

## Добрый день, Алена!

Итак, начнем нашу учебу. Я выслал Вам почтовой бандеролью свою книгу по прикладной статистике. После ее прошу Вас послать электронное письмо с подтверждением получения. Мы будем ссылаться в наших лекциях на те или иные страницы и задачи из этой книги. Главная Ваша цель - освоить всю Программу обучения, не отвлекаясь на второстепенные детали. В наших лекциях будут кроме текста еще и графические элементы, на которых будут панели пакета STATISTICA графические результаты анализа. Для удобства мы будем в тексте выделять цветом предложения заимствованные из Программы.

### Индивидуальная Программа подготовки по биостатистике

В этой лекции мы рассмотрим первый раздел Программы.

#### 1. Раздел "Основы проверки статистических гипотез"

Понятие статистической гипотезы. Ограниченность сдвиговой парадигмы отечественной экспериментальной биомедицины. Понятие о доверительной вероятности и уровне значимости. Ошибки первого и второго рода. Нулевая и альтернативная гипотеза; односторонние и двусторонние гипотезы. Основные этапы проверки гипотезы. Основная гипотеза о проверке нормальности распределения. Критерии Колмогорова-Смирнова и Шапиро-Уилки для проверки основной гипотезы. Применение графического способа оценки нормальности распределения. Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений. Типичные ошибки использования t-критерия Стьюдента при анализе биомедицинских данных. Проверка гипотезы о равенстве дисперсий двух нормальных распределений. Проверка статистических гипотез в пакете STATISTICA.

Итак, начнем с **понятия статистической гипотезы**. Прежде всего отметим, что в реальных ситуациях исследователь располагает достаточно ограниченными возможностями. Мало времени, денег, нет других ресурсов и т.д. По этой причине мы ограничены в сборе информации об интересующих нас объектах. Иными словами, мы можем собрать и проанализировать довольно ограниченное число интересующих нас объектов (наблюдений). Эти объекты будем называть выборкой, подразумевая, что объекты "выбираются" из гораздо большего числа возможных наблюдений. Ту совокупность исследуемых нами объектов, откуда "комплектуется" выборка, будем назвать генеральной совокупностью (ГС). К примеру, Вы взяли свои выборки из ГС содержащего весьма большое количество элементов. В случаях с такими ГС с ними можно обращаться как с ГС, имеющими бесконечно большое количество элементов. Это следует из так называемого закона больших чисел, главный вывод которого заключается в том, что с увеличением объема выборки значения выборочных характеристик (среднее, дисперсия и т.д.) приближаются к значениям таковых в **ГС (см. п.3.3.1 моей книги)**. Очевидно, что ограниченность имеющейся в выборке информации не позволит нам получить выводы абсолютно надежные. Это означает, что мы в принципе не можем получить на основании ограниченной информации абсолютно надежные заключения. Тем не менее, другого выхода нет, и мы

## 2Лекция 1 - учебный материал Н-вой Алене Николаевне.

вынуждены опираясь на ограниченные выборки делать по ним выводы. Все такие выводы имеют разную степень надежности. Степень надежности наших выводов зависит не только от объема выборки. Естественно, что надежность возрастает с увеличением объема наблюдений в выборке (объема выборки). Однако, даже имея большой объем наблюдений можно получить ненадежные, а то и ложные выводы. Так же как неумелый повар из добротных продуктов может приготовить несъедобное блюдо, так и специалист не умеющий выбрать для решения своих задач адекватный метод анализа рискует получить неверный вывод.

Какие же выводы можно получать на основании выборок? Во-первых, нужно усвоить, что все выводы делаются относительно предположений, которые принято называть ГИПОТЕЗАМИ, причем не просто гипотезами, а **СТАТИСТИЧЕСКИМИ ГИПОТЕЗАМИ** (см. Главу 4 в моей книге). Во всех статистических гипотезах формулируются те или иные предположения относительно параметров ГС, например, относительно генерального среднего, которое называют также математическим ожиданием. Это могут быть гипотезы относительно дисперсии (также генеральной), коэффициентов корреляции и многих, многих других параметров. В математической статистике принято генеральные параметры (параметры генеральной совокупности) обозначать греческими буквами, а аналогичные им параметры полученные из выборки (выборочные параметры) обозначать соответствующими латинскими буквами. Например, генеральное среднее (математическое ожидание) обозначают греческой буквой "мю" -  $\mu$ , а выборочное среднее обозначают латинской буквой М, либо буквой X с чертой сверху. Дисперсию генеральную обозначают  $\sigma^2$  а выборочную дисперсию -  $s^2$ .

