M + \sigma$ *и* $M - \sigma$ от среднего значения на графике нормальной кривой расположены абсциссы ее двух точек перегиба, которые показывают переход от **типичных** величин вариант совокупности к **нетипичным**, хотя и принадлежащих еще к данной совокупности.

В интервале **нормы**, между абсциссами, от $M - \sigma$ до $M + \sigma$ находится 68,27% всей площади нормального распределения, т. е. вариант, или дат совокупности; между $M - 2\sigma$ и $M+2\sigma$ заключается 95,45% дат от всего объема и в интервале от $M - 3\sigma$ до $M+3\sigma$ лежит 99,73% от всего объема нормально распределенной совокупности.

Биномиальное распределение

Относится к дискретным величинам, то есть к тем, которые могут быть представлены только целыми числами. Например, глазков в картофелине может быть только целое число и т.д.

В общем виде.

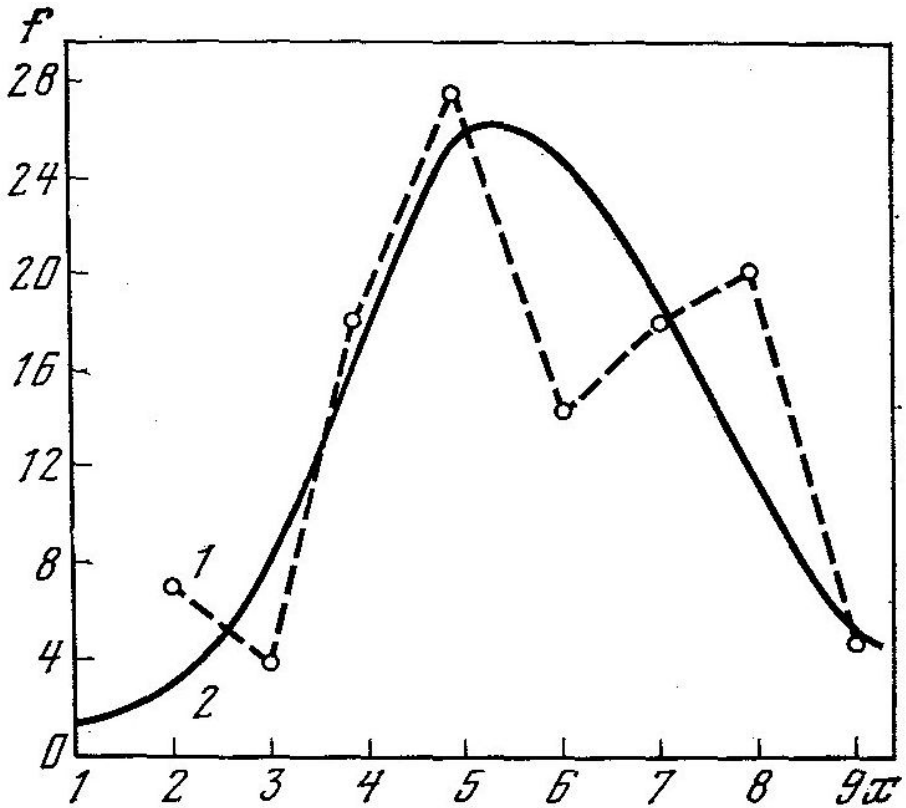
$$f' = \frac{N_n c \psi(t)}{\sigma}$$

Где f частоты,

$N_п$ – число проб,

t – нормированное отклонение, (x-M)/σ,

c – классовый интервал.



Характеристики биномиального распределения

Во многом близко к нормальному. Отличие состоит лишь в том, что оно характеризует поведение *дискретных признаков*, *выраженных целыми числами*.

Как правило, для описания биологических признаков подходит симметричное биномиальное распределение, у которого дисперсия много меньше средней.

Выборка при биномиальном распределении обычно образуется, когда берут $\boldsymbol{N_n}$ проб одинакового объема, равного $\boldsymbol{n}$

Вероятность появления события постоянна для кажлой пробы (лист растения либо заразится грибом, либо нет)

Два исхода – поэтому бином

Распределение Пуассона

Частный случай биномиального распределения:
Вариант описания стохастического поведения *дискретных количественных признаков* для случаев, когда *вероятность элементарных альтернативных событий неодинакова*, одно из них наблюдается заметно чаще другого ($p << q$).

Закон Пуассона описывает редкие события, происходящих 1, 2, 3 и т. д. раз на сотни и тысячи обычных событий.

