**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ**

**Федеральное государственное автономное образовательное учреждение высшего профессионального образования Национальный исследовательский университет «Высшая школа экономики»**

**Московский институт электроники и математики Национального исследовательского университета «Высшая школа экономики»**

**Кафедра высшей математики**

# ЗАДАЧИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ И ИХ РЕШЕНИЕ С ИСПОЛЬЗОВАНИЕМ ПРИЛОЖЕНИЯ Microsoft Excel

**Учебно-методическое пособие для выполнения курсовой работы по дисциплине «Теория вероятностей и математическая статистика»**

**Москва 2013**

Составитель ст. преподаватель Ю.Б. Гришунина

Учебно-методическое пособие содержит теоретические сведения, необходимые для выполнения курсовой работы: определения, формулировки теорем, основные формулы. Для каждого задания приведена постановка задачи, указан метод ее решения с подробными комментариями, перечислены функции Excel, которые рекомендуется применять для построения графиков и выполнения вычислений, проведен анализ конкретных статистических данных. Предназначено для студентов всех специальностей, изучающих дисциплину «Теория вероятностей и математическая статистика».

УДК 519.2

Задачи математической статистики и их решение с использованием приложения Microsoft Excel. Учебно-методическое пособие для выполнения курсовой работы по дисциплине «Теория вероятностей и математическая статистика»./Моск. ин-т электроники и математики Национального исследовательского университета «Высшая школа экономики»; Сост. Ю.Б.Гришунина. М., 2013. – 32 с.

Табл. 7. Библиогр.: 6 назв.

ISBN 978-5-94506-311-2

**Учебное издание**

**Задачи математической статистики и их решение с использованием приложения Microsoft Excel**

**Составитель ГРИШУНИНА Юлия Борисовна**

Редактор С.П.Клышинская
Технический редактор О.Г.Завьялова

Подписано в печать 19.02.2013. Формат 60×84/16. Бумага офсетная.
Печать – ризография. Усл. печ. л. 2,0. Уч.-изд.л. 1,8. Тираж 30 экз. Изд.№ 15. Заказ *63*.
Бесплатно.
Московский институт электроники и математики Национального исследовательского университета «Высшая школа экономики».
109028, Москва, Б. Трехсвятительский пер., 3.
Редакционно-издательский отдел Московского института электроники и математики Национального исследовательского университета «Высшая школа экономики».
Участок МИЭМ типографии НИУ ВШЭ.
113054 Москва, ул. М.Пионерская, 12.



## Введение

Задачи математической статистики можно условно разделить на три группы:

- непараметрические задачи;
- параметрические задачи;
- проверка статистических гипотез.

Непараметрические задачи имеют место в случаях, когда нет информации о виде распределения генеральной совокупности; тогда возникает проблема оценки функции распределения или плотности распределения.

Если известны вид распределения и область возможных значений параметров и необходимо оценить неизвестные параметры распределения – это параметрические задачи; при этом можно поставить задачу точечного или интервального оценивания.

Проверка статистических гипотез – это проверка различных предположений о вероятностных закономерностях изучаемого явления или процесса на основе статистических данных, т.е. результатов наблюдений или экспериментов; эти задачи могут быть как непараметрическими (например, гипотезы о виде теоретического распределения, о независимости выборок), так и параметрическими (гипотезы о значениях параметров вероятностных моделей, о равенстве средних, дисперсий и т.д.).

Содержание курсовой работы включает в себя все перечисленные задачи математической статистики и состоит из следующих заданий:

1. Смоделировать выборку объема n=30 из генеральной совокупности с заданной теоретической функцией распределения (распределение Вейбулла или логнормальное распределение с заданными параметрами).
2. Построить график эмпирической функции распределения; построить гистограмму и полигон частот. Сравнить построенные графики с соответствующими графиками теоретической функции распределения и плотности распределения.
3. Построить точечную оценку неизвестного параметра заданного распределения:
   а) методом моментов;
   б) методом максимального правдоподобия.
   Сравнить полученные оценки с истинным значением параметра.
4. Построить доверительный интервал надежности $1-\gamma$ для неизвестного математического ожидания:
   а) считая дисперсию известной;
   б) считая дисперсию неизвестной.
   Сравнить точность полученных интервалов.
5. Используя критерий Пирсона, на заданном уровне значимости $\gamma$ проверить, согласуется ли гипотеза о виде теоретического распределения с представленной выборкой.

