

## ЛАБОРАТОРНАЯ РАБОТА № 2

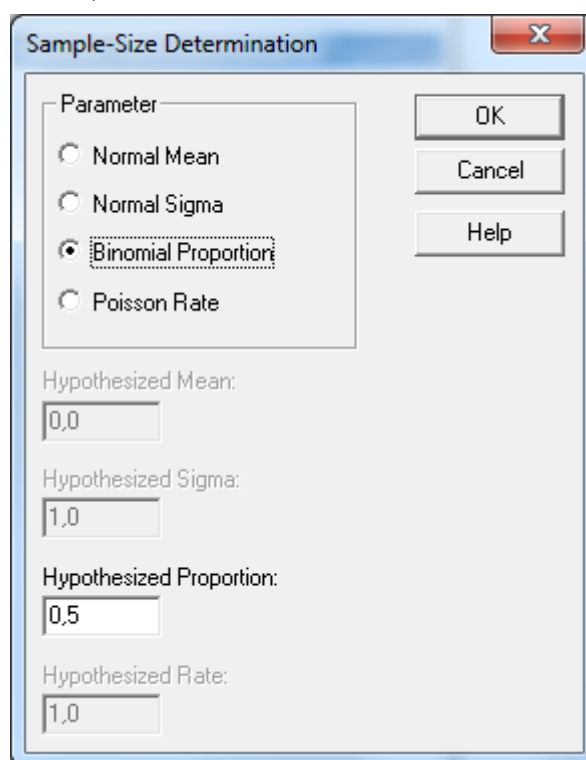
### СВЯЗЬ СТАТИСТИКИ С ТЕОРИЕЙ ВЕРОЯТНОСТЕЙ

#### Определение объема выборки

Для проведения статистических исследований бывает необходимо знать, выборку какого объема мы должны использовать. Понятно, что если оценка состоятельная, то чем большего объема выборка, тем лучше результат. Но увеличение объема выборки обычно приводит к увеличению стоимости исследования.

Рассмотрим задачу: Одна из компаний, проводящих социологические исследования, перед очередными президентскими выборами, на основании опроса 200 избирателей выдала прогноз, что с вероятностью 0,95 шансы Барака Обамы и Митта Ромни равны. При этом утверждается, что оценка погрешности такого прогноза не более 0,03. Можем ли мы доверять этому результату? Каков должен был быть объем выборки, чтобы этому утверждению можно было доверять?

Для решения задачи выберите в меню *Tools, Sample Size Determination, One Sample*. Раскроется окно, в котором надо выбрать тип распределения (в нашем случае биномиальное)



Гипотетическая вероятность у нас 0,5. Нажмите ОК, раскроется окно

**Sample-Size Determination Options**

**Control**

☒ Absolute Error  
+/- 0,03

☐ Relative Error  
+/- 10,0 %

☐ Power  
95,0 %

Difference to Detect:  
0,05

☐ Sample Size  
30

**Confidence Level:**  
95,0 %

**Alternative Hypothesis**

☒ Not Equal  
☐ Less Than  
☐ Greater Than

**Sigma**

☒ To Be Estimated  
☐ Known

OK  
Cancel  
Help

В этом окне надо ввести уровень ошибки (0,03), нажмите ОК. Раскроется результат

### Sample-Size Determination

Parameter to be estimated: binomial parameter  
Desired tolerance: +/- 0,03 when proportion = 0,5  
Confidence level: 95,0%

The required sample size is n=1098 observations.

#### **The StatAdvisor**

This procedure determines the sample size required when estimating the proportion of a binomial distribution. 1098 observations are required to estimate theta to within +/-0,03 (assuming theta is around 0,5) with 95,0% confidence.

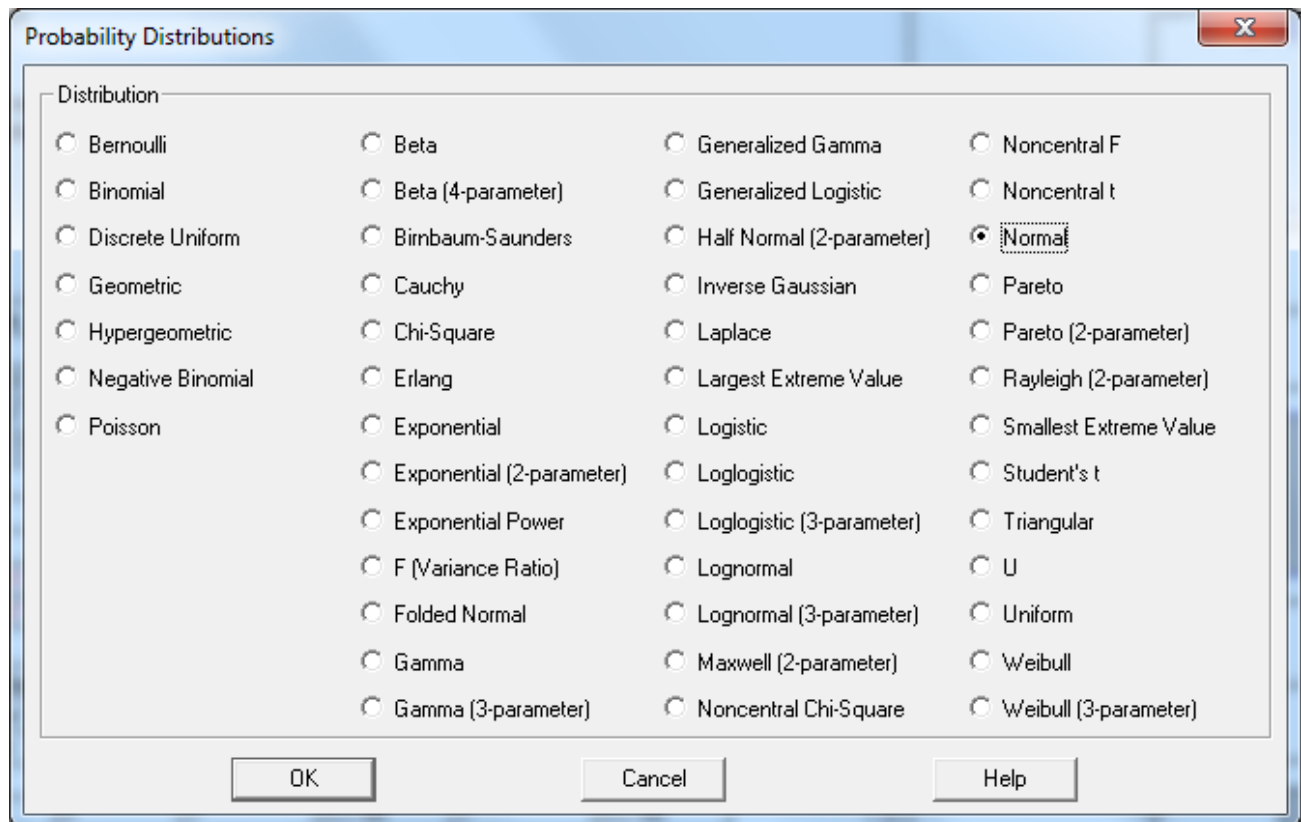
Т.е. требуемый объем выборки – 1098 наблюдений. Только выборка такого объема гарантирует нам требуемую точность.