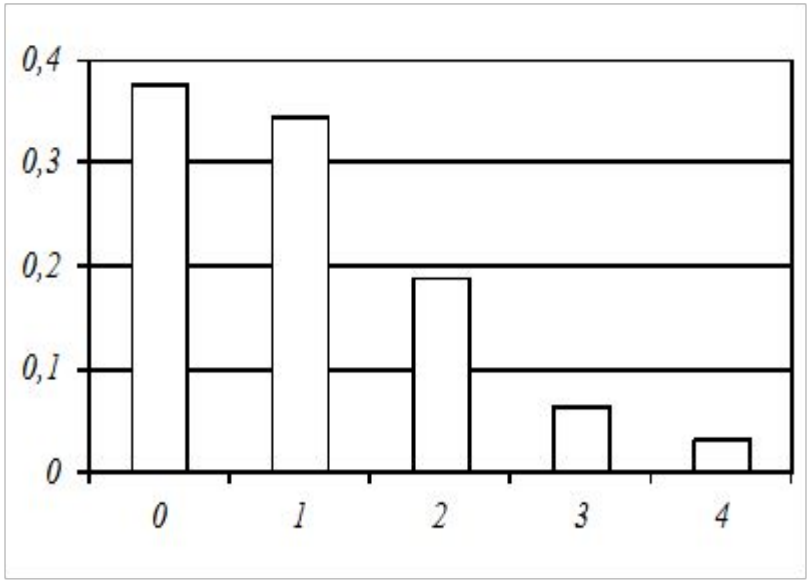
Примеры таких явлений – частота нарушений хромосомного аппарата на каждую тысячу митозов, встречаемость семян сорняка в большой серии навесок семян культурного растения, число повторных попаданий животных в ловушки.

Пример распределения Пуассона

Распределение Пуассона резко асимметрично, причем *дисперсия равна средней арифметической*, что может служить критерием для оценки характера распределения изучаемого признака .

Пример. В течение одного года пометили кольцами и выпустили на волю 32 птицы. В последующие пять лет часть из них отлавливали повторно: 7 экз. по одному разу, 7 – по два, 2 – по три, 1 экз. – четыре раза, 15 экз. окольцованных птиц повторно не попадались: $\bar{x} \cdot a$

Число повторных отловов, *x*	Число отловленных животных, *a*	Число случаев повторного отлова, x*a
0	**15**	**0**
1	**7**	**7**
2	**7**	**14**
3	**2**	**6**
4	**1**	**4**
n	**32**	**31**



Расчеты показали, что средняя арифметическая (*M*) примерно равна дисперсии (σ^2)

$$M = \frac{\sum x}{n} = \frac{31}{32} = m \cdot p = 4 \cdot 0.242 = \mathbf{0.968}$$ экз

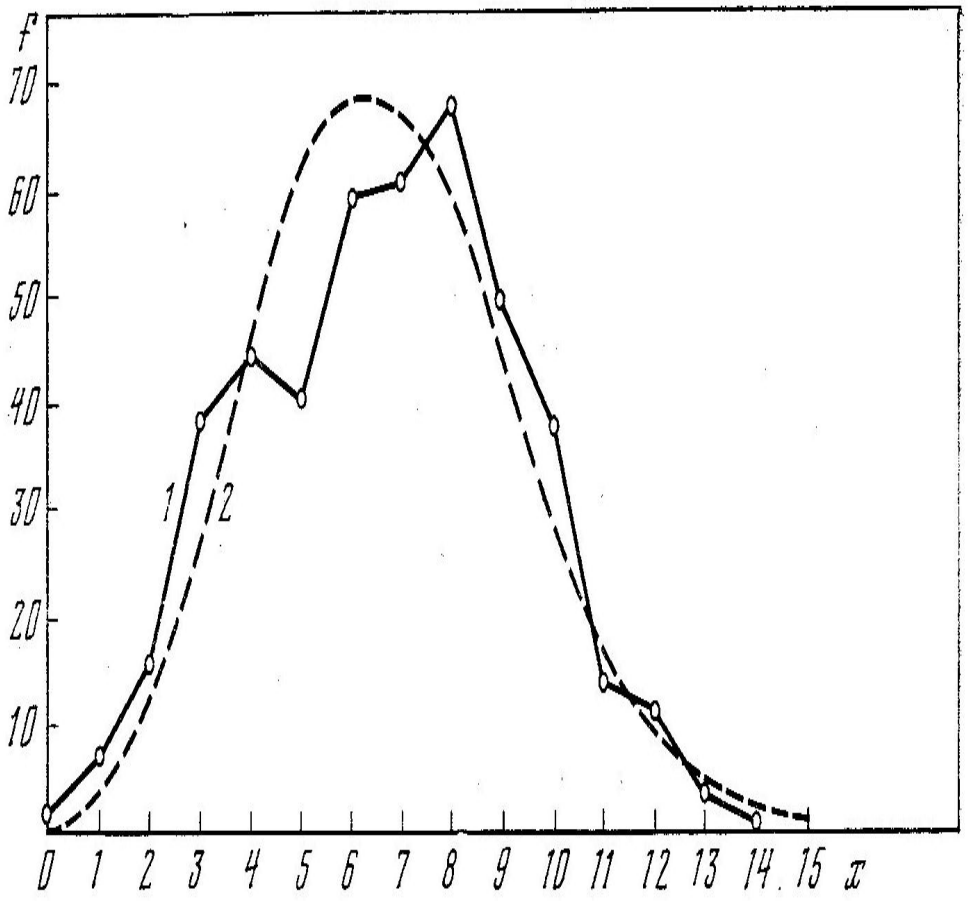
$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}} = \sqrt{\frac{69 - \frac{(32)^2}{32}}{(32-1)}} =$$ 1.121 экз., $\sigma^2 =$ **1.257**,