Основные цели данной курсовой работы:

- усвоение терминологии, используемой в математической статистике;

- углубленная проработка определений, постановок задач математической статистики и методов их решения;

- закрепление знаний по следующим, ранее изученным, разделам курса: случайные величины, способы их задания и числовые характеристики, предельные теоремы теории вероятностей и др.;

- получение опыта работы со статистическими данными, их обработки и анализа;

- приобретение умения интерпретировать полученные результаты;

- совершенствование навыков работы в Excel, в частности, в разделах меню СТАТИСТИЧЕСКИЕ ФУНКЦИИ и МАСТЕР ДИАГРАММ.

## Сведения о распределении Вейбулла и логнормальном распределении

Необходимые для выполнения курсовой работы сведения о распределении Вейбулла и логнормальном распределении приведены в *таблице 1*.

*Таблица 1*

| | Распределение Вейбулла $W(r,\lambda)$ | Логнормальное распределение $LogN(\mu,\sigma)$ |
|---|---|---|
| Плотность распределения | $\begin{cases} 0, x \le 0 \\ \frac{r}{\lambda^r} x^{r-1} e^{-\left(\frac{x}{\lambda}\right)^r}, x > 0 \end{cases}$ | $\begin{cases} 0, x \le 0 \\ \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0 \end{cases}$ |
| Функция распределения | $\begin{cases} 0, x \le 0 \\ 1-e^{-\left(\frac{x}{\lambda}\right)^r}, x > 0 \end{cases}$ | $\begin{cases} 0, x \le 0 \\ \int_{-\infty}^{\ln x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, x > 0 \end{cases}$ |
| Математическое ожидание | $\lambda\Gamma\left(1+\frac{1}{r}\right)$, Г - гамма-функция | $e^{\mu+\frac{\sigma^2}{2}}$ |
| Дисперсия | $\lambda^2\left(\Gamma\left(1+\frac{2}{r}\right)-\Gamma^2\left(1+\frac{1}{r}\right)\right)$ | $\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}$ |

## Задание 1. Моделирование выборки из генеральной совокупности с заданной теоретической функцией распределения

*Определение 1. Генеральной совокупностью называется вероятностное пространство* $(\Omega, F, P)$ *и определенная на этом пространстве случайная величина X.*

Случайную величину *X*, ее функцию распределения *F(x)=P(X<x)*, ее числовые характеристики и другие параметры будем называть теоретическими; все перечисленные элементы являются составными частями математической модели изучаемого явления или процесса. Таким образом, генеральная



совокупность состоит из тех объектов, которые подлежат наблюдению и исследованию, и относительно которых требуется сделать выводы при анализе конкретной проблемы.

*Определение 2. Выборкой объема n* $(X_1,...,X_n)$ *называется последовательность n независимых одинаково распределенных случайных величин, распределение каждой из которых совпадает с теоретическим распределением F(x).*

Иными словами, выборка - это результат *n* последовательных независимых наблюдений над теоретической случайной величиной *X*.

Методика моделирования выборки основана на следующем *утверждении*:

*Утверждение 1. Пусть X – непрерывная случайная величина, и ее функция распределения F(x)=P(X<x) монотонно возрастает. Тогда случайная величина Y= F(X) имеет равномерное распределение на отрезке [0;1].*

Доказательство. Напомним, что если случайная величина равномерно распределена на отрезке *[a;b]*, то ее функция распределения имеет следующий вид: $G(y)=\begin{cases} 0, y \le a \\ \frac{y-a}{b-a}, a < y \le b \\ 1, y > b \end{cases}$. Очевидно, для отрезка *[0;1]* $G(y)=\begin{cases} 0, y \le 0 \\ y, 0 < y \le 1 \\ 1, y > 1 \end{cases}$.

Найдем функцию распределения случайной величины *Y= F(X)*. При $y \le 0$ $G_Y(y)=P(Y<y)=P(F(X)<y)=0$, поскольку *F(X)* – это вероятность, и она не может принимать отрицательные значения. Аналогично, если $y>1$, то $G_Y(y)=1$, т.к. вероятность всегда $\le 1$.

