1. ТОЧЕЧНЫЕ И ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ

При проведении биологических исследований могут быть изучены все объекты массива или только их часть. В первом случае исследования называют полными, или сплошными, во втором случае — частичными, или выборочными. В математической статистике весь массив объектов одной категории называют генеральной совокупностью. Изучение генеральной совокупности проводят редко. В большинстве случаев изучается часть генеральной совокупности, называемая выборочной совокупностью, или выборкой. Выборка должна соответствовать следующим условиям:

- сформирована по принципу случайного отбора (рандоминизированно);
- доступна для изучения, объем выборки может быть любым, он определяется задачами исследования;
- характеризовать всю генеральную совокупность; группы, выделенные не для характеристики всей генеральной совокупности (например, на выставку), не могут быть использованы в качестве выборки.

Важнейшим требованием к выборке является ее репрезентативность, то есть правильная представимость в ней пропорций генеральной совокупности [1], [9], [12], [14], [19].

Числовые показатели, характеризующие генеральную совокупность, называют генеральными параметрами, а числовые показатели, характеризующие выборку, называют выборочными характеристиками, или статистиками.

Выборочные характеристики являются приближенными оценками генеральных параметров. Это случайные величины, варьирующие вокруг своих параметров. Оценки генеральных параметров по выборочным характеристикам могут быть точечными и интервальными.

Точечные оценки генеральных параметров — это числа, вычисляемые по случайной выборке.

Интервальные оценки генеральных параметров — значения, в пределах которых с заданной доверительной вероятностью находится генеральный параметр.

Точечные и интервальные оценки генеральных параметров в программе STATISTICA проводятся на основе методов описательные статистики (Descriptive statistics). В программе STATISTICA эти методы реализованы в разделе Основные статистики/Таблицы (Basic Statistics/Tables), меню Statistics.

Значения переменных для анализа в электронную таблицу STATISTICA загружают из приложения или вводят с клавиатуры.

Для ввода данных в электронную таблицу STATISTICA, подготовленных в каком-либо другом приложении, можно воспользоваться одним из способов: буфером обмена, технологией динамического обмена данными, средствами импорта файлов. Буфер обмена — самый быстрый и простой путь ввода данных из прикладных программ Windows. Для реализации этого способа необходимо: в исходном материале выделить данные, которые необходимо скопировать; в меню Правка (Edit) выбрать команду Копировать, данные будут скопированы в буфер обмена; перейти в электронную таблицу STATISTICA и установить указатель там, где следует скопировать данные, затем нажать кнопку мыши; в меню Правка (Edit) выбрать команду Вставка (Paste), данные будут скопированы в направлении вправо и вниз от места, обозначенного курсором.

Иногда необходимо установить связь между данными из какого-либо приложения (источника или сервера), например Excel, и таблицей STATISTICA (клиентский файл) таким образом, чтобы при изменении данных в сервере соответствующие изменения произошли в таблице STATISTICA — клиенте. Связи такого типа в STATISTICA устанавливаются при помощи процедуры динамического обмена данными (DDE) из меню Правка (Edit). Для создания связи нужно активизировать кнопку Новая связь, откроется окно. В поле DDE связь пишется инструкция связи (обслуживание, разделы, элементы), которая связывает ячейки электронной таблицы источника с ячейками в электронной таблице STATISTICA. После напи-

сания инструкции связи и нажатия ОК в таблице STATISTICA (клиенте) появятся элементы из соответствующего источника (сервера).

Импорт файлов реализован при помощи команды Получение внешних данных в меню Данные (Data). Эта команда формирует запросы из других баз данных. Программа STATISTICA позволяет обращаться к наиболее распространенным базам данных (БД): Oracle, MS SQL Server, Sybase, MS Access, Fox Pro и др. Для доступа к данным используется драйвер ODBC (Open Data Base Connectivity — совместимость открытых баз данных), который позволяет приложению обращаться к БД на языке SQL. Запросы дают возможность выбрать из таблиц БД необходимые для статистического анализа данные и сохранить их в программе STATISTICA. Параметры подключения проходят тестирование. Если параметры подключения указаны верно, нажатием кнопки ОК производится подключение к базе данных и импорт данных в программу STATISTICA. После импорта данным запроса присваивается имя (чтобы сохранить для дальнейшего использования, запросы сохраняются в файлах с расширением *sqy), и данные запроса передаются в таблицу STATISTICA [6], [21].