$$\boldsymbol{\sigma^2 \approx M}$$

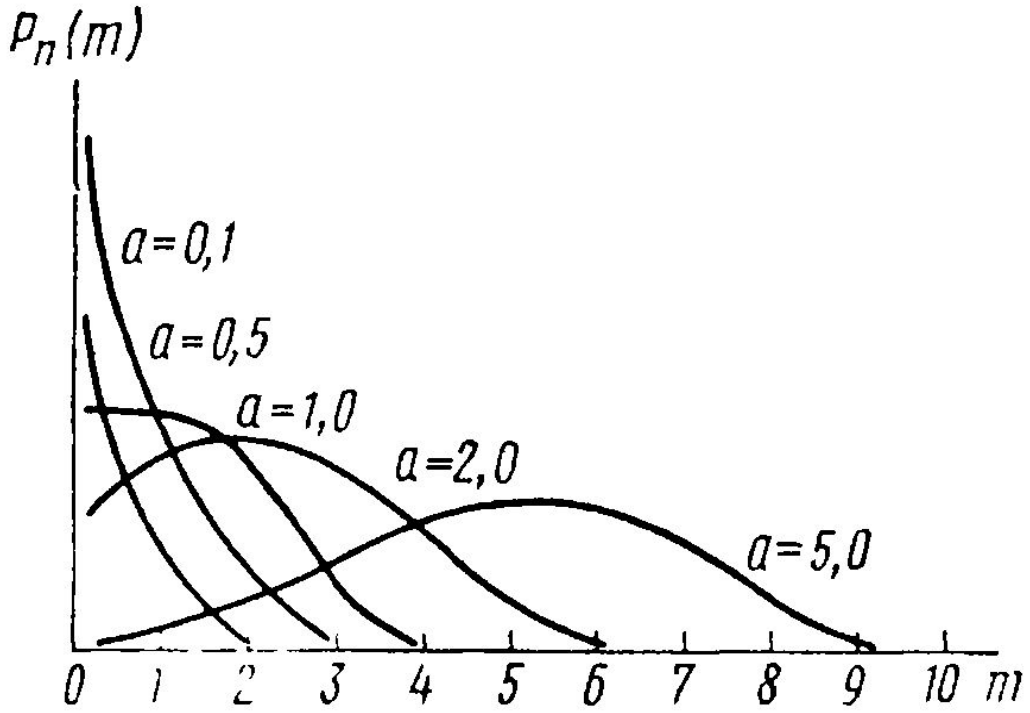
Распределение Пуассона

$$f' = \frac{M^x}{x!} N_n e^{-M},$$

где f' — теоретические частоты распределения Пуассона, т. е. число проб, обладающих той или иной долей наблюдаемого признака; *x* — варианты, отдельные значения наблюдаемого признака; *x*! — (икс-факториал) обозначает произведение ряда натуральных чисел, например: 3! = 1-2-3 = 6; *M* — средняя арифметическая данного ряда; *Nп* — общее число проб



При возрастании произведения np - (вероятная частота ожидаемого события) распределение Пуассона стремится к нормальному



Оценка сильно отклоняющихся вариант

Относится ли данная варианта вместе с другими вариантами изучаемой выборки к одной и той же генеральной совокупности или – к разным?

Сформировано ли данное значение варианты под действием тех же доминирующих и случайных факторов, что и все остальные варианты данной выборки, или это были иные факторы?

2 возможных ответа:

- 1. Факторы те же, т. е. все варианты взяты из одной и той же генеральной совокупности.
- 2. Факторы иные, т. е. особенная варианта и выборка порознь взяты из разных генеральных совокупностей

Ответ можно получить с использованием свойств нормального распределения

Если все варианты были взяты из одной генеральной совокупности, они должны отличаться друг от друга только в силу случайных причин и (с вероятностью $P = 0.95$) находиться в диапазоне $\boldsymbol{M \pm 2\,\sigma}$.