Теперь Алена давайте поговорим относительно **ограниченности сдвиговой парадигмы отечественной экспериментальной биомедицины**. Один из основных принципов системного анализа, который является методологической базой любых исследований, является совместное использование процедур декомпозиции и агрегирования при изучении сложных систем. Живые организмы являются самыми сложными системами. Как и все сложные системы, они имеют в своей структуре подсистемы. Еще Рене Декарт писал: "Расчлените каждую изучаемую вами задачу на столько частей, сколько требуется, чтобы их было легко решить". Успех и значимость аналитического метода заключается не только, и даже не столько в том, что он позволяет сложное расчленить на менее сложные части, а в том, что при правильном и квалифицированном воссоединении этих простых частей воедино, они вновь способны образовать единое целое. Очевидно, что это возможно далеко не всегда, и во многом определяется профессионализмом аналитика. Более того, при объединении частей в новое целое может возникнуть нечто качественно новое, такое, чего не было и не могло быть без такого объединения. Такое появление новых свойств системы носит название эмерджентности. Фактически это свойство есть проявление известного закона перехода количества в качество. При этом, чем больше отличаются свойства совокупности от свойств исходных элементов, тем выше уровень организации системы. В кибернетике было показано, что чем выше степень согласованности, коррелированности изменения

### 3Лекция 1 - учебный материал Н-вой Алене Николаевне.

параметров отдельных частей системы, тем больше спектр реакций и выбора данной системы на изменяющиеся условия. В полной мере высказанные тезисы справедливы и при анализе многомерных живых систем. Конечно, добиться полного изучения всех подсистем и существующих в них и между ними взаимосвязей вряд ли при нынешнем уровне развития науки и техники возможно, тем не менее различные исследовательские техники позволяют исследовать довольно глубоко достаточно сложные многомерные системы. Примером этого могут служить успехи в деле изучения генома.

Однако, понимая что живые организмы по определению являются многомерными системами, мы должны понимать и ограниченность наших возможностей. Поэтому первое, что необходимо сделать, изучая живые многомерные системы, сразу после формулировки целей исследования сформулировать задачи, решение которых позволит достичь этих целей. На этапе формулировки задач следует перечислять конкретные признаки (переменные), которые необходимы при решении данной задачи. Вполне возможно, что первоначальный перечень этих признаков окажется далек от конечного списка, который сформируется после 2-3 итераций и корректив этих задач. Однако важно этот перечень признаков иметь, с тем, чтобы оценить реальность, возможность сбора необходимой информации - имеется в виду возможность получения необходимой выборки наблюдений и регистрации, измерения признаков. Процедура формулировки задач сопровождается и выбором методов решения этих задач. Так, если одна из задач предполагает проверку гипотез о равенстве средних значений количественного признака в двух или более группах, то декомпозирую эту задачу, мы приходим к необходимости формулировки и задач более низкого уровня, в частности для выбора метода проверки гипотезы (статистического критерия), необходимо произвести проверку нормальности распределения, а также ряд других допущений метода, предполагаемого использовать при решении этой задачи.

Анализ большого количества публикаций (только за последние три года нами было проанализировано около 2 тысяч публикаций по биомедицинской тематике выполненных в России и за рубежом) позволяют сформулировать доминирующую в российской биомедицинской науке статистическую парадигму, которая проявляет себя как латентная, скрытая закономерность только при анализе достаточно большого количества публикаций.

(Более подробно результаты этого исследования представлены в статье "НАУКОМЕТРИЧЕСКИЙ АНАЛИЗ СТАТИСТИЧЕСКОЙ ПАРАДИГМЫ ЭКСПЕРИМЕНТАЛЬНОЙ БИОМЕДИЦИНЫ (ПО МАТЕРИАЛАМ ПУБЛИКАЦИЙ) опубликованной в БИОМЕТРИКЕ и ряде печатных изданий).

Суть этой парадигмы, названной нами СДВИГОВОЙ, заключается в доминировании у отечественных исследователей в области биологии и медицины представления о том, что основное (а возможно и единственное!) различие между группами сравнения заключается в тривиальном, механическом сдвиге среднего значения исследуемой переменной. При этом игнорируются все остальные не менее важные параметры распределения признака, такие как меры рассеяния (дисперсия, размах и т.д.) и меры формы (эксцесс и асимметрия), корреляции между признаками и т.д. Более того, игнорируются возможные и весьма важные изменения законов распределения вероятностей в сравниваемых группах, изменения структуры связей между объектами исследования. Можно

#### 4Лекция 1 - учебный материал Н-вой Алене Николаевне.

идентифицировать такой подход как одномерный, механистический взгляд на сугубо многомерные взаимодействующие системы.