Пусть мы решили, что нас устраивает погрешность ошибки 5%. Сколько людей мы должны опросить в этом случае? Результат покажите преподавателю.

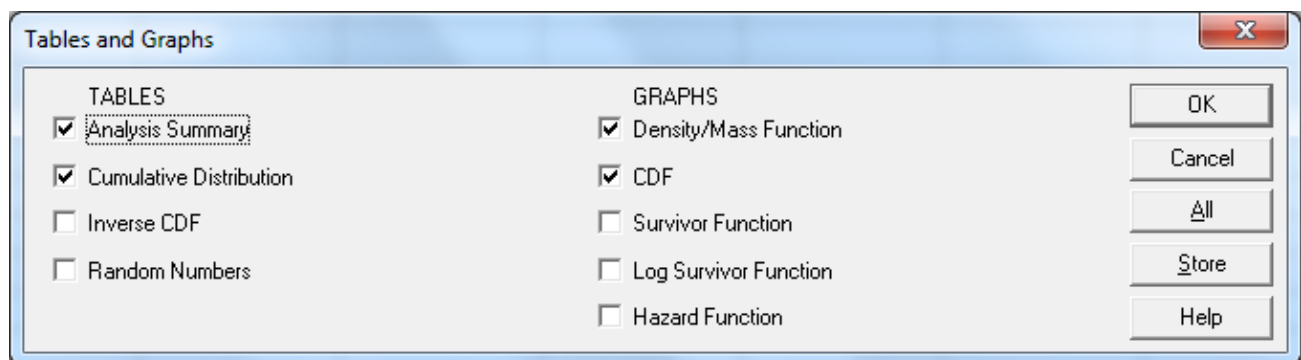
## **Генератор случайных чисел**

Для получения нужных нам на занятиях выборок не обязательно организовывать и проводить специальное исследование, можно поступить проще. В пакете *Statgraphics* есть генератор случайных чисел, позволяющий имитировать случайный выбор из генеральных совокупностей различной природы. Воспользуемся им сейчас.

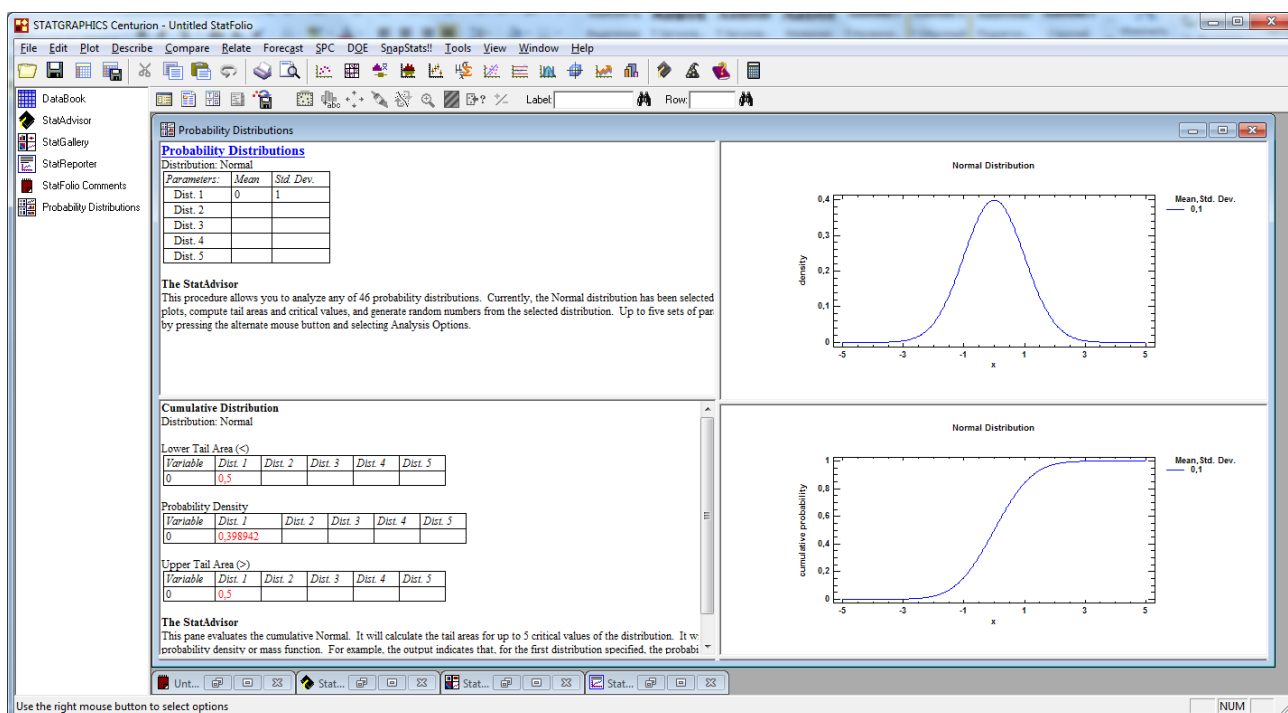
Выберите в меню **Plot, Probability Distributions**. Раскроется следующее диалоговое окно