Эта величина, ***нормированное отклонение***, и служит безразмерной характеристикой отклонения варианты от средней арифметической:

$$t = \frac{x - M}{\sigma} \sim t_{табл}$$

,

где t – критерий выпада (исключения),
x – выделяющееся значение признака,
M – средняя величина для группы вариант,
$t_{табл.}$ – стандартные значения критерия выпадов, определяемые свойствами нормального распределения, их можно найти по таблице

Для больших выборок пользуются значением $t_{табл.} = 2$ при $P = 0.95$,

Значение критерия *t* для отбраковки «выскакивающих» вариант с известными параметрами распределения

n	α			n	α		
	0.05	**0.01**	**0.001**		**0.05**	**0.01**	**0.001**
5	3.04	5.04	9.43	20	2.15	2.93	3.98
6	2.78	4.36	7.41	25	2.11	2.85	3.82
7	2.62	3.96	6.37	30	2.08	2.80	3.72
8	2.51	3.71	5.73	35	2.06	2.77	3.65
9	2.43	3.54	5.31	40	2.05	2.74	3.60
10	2.37	3.41	5.01	45	2.04	2.72	3.57
11	2.33	3.31	4.79	50	2.03	2.71	3.53
12	2.29	3.23	4.62	60	2.02	2.68	3.49
13	2.26	3.17	4.48	70	2.01	2.67	3.46
14	2.24	3.12	4.37	80	2.00	2.66	3.44
15	2.22	3.08	4.28	90	2.00	2.65	3.42
16	2.20	3.04	4.20	100	1.99	2.64	3.41
17	2.18	3.01	4.13	0	1.96	2.58	3.29
18	2.17	2.98	4.07				

Когда параметры распределения неизвестны, можно использовать сравнение различий максимальной и минимальной вариант, «размах» значений ряда. Для этого существуют два критерия, для максимальной и минимальной вариант

Имеется ранжированный ряд, где представлена высота растений одного вида (в см)

82 **77** 74 74 73 66 64 63 63 62 54 **44** **43**

$$\vartheta_n = \frac{x_N - x_{N-1}}{x_N - x_2} \qquad \vartheta_n = \frac{82-77}{82-44} = 0{,}132$$

Для максимальной

Табличное значение критерия для N = 13 составляет **0,52 > 0,13**, т. е. больше, чем вычисленная величина. Варианту нельзя исключать из выборки.

$$\vartheta_1 = \frac{x_2 - x_1}{x_{N-1} - x_1} \qquad \vartheta_n = \frac{44-43}{77-43} = 0{,}029$$

Для минимальной

Полученное значение меньше табличного **0,029<0,520**, поэтому данное значение отбрасывать также не стоит.

Средняя арифметическая, среднеквадратическое отклонение, ошибка средней, достоверность

Насколько статистические оценки совпадают с истинными, свойствами генеральной совокупности?

Для вычисления статистической ошибки выборочной средней *M* используется формула

$$m = \pm \frac{\sigma}{\sqrt{n}}$$

Стандартное отклонение отражает разброс всех вариантов относительно средней, а стандартная ошибка показывает пределы, в которых, с известной вероятностью, может располагаться средняя величина.

В интервале $M \pm 1m$ средняя величина генеральной совокупности может находиться с вероятностью 68.3 %, в интервале $M \pm 2m$ - с вероятностью 95.5 %, а в пределах $M \pm 3m$ - с вероятностью 99.7 %.

Метод нахождения доверительных интервалов **в случае анализа небольших выборок** найден английским статистиком Госсетом, известном под псевдонимом **Стьюдент**

Величина ***t*** показывает, во сколько раз необходимо увеличить стандартную ошибку выборочного статистического параметра для того, что бы при определенном уровне вероятности судить о тех пределах, в которых располагается генеральное значение.

Величина *t* напрямую зависит лишь от уровня вероятности ***P*** и числа степеней свободы **n,** которое равно глубине выборки -1. (**объем выборки без числа ограничений)**

В большинстве биологических исследований принимают *P*=0.95 (то есть 95 случаев из 100), в наиболее ответственных случаях - 0.99 или 0.999

Сравнение средних величин

В биологических экспериментах особое значение имеют различия, на основании которых судят об эффективности действия тех или иных факторов, например, по разности между опытной и контрольной группами делают заключение о результатах опыта.

Важно оценить статистическую *достоверность разности*, т. е. определить, можно ли данное различие считать закономерным, *характерным для всей генеральной совокупности* и рассматривать его как результат действия особенных факторов, или же оно случайно и является следствием недостаточного количества данных и в следующих опытах может не проявиться

Обнаружение достоверных отличий статистических параметров – первый шаг к познанию новых биологических закономерностей, причем количественно доказанных

Критерии достоверности отличий

Сравнения выборочных средних – это вопрос о том, действовал ли при составлении одной из выборок новый систематический фактор по сравнению с другой выборкой

Отличия между средними могут иметь два противоположных источника:
1. Обе выборки взяты из одной генеральной совокупности, но средние отличаются в силу ошибки репрезентативности.
2. Выборки взяты из разных генеральных совокупностей, отличие средних вызвано, в основном, действием разных доминирующих факторов (а также и случайно).