Исходные данные для статистического анализа в программе STATISTICA организованы в виде таблицы (рис. 1).

Электронная таблица состоит из строк и столбцов. В отличие от обычных электронных таблиц, в которых строки и столбцы равноправны, в STATISTICA они имеют разные смысловые значения. Столбцы таблицы называются переменными (Variables), представляют собой наблюдаемые величины. В электронной таблице пользователь может задать спецификации переменных: формат отображения (например, число десятичных знаков), коды пропущенных значений (при хранении данных STATISTICA приписывает пропущенным наблюдениям по умолчанию код — 9999, пользователь может установить значение этого кода для каждой конкретной переменной; способ обработки пропущенных данных определяется после выбора метода статистического анализа), длинные имена переменных, комментарии для отдельных значений, формулы, которые можно использовать для преобразования каждой переменной.

Окно спецификаций переменной вызывается двойным щелчком на имени переменной в таблице исходных данных.

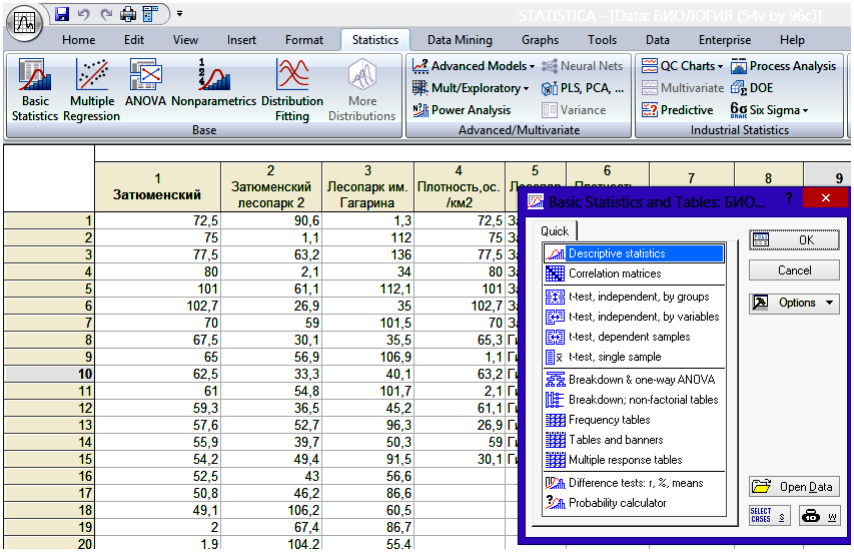

Рис. 1. Электронная таблица программы STATISTICA и методы раздела Основные статистики/Таблицы (Basic Statistics/Tables)

Результаты наблюдений записываются в строках таблицы (Cases). Нулевой столбец, в котором по умолчанию указаны номера наблюдений, при необходимости может быть изменен на имена случаев либо даты наблюдений.

Для удобной работы с переменными, принимающими текстовые значения, реализован так называемый механизм двойной записи, согласно которому каждому текстовому значению переменной в спецификации ставится в соответствие некоторое число. Это соответствие может быть установлено автоматически (самой системой при вводе данных) или определено пользователем. При работе с данными всегда можно переключиться с текстовой на числовую форму записи исходных данных.

1.1. Точечные оценки параметров

Рассмотрим применение методов описательной статистики (Descriptive statistics) для характеристики статистических совокупностей.

Пример 1. Приведены показатели плотности птиц (особей/км^2) в лесопарке «Затюменский» (рекреационная нагрузка 21 чел./ч).

