Пособие для ВФ по математической статистике.

Составитель Зайцев А.А.
Практические работы по математической статистике в среде MS EXCEL. Учебно-методическое пособие по курсу “Экономико-математические методы и моделирование”. М.,МИИГАиК, 2009г., 22стр.

Учебно-методическое пособие составлено в соответствии с утверждённой программой курса “Экономико-математические методы и моделирование” для студентов вечернего факультета (специальность “Землеустройство и кадастры”), рекомендуемого кафедрой вычислительной техники и автоматизированной обработки аэрокосмической информации.

Оно содержит краткое теоретическое введение и практические задания по математической статистике с подробными описаниями и примерами.

Библиография: 5 названий.

Рецензенты:

Оглавление.

Предисловие.

При изучении курса “Экономико-математические методы и моделирование” студентам предлагается выполнить практические работы на ПК по математической статистике (в среде MS EXCEL). Это связано с тем, что до этого студентам читается курс теории вероятностей с элементами математической статистики, но на статистику отводится очень мало часов. При этом практические занятия на ПК отсутствуют, хотя для статистики они очень желательны.

Поэтому в данном пособии мы очень кратко напоминаем необходимые теоретические понятия из математической статистики, отсылая за подробностями к литературе. Затем подробно описываем содержание предлагаемых практических работ, даём примеры их выполнения и варианты самостоятельных заданий.

Мы предполагаем, что в предыдущем обучении студенты в основном освоили работу в EXCEL и даём пояснения лишь к его конструкциям, специфическим для задач математической статистики. Укажем также, что на ПК должен быть установлен прикладной Пакет анализа, доступный через команду Анализ данных меню Сервис. Если эта команда отсутствует в меню, в меню Сервис/Надстройки необходимо активизировать (т.е. поставить галочку) пункт Пакет анализа.

Теоретическое введение.

Случайные величины.

Случайная величина (с.в.) есть величина, которая может принимать свои значения с заранее заданными вероятностями. Дискретная случайная величина X (д.с.в.) задаётся своими значениями $x_1, x_2, \ldots, x_n$ и их вероятностями $p_1, p_2, \ldots, p_n$, причём $p_1 + p_2 + \ldots + p_n = 1$ (говорят, что $x_1, x_2, \ldots, x_n, p_1, p_2, \ldots, p_n$ задают закон распределения). Непрерывная случайная величина (н.с.в.) X задаётся непрерывной функцией $F(x)$, которая задаёт вероятности $P(X < x) = F(x)$. Функция $F(x)$ называется функцией распределения с.в. и имеет смысл и для д.с.в.: $F(x) = P(X < x) = \sum_{x_i < x} p_i$. Функция распределения (ф.р.) полностью задаёт с.в.; она принимает значения из отрезка [0;1], определена при всех вещественных x, монотонно возрастает и непрерывна слева. Если $F(x)$ дифференцируема, то её производная $f(x) = F'(x)$ называется плотностью распределения. $f(x)$ принимает только неотрицательные значения, $F(x) = \int_{-\infty}^{x} f(t)dt$, а интеграл $\int_a^b f(x)dx$ равен вероятности того, что с.в. X примет значение из интервала $(a;b)$: $\int_a^b f(x)dx = P(a < X < b)$. В частности, $\int_{-\infty}^{\infty} f(x)dx = 1.$

Говорят, что с.в. имеет распределение некоторого вида, если её ф.р. (или плотность, или закон распределения) имеет определённый вид (часто встречающийся в математической статистике) . Например, для д.с.в. формула $P(X = k) = C_n^k p^k q^{n-k}$ (где n, p - параметры, $0 < p < 1$, $q = 1 - p$, n - натуральное число, $k = 0,1,\ldots,n$) характеризует биномиальное распределение, а формула $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-a)^2}{2\sigma^2}\right)$ задаёт плотность нормального распределения с параметрами a и σ>0. Величины, связанные с такими распределениями, обычно можно вычислять, используя раздел Статистические функции EXCEL.

Для наглядного изображения функций в математической статистике, кроме обычных в математике графиков, используются гистограммы (столбчатые диаграммы). Это удобно, если аргумент функции принимает конечное число значений.

Выборка и анализ с.в. по выборке.

Пусть имеется некоторое множество исследуемых объектов (генеральная совокупность), каждый из которых характеризуется количественным признаком (числом). Будем наугад выбирать объекты из генеральной совокупности и смотреть значение количественного признака на этом объекте. Получим с.в. X с некоторым законом распределения (определяемым генеральной совокупностью). Задача математической статистики – описать этот закон распределения , зная значения $x_1,\ldots,x_n$ количественного признака на выбранных элементах $\xi_1,\ldots,\xi_n$ генеральной совокупности (по выборке). Набор $x_1,\ldots,x_n$ называют реализацией выборки; поскольку элементы $\xi_1,\ldots,\xi_n$ интересуют нас только с точки зрения их количественного признака, мы для простоты будем называть $x_1,\ldots,x_n$ просто выборкой. Если расположить элементы выборки по возрастанию, то полученную числовую последовательность называют вариационным рядом.