Исходно предполагается (Но): «достоверных отличий между средними нет»

Критерий Стьюдента

.Поскольку выборочные средние имеют нормальное распределение, критерий отличия двух выборочных средних также базируется на *свойствах нормального распределения*: в границах $M_{общ.} \pm 1.96 \cdot m$ (или приблизительно $M_{общ.} \pm 2 \cdot m$) выборочные средние арифметические отличаются от общей (генеральной) средней по случайным причинам.

$$t = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}} \sim t_{(\alpha,\, df)}.$$

Полученное значение критерия *t* Стьюдента сравнивают с табличным при выбранном уровне значимости (обычно для $\alpha = 0.05$) и числе степеней свободы (*объемы выборок без числа ограничений*, $df = n_1 + n_2 - 2$).

Если полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны. Это говорит о том, что различия случайны, определенного вывода сделать нельзя, нулевая гипотеза остается неопровергнутой.

Мера варьирования величины – σ, (сигма), коэффициент вариации

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

Чем больше случайных факторов, чем они сильнее, тем дальше разбросаны варианты вокруг средней и тем больше среднее квадратичное отклонение.

Термин «случайное» - синоним слова «неизвестное», «неподконтрольное». Пока мы каким-либо способом не выразим интенсивность фактора (группировкой, градацией, числом), до тех пор он останется фактором, вызывающим случайную изменчивость.

«Именованность» - недостаток среднего квадратического отклонения, как мерила изменчивости признаков устраняется, если выразить этот показатель в процентах от величины средней арифметической данного распределения, Полученный таким образом показатель называется коэффициентом вариации

$$V = \frac{\sigma}{\bar{x}}$$

Если коэффициент вариации больше 33%, выборка неоднородна

Оценка репрезентативности выборки

В практике биометрического анализа используется относительная ошибка измерений – «показатель точности опыта» – отношение ошибки средней к самой средней арифметической, выраженное в процентах:

$$\varepsilon = \frac{m}{M} \cdot 100\%$$

Чем точнее определена средняя, тем меньше будет ε, и наоборот. Точность считается хорошей, если ε меньше 3%, и удовлетворительной при $3 < \varepsilon < 5\%$

Оптимальный объем выборки

Для непрерывных признаков метод состоит в том, чтобы, используя известные соотношения между средней, стандартным отклонением, ошибкой средней, плотностью вероятности распределения Стьюдента, найти число степеней свободы, соответствующее доверительному интервалу для средней при уровне значимости $\alpha = 0.05$

$$n = \left(\frac{t \cdot CV}{\varepsilon} \right)^2$$

Где CV – приблизительное значение коэффициента вариации (%),
ε – планируемая точность оценки (погрешности) (%).

n – объем выборки,
t – граничное значение из таблицы распределения Стьюдента (таблица), соответствующее принятому уровню значимости при планируемом объеме выборки,

Пример оценки объема выборки

Рассчитаем необходимый объем условной выборки, обеспечивающий хорошую точность ε = 3%, для уровня значимости α = 0.05 (t = 1.98, для $df \approx 100$) и для коэффициента вариации CV = 12% (такова относительная изменчивость многих размерно-весовых признаков животных):