72,5	75,0	77,5	80,0	101,0	102,7	70,0	67,5	65,0	62,5	61,0
59,3	57,6	55,9	54,2	52,5	50,8	49,1	2,0	1,9	47,4	45,7
44,0	42,3	39,6	37,1	34,6	32,1	29,6	27,1	24,6	22,1	

Для выбора из электронной таблицы переменной плотность птиц (особей/км^2) в лесопарке «Затюменский» надо нажать кнопку Variables и в открывшемся диалоговом окне активизировать исследуемую переменную (рис. 2).

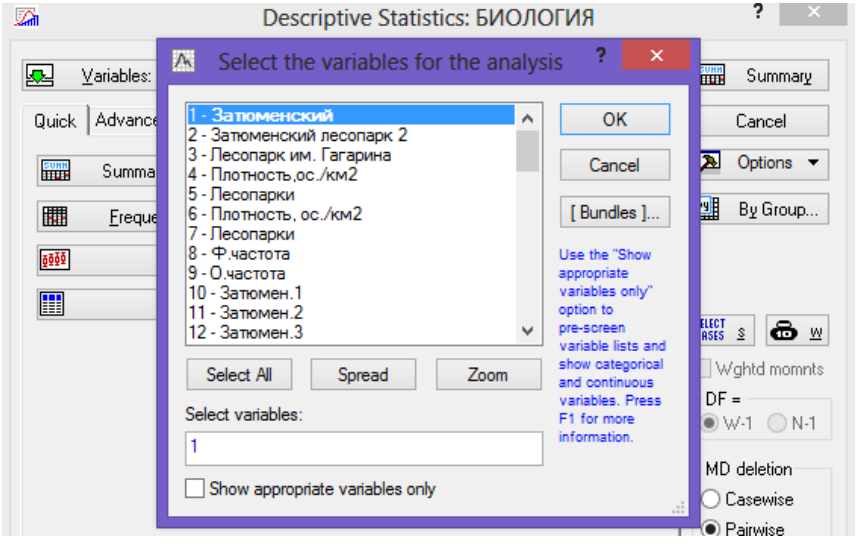

Рис. 2. Выбор переменной (переменных) для анализа

На вкладке Quick (или Summary/Descriptive statistics) программа отражает результаты определения основных статистических показателей:

Среднее арифметическое (Mean, $\overline{X}$) $\overline{X}$ = 51,38 особей/км². Показатель средней плотности птиц в лесопарке «Затюменский» составляет 51,38 особей/ км^2.

Минимум и максимум (Minimum & Maximum): min = 1,9; max = 102,7.

Среднее квадратическое отклонение (Standard Deviations, S_x) S_x = ±23,73 особей/км^2.

Среднее квадратическое отклонение — величина, показывающая среднее отклонение вариант от среднего значения. Варианта — числовое значение отдельного объекта.

Количество (Valid, N) N = 32.

Статистические показатели для полного анализа выборочной совокупности выбираются на вкладке Advanced установлением флажков напротив соответствующих статистик. При помощи кнопки Select all stats можно выбрать все статистики. Они разделены на три группы (рис. 3).