Функция распределения и плотность с.в. X называются теоретическими, а величины, вычисленные по выборке $x_1,\ldots,x_n$ (и характеризующие её) – эмпирическими. К ним относятся выборочное среднее $x_в = \dfrac{x_1+\ldots+x_n}{n}$, выборочная дисперсия $D_в = \dfrac{\sum (x_в - x_i)^2}{n}$, выборочное среднее квадратическое отклонение $\sigma_в = \sqrt{D_в}$, размах $(x_{\max} - x_{\min})$ (здесь $x_{\max}$ и $x_{\min}$ - наибольшее и наименьшее среди чисел $x_1,\ldots,x_n$), коэффициент асимметрии и эксцесс. Кроме того, выборку можно характеризовать относительными частотами и эмпирической функцией распределения (аналогами плотности и ф.р.).

Объясним эти понятия подробнее. Разобьём интервал $(x_{\min}; x_{\max})$ на несколько меньших интервалов точками $a_0 = x_{\min}, a_1, \ldots, a_l = x_{\max}$, считая $a_0 < a_1 < \ldots < a_l$. Каждое число x_j из выборки попадёт в некоторый интервал (числа x_j, совпавшие с границами интервалов, отнесём к любому из них). Тогда можно найти частоты n_i, т.е. количество чисел из выборки, попавших в интервал (a_{i-1}, a_i). Величины n_i/n называются относительными частотами, а функция $F^*(x) = \sum_{x_i < x} (n_i/n)$ - эмпирической функцией распределения. При достаточно больших значениях n (хотя бы несколько десятков) $F^*(x)$ хорошо приближает функцию распределения $F(x)$, а функция $f^*(x) = \dfrac{1}{a_i - a_{i-1}} \cdot \left(\dfrac{n_i}{n}\right)$ (если $x \in (x_{i-1}, x_i), 1 \le i \le n$) - плотность распределения $f(x)$.

Случайные числа и моделирование выборки с заданным законом распределения.

При исследовании генеральной совокупности с известным законом распределения полезно иметь способ формировать выборку $x_1,\ldots,x_n$ из неё с помощью некоторого алгоритма на ПК. Оказывается, это можно сделать в два шага.

Шаг 1. Обозначим через R непрерывную с.в., распределённую равномерно в интервале $(0;1)$ (т.е. плотность с.в. R принимает значение 1 для $x \in (0;1)$ и 0 вне этого интервала). Случайными числами называют последовательность значений с.в. R. Другими словами, это

такая последовательность $r_1, r_2, r_3, \ldots$ чисел из интервала $(0;1)$, что количество n_{ab} чисел, попавших в любой подинтервал $(a;b) \subset (0;1)$, среди первых N членов последовательности, составляет долю, равную $(b-a)$ (точнее, $\lim\limits_{n\to\infty} \frac{n_{ab}}{N} = b-a$). Эквивалентное описание: $P(r_j < c) = c$ для любого $c \in (0;1)$ и любого j. Смоделировать на ПК в точности такую последовательность невозможно из-за ограниченности числа разрядов чисел, с которыми работает ПК. Однако можно смоделировать квазиравномерную случайную величину (ограничиваясь имеющимися на ПК разрядами), которая вполне удовлетворительно заменяет равномерную с.в. в практических задачах. В системе MS EXCEL квазислучайную последовательность можно получить обращением к оператору ГЕНЕРАЦИЯ СЛУЧАЙНЫХ ЧИСЕЛ в разделе АНАЛИЗ ДАННЫХ панели СЕРВИС.

Шаг 2. Пусть $r_1, r_2, r_3, \ldots$ последовательность случайных чисел (равномерно распределённая на (0;1)), и требуется построить последовательность $x_1, x_2, x_3, \ldots$, распределённую на числовой прямой в соответствии с некоторой ф.р. $F(x)$. Ограничимся случаем, когда $F(x)$ строго монотонна и непрерывна. Тогда для всякого r_j однозначно определён такой x_j, что $F(x_j) = r_j$. Пусть c - любое число. Имеем $P(x_j < c) = P(F(x_j) < F(c))$ в силу монотонности F, или $P(x_j < c) = P(r_j < F(c))$. Но $r_1, r_2, r_3, \ldots$ равномерно распределены на (0;1), поэтому $P(r_j < F(c)) = F(c)$. Значит, $P(x_j < c) = F(c)$ при любом c. Но это и означает, что $F(x)$ является функцией распределения для последовательности $x_1, x_2, x_3, \ldots$.

Таким образом, если $r_1, r_2, r_3, \ldots$ равномерно распределены на (0;1), то $x_j = F^{-1}(r_j)$ (здесь F^{-1} - символ обратной для F функции, т.е. такой, что $F^{-1}(y) = x \Leftrightarrow F(x) = y$) распределены на $-\infty < x < \infty$ в соответствии с ф.р. $F(x)$. Функции, обратные к часто используемым ф.р., также содержатся в разделе Статистические функции раздела функций EXCEL (либо легко вычисляются непосредственно).

Оценка параметров с.в. по выборке.