В данном окне можно выбрать одно из возможных распределений, выберите **Normal**, нажмите кнопку ОК, раскроется окно



Нажмите ОК. Перед Вами раскрылось окно **Probability Distribution**, по умолчанию сгенерировалось распределение с параметрами  $N(0,1)$ .




Если вы хотите получить распределение с другими параметрами, щелкните в этом окне правой кнопкой, в появившемся контекстном меню выберите *Analysis Options*, раскроется окно.

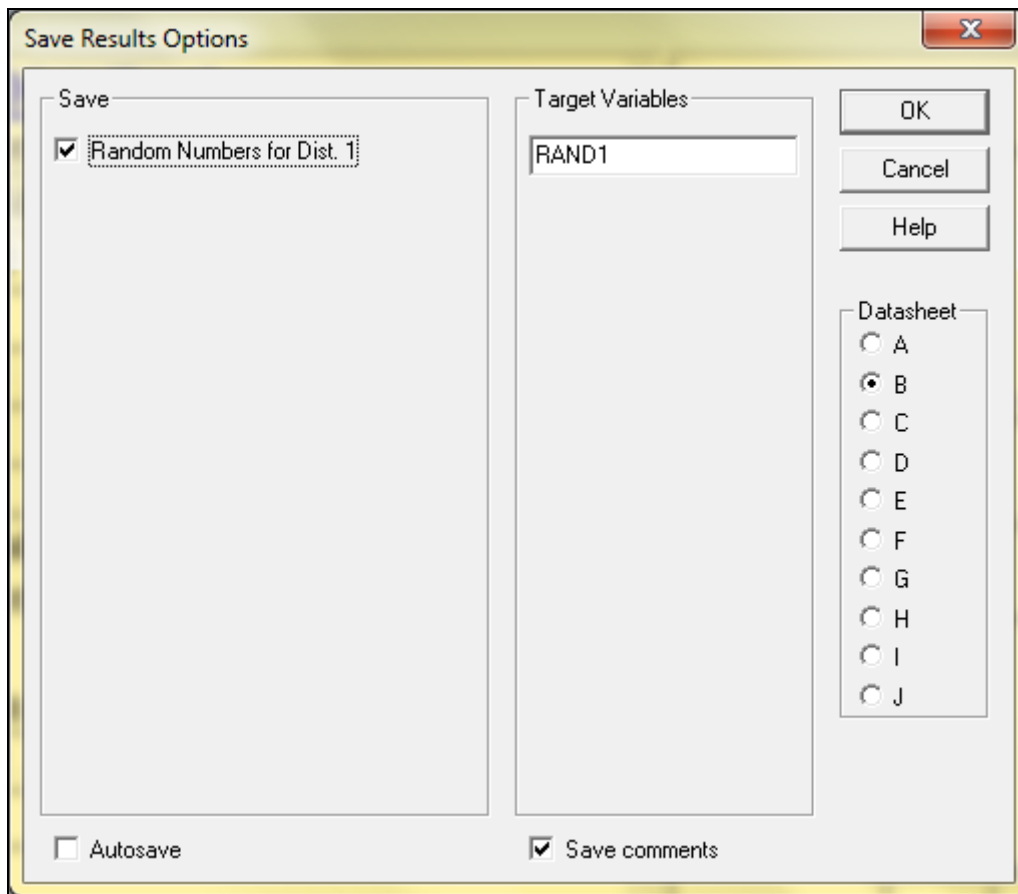
**Normal Options**


Mean	Std. Dev.
0.0	1.0


OK  
Cancel  
Help

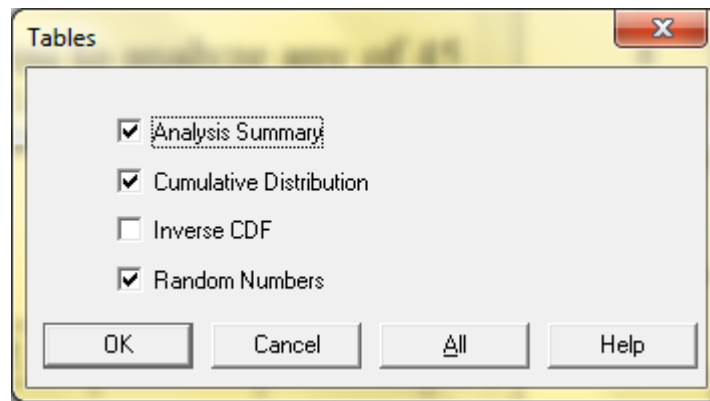
В этом окне укажите параметры распределения (т. е. среднее и стандартное отклонение), нажмите кнопку ОК. Если вы хотите вывести на экран несколько выборок из нормально распределенных генеральных совокупностей, можно набрать в этом окне их параметры.

Для того чтобы данные появились в таблице, нажмите кнопку *Save results* , раскроется диалоговое окно, в котором нужно включить *Random Numbers for Dist.1*.