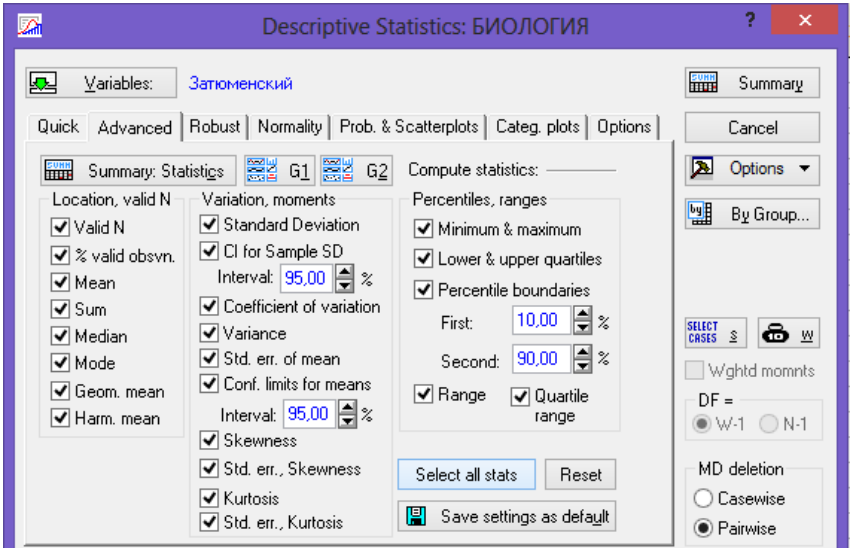

Рис. 3. Статистики для характеристики совокупностей

1. Показатели положения (location) (рис. 4).

Количество (Valid) N = 32; % обработанных значений (% valid obvn.).

Variable	Descriptive Statistics (БИОЛОГИЯ)							
	Valid N	% Valid obs.	Mean	Geometric Mean	Harmonic Mean	Median	Mode	Frequency of Mode
Затюменский	32	33,33333	51,38125	41,51246	19,23983	51,65000	Multiple	1

Рис. 4. Показатели положения

Среднее арифметическое (Mean) $\overline{X}$ = 51,38 особей/км².

Медиана (Median) M_e = 51,65 особей/км². Медиана — это значение, которое делит выборку на две равные части.

Мода (Mode) — значение, наиболее часто встречающейся варианты в данной совокупности. Среди показателей плотности птиц лесопарка «Затюменский» нет повторяющихся значений.

Среднее геометрическое (Geom. mean, $\overline{X_g}$) определяется при оценке средних темпов изменения величины переменной за определенные промежутки времени.

Среднее гармоническое (Harm. mean, $\overline{X_h}$) определяется при работе с переменными величинами, изменяющимися во времени.

2. Показатели изменчивости (variation), моментные характеристики (moments) (рис. 5).

Variable	Descriptive Statistics (БИОЛОГИЯ)							
	Variance	Std.Dev.	Coef.Var.	Standard Error	Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis
Затюменский	562,9700	23,72699	46,17830	4,194379	0,023372	0,414457	0,262514	0,809371

Рис. 5. Показатели изменчивости, моментные характеристики

Дисперсия (Variance) S_x^2 = 562,63 особей/км².

Среднее квадратическое отклонение (Standard Deviations) S_x = ±23,72 особей/км².

Коэффициент вариации Cv = 46,1%.

Ошибка репрезентативности для среднего арифметического (стандартная ошибка) $S_{\bar{X}} = \pm 4{,}19$ особей/км².

Коэффициент асимметрии (Skewness, *As*) — показатель, характеризующий симметричность распределения. При нормальном распределении коэффициент асимметрии равен нулю. Если коэффициент асимметрии существенно отличается от нуля, то распределение несимметрично. Определяется коэффициент асимметрии по формуле:

$$As = \frac{\sum (x - \bar{X})^3}{n \cdot S_x^{\,3}}.$$

Стандартная ошибка асимметрии (Standard errow of Skewn., S_{A_s}):

$$S_{A_s} \sqrt{\frac{6}{n+3}}.$$

Коэффициент эксцесса (Kurtosis, *Ex*) характеризует особенность распределения вариант выборки около своего центра. Определяется коэффициент эксцесса по формуле:

$$Ex = \frac{\sum (x - \bar{X})^4}{n \cdot S_x^4} - 3.$$

Стандартная ошибка эксцесса (Standard errow of Kurtosis, S_{E_x}):

$$S_{E_x} = 2 \cdot \sqrt{\frac{6}{n+5}}.$$

Для нормального распределения коэффициент эксцесса, так же как и коэффициент асимметрии, равен нулю. Если коэффициенты асимметрии и эксцесса превосходят критические (стандартные) значения, приведенные в табл. 1, 2, гипотеза о нормальности распределения не принимается, формулируется вывод о наличии у распределения значимой асимметрии или эксцесса.