Мы упоминали, что ф.р. полностью определяет с.в. (и определяется ею). Но в задачах статистики редко известна ф.р.; обычно известна лишь часть параметров, характеризующих с.в. (например, вид функции распределения, но с неизвестными коэффициентами, или её математическое ожидание (м.о.), или дисперсия и т.д.), а остальные надо уточнить, зная некоторую выборку, построенную по этой с.в. Примерами таких уточнений являются формулы для выборочного среднего $x_в$, выборочной дисперсии $D_в$, выборочного среднего квадратического отклонения (с.к.о.) σ_u, эмпирической ф.р., приведённые выше. Значения $x_в, D_в, \sigma_в$ дают некоторые разумные приближения (точечные оценки) к истинным значениям м.о., дисперсии и с.к.о. исследуемой с.в., а эмпирическая ф.р. может подсказать тип ф.р.с.в. Хотя другая выборка может дать несколько другие значения тех же величин, т.е. $x_в, D_в, \sigma_в$ сами являются случайными величинами, если аргументом считать выборку.

Как же эффективнее использовать информацию, заключённую в выборке? Идея состоит в следующем. Среди всех выборок (отвечающих исследуемой с.в.) есть “плохие”, оценки по которым дают далёкие от истинных значения для искомого параметра, и”хорошие”, дающие близкие к истинному значения. Допустим, что доля “хороших” выборок среди всех выборок есть γ (например, при $\gamma = 0.9$ это значит, что “хороших” выборок 90% от всех). Тогда можно сказать, что с вероятностью γ мы имеем дело с “хорошей” выборкой и, следовательно, получим близкое к истинному значение параметра. Оказывается, во многих важных случаях удаётся по величине γ вычислить границы интервала, в котором находятся все близкие к истинному значению параметра выборочные оценки. Этот интервал

называется доверительным интервалом, а γ- доверительной вероятностью. Процесс нахождения доверительного интервала по доверительной вероятности называется интервальным оцениванием.

Критерий χ^2 Пирсона.

Рассмотрим задачу определения по выборочным данным теоретической плотности распределения с.в. По выборке мы можем построить эмпирическую плотность распределения $f^*(x)$ и её гистограмму, которая, как отмечено выше, хорошо приближает теоретическую плотность $f(x)$. Это позволяет составить гипотезу о виде искомого распределения (возможно, без точных значений некоторых параметров). Далее, в качестве значений этих параметров можно принять их точечные оценки по выборке. Теперь надо проверить, насколько эмпирическая плотность близка к построенной гипотетической.

Степень этой близости оценивается величиной $\chi^2_{набл} = \sum_{i=1}^{l} \frac{(n_i - n_i')^2}{n_i'}$, где i пробегает номера всех интервалов разбиения, n_i- частоты этих интервалов (см. выше), $n_i' = np_i$, где p_i- вероятность попадания с.в. в i-й интервал разбиения, сосчитанная через теоретическую (гипотетическую) плотность.

Величина $\chi^2_{набл} > 0$ есть с.в., зависящая от выборки, и её ф.р. найдена теоретически. Она называется ф.р. χ^2 (читается “хи квадрат”) с k степенями свободы, где $k = l - 1 - r$, r- число оцененных по выборке параметров.

Чем больше значение $\chi^2_{набл}$, тем сильнее эмпирическое распределение отличается от теоретического, и тем меньше оснований принять построенную гипотезу. Выберем некоторое значение $\alpha \in (0;1)$ (уровень значимости; например, $\alpha = 0.05$) и такое $\chi^2_{кр}$ (критическое значение), что $P(\chi^2_{набл} > \chi^2_{кр}) = \alpha$. Договоримся принимать гипотезу, если окажется $\chi^2_{набл} < \chi^2_{кр}$, и отвергать в противном случае. Это соответствует тому, что мы допускаем небольшие расхождения эмпирических и теоретических частот, но отвергаем гипотезу, если эти расхождения достигают величины, вероятность которых меньше α.

Задание №1. *Построение графиков и гистограмм.*

А) Дискретные распределения.

Построить график функции распределения и гистограмму вероятностей заданного распределения.

В) Непрерывные распределения.

Построить графики функции распределения и плотности заданного распределения.

Последовательность выполнения примерного задания.

А) Пусть дано биномиальное распределение с параметрами $n = 15,\ p = 0.6$.

Графики будем строить по точкам. Зададим несколько значений случайной величины $X: x_i = 0,1,2,\ldots,15,$ поместив их в одном столбце. Соседние столбцы отведем для значений функции распределения $F(x_i)$ и вероятностей p_i для каждого значения X.

Функция распределения и вероятности вычисляются с помощью статистической функции БИНОМРАСП с параметрами: “число испытаний” – 15, “вероятность успеха” – 0,6,

“интегральная” – ИСТИНА (для функции распределения) или ЛОЖЬ (для вероятностей). Параметр “число успехов” – это значение случайной величины X. При заполнении столбцов удобно использовать операцию “протаскивания”.

В меню выбираем блок “Вставка”, затем “Диаграмма”. Для функции распределения используем “точечную” диаграмму, для гистограммы вероятностей – “гистограмму”.

В) Пусть дано нормальное распределение с параметрами $a = 15$ и $\sigma = 1{,}25$.