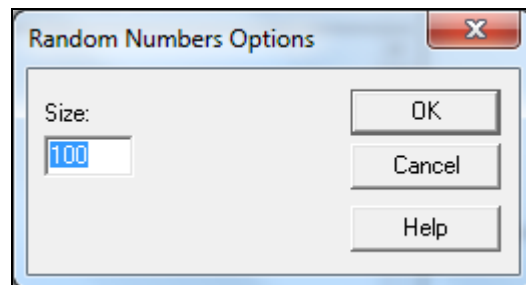


Если вы не изменили имя *RAND1* в поле **Target Variables**, то столбец данных будет называться *RAND1*. Если же вы хотите изменить имя столбца, сделайте это. Обратите внимание, в каком листе вы сохраняете результат. Затем нажмите кнопку ОК. Вы можете посмотреть, что получилось, раскрыв окно, которое пока имеет имя *Untitled*, и выбрав соответствующий лист. Его можно раскрыть, например, выбрав в строке меню *Window* либо выбрав в левой части окна иконку . Вернитесь теперь в окно с результатами *Probability Distributions*.

По умолчанию объем выборки будет 100. Допустим, вы хотите использовать выборку другого объема. Для этого нажмите кнопку **Table and graphs** , выберите в раскрывшемся окне *Random Numbers*,



нажмите кнопку ОК, в раскрывшемся окне *Random Numbers* щелкните правой кнопкой, в контекстном меню выберите *Pane Options*. Перед вами раскроется окно,

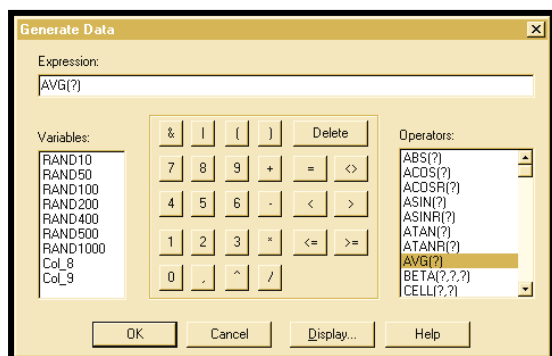


в нем задайте объем выборки. Нажмите кнопку ОК. Если есть необходимость в изменении параметров распределения — щелкните правой кнопкой, выберите *Analyses Options*. Для сохранения новой выборки нажмите кнопку *Save Results*, введите новое имя выборки (если сохранить старое, то новая выборка перекроет старую). Выполните операцию генерирования выборки.

Выясним, что происходит с частотой при увеличении  $n$ .

1. Выберите в строке меню *Plot, Probability Distributions*. Сгенерируйте выборку из распределения *Bernoulli* размером 10 (обратите внимание на задаваемую вами ожидаемую вероятность, она должна быть равной 0,5). Таким же образом сгенерируйте еще пять выборок с объемами  $n=50, 100, 200, 500, 1000$  (во всех выборках  $p=0,5$ ). Назовите столбцы *RAND10, RAND50, RAND100, RAND200, RAND500, RAND1000*. Посмотрите, какие выборки получились. *Обязательно контролируйте объем выборок!* (Если объем выборки неверный – повторите генерирование). Вы должны обнаружить, что выборки состоят из нулей и единиц. Измените тип чисел на *Integer* (вспоминайте, как это сделать, посмотрите работу № 1). Можно представить себе, что эти выборки были получе-

ны в результате бросания монетки: так, первая выборка есть серия из 10 бросаний, в некоторых из них выпал герб (1), в других — не выпал (0).



Посмотрим, что происходит с частотой появления герба при увеличении объема выборки. Для бернуллиевского распределения эта частота совпадает с выборочным средним (почему?). Щелкните **правой** кнопкой по названию свободного

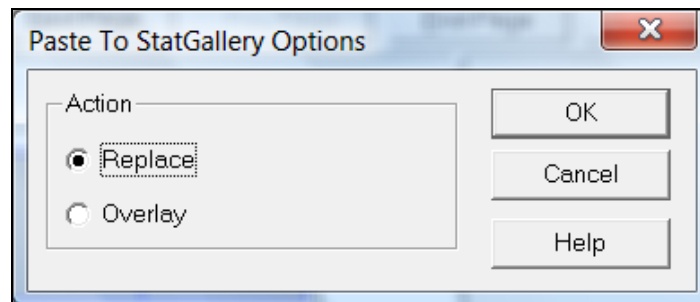
столбца, в появившемся контекстном меню выберите **Generate Data**, щелкните левой кнопкой. Раскроется окно **Generate Data**, в нем в правой части в поле **Operators** выберите **AVG(?)** (среднее), щелкните дважды, название перейдет в поле **Expression**. Затем щелкните в этом поле, сотрите в скобках вопросительный знак, щелкните дважды в поле **Variables** по названию нужного вам столбца, например, RAND10. Прodelайте это для каждой из выборок. В первой строке вашей таблицы будут выписаны частоты для каждой из выборок. Хотелось бы представить результаты графически, пока вы не умеете этого делать в *Statgraphics*, поэтому нарисуйте в тетради график изменения частоты в зависимости от объема выборки. Проанализируйте полученные результаты. К чему должна приближаться относительная частота? На основании какой теоремы из курса теории вероятностей мы могли бы предсказать вид полученного графика? Обсудите результат с преподавателем.

Посмотрим, что происходит с полигоном (гистограммой) при увеличении объема выборки  $n$ .