В выборке (пример 1) значимая асимметрия и эксцесс не наблюдаются. Коэффициент асимметрии (*As* = 0,023) и коэффициент эксцесса (*Ex* = 0,26) меньше стандартных значений, приведенных в табл. 1, 2.

Таблица 1

Критические значения коэффициента асимметрии, *As*

Объем выборки	*Уровни значимости*		*Объем выборки*	*Уровни значимости*	
	0,05	*0,01*		*0,05*	*0,01*
25	0,711	1,061	250	0,251	0,360
30	0,611	0,982	300	0,230	0,329
35	0,621	0,921	350	0,213	0,305
40	0,587	0,869	400	0,200	0,285
45	0,558	0,825	450	0,188	0,269
50	0,533	0,787	500	0,179	0,255
60	0,492	0,723	550	0,171	0,243
70	0,459	0,673	600	0,163	0,233
80	0,432	0,631	650	0,157	0,224
90	0,409	0,596	700	0,151	0,215
100	0,389	0,567	750	0,146	0,208
125	0,350	0,508	800	0,142	0,202
150	0,321	0,464	850	0,138	0,196
175	0,298	0,430	900	0,134	0,190
200	0,280	0,403	950	0,130	0,185

Таблица 2

Критические значения коэффициента эксцесса, *Ex*

Объем выборки	*Уровни значимости*	
	0,05	*0,01*
1	*2*	*3*
11	0,907	0,936
16	0,888	0,914
21	0,877	0,900
26	0,869	0,890
31	0,863	0,883
36	0,858	0,877
41	0,854	0,872

Окончание табл. 2

1	*2*	*3*
46	0,851	0,868
51	0,848	0,865
61	0,843	0,859
71	0,840	0,855
81	0,840	0,855
91	0,835	0,848
101	0,834	0,846
201	0,823	0,832

3. Процентили, размахи (percentiles, ranges) (рис. 6).

	Descriptive Statistics (БИОЛОГИЯ)							
Variable	**Minimum**	Maximum	Lower Quartile	Upper Quartile	Percentile 10,00000	Percentile 90,00000	Range	Quartile Range
Затюменский	1,900000	102,7000	35,85000	66,25000	24,60000	77,50000	100,8000	30,40000

Рис. 6. Процентили и размахи

Минимум и максимум (Minimum & Maximum): min = 1,9; max = 102,7.

Минимальная и максимальная квартили (Lower & upper quartiles, $P_{25}; P_{75}$). $P_{25} = 35{,}85$; $P_{75} = 66{,}25$. Квартиль — значение переменной, ниже которого находится часть (25% и 75%) выборки.

Размах (Range) — разность между максимальным и минимальным значениями выборки.

Квартильный размах (Quartiles range) — разность значений верхней и нижней квартилей.

Программа STATISTICA позволяет задать определение значения процентилей. В практике обычно используют процентили: P_3, P_{97}; P_{10}, P_{90}.

Для анализа изменчивости переменных предусмотрено построение графиков на вкладке Box & Whisker. Выбор показателей для построения графиков проводится на вкладке Options.

Показатели для оценки изменчивости на графике:

- медиана / квартиль / размах;
- среднее арифметическое / стандартная ошибка / среднее квадратическое отклонение;
- среднее арифметическое / среднее квадратическое отклонение / 1,96 · среднее квадратическое отклонение;
- среднее арифметическое / стандартная ошибка / 1,96 · стандартная ошибка.

1.2. Робастная оценка в программе STATISTICA

Статистический метод, способный действовать в условиях выбросов (анг. outlier), называют робастным. Выбросами в статистике считают значения, выделяющиеся из общей выборки. Причины выбросов бывают разные (ошибки измерения; необычная природа входных данных; выбросы могут быть частью распределения, при нормальном распределении (это распределение будет рассмотрено в разделе 2) каждое 22-е измерение выходит из интервала ± две сигмы, каждое 370-е измерение — из интервала ± три сигмы). Определяются выбросы на основе различных методов.