Для иллюстрации ущербности такой парадигмы приведем такой искусственный пример. Предположим, что мы поставили задачу сравнить между собой температуру тела персонала больницы с температурой тела всех пациентов этой больницы. Проверка (статистически вполне корректная) показала, что средние температуры двух сравниваемых групп статистически значимо не различаются. Между тем эти две группы имеют принципиальное различие в вариации этого показателя. Действительно, если принять во внимание, что персонал больницы исполняет свои обязанности будучи здоровым, то разброс температуры персонала будет сравнительно мал. Иное дело пациенты этой больницы, среди которых будут находиться больные с повышенной температурой, выздоравливающие пациенты с нормальной температурой, и те, кто недавно поступил в морг. В результате средняя температура пациентов будет равна средней температуре персонала, однако существенное различие групп по этому показателю будет заключаться в минимальных и максимальных значениях. Несмотря на искусственный характер этого примера, в реальных исследованиях такая ситуация встречается достаточно часто. Причем именно эти минимальные и максимальные значения, как правило, и несут нередко важнейшую информацию о сравниваемых группах. Вот почему при сравнении двух совокупностей между собой ни в коем случае нельзя останавливаться только на сравнении одних средних, забывая при этом все остальные характеристики. Конечно, такой подход намного увеличивает число решаемых при этом задач. Но этот же подход при решении этих задач дает соответственно и большее количество ответов, которые являются информацией к размышлению и материалом для формулирования выводов о механизмах изучаемых явлений.

Вернемся к проверке статистических гипотез. Речь здесь идет о том, что в подавляющем большинстве публикаций (диссертации, статьи, монографии и т.д.) авторы производят так называемую "проверку достоверности различий средних". Под этим подразумевается проверка двух статистических гипотез, одна из которых называется нулевой, а другая, конкурирующая с первой, называется альтернативной. Эти гипотезы принято записывать следующим образом:  $H_0: \mu_1 = \mu_2$  при  $H_1: \mu_1 \neq \mu_2$ . Первая буква H взята от латинского слова "Гипотеза", индексы 1 и 2 относятся к первой и второй генеральной совокупности. Первую гипотезу называют нулевой, поскольку ее (гипотезу) можно записать и несколько иначе, а именно в таком виде:  $H_0: \mu_1 - \mu_2 = 0$ . Фактически первая гипотеза утверждает, что генеральные средние в первой и второй совокупностях (генеральных же) равны между собой. Тогда как вторая гипотеза утверждает обратное. Причем альтернативные гипотезы могут быть двух типов - односторонние и двусторонние. Более подробно об этом почитайте в моей книге.

Теперь давайте обсудим понятия **доверительной вероятности и уровня значимости**. На первый взгляд это довольно простые понятия, особенно первое. Действительно, доверительная вероятность - это степень нашего доверия к чему-то. В нашем контексте - доверия к выводам. В частности, к выводам статистическим. Для лучшего понимания этого термина я прошу Вас почитать внимательно в моей

## 5 Лекция 1 - учебный материал Н-вой Алёне Николаевне.

книге раздел о построении доверительных интервалов (см п. 3.4). Нам же с Вами нужно внимательно обсудить второй термин, а именно то самое "p", которое можно встретить в каждой статье где есть хоть какой-то анализ биомедицинских данных. Главный смысл этой величины связан с ситуацией, которая более всего импонирует исследователю, а именно с тем случаем когда мы отвергаем нулевую гипотезу и делаем утверждение о неравенстве (чего-либо: средних, дисперсий, законов распределения и т.д.). Однако поскольку свои выводы мы делаем на основе довольно ограниченной информации, то наши выводы не абсолютны. В принципе возможны следующие 4 ситуации.

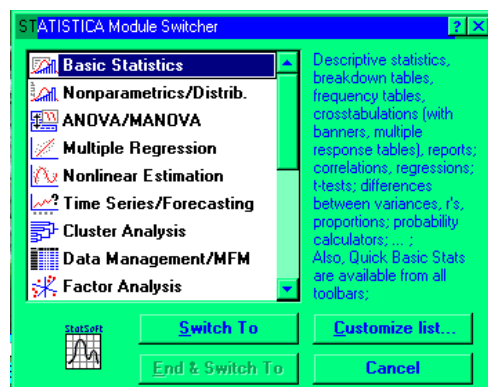
Предположим, что действительно нулевая гипотеза верна. В этом случае мы можем получить 2 вывода: 1) - да, нулевая гипотеза верна, 2) нет, верна альтернативная гипотеза. Аналогичные выводы мы можем получить и для того случая, когда на самом деле верна альтернативная гипотеза. Конечно, было бы идеально чтобы наши выводы не имели ошибок, т.е. в первом случае всегда делался вывод о том, что верна нулевая гипотеза, а во втором случае, что верна альтернативная гипотеза! Но, увы, это практически недостижимо! Всегда есть вероятность того, что мы принимаем ошибочный вывод. Но в этом случае хотелось бы, чтобы наша уверенность в том, что мы приняли верное решение (вывод) была больше чем вероятность нашей ошибки. И вот именно на этом этапе мы и имеем дело с уровнем значимости "p". Итак, каков же смысл величины "p"? Например, мы выполнили проверку каких то конкурирующих статистических гипотез. При такой проверке всегда вычисляется значение определенного, конкретного статистического критерия. Например, критерия Стьюдента, Фишера, Пирсона и т.д. Для конкретного значения такого статистического критерия и вычисляется и значение уровня значимости "p".