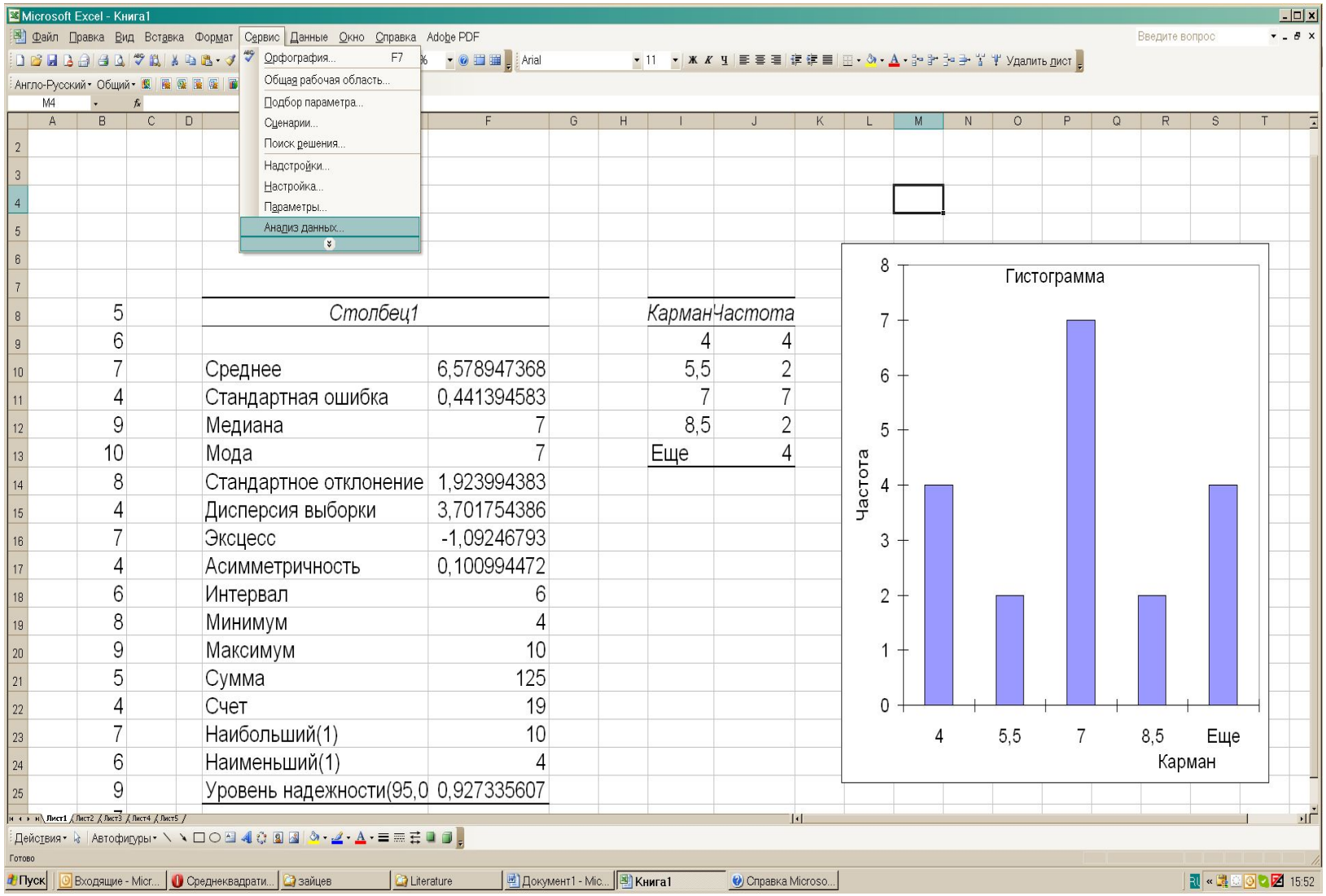
$$n = \left(\frac{1.98 \cdot 12}{3} \right)^2 = 62.726 \approx 63 \text{ экз}$$

Несколько примеров

В процессе анализа данных, как правило, присутствуют следующие основные этапы:

1. Ввод данных
2. Преобразование данных
3. Визуализация данных
4. Статистический анализ
5. Представление результатов