Как и для дискретного распределения, графики будем строить по точкам. В одном столбце зададим значения случайной величины X (подумайте сами, какие значения x случайной величины лучше выбрать), в соседних столбцах разместим значения функции распределения $F(x)$ и плотности распределения $f(x)$, вычисленные с помощью статистической функции НОРМРАСП с параметрами: “среднее” – 15, “стандартное откл.” – 1,25, “интегральная” – ИСТИНА (для функции распределения) или ЛОЖЬ (для плотности распределения). Параметр “ x ” – это значение случайной величины X. Графики строим, используя точечную диаграмму.

Задание №2. *Исследование случайной величины по выборке.*

Смоделировать выборку, подчиняющуюся заданному закону распределения, используя случайные числа. Построить соответствующую этой выборке эмпирическую функцию распределения, её график (график накопленных частот), гистограмму частот и гистограмму относительных частот. Построить график эмпирической плотности распределения. Определить эмпирические характеристики построенной выборки: выборочное среднее, выборочную дисперсию, среднеквадратическое отклонение, размах, коэффициент асимметрии, коэффициент эксцесса. Построить графики теоретической функции и плотности исходного распределения с заданными параметрами и сравнить их с аналогичными графиками эмпирических величин. Вычислить математическое ожидание и дисперсию теоретического распределения и сравнить их с соответствующими эмпирическими величинами.

Последовательность выполнения примерного задания.

Выполним задание для гамма-распределения с параметрами $\alpha = 12, \beta = 30$ при $n = 30$.

1. Моделирование случайных чисел $x_i\,(i = 1,2,\ldots,30)$ с заданным законом распределения:

a) в первом столбце рабочего листа размещаем последовательность y_i случайных чисел, равномерно распределённых на отрезке $[0;1]$ (она строится обращением к панелям СЕРВИС→АНАЛИЗ ДАННЫХ→ГЕНЕРАЦИЯ СЛУЧАЙНЫХ ЧИСЕЛ);

b) в соседнем столбце разместим искомую последовательность $x_i\,(i = 1,2,\ldots,30)$, вычисленную по правилу $x_i =$ ГАММАОБР$(y_i;\alpha;\beta)$ (здесь ГАММАОБР$(y_i;\alpha;\beta)$ - функция, обратная к функции гамма-распределения).

2. Формирование вариационного ряда, разбиение на интервалы и определение частот:

a) полученные значения x_i копируем в соседний столбец и сортируем числа в нём по возрастанию (т.е. выделяем столбец, обращаемся к панелям ДАННЫЕ→СОРТИРОВКА (сортируем в пределах указанного диапазона) и выбираем в списке ПО ВОЗРАСТАНИЮ); получаем вариационный ряд (члены которого по-прежнему обозначим x_i);

b) вычисляем (рекомендуемое формулой Старджеса) число интервалов $L = 1 + 3{,}322 \cdot LOG(n)$ (с округлением до целого числа);

c) определяем ширину интервала группировки $h = (x_{\max} - x_{\min})/L$ (здесь $x_{\max}$ и $x_{\min}$ – наибольшее и наименьшее среди чисел x_i);

d) формируем столбец правых границ интервалов группировки (т.е. столбец чисел $x_{\min} + h, x_{\min} + 2h, x_{\min} + 3h, \ldots, x_{\max}$); аналогично получаем столбец левых границ $x_{\min}$, $x_{\min} + h, \ldots, x_{\max} - h$);

e) определяем частоты n_i (количества элементов выборки, попавших в каждый интервал):
для этого входим в СЕРВИС→АНАЛИЗ ДАННЫХ→ГИСТОГРАММА, указываем в диалоге (через выделение рамкой) входной интервал (т.е. значения x_i), интервал карманов (т.е. правых границ интервала группировки) и выходной интервал (с тем же количеством ячеек, что интервал карманов); после ОК получаем в выходном интервале искомые частоты (сами подумайте, что можно сделать, если при сортировке значений с.в. по интервалам возникли интервалы, содержащие только одно-два значения).

3. Построение эмпирической функции распределения, её графика и гистограмм частот и относительных частот:

a) удобно расположить один за другим столбцы левых и правых границ интервалов, столбец частот, столбец относительных частот n_i/n и столбец значений эмпирической функции распределения $F^*(x)$ (её значения на данном интервале получаются суммированием значений относительных частот на предшествующих интервалах, включая данный); полезно также вычислить и добавить к полученной таблице столбец, содержащий середины интервалов группировки;

b) строим график и гистограммы так же, как в задании №1, используя в качестве аргументов середины интервалов группировки (это позволит легко сравнивать эмпирические графики и гистограммы с теоретическими).

4. Построение графика эмпирической плотности распределения:
значение эмпирической плотности на интервале есть отношение относительной частоты n_i / n этого интервала к его длине h; в качестве аргументов используем середины интервалов.

5. Определение эмпирических характеристик выборки:
эти величины вычисляются с помощью пакета статистических функций (см. функции СРЗНАЧ, ДИСП), либо легко выражаются через такие функции.