Итак, предположим, что мы получили значение "p" равное 0,001. Что это означает? А означает это следующее. Мы отклонили нулевую гипотезу и приняли альтернативную, но есть вероятность равная 0,001 что мы ошибочно сделали такое заключение о том, что верна альтернативная гипотеза, тогда как на самом деле верна нулевая гипотеза. **Итак, "p" - это вероятность ошибочно отвергнуть нулевую гипотезу при отсутствии различий. Если это утверждение еще упростить, то "p" - это вероятность справедливости нулевой гипотезы при условии ее отвержения.** Нередко эту вероятность называют вероятностью ошибки первого рода. Для случая же когда действительно верна альтернативная гипотеза, а делается вывод и том, что различия нет, называется ошибкой второго рода. К сожалению, в используемых популярных пакетах вероятность ошибки второго рода специально не вычисляется. **Основные этапы проверки гипотезы.** 1). Нужно сформулировать нулевую и альтернативную гипотезы. 2) Выбрать статистический критерий для проверки таких гипотез. Это очень ответственный этап, и нужно знать, в каких случаях можно применять те или иные критерии. Кстати, часть гипотез можно проверять не только одним статистическим критерием, а несколькими. Понятно, что разные критерии могут при этом давать и разную надежность получаемых выводов. Некоторые критерии могут применяться только в том случае, если известно, что переменные подчиняются нормальному закону распределению. Эти критерии называют параметрическими, в отличие от других

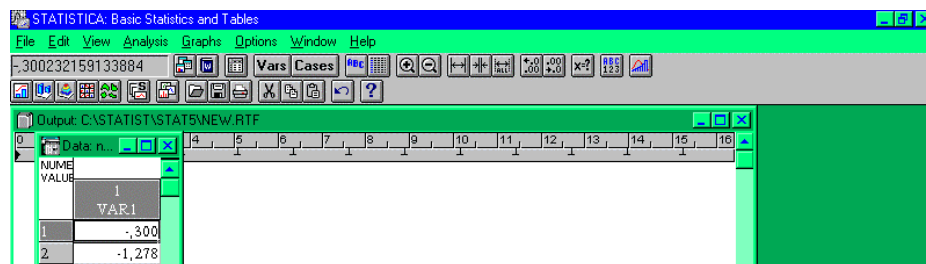
## БЛекция 1 - учебный материал Н-вой Алене Николаевне.

критериев, непараметрических, для которых нормальность распределения не обязательна. Рассмотрим, как же проверять нормальность распределения в пакете STATISTICA. Гипотезу о нормальности принято называть также **основной гипотезой**. В пакете можно использовать **критерий Колмогорова-Смирнова и Шапиро-Уилки для проверки основной гипотезы**. Используем для этой цели массив данных

под именем NORMAL. Для работы с пакетом STATISTICA рекомендую Вам создать на экране PC (на "рабочем столе") ярлык для программы из пакета STATISTICA с именем STA\_WIN.EXE. Итак, предположим что Вы, Алена, уже создали такой ярлык на рабочем столе и щелкнули по нему 2 раза мышкой. После этого на экране появится следующее меню (см. рисунок слева).



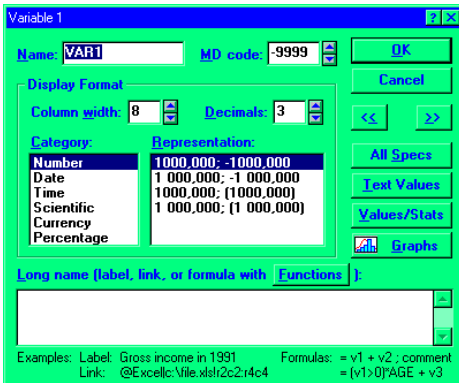
Выберите из него самое верхний модуль Basic Statistics (Основные статистики) и щелкните по нему мышкой. В результате появится новое окно, часть которого я привожу ниже (см. слева).



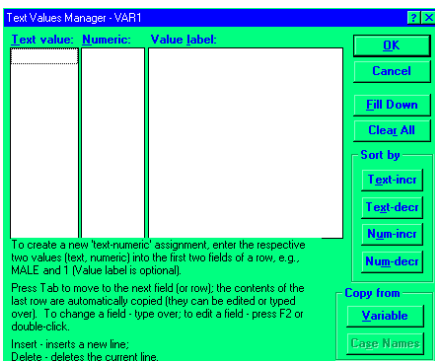
Пощелкав по отдельным меню (File, Edit, View, Analysis, Graph, Options, Window, Help) можно получить некоторое представление о назначении этих