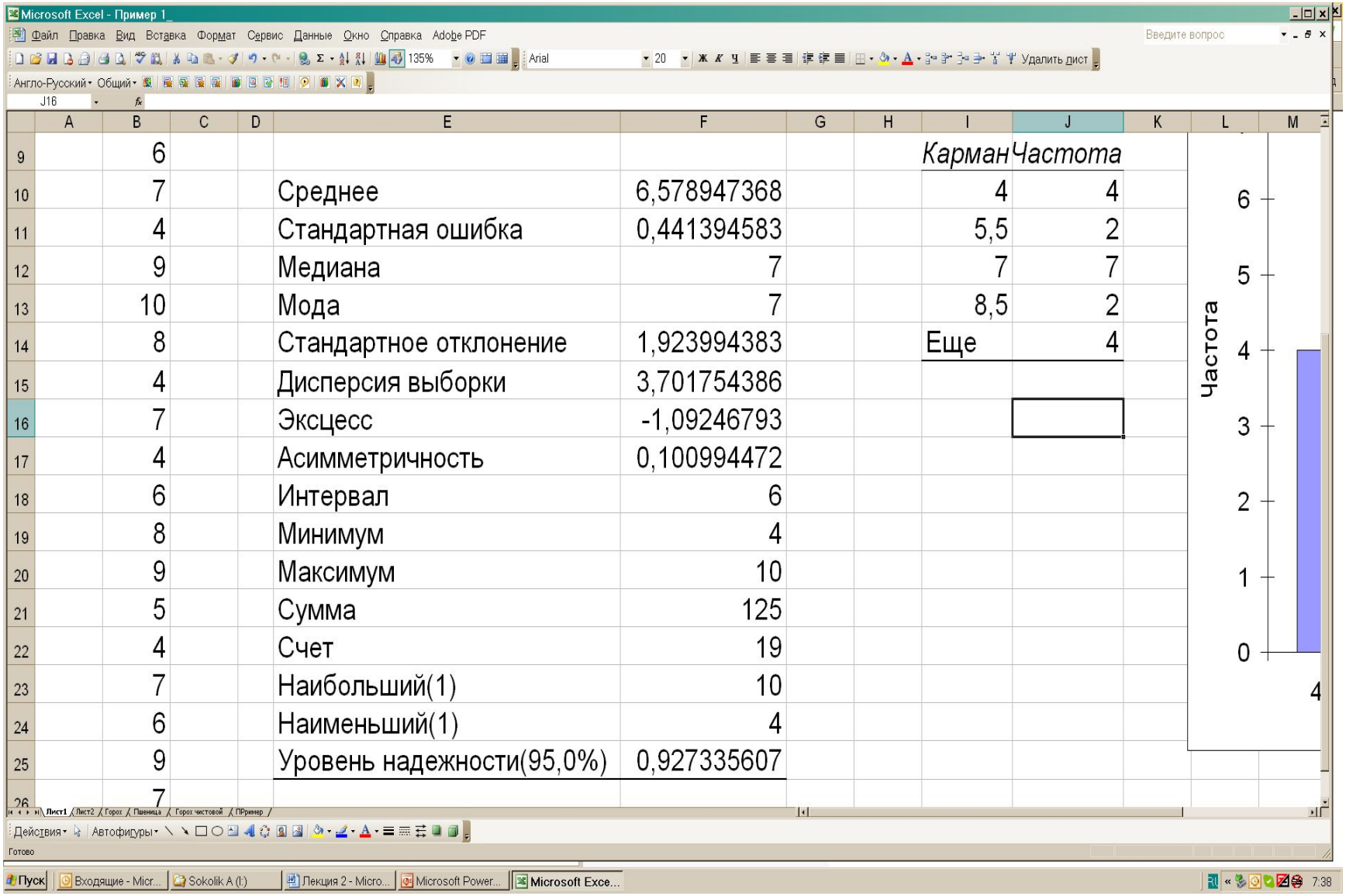
Что позволяет программа Excel



	Столбец1	
Среднее	6,578947368	
Стандартная ошибка	0,441394583	
Медиана	7	
Мода	7	
Стандартное отклонение	1,923994383	
Дисперсия выборки	3,701754386	
Эксцесс	-1,09246793	
Асимметричность	0,100994472	
Интервал	6	
Минимум	4	
Максимум	10	
Сумма	125	
Счет	19	
Наибольший(1)	10	
Наименьший(1)	4	
Уровень надежности(95,0	0,927335607	