6. Построение графиков теоретической функции и плотности исходного распределения использует технологию построения графиков, описанную в задании №1. Для сравнение их с аналогичными графиками эмпирических величин удобно построить два графика (теоретический и эмпирический) на одной диаграмме. Для выполнения этого после задания параметров первого графика (значений X и значений Y) нужно в окне диаграммы перейти в раздел РЯД и использовать клавишу ДОБАВИТЬ, после чего появится строка РЯД2 и строки « значения Y » и «значения X », в которых надо сослаться на параметры второго графика.

Математическое ожидание и дисперсия теоретического распределения вычисляются через параметры, их задающие, по формулам, имеющимся в разделе «Варианты задания».

Задание №3.. *Интервальное оценивание*

По заданной выборке при известной доверительной вероятности γ построить доверительные интервалы:

а) для математического ожидания генеральной совокупности при известной дисперсии;

b) для математического ожидания при неизвестной дисперсии;
c) для дисперсии генеральной совокупности при известном математическом ожидании;
d) для дисперсии при неизвестном математическом ожидании.
Считать, что генеральная совокупность имеет нормальный закон распределения.

Последовательность выполнения примерного задания.

Выполним задание для выборки объёма $n = 20$ с элементами 29,23,25,24,9, 18,36,27,33,20, 12,40,29,22,30, 9,39,5,20,59. Пусть задано $\gamma = 0{,}95$, а также в случае a) дисперсия $\sigma^2 = 100$, в случае c) математическое ожидание $M(X) = 25$.

Прежде всего с помощью статистических функций СРЗНАЧ и СТАНДОТКЛОН вычислим параметры выборки – выборочное среднее $x_в$ и выборочное среднее квадратическое отклонение (несмещённое!) s . Теперь последовательно построим доверительные интервалы по пунктам задания.

a) Из теории известно, что доверительный интервал есть $(x_в - t \cdot \sigma / \sqrt{n},\ x_в + t \cdot \sigma / \sqrt{n})$, где t вычисляется через функцию Лапласа $\Phi(t)$ из соотношения $\Phi(t) = \gamma / 2$. EXCEL не содержит стандартной функции Лапласа, но включает функцию $\Phi(t) + 0{,}5$ (вычисляется как НОРМСТРАСП(t)) и обратную к ней функцию, которая вычисляется с помощью НОРМСТОБР. Поэтому можно вычислять t =НОРМСТОБР($\gamma / 2 + 0{,}5$).

b) В этом случае доверительный интервал есть $(x_в - t_\gamma \cdot s / \sqrt{n}, x_в + t_\gamma \cdot s / \sqrt{n})$, где t_γ вычисляется через обратное распределение Стьюдента по формуле t_γ =СТЬЮДРАСПОБР(1-$\gamma, n-1$).

c) Здесь доверительный интервал определяется неравенствами $n \cdot \sigma_в^2 / \chi_2^2 < \sigma^2 < n \cdot \sigma_в^2 / \chi_1^2$, где $\sigma_в^2$ – выборочная дисперсия $\sigma_в^2 = (\sum_{i=1}^{n} (x_i - M(X))^2) / n$ с заданным по условию $M(X)$, а χ_1^2 и χ_2^2 выбираются из условий $P(\chi^2 > \chi_2^2) = (1-\gamma)/2$ и $P(\chi^2 > \chi_1^2) = (1+\gamma)/2$, где случайная величина χ^2 имеет распределение χ^2 с n степенями свободы. В EXCEL они находятся по формулам χ_2^2=ХИ2ОБР((1-γ)/2,n) и χ_1^2=ХИ2ОБР((1+γ)/2,n), а при вычислении $\sigma_в^2$ удобно использовать функции СУММКВРАЗН или СУММКВ.

d) Доверительный интервал задаётся неравенствами $(n-1) \cdot s^2 / \chi_2^2 < \sigma^2 < (n-1) \cdot s^2 / \chi_1^2$, но значения χ_1^2 и χ_2^2 находятся (в отличие от п.c)) при (n-1) степенях свободы. Подчеркнём также, что выборочная дисперсия вычисляется как s^2, поскольку M(X) не задано.

Задание №4. *Проверка статистической гипотезы о законе распределения с помощью критерия χ^2 («хи квадрат») Пирсона.*

Дана выборка x_i $(i = 1,2\ldots, n)$ значений случайной величины X. Требуется

a) По выборке x_i построить гистограмму частот и выдвинуть гипотезу о законе распределения случайной величины X.
b) Проверить выдвинутую гипотезу с помощью критерия χ^2 Пирсона при уровне значимости α=0.05.

Последовательность выполнения примерного задания.