ниспадающих меню. Ниже краткое описание назначения каждого из этих меню. File - как и во всех Windows-приложениях служит для открытия и сохранения **активных файлов** (т.е. сохранения того файла, с которым Вы в данный момент работаете). Здесь же можно производить операции импорта и экспорта файлов, открытия графиков, печати файлов и т.д. В нижней части меню приводится список тех файлов, с которыми Вы работали в последнее время. Меню Edit служит для редактирования файлов. В частности очень полезны здесь возможности редактирования переменных (variables) и наблюдений (cases). Их можно добавлять, удалять, изменять их и т.д. Следующее меню View - просмотр. В верхней строке этого меню Text Value - это возможность выбора просмотра числовых или текстовых значений переменных. Смысл этой опции следующий. Если мы в нашем массиве данных обозначаем в признаке А 4 сорта картофеля (А1-А4; 1-среднеспелый, 2-среднеранний, 3-среднеранний), то можно видеть либо числа 1, 2, 3 и 4, либо же присвоить каждой из градаций этого признака словесное описание: для 1-среднеспелый, для 2 - среднеранний и т.д. Для перехода от чисел к текстовым обозначениям градаций таких качественных признаков можно использовать кнопку "ABC" которая находится во втором ряду верхнего меню вслед за кнопками "Vars" и "Cases". Ниже приведен рисунок с частью этого меню, где видны эти кнопки.





Ясно что текстовые метки, обозначения этих градация сами собой не возникнут, их надо внести, записать для каждой градации конкретного признака. Для этой цели можно перейти к этой возможности щелкнув два раза по прямоугольнику с названием переменной. Например, в нашем массиве NORMAL всего одна переменная VAR1. Если мы щелкнем 2 раза мышкой по этому серому прямоугольнику, где написано VAR1, то получим следующее меню, которое приведено слева.

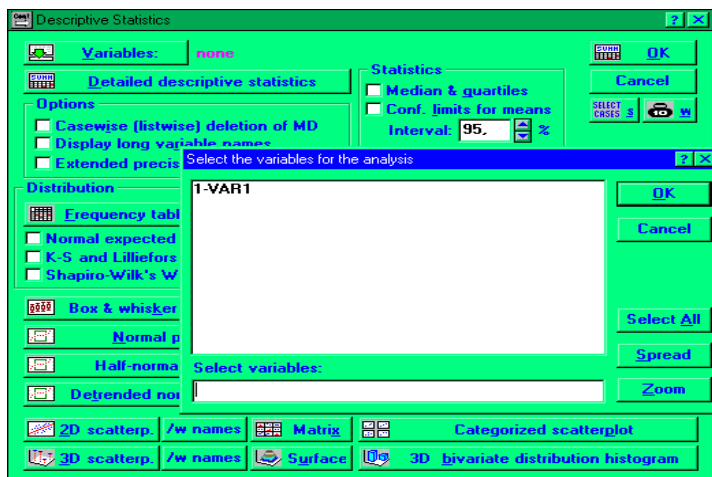


Далее выбираем справа на это меню кнопку Text Value и получим опять таки еще одно меню, которое приведено ниже.

Вводя числовые и текстовые значения в отдельных строках (а число таких строк будет равно числу градация признака, например, для сортов картофеля будет 4 строки) мы и зададим текстовые значения для отдельных градация признака. Вы можете

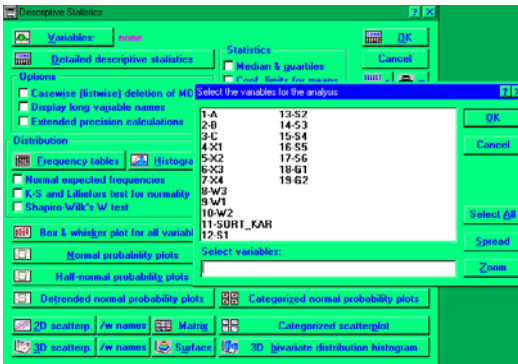
потренироваться в редактировании этих градаций в Вашем массиве, который я высылаю Вам, но предварительно создайте страховую копию, и храните ее отдельно, чтобы в том случае, когда учебная копия файла будет испорчена, можно было бы ее уничтожить (удалить), и скопировать на ее место страховую копию.

Следующее меню самое главное - Analysis. Здесь мы выбираем нужные нам процедуры статистического анализа. Щелкнув по двум первым строчкам меню перейдем на меню с набором групп методов: дескриптивные статистики, корреляционные матрицы, t-тест для независимых выборок и т.д. Для нашего случая (Основная гипотеза о проверке нормальности распределения. Критерии Колмогорова-Смирнова и Шапиро-Уилки для проверки основной гипотезы.) нам необходимо выбрать в этом меню строку "Descriptive statistics" щелкнув по ней и затем по кнопке "OK" Появится новое меню в верхней левой части которого будет кнопка "Variables" правее которой будет надпись none (не



указаны переменные). Щелкнув по кнопке "Variables" мы получим поверх текущего меню еще одно меню, в итоге все это будет выглядеть так (см. рисунок слева).

В последнем меню сверху по английски написано "Выберите переменные для анализа". Поскольку в нашем массиве NORMAL всего одна переменная, то в этом списке есть лишь 1-VAR1.