Сгенерируйте (тип датчика — **Normal**) выборку объема  $n=1000$ . Сохраните ее с именем  $N1000$ . Постройте полигон для выборки, причем сделайте его относительным (**relative**). Далее будем строить выборки этого распределения с разными объемами, и сохранять их с именами  $N500$ ,  $N100$ ,  $N20$ . Сделаем так, чтобы полигоны для всех выборок можно было просмотреть на одном экране. Щелкните снова правой кнопкой по графику, выберите в контекстном меню

***Copy Pane to Gallery***, раскроется окно, щелкните в левом верхнем углу правой кнопкой, выберите ***Paste***.

Сгенерируйте вторую выборку объемом  $n=500$ . Постройте для нее полигон, щелкните дважды по графику, чтобы раскрыть во все окно, затем щелкните по полигону правой кнопкой, если он будет того же цвета, смените его цвет заполнения (пусть линия будет желтой). Затем снова выберите ***Copy Pane to Gallery***, в раскрывшемся окне щелкните правой кнопкой, выберите ***Paste***, в раскрывшемся окне выберите ***Overlay*** (наложение), щелкните ОК.



Разверните в нижней части экрана ***StatGallery***. На раскрывшемся экране увидите два полигона, наложенные друг на друга. Сверните экран ***StatGallery*** снова. Самостоятельно постройте полигоны для выборок объема  $n=100$  (зеленая линия) и  $n=20$  (красная линия). Все четыре полигона должны быть на одном графике. *Покажите результат преподавателю.*

Чтобы убедиться, что не всегда полигон при увеличении объема выборки приближается к нормальной кривой Гаусса, сгенерируйте выборку объема 1000 из генеральной совокупности с распределением ***Uniform*** (равномерное распределение), назовите ее ***U1000***. Постройте для этой выборки гистограмму частот с аппроксимирующей кривой типа ***Normal***. Для того чтобы это сделать, выберите в меню ***Describe, Distribution Fitting, Fitting Uncensored Data***. Выберите переменную ***U1000***, нажмите ОК. Насколько хорошо гистограмма укладывается на нормальную кривую? Нажмите кнопку  ***Input dialog***, выберите столбец ***N1000***. Посмотрите, как изменится график. *Объясните результат преподавателю.*



Выясним, как изменяются графики плотности для различных распределений при изменении их параметров.

В пакете *Statgraphics* вы можете построить графики плотностей (и функций распределений) для некоторых типичных распределений. До сих пор Вы строили их **статистические аналоги** – гистограмму и полигон. Давайте посмотрим, как будут меняться графики плотности некоторых известных вам распределений при изменении параметров. Постройте в одном окне графики нормального распределения  $N(0,1)$ ,  $N(0,2)$ ,  $N(6,1)$ . Для этого в строке меню выберите **Plot**, затем **Probability Distributions**, затем **Normal**, **OK**. Щелкните **правой** кнопкой, выберите **Analysis Options**, сгенерируйте сразу три нужные выборки.



Чем отличается график распределения  $N(0,2)$  от  $N(0,1)$ ,  $N(6,1)$  от  $N(0,1)$ ? *Покажите результат преподавателю.*

Повторите то же самое для распределения  $\chi^2$ -квадрат (*Chi-Squared*). Задайте число степеней свободы 4, 10, 50. Что происходит с графиками с увеличением числа степеней свободы? *Покажите результат преподавателю.*

Повторите все для распределения Стьюдента (*Student's t*). Задайте числа степеней свободы 1, 5, 30. Что происходит с графиками с увеличением числа степеней свободы? *Покажите результат преподавателю.*

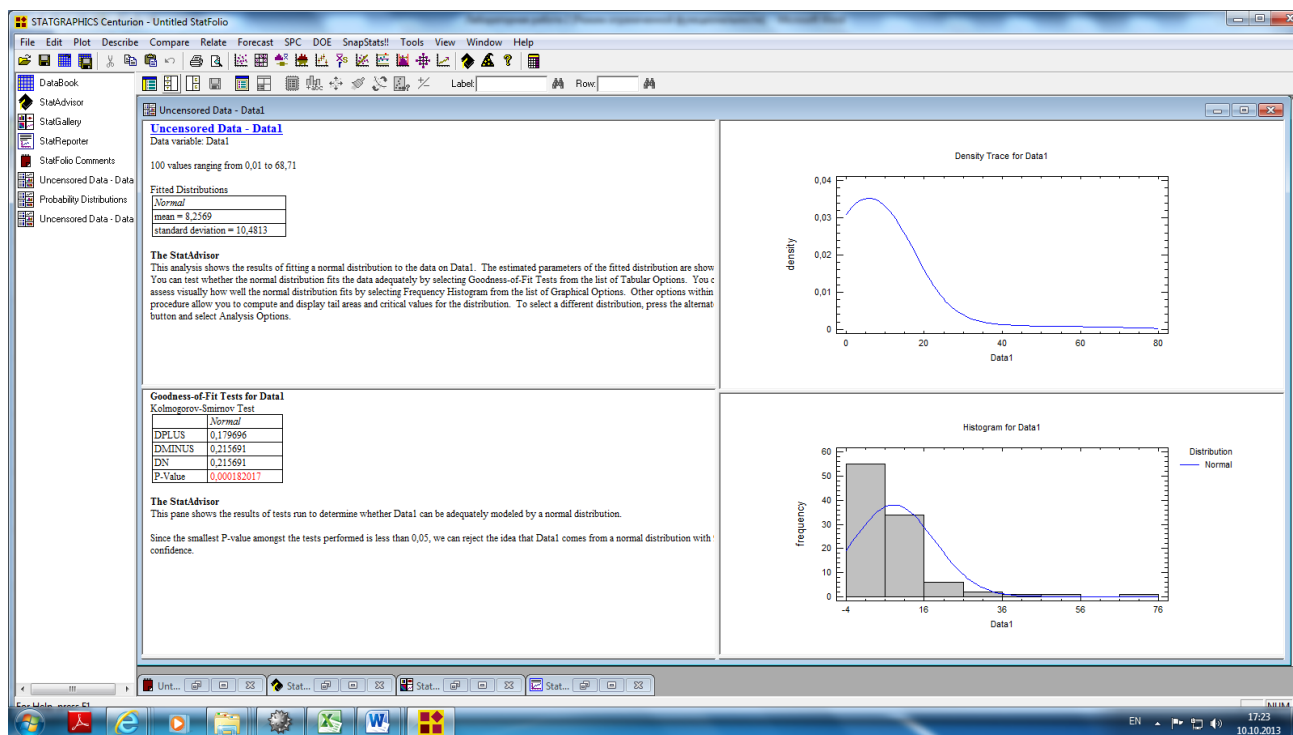
То же самое для распределения Фишера *F(Variance Ratio)*. В поле **Numerator df** (число степеней свободы числителя) задайте параметры 6, 12, 6. В поле **Denominator df** (число степеней свободы знаменателя) задайте значения параметра 6, 6, 60. *Покажите результат преподавателю.*