Пусть выборка x_i объёма n =50 задана таблицей

102	132	117	94	123	146	140	114	134	128
94	75	102	88	129	110	121	102	96	136
65	121	102	94	139	116	85	110	113	107
115	100	121	82	93	131	164	95	98	105
101	114	86	95	122	97	106	114	128	103

a) По выборке x_i строим интервальный вариационный ряд и гистограмму частот (см. об этом работу №2, Последовательность выполнения примерного задания, п.2; отметим только, что сортировку массива x_i по возрастанию здесь можно опустить).При этом будут определены число интервалов L, ширина интервала h, столбцы левых и правых границ интервалов и их частоты n_i. В рассматриваемом примере конфигурация гистограммы даёт основание выдвинуть гипотезу о нормальном распределении.

b) Проверим эту гипотезу. Так как частота попадания в каждый интервал должна быть не менее 5, то первый интервал присоединим ко второму, а последний к предпоследнему. Заново перенумеруем интервалы. Таким образом, количество интервалов L становится равным 5, а ширина первого и последнего в два раза больше первоначальной ширины h. Интервальный ряд представим в виде дискретного ряда распределения, рассчитав среднее арифметическое концов интервала x_i^*. Разместим значения x_i^* в столбце рядом с частотами. По полученному дискретному ряду распределения вычислим оценки параметров предполагаемого распределения: среднее $x_{cp}^* = (\sum_{i=1}^{L} x_i^* \cdot n_i)/n$ и среднее квадратическое отклонение $\sigma^* = \sqrt{\sum_{i=1}^{L}((x_i^*)^2 \cdot n_i/n) - (x_{cp}^*)^2}$. В EXCEL для вычисления таких сумм удобно использовать функцию СУММПРОИЗВ.

Вычислим наблюдаемое значение критерия Пирсона $\chi^2_{набл}$. Для этого найдём сначала теоретические вероятности $p_i = P(x_{iлев} \le X \le x_{iправ})$ попадания в i-й интервал, которые для нормального распределения считаются по формулам $p_i = \Phi((x_{iправ} - x_{ср}^*)/\sigma^*) - \Phi((x_{iлев} - x_{ср}^*)/\sigma^*)$, где $\Phi(\cdot)$ – функция Лапласа (см. Задание №3, Пример выполнения задания, п.а)), причём левую границу первого интервала и правую границу последнего примем равными $-\infty$ и $+\infty$ соответственно, взяв за бесконечность какое-либо большое число. Затем определяем теоретические частоты по формуле $n_i' = n \cdot p_i$, отношения $((n_i - n_i')^2)/n_i'$ и сумму этих отношений, которая и есть $\chi^2_{набл}$.

Найдём критическое значение критерия $\chi^2_{кр}$, т.е. процентную точку χ^2-распределения уровня $\alpha \cdot 100\%$ с 2 степенями свободы. Это делается с помощью статистической функции ХИ2ОБР. Так как $\chi^2_{набл} < \chi^2_{кр}$, то гипотеза о нормальном распределении принимается.

Варианты заданий.

К заданию №1.

Варианты биномиального распределения

Вар.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
N	20	21	22	23	24	25	20	21	22	23	24	25	20	21	22
p	0.1	0.2	0.3	0.4	0/5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5

Вар.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
N	20	21	22	23	24	25	20	21	22	23	24	25	20	21	22
p	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5	0.1

Варианты нормального распределения

Вар.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a	11.5	15.2	12.8	19	9.6	14.3	17.1	13.6	15.8	12.2	19.7	16.6	14.9	8.9	13.1
σ	0.8	1.2	1.1	0.9	2.4	1.3	1.6	1.4	2.1	2.8	1.5	1.7	1.9	2.3	2.5

Вар.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
a	11.5	15.2	12.8	19	9.6	14.3	17.1	13.6	15.8	12.2	19.7	16.6	14.9	8.9	13.1
σ	1.2	1.1	0.9	2.4	1.3	1.6	1.4	2,1	2.8	1.5	1.7	1.9	2.3	2.5	1.2

К заданию №2.

Законы распределения и их параметры.
Объём выборки n везде принят равным 50.

№ варианта	Распределение, M и D	Параметры	Обратная к ф.р , плотность, ф.р..
1.1 1.2 1.3 1.4	Нормальное, M=a, D=σ^2	a=3500,σ=1000 a=4000,σ=1500 a=3800,σ=1000 a=4200,σ=1600	НОРМОБР, НОРМРАСП(…ЛОЖЬ), НОРМРАСП(…ИСТИНА)
2.1 2.2 2.3 2.4	Логнормальное, M= $\exp(a-(\sigma^2/2))$ D= $\exp(2a+\sigma^2)\cdot(\exp(\sigma^2)-1)$	a=8, σ=1 a=8.5, σ=0.5 a=7.5, σ=0.8 a=7, σ=0.6	ЛОГНОРМОБР, НОРМРАСП(LN(X);…;ЛОЖЬ)/X НОРМРАСП(LN(X);a; σ ;ИСТИНА)
3.1 3.2 3.3 3.4	Экспоненциальное (показательное), M=1/λ, D=1/λ^2	λ=0.00025 λ=0.0002 λ=0.0005 λ=0.0004	По формулам $x_j=(-1/\lambda)\cdot\ln(1-y_j)$, $\lambda\cdot\exp(-\lambda x)$, 1-$\exp(-\lambda x)$ (при x>0)
4.1 4.2 4.3 4.4	Гамма-распределение, M= $\alpha\cdot\beta$, D= $\alpha\cdot\beta^2$	α=10, β=400 α=11, β=350 α=9, β=450 α=8, β=500	ГАММАОБР, ГАММАРАСП(x, α, β, ЛОЖЬ), ГАММАРАСП(x, α, β, ИСТИНА)