Для случая Вашего массива это меню будет выглядеть иначе (см. слева)

Для анализа переменных можно выбрать либо одну переменную, щелкнув по ней мышкой, либо несколько. Если эти переменные следуют в списке последовательно друг за другом, то нажмите мышку на первой переменной и не отпуская "тащите" ее ниже до последней переменной. Если же

список нужных переменных не последовательный, то можно "закрасить" первую часть списка, затем отпустить мышку, далее нажать клавишу Control (CTRL) обычно я это делаю нажимая левую клавишу CTRL, перевести курсор мышки на следующий признак и удерживая клавишу CTRL нажатой, продолжить закрасивать другие необходимые нам признаки (переменные) в этом списке. Можно

заказать список переменных и иначе. Для этого записать в нижней части меню с надписью "Select variables" номера нужных нам переменных. Например я



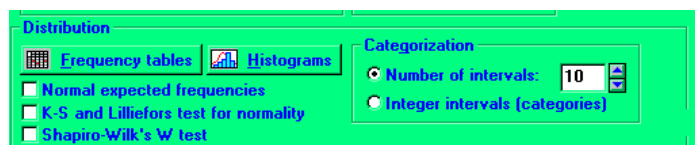
выбрал следующие переменные: 1-3 8-9 16-17. Тогда я запишу эти номера в строке таким образом:



Для массива NORMAL выберем только одну переменную VAR1. После выбора переменной щелкнем по "ОК", меню выбора переменных закроется, и далее будем выбирать опции для проверки нормальности распределения нашего выбранного признака. В средней части нашего меню есть область Distribution (Распределение). Именно здесь мы и укажем как производить проверку нормальности

распределения нашего признака. Кнопка "Frequency tables" дает нам таблицу частот фактических, и ожидаемых вычисленных из предположения, что в

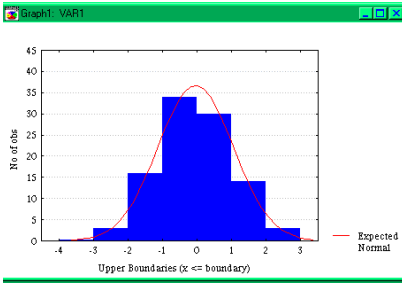
генеральной совокупности, откуда изъята выборка, данный признак подчиняется нормальному распределению. Щелкнув по этой кнопке, мы получим таблицу, которая приведена на следующей



BASIC STATS	Count	Cumul Count	Percent of Val	Cumul of Val	% of a Cases	Cumul. of All
-4,000 < x <= -3,000	0	0	0,0	0,	0,0	0,
-3,000 < x <= -2,000	3	3	3,0	3,	3,0	3,
-2,000 < x <= -1,000	16	19	16,0	19,	16,0	19,
-1,000 < x <= 0,0000	34	53	34,0	53,	34,0	53,
0,0000 < x <= 1,0000	30	83	30,0	83,	30,0	83,
1,0000 < x <= 2,0000	14	97	14,0	97,	14,0	97,
2,0000 < x <= 3,0000	3	100	3,0	100,	3,0	100,
Missing	0	100	0,0		0,0	100,

Обратите внимание, что ожидаемых частот, вычисленных из предположения о нормальности, нет, поскольку мы такую опцию не заказали. В нижней строке Missing (пропущенные, не измеренные значения) показано, что в нашем случае пропусков не было. Если щелкнуть кнопку Histograms, то мы получим гистограмму распределения наших значений признака по интервалам (подробнее о построении гистограмм см. мою книгу).

Обратите внимание, что красной кривой линией обозначено теоретическая кривая нормального распределения для данного массива, но нет никаких статистических критериев и отвечающих им достигнутых уровней значимости. Полагаю, что Вы уже догадались, почему их нет. Конечно, их нет, поскольку мы их и не заказали ранее. Итак, закажем проверку нормальности с помощью статистических критериев. Для этого на нашем меню (см. слева) щелкнем мышкой в маленьких окошечках рядом с "Normal expected frequencies", "K-S and Lilliefors test for normality", "Shapiro-Wilk's W test". После этого часть нашего меню станет выглядеть так (см. ниже):



**Distribution**

Frequency tables Histograms

Normal expected frequencies

K-S and Lilliefors test for normality

Shapiro-Wilk's W test

**Categorization**

Number of intervals: 10

Integer intervals [categories]