Убедимся теперь, что распределения Стьюдента и  $\chi^2$ –Пирсона действительно асимптотически нормальны. Постройте графики функций распределений  $N(0,1)$ ,  $t(30)$ ,  $\chi^2(50)$ ,  $N(50,10)$ . Почему нормальное распределение берется с такими параметрами? Теперь сделайте так, чтобы в окне *StatGallery* в левой части экрана были наложены друг на друга графики  $N(0,1)$ ,  $t(30)$ , а в правой —  $\chi^2(50)$ ,  $N(50,10)$ . В итоге вы должны увидеть, что распределение  $t(30)$  практически не отличается от  $N(0,1)$ , а  $\chi^2(50)$  мало отличается от  $N(50,10)$ . В статистике часто используется этот факт: распределение Стьюдента и распределение  $\chi^2$  являются асимптотически нормальными, т. е.  $t(30) \cong N(0,1)$ , а  $\chi^2(v) \cong N(v, \sqrt{2v})$  при  $v \geq 50$ . *Покажите результат преподавателю.*

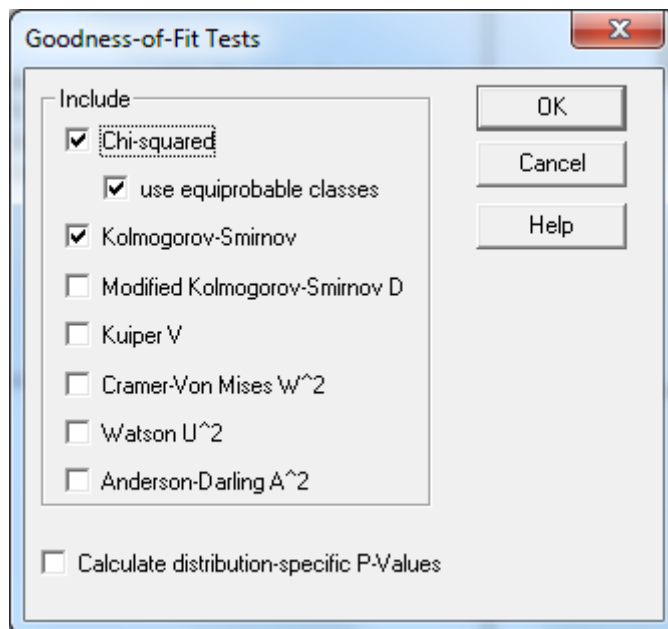
В некоторых случаях нам необходимо узнать распределение генеральной совокупности, из которой взята выборка. Рассмотрим следующую задачу:

**Задача.** Измерялось время обращения к некоторой базе данных. Данные находятся в файле **Очередь к БД.sf6**. определить тип распределения генеральной совокупности.

В меню выберите *Distribution Fitting*, затем *Fitting Uncensored Data*. Введите в окне выборку, раскроется окно



в этом окне мы видим тест Колмогорова-Смирнова для проверки нормальности распределения. Видно, что  $p\text{-value}=0,00018 < 0,05$ ; следовательно, мы отвергаем гипотезу о нормальности распределения. Воспользуемся теперь критерием  $\chi^2$ , для этого щелкните в окне **Goodness-of-Fit Tests** правой кнопкой, выберите **Pane Options**, раскроется окно



Выберите в нем еще и распределение **Chi-squared**.

#### Goodness-of-Fit Tests for Data1

##### Chi-Squared Test

	<i>Lower</i>	<i>Upper</i>	<i>Observed</i>	<i>Expected</i>	
	<i>Limit</i>	<i>Limit</i>	<i>Frequency</i>	<i>Frequency</i>	<i>Chi-Squared</i>
at or below		-9,89315	0	4,17	4,17
	-9,89315	-6,23864	0	4,17	4,17
	-6,23864	-3,80023	0	4,17	4,17
	-3,80023	-1,88292	0	4,17	4,17
	-1,88292	-0,256184	0	4,17	4,17
	-0,256184	1,18738	18	4,17	45,93
	1,18738	2,50768	10	4,17	8,17
	2,50768	3,74232	10	4,17	8,17
	3,74232	4,91714	8	4,17	3,53
	4,91714	6,05133	9	4,17	5,61
	6,05133	7,16019	3	4,17	0,33
	7,16019	8,2569	7	4,17	1,93
	8,2569	9,35361	5	4,17	0,17
	9,35361	10,4625	6	4,17	0,81
	10,4625	11,5967	2	4,17	1,13
	11,5967	12,7715	4	4,17	0,01
	12,7715	14,0061	5	4,17	0,17
	14,0061	15,3264	2	4,17	1,13
	15,3264	16,77	1	4,17	2,41
	16,77	18,3967	2	4,17	1,13
	18,3967	20,314	1	4,17	2,41
	20,314	22,7524	1	4,17	2,41
	22,7524	26,4069	1	4,17	2,41
above	26,4069		5	4,17	0,17