№ варианта	Распределение, M и D	Параметры	Обратная к ф.р , плотность, ф.р.
5.1 5.2 5.3 5.4	Нормальное, M=a, D=σ^2	a=3500,σ=2000 a=4000,σ=3000 a=3800,σ=2000 a=4200,σ=3200	НОРМОБР. НОРМРАСП(…ЛОЖЬ), НОРМРАСП(…ИСТИНА)
6.1 6.2	Логнормальное,	a=8, σ=2 a=8.5, σ=1	ЛОГНОРМОБР, НОРМРАСП(LN(X);…;ЛОЖЬ)/X

6.3 6.4	M= $\exp(a-(\sigma^2/2))$ D= $\exp(2a+\sigma^2)\cdot(\exp(\sigma^2)-1)$	a=7.5, σ=1,6 a=7, σ=1,2	НОРМРАСП(LN(X);a; σ ;ИСТИНА)
7.1 7.2 7.3 7.4	Экспоненциальное (показательное), M=1/λ, D=$1/\lambda^2$	λ=0.0005 λ=0.0004 λ=0.001 λ=0.0008	По формулам $x_j=(-1/\lambda)\cdot\ln(1-y_j)$, $\lambda * \exp(-\lambda x)$, $1-\exp(-\lambda x)$ (при x>0)
8.1 8.2 8.3 8.4	Гамма-распределение, M= $\alpha\cdot\beta$, D= $\alpha\cdot\beta^2$	α=10, β=800 α=11, β=700 α=9, β=900 α=8, β=1000	ГАММАОБР, ГАММАРАСП(x; α; β; ЛОЖЬ), ГАММАРАСП(x; α; β; ИСТИНА)

К заданию №3.

№№ вариантов	Выборка (одна для трёх вариантов)	Известная дисперсия D	Известное м.о. M	Доверит. вероятность γ
1.1, 1.2, 1.3	9.27 9.93 9.42 9.01 8.39 7.98 9.26 10.18 8.23 8.03 9.29 7.94 9.86 9.08 9.12 8.78	0.25	9	0.9 (вар.1.1) 0.95 (вар.1.2) 0.99 (вар.1.3)
2.1, 2.2, 2.3	99.7 98.7 100.2 101.3 101.2 101.7 97.8 99.8 101.1 98.9 99.3 98.3 98.2 99 99.2 97.9 99.4 99.6 100.1 99.6	1	100	0.9 (вар.2.1) 0.95 (вар.2.2) 0.99 (вар.2.3)
3.1, 3.2, 3.3	442 382 409 391 245 333 512 422 484 350 273 559 447 377 451 247 545 210 358 743 393 402 133 349	10000	400	0.9 (вар.3.1) 0.95 (вар.3.2) 0.99 (вар.3.3)
4.1, 4.2, 4.3	214 187 215 179 199 177 192 232 200 184 222 222 211 211 203 207 220 183 174 180	800	200	0.9 (вар.4.1) 0.95 (вар.4.2) 0.99 (вар.4.3)
5.1, 5.2, 5.3	2757 2746 9623 3409 2298 2420 3355 3379 3344 789 2245 2046 2270 3279 2883 4289 945 2847 3103 2456	1000000	2530	0.9 (вар.5.1) 0.95 (вар.5.2) 0.99 (вар.5.3)
6.1, 6.2, 6.3	35.36 22.12 28.63 27.74 26.46 30.09 35.70 31.75 31.11 28.13 34.98 31.78 30.08 38.45 31.16 22.47 35.59 25.75 28.82 31.42 27.82 30.62 25.15 32.37 30.88	16	30	0.9 (вар.6.1) 0.95 (вар.6.2) 0.99 (вар.6.3)
7.1, 7.2, 7.3	111 104 116 103 114 117 110 109 104 109 100 113 117 108 116 111 112 116 106 109	25	110	0.9 (вар.7.1) 0.95 (вар.7.2) 0.99 (вар.7.3)

8.1, 8.2, 8.3	24050 32500 31100 31450 30100 31450 29850 28100 29500 26600 25550 22800 26300 22400 30000 24900	9	28000	0.9 (вар.8.1) 0.95 (вар.8.2) 0.99 (вар.8.3)
9.1, 9.2, 9.3	15.0 14.1 16.0 13.4 14.0 15.7 13.6 12.5 16.1 13.4 15.8 13.7 14.4 12.6 15.0 14.3 13.8 15.4 16.4 15.4	1	14	0.9 (вар.9.1) 0.95 (вар.9.2) 0.99 (вар.9.3)
10.1, 10.2, 10.3	486 537 647 625 739 618 673 510 778 730 650 649 608 538 600 535 572 751 682 558	10000	600	0.9 (вар.10.1) 0.95 (вар.10.2) 0.99 (вар.10.3)

К заданию №4.

Исходным данным для варианта является выборка, смоделированная по некоторому закону распределения (например, закон распределения и соответствующая выборка для варианта берутся те же, что в Задании №2).

Требуется проверить с помощью критерия Пирсона, можно ли для неё принять гипотезу о законе распределения, с помощью которого она была построена (но с параметрами, найденными по выборке), при уровнях значимости $\alpha_1 = 0{,}05$; $\alpha_2 = 0{,}01$; $\alpha_3 = 0{,}1$.