Простейший метод основан на межквартильном расстоянии. Все значения, которые не попадают в диапазон $[(x_{25} - 1{,}5 \cdot (x_{75} - x_{25}))]$, $[(x_{75} + 1{,}5 \cdot (x_{75} - x_{25}))]$, считаются выбросами. Минимальное значение плотности птиц 1,9 особей/км^2 (пример 1) и максимальное значение плотности попадают в диапазон $[(x_{25} - 1{,}5 \cdot (x_{75} - x_{25}))]$, $[(x_{75} + 1{,}5 \cdot (x_{75} - x_{25}))]$.

Для проведения устойчивой оценки программа STATISTICA определяет:

- усеченное среднее (trimmed mean) — среднее значение после удаления выбросов;
- винсоризованное среднее (winsorized mean) — среднее значение после замены выбросов процентилью, по которой сделано усечение;
- критерий Граббса для выбросов (Grubbs test for outliers) (рис. 7).

Variable	Descriptive Statistics (БИОЛОГИЯ)								
	Valid N	Mean	Trimmed mean 5,0000%	Winsorized mean 5,0000%	Grubbs Test Statistic	p-value	Minimum	Maximum	Std.Dev.
Затюменский	32	51,38125	51,30714	51,27500	2,162885	0,812392	1,900000	102,7000	23,72699

Рис. 7. Робастная оценка в системе STATISTICA

Критерий Граббса (T) определяется по формуле:

$$T = \left(x_i - \overline{X}\right) : S_x ,$$

где x_i — текущее значение выборки; $\overline{X}$ — среднее арифметическое; S_x — среднее квадратическое отклонение.

Среднее арифметическое, усеченное среднее, винсоризованное среднее имеют примерно одинаковые значения. Критерий Граббса для выделяющегося значения (102,7) из выборки имеет уровень значимости 0,8123 (0,8123 больше 0,05). Критерий Граббса не превышает критическое значение 2,938 (табл. 3). Выделяющееся значение (102,7) не является выбросом.

Таблица 3

Критические значения для критерия Граббса

№	*Одно наибольшее или одно наименьшее значение при уровне значимости*	
	0,01	*0,05*
1	*2*	*3*
3	1,155	1,155
4	1,496	1,481
5	1,764	1,715
6	1,973	1,887
7	2,131	2,020
8	2,274	2,126
9	2,387	2,215
10	2,482	2,290
11	2,564	2,355

Окончание табл. 3

1	*2*	*3*
12	2,636	2,412
13	2,699	2,462
14	2,755	2,507
16	2,852	2,585
18	2,932	2,651
20	3,001	2,709
22	3,060	2,758
24	3,112	2,802
26	3,157	2,841
28	3,199	2,876
30	3,236	2,908
32	3,270	2,938
34	3,301	2,965
36	3,330	2,991
38	3,356	3,014
40	3,381	3,036

При оценке выбросов наряду с критерием Граббса принято определять критерий Шовене, критерий Пирса, Q-тест Диксона.

Статистические характеристики, полученные на материале выборок, являются случайными величинами, варьирующими вокруг своих генеральных параметров. Такие выборочные характеристики рассматриваются как приближенные значения или точечные оценки соответствующих генеральных параметров. Выборочное среднее ($\overline{X}$) является оценкой генерального среднего (μ), выборочная дисперсия является (S_x^2) — оценкой генеральной дисперсии (σ_x^2), среднее квадратическое отклонение (S_x) — оценкой стандартного отклонения (σ_x), характеризующего генеральную совокупность.

Имея множество выборок из одной генеральной совокупности, можно получить достаточно точную величину генерального параметра. Для того чтобы по одной выборке оценить генеральные параметры, требуется определить:

1) ошибку репрезентативности (статистическую ошибку) — величину отклонения выборочного показателя от его генерального параметра;

2) показатель точности (*Cs*);

3) доверительный интервал — область, в которой с определенной вероятностью находится величина генерального параметра.