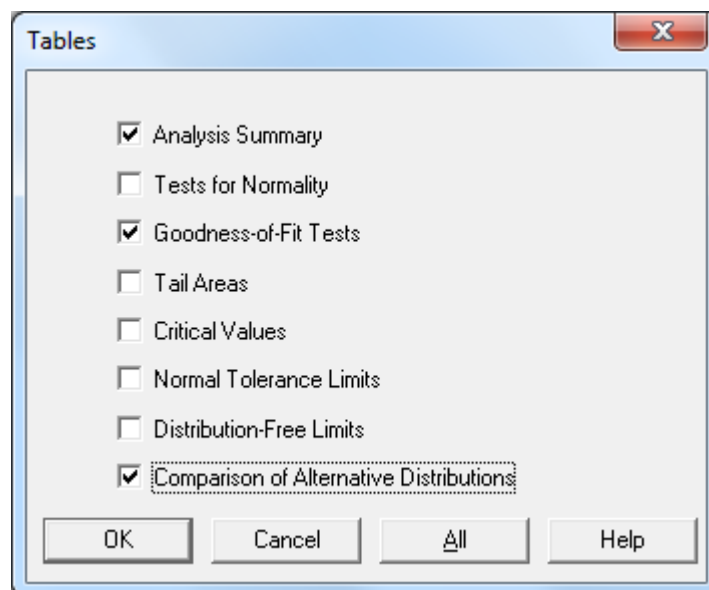
Chi-Squared = 108,8 with 21 d.f. P-Value = 7,76046E-14

#### Kolmogorov-Smirnov Test

	<i>Normal</i>
DPLUS	0,179696
DMINUS	0,215691
DN	0,215691
P-Value	0,000182017

Результат P-Value = 7,76046E-14 < 0,05 говорит о том, что критерий  $\chi^2$  также отвергает нормальность распределения. Попробуем подобрать распределение.

Нажмите кнопку  **Tables and graphs**, выберите



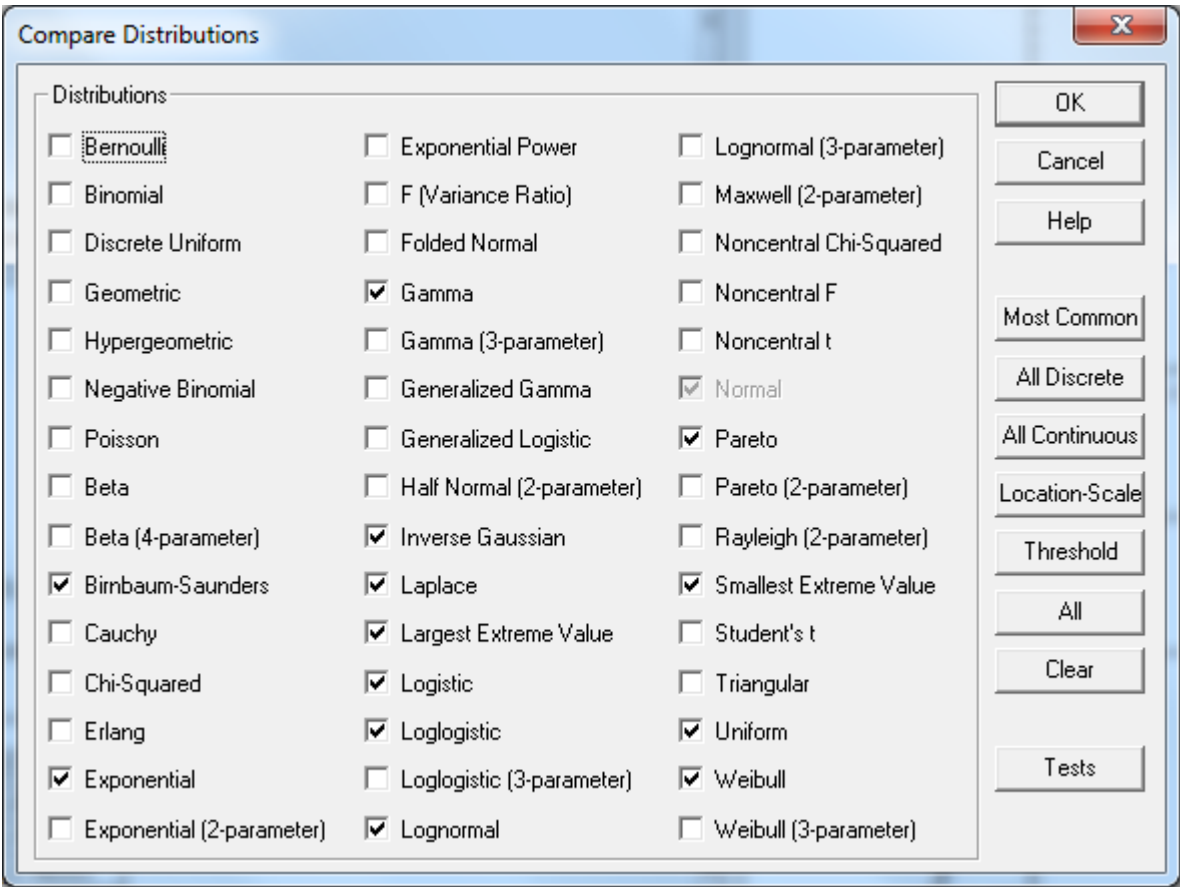
**Comparison of Alternative Distributions**, раскроется окно с результатами