BASIC STATS

K-S d=,05455, p> .20; Lilliefors p> .20  
Shapiro-Wilk W=,97882, p<,4662

Category	Count	Cumu Count	Perce of Va	Cumul of Va	% of Case	Cumul of Al	Expec Coun	Cumu Expec	Perce Expec	Cumul Expec
-4,000 < x <= -3	0	0	0,	0,	0,	0,	,	,	,	,
-3,000 < x <= -2	3	3	3,	3,	3,	3,	3,	4,	3,	4,
-2,000 < x <= -1	16	19	16,	19,	16,	19,	15,	19,	15,	19,
-1,000 < x <= 0	34	53	34,	53,	34,	53,	33,	51,	33,	51,
0,0000 < x <= 1	30	83	30,	83,	30,	83,	32,	83,	32,	83,
1,0000 < x <= 2	14	97	14,	97,	14,	97,	14,	97,	14,	97,
2,0000 < x <= 3	3	100	3,	1E2	3,	1E2	3,	1E2	3,	1E2
Missing	0	100	0,		0,	1E2				

Т.е. в этих окошечках появились "галочки", что означает, что данные опции стали активны. Далее

щелкнем по кнопке "Frequency tables" и посмотрим, что же мы получим в этот раз (см. слева). Как видим, теперь кроме фактических частот появились и частоты ожидаемые, вычисленные в

предположении нормальности. Далее, в верхней части таблицы, и соответственно в выходном (output-файле, который по умолчанию формируется в rtf-формате) файле, появились следующие строки: K-S d=,05455, p> .20; Lilliefors p> .20 Shapiro-Wilk W=,97882, p<,4662 . Это означает, что в тесте Колмогорова-Смирнова максимальное расстояние dmax между фактической и теоретической функциями распределениями d=,05455.

(Более подробно о применении этого критерия см. в **БИОМЕТРИКЕ** нашу статью "Критерий Колмогорова-Смирнова: особенности применения").

Этой величине отвечает достигнутый уровень значимости "p"=0,20, такое же значение получено и при использовании так называемой поправки Лиллиефорса. Значение статистики Шапиро-Уилки оказалось равным 0,97882 с уровнем значимости "p"=0,4662. Поскольку все уровни значимости более 5%, то мы принимаем нулевую гипотезу о том, что распределение признака VAR1 подчиняется нормальному закону.

**Алена! Я предлагаю Вам провести проверку нормальности распределения аналогичным образом для всех количественных признаков в Вашем массиве. Сделайте это отдельно по сортам картофеля и фенофазам. Для этой цели вначале скопируйте ваш массив несколько раз с разными именами (по сортам и фенофазам). Затем в каждом из массивов удалите лишнее и оставьте только нужные строки (наблюдения). В своем ответном письме, которое Вы подготовите в течение недели, Вы Алена, подробно сообщите мне о результатах этой проверки. В следующих лекциях мы узнаем как это выполнить не создавая отдельные массивы.**

Рассмотрим теперь следующий пункт Программы: Применение графического способа оценки

Box & whisker plot for all variables Categorized box & whisker plots

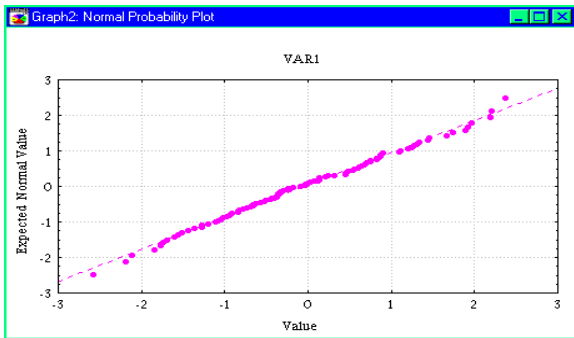
Normal probability plots Categorized means (interaction) plots

Half-normal probability plots Categorized histograms

нормальности распределения. В нижней части меню есть кнопка "Normal probability plots" (Здесь она посередине слева). Если Вы, Алена,

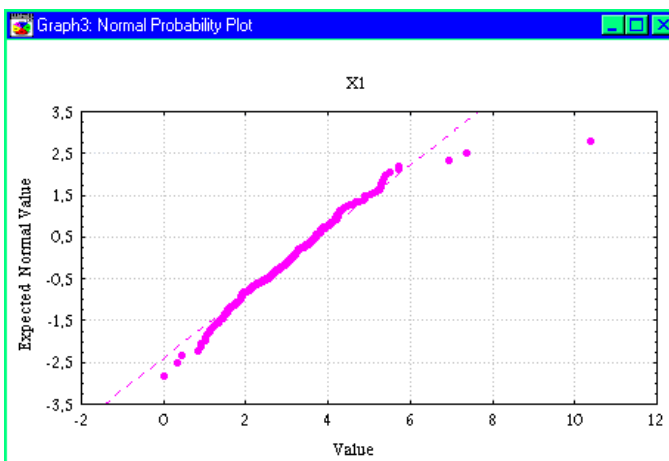
## 10 Лекция 1 - учебный материал Н-вой Алёне Николаевне.