Осталось рассмотреть случай $0<y\le 1$. Заметим, что поскольку функция *F(x)* непрерывна и монотонно возрастает, то у нее существует обратная функция, которая также является монотонно возрастающей. Поэтому $G_Y(y)=P(Y<y)=P(F(X)<y)=P(F^{-1}(F(X))<F^{-1}(y))=P(X<F^{-1}(y))=$ $=F^{-1}(F(y))=y$, что и треб.

Из доказанного утверждения следует, что если *Y= F(X)* – реализация случайной величины, имеющей равномерное распределение на отрезке *[0;1]*, то $X=F^{-1}(Y)$ - это реализация случайной величины, имеющей распределение *F(x)*. Поэтому для того, чтобы смоделировать выборку из генеральной совокупности с заданным теоретическим распределением, нужно сначала получить выборку $(Y_1,...,Y_n)$ из генеральной совокупности с равномерным распределением на отрезке *[0;1]*, а затем, подставляя полученные числа в формулу обратной функции, вычислить значения $(X_1,...,X_n)$, которые и будут являться реализациями случайной величины с заданным распределением *F(x)*.

Поскольку функции распределения случайных величин, имеющих распределение Вейбулла и логнормальное распределение, непрерывны и монотонно возрастают на интервале $(0;\infty)$, то для моделирования соответствующих выборок можно использовать предложенный алгоритм.

Случайные числа из отрезка *[0;1]* генерируются в Excel с помощью функции СЛЧИС() в Меню/Вставка/Функция/Математические; аргументов у данной функции нет.

*Замечание 1.* Функция СЛЧИС() обладает следующим свойством: ее значение пересчитывается при каждом обращении к Excel, что приводит к изменению исходных данных в процессе выполнения работы. Чтобы этого не происходило, рекомендуется сохранить результаты, полученные при первом обращении к данной функции, как значения (константы). Для этого используется СПЕЦИАЛЬНАЯ ВСТАВКА в Меню/Правка/Специальная вставка/Значения.

Для распределения Вейбулла формула обратной функции легко выводится с помощью простых аналитических преобразований:

$Y_i = F(X_i) = 1 - e^{-\left(\frac{X_i}{\lambda}\right)^r}$, отсюда $X_i = \lambda(-\ln(1-Y_i))^{\frac{1}{r}}$.

Для логнормального распределения значения $X_i$ вычисляются при помощи функции ЛОГНОРМОБР($Y_i, \mu, \sigma$) в Меню/Вставка/Функция/ /Статистические.

В качестве примера рассмотрим выполнение заданий курсовой работы для распределения Вейбулла с параметрами $r = 1$ и $\lambda = 1$.

*Замечание 2.* Если $r = 1$, то распределение Вейбулла – это экспоненциальное распределение с параметром $\frac{1}{\lambda}$ $(exp(\frac{1}{\lambda}))$.

При подстановке $r = 1$ и $\lambda = 1$ в формулу для обратной функции получаем, что $X_i = -\ln(1 - Y_i)$.