Затем оставим в исходной выборке только первые 30 элементов и проведём тот же анализ для этой новой выборки объёма n=30.

Сравниваем значения $\chi^2_{кр}$, полученные в обоих случаях.

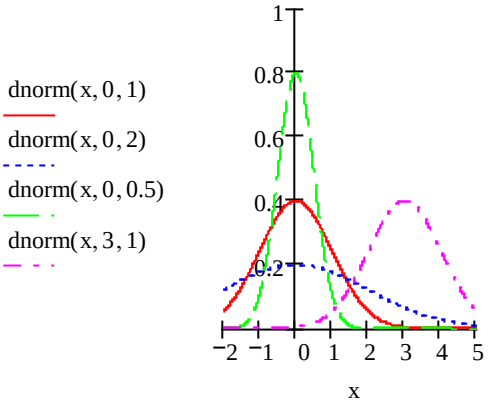
Замечание. Исходную выборку можно получить, используя датчик случайных чисел с заданным законом распределения, имеющийся в MATHCAD (функции rnorm, rexp, rgamma, rlnorm и т.п.). Возможно также появление таких датчиков в новых версиях EXCEL.

Литература.

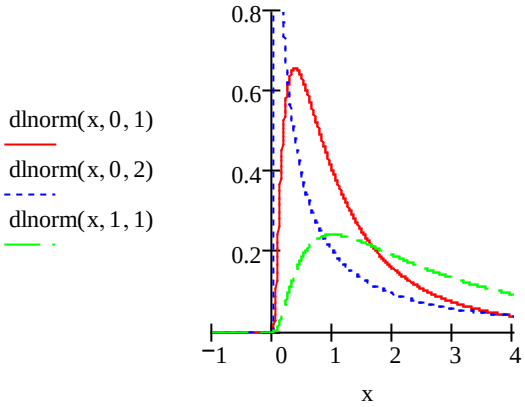
1. Гмурман В.Е. Теория вероятностей и математическая статистика. 2003. 479с.
2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике.: Учебное пособие. М.: Высшая школа, 2003,404с.
3. Плис А.И., Сливина Н.А. MATHCAD. Математический практикум. Учеб. пособие.
4. Мишин И.В. Теория вероятностей и математическая статистика. Учеб. пособие. М.: МИИГАиК, 2008, 107с.
5. Бородин А.Н. Элементарный курс теории вероятностей и математической статистики. СПб.: 2003.
6. Вентцель Е.С., Овчаров А.А. Теория вероятностей и её инженерные приложения. 1988.
7. Письменный Д.Т. Конспект лекций по теории вероятностей и математической статистике. М., Айрис Пресс, 2006.

Приложение. Графики плотностей вероятностей некоторых распределений.

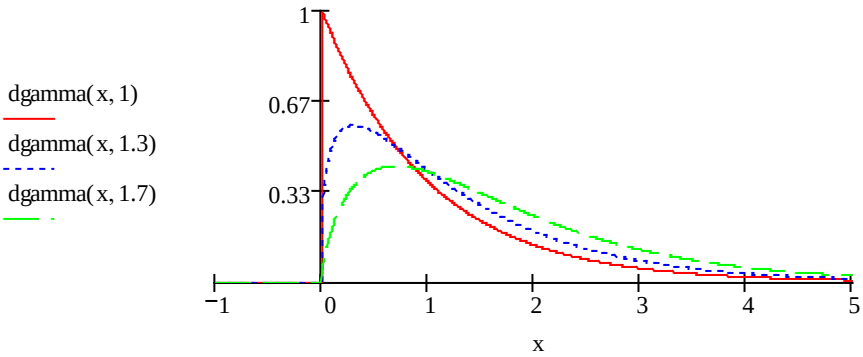
1.Плотность нормального распределения dnorm(x,a,σ)



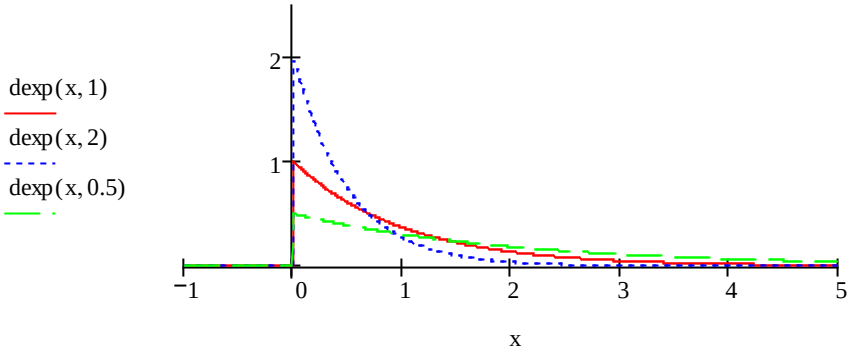
2.Плотность логнормального распределения
dlnorm(x,a,σ)



3.Плотность стандартного гамма-распределения dgamma(x,s)
(то же, что гамма-распределение, описанное в EXCEL, при α =s, β=1)



4.Плотность экспоненциального распределения dexp(x,λ)



5.Плотность "хи-квадрат" распределения с k степенями свободы dchisq(x,k)