Карман	Частота
4	4
5,5	2
7	7
8,5	2
Еще	4

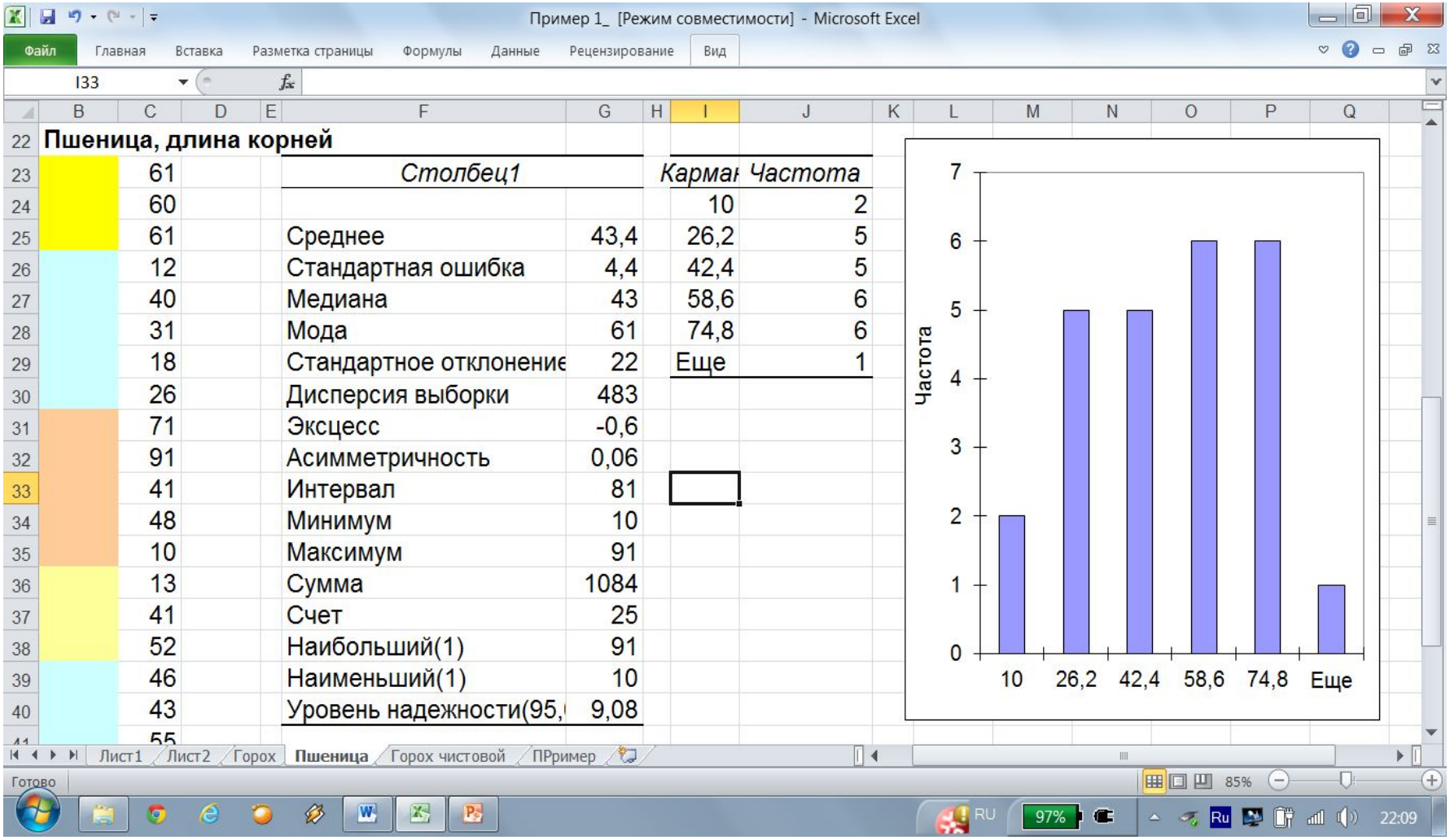
Статистические показатели



6			
7	Среднее	6,578947368	
4	Стандартная ошибка	0,441394583	
9	Медиана	7	
10	Мода	7	
8	Стандартное отклонение	1,923994383	
4	Дисперсия выборки	3,701754386	
7	Эксцесс	-1,09246793	
4	Асимметричность	0,100994472	
6	Интервал	6	
8	Минимум	4	
9	Максимум	10	
5	Сумма	125	
4	Счет	19	
7	Наибольший(1)	10	
6	Наименьший(1)	4	
9	Уровень надежности(95,0%)	0,927335607	
7			

Карман	Частота
4	4
5,5	2
7	7
8,5	2
Еще	4

Длина корней проростков пшеницы



Длина корней проростков гороха

Пример 1 [Режим совместимости] - Microsoft Excel

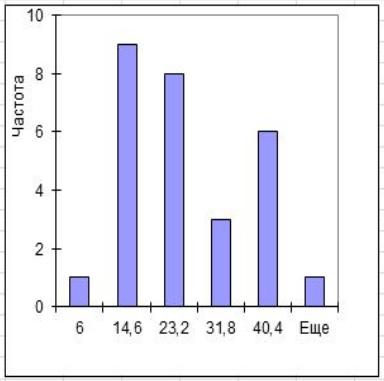
Файл | Главная | Вставка | Разметка страницы | Формулы | Данные | Рецензирование | Вид

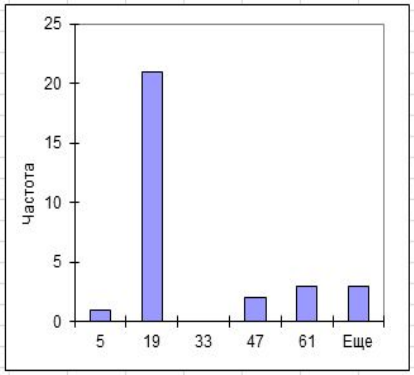
Y19

Длина корней 14 дневных проростков гороха

Контроль	Блокатор ионного транспорта			Карман	Частота
36	17			6	1
36	15	Среднее	20,71	14,6	9
25	11	Стандартная ошибка	2,16	23,2	8
10	10	Медиана	16,50	31,8	3
36	11	Мода	36,00	40,4	6
15	6	Стандартное отклон	11,41	Еще	1
18	49	Дисперсия выборки	130,29		
12	10	Эксцесс	-0,30		
11	10	Асимметричность	0,84		
16	13	Интервал	43,00		
12	75	Минимум	6,00		
32	10	Максимум	49,00		
8	8	Сумма	580,00		
19	57	Счет	28,00		
29	9	Наибольший(1)	49,00		
20	8	Наименьший(1)	6,00		
24	7	Уровень надежности	4,43		

				Карман	Частота
16	65			5	1
12	11	Среднее	22,80	19	21
35	8	Стандартная ошибка	4,04	33	0
	34	Медиана	11,50	47	2
	15	Мода	10,00	61	3
	11	Стандартное отклон	22,13	Еще	3
	10	Дисперсия выборки	489,61		
	12	Эксцесс	0,19		
	60	Асимметричность	1,35		
	13	Интервал	70,00		
	15	Минимум	5,00		
	14	Максимум	75,00		
	14	Сумма	684,00		
	55	Счет	30,00		
	10	Наибольший(1)	75,00		
	7	Наименьший(1)	5,00		
	42	Уровень надежности	8,26		
	8				
	5				





Лист1 | Лист2 | Горох | Пшеница | Горох чистовой

Готово

52%

RU 97% 22:25