почитаете в моей и других книгах относительно функции плотности нормального распределения, то вспомните, что в выражении для этой плотности есть экспонента. Если прологарифмировать это выражение, то эта экспонента даст нам выражение для прямой линии. Именно это свойство и используется при графическом методе проверки нормальности распределения. В данных осях значения



переменной близкой к нормальному распределению будут расположены вблизи прямой линии, как на рисунке слева, полученном после нажатия на кнопку "**Normal probability plots**". Для пояснения скажу Вам Алёна, что данный массив из 100 чисел я сам сгенерировал (создал) в пакете Excel задав нормальное распределение с параметрами "среднее=0", "стандартное отклонение=1".

Поэтому как по вертикали, так и по горизонтали

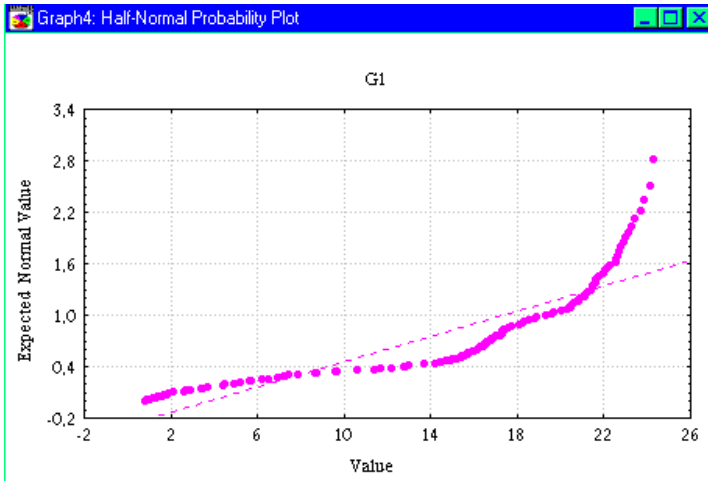


интервал изменения от -3 до +3. Давайте посмотрим как будет выглядеть аналогичный график (см. слева) для одного из признаков (переменных) Вашего массива, например признака X1- Количество хлорофилла "а" в листе, (мг CO<sub>2</sub>/дм<sup>2</sup>). Обратите внимание, что в верхней правой части графика есть три точки расположенные далеко от прямой. Особенно сильно выделяется одна точка. Тогда как

большинство точек расположены либо на самой прямой нормального распределения, либо в непосредственной близости от нее. Нетрудно определить что же это за наблюдения: это те образцы, для которых значения X1 (хлорофилла "а") больше 6. Точнее для двух точек это значения порядка 7 (от 6 до 8), тогда как для самой удаленной точки это значение порядка 10 с небольшим, точнее 10,3962 (найти это значение можно в самом массиве в строке с номером 60). Итак, можно сделать вывод о том, что данное наблюдение является аномальным, выбросом, по сравнению с остальным массивом. И по этой причине имеет смысл его удалить из массива. Т.о. Алёна, графический способ полезен еще и как метод оценки аномальных выбросов. Очень важный вывод, который следует запомнить, заключается еще и в том, что эти аномальные точки (наблюдения, опыты, эксперименты - в нашем случае эти слова есть синонимы) **удалены** от основного массива точек. Т.е. наблюдается разрыв между ними и прямой (или кривой) на которой расположен основной массив точек. Это-то и позволяет прийти к выводу о том, что основной массив подчиняется нормальному закону, и лишь несколько наблюдений являются аномальными.

## 11 Лекция 1 - учебный материал Н-вой Алене Николаевне.

А теперь посмотрим как выглядит аналогичная картинка для другого признака - G1 - ширина (см) (рисунок ниже).



Здесь мы уже не видим выбросов точек действительно удаленных от непрерывной кривой. Но точки расположены не на прямой, а на кривой. Причем эту кривую можно условно разделить на несколько частей. К примеру, слева, в интервале G1 от самых минимальных до примерно до 14 см, точки расположены на прямой линии, не на пунктирной, отвечающей

нормальному распределению для всего массива, но все же на прямой. Далее мы видим среднюю часть, состоящую из двух вогнутых кривых (интервалы 14-18 и 18-21,5). И наконец последняя правая верхняя часть (интервал от 21,5 до максимальных значений). Это позволяет сделать вывод о том, что имеет смысл разделить на сорта и фенотипы и тогда возможно для отдельных подматриц (подмассивов) мы увидим нормальность распределения для каждого случая.

**Проверка гипотезы о равенстве математических ожиданий двух нормальных распределений. Типичные ошибки использования t-критерия Стьюдента при анализе биомедицинских данных. Проверка гипотезы о равенстве дисперсий двух нормальных распределений.**

Для проверки гипотезы о равенстве генеральных средних двух нормальных распределений можно использовать несколько критериев. Самый популярный из них, это t-критерий Стьюдента. К сожалению, в силу своей популярности и простоты вычисления он чаще всего используется ошибочно. Если Вы внимательно читали материалы раздела "Кунсткамера" в "БИОМЕТРИКЕ", то обратили внимание, что эти ошибки типичны...

---

**Полностью данная лекция занимала более 30 страниц текста...**