#### Comparison of Alternative Distributions

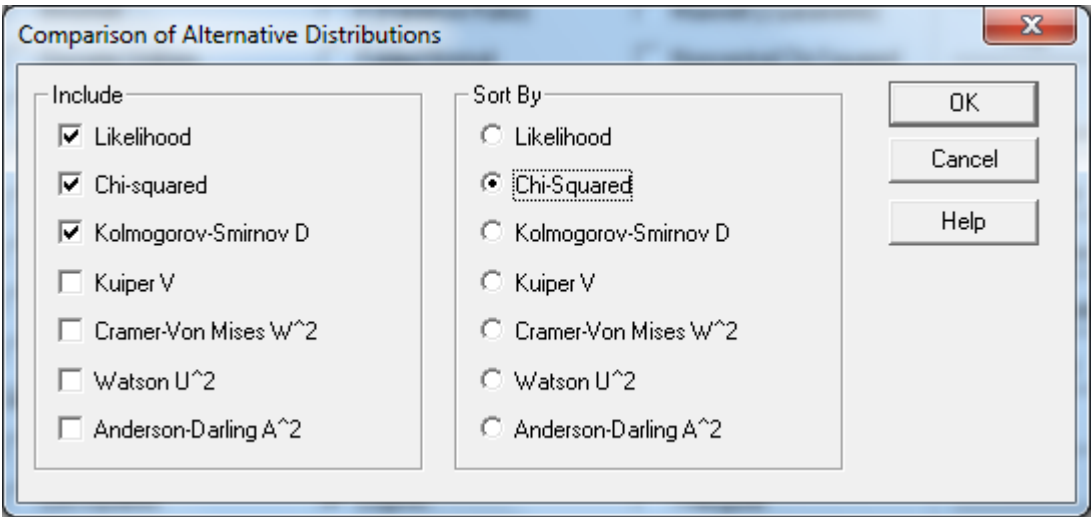
<i>Distribution</i>	<i>Est. Parameters</i>	<i>Log Likelihood</i>	<i>KS D</i>
Weibull	2	-308,671	0,0662112
Gamma	2	-308,793	0,0673059
Exponential	1	-311,105	0,0670094
Loglogistic	2	-315,851	0,0811315
Lognormal	2	-320,735	0,130128
Largest Extreme Value	2	-335,432	0,113865
Birnbaum-Saunders	2	-339,5	0,346297
Laplace	2	-348,682	0,209302
Logistic	2	-354,195	0,18223
Normal	2	-376,353	0,215691
Inverse Gaussian	2	-381,916	0,456014
Uniform	2	-422,975	0,674323
Pareto	1	-1,E9	0,26126
Smallest Extreme Value	<no fit>		

Здесь выводятся на экран результаты метода максимального правдоподобия (*Log Likelihood*) и максимальное отклонение из метода Колмогорова-Смирнова (*KS D*). Результаты упорядочены по результатам метода максимального правдоподобия. Можно вывести столбец со значениями P-Value для критерия  $\chi^2$  и упо-

рядочить по этим результатам. Для этого щелкните правой кнопкой, выберите *Pane Options*, раскроется окно



Нажмите кнопку *Tests*, раскроется окно



в окне в левой части окна *Include* выберите *Chi-squared*, а в части *Sort By* тоже *Chi-squared*. Появится окно с результатами

Comparison of Alternative Distributions				
Distribution	Est. Parameters	Log Likelihood	Chi-Squared P	KS D
Weibull	2	-308,671	0,703807	0,0662112
Gamma	2	-308,793	0,491066	0,0673059
Exponential	1	-311,105	0,306734	0,0670094

Loglogistic	2	-315,851	0,147973	0,0811315
Lognormal	2	-320,735	0,0341014	0,130128
Largest Extreme Value	2	-335,432	0,000289407	0,113865
Birnbaum-Saunders	2	-339,5	3,61789E-7	0,346297
Logistic	2	-354,195	3,03123E-7	0,18223
Laplace	2	-348,682	4,08922E-11	0,209302
Normal	2	-376,353	7,76046E-14	0,215691
Pareto	1	-1,E9	2,22045E-16	0,26126
Uniform	2	-422,975	0,0	0,674323
Inverse Gaussian	2	-381,916	0,0	0,456014
Smallest Extreme Value	<no fit>			

Как видите, пакет предлагает распределение Вейбулла (Weibull), впрочем, для распределений с одним параметром предлагается показательное распределение (Exponential).

### **Задание.**

1. Сгенерировать три выборки объемом 1000, каждая из которых имеет биномиальное распределение

- 1) с параметрами  $p=0,5$  ;  $N(\text{trial})=100$
- 2) с параметрами  $p=0,01$  ;  $N(\text{trial})=100$
- 3) с параметрами  $p=0,99$  ;  $N(\text{trial})=100$

2. Постройте гистограммы для каждой из выборок с наложенной нормальной плотностью.

А) Для какой из выборок гистограмма «похожа» на нормальную кривую? Почему это можно было ожидать (вспомните предельные теоремы из теории вероятностей).

Б) На какое распределение должна быть «похожа» гистограмма для второго распределения? Наложите это распределение на гистограмму.

В) Почему нормальная аппроксимация дает плохой результат для третьей выборки?

Сформируйте новую выборку, в которой элементы третьей выборки вычитаются из ста. На что будет похожа гистограмма такой выборки?

### **ВОПРОСЫ**

1. Каков содержательный смысл распределения Бернулли?
2. Почему для распределения Бернулли выборочное среднее совпадает с частотой?
3. Почему в приложениях чаще других встречается нормальное распределение?
4. Что происходит с частотой появления единицы для распределения Бернулли при увеличении объема выборки? В какой теореме из теории вероятностей обосновывается полученное утверждение?
5. Каков содержательный смысл параметров биномиального распределения?

6. Какой смысл имеют параметры  $a$ ,  $b$  в распределении  $N(a,b)$ ?
7. Что происходит с графиком плотности нормального распределения, если увеличивать математическое ожидание? Дисперсию?
8. Что происходит с графиком плотности распределения Стюдента при увеличении числа степеней свободы?
9. Что происходит с графиком плотности распределения  $\chi^2$  при увеличении числа степеней свободы?