Результаты вычислений приведены в *таблице 2.*

*Таблица 2*

| Yi | Yi(coxp) | Xi=λ*(-ln(1-Yi))^(1/r) | Вариационный ряд |
|---|---|---|---|
| 0,347971 | 0,847381 | 1,879811142 | 0,019239271 |
| 0,241207 | 0,778783 | 1,50861118 | 0,10697888 |
| 0,637412 | 0,279592 | 0,3279382 | 0,110407836 |
| 0,786839 | 0,604649 | 0,927980224 | 0,148217048 |
| 0,763542 | 0,247573 | 0,284450738 | 0,263454902 |
| 0,430263 | 0,019055 | 0,019239271 | 0,284450738 |
| 0,173475 | 0,231608 | 0,263454902 | 0,3279382 |
| 0,702497 | 0,137756 | 0,148217048 | 0,417126872 |
| 0,569435 | 0,9587 | 3,186887282 | 0,615129532 |
| 0,411171 | 0,903895 | 2,342313608 | 0,617270619 |
| 0,219776 | 0,459429 | 0,615129532 | 0,683307635 |
| 0,447608 | 0,850001 | 1,897124181 | 0,690556227 |
| 0,15604 | 0,104531 | 0,110407836 | 0,745317271 |
| 0,867563 | 0,101455 | 0,10697888 | 0,921818008 |
| 0,905902 | 0,980393 | 3,931873824 | 0,927980224 |
| 0,393889 | 0,341063 | 0,417126872 | 0,950930092 |
| 0,739732 | 0,817401 | 1,700465321 | 0,970320123 |
| 0,153381 | 0,613619 | 0,950930092 | 0,994894594 |
| 0,970348 | 0,498703 | 0,690556227 | 1,50861118 |
| 0,221766 | 0,460585 | 0,617270619 | 1,662138304 |
| 0,966977 | 0,621038 | 0,970320123 | 1,700465321 |
| 0,989231 | 0,835046 | 1,802090764 | 1,717636578 |
| 0,974033 | 0,525416 | 0,745317271 | 1,802090764 |
| 0,145457 | 0,984403 | 4,160674435 | 1,879811142 |
| 0,947302 | 0,495056 | 0,683307635 | 1,897124181 |
| 0,003433 | 0,810267 | 1,662138304 | 2,342313608 |
| 0,757656 | 0,914615 | 2,460582537 | 2,460582537 |
| 0,780517 | 0,602205 | 0,921818008 | 3,186887282 |
| 0,012167 | 0,630238 | 0,994894594 | 3,931873824 |
| 0,295719 | 0,82051 | 1,717636578 | 4,160674435 |



Столбец 1 (Yi) – это значения функции СЛЧИС(), вычисленные автоматически при повторном обращении к Excel; столбец 2 (Yi(coxp)) – это значения функции СЛЧИС(), вычисленные и сохраненные при первом обращении к Excel; столбец 3 – это выборка из генеральной совокупности с теоретическим распределением Вейбулла $W(1,1)$.

Для удобства дальнейшей работы рекомендуется построить вариационный ряд $(X_1^*, ..., X_n^*)$ (столбец 4), т.е. расположить полученные данные в порядке возрастания. Для этого их надо сначала скопировать как значения из столбца 3 с помощью функции СПЕЦИАЛЬНАЯ ВСТАВКА в Меню/Правка/Специальная вставка/Значения, а затем упорядочить, используя функцию СОРТИРОВКА в Меню/Данные/Сортировка/По возрастанию.

## Задание 2. Эмпирическая функция распределения. Гистограмма и полигон частот

### Эмпирическая функция распределения

Теоретическая функция распределения $F(x)=P(X<x)$ определяет вероятность события $\{X<x\}$. Согласно статистическому определению вероятности относительная частота события в $n$ независимых испытаниях, т.е. доля тех испытаний, в которых данное событие произошло, является оценкой для его вероятности; под оценкой понимается некоторая функция, зависящая от результатов наблюдений, значение которой мало отличается от истинного значения величины, которое требуется оценить. Поэтому оценкой для функции распределения является относительная частота события $\{X<x\}$ в $n$ испытаниях, которая называется эмпирической функцией распределения и вычисляется как отношение числа испытаний, в которых произошло событие $\{X<x\}$, к общему числу испытаний $n$. По *определению 2* каждое выборочное значение $X_i$ является реализацией теоретической случайной величины $X$, поэтому наступление события $\{X<x\}$ в $i$-м испытании означает, что $X_i \in (-\infty; x)$, а число испытаний, в которых произошло событие $\{X<x\}$, равно числу выборочных значений, меньших $x$. Таким образом, эмпирическая функция распределения определяется следующим образом:

*Определение 3. Эмпирической функцией распределения* $\hat{F}_n(x)$ *называется функция* $\hat{F}_n(x)=\frac{\nu(x)}{n}$, *где* $\nu(x)$ *- число точек вариационного ряда, меньших x; n – объем выборки.*

Из этого определения следует, что, если в выборке нет повторяющихся значений, то эмпирическая функция распределения – это ступенчатая функция со скачками, равными по величине $\frac{1}{n}$, в точках вариационного ряда. Такая функция распределения соответствует распределению дискретной случайной величины, которая принимает каждое из значений $X_1,...,X_n$ с вероятностью $\frac{1}{n}$. Отметим также, что эмпирическая функция распределения обладает всеми свойствами функции распределения.

*Замечание 3.* Если некоторое значение повторяется в выборке *k* раз, то скачок эмпирической функции распределения в соответствующей точке вариационного ряда равен $\frac{k}{n}$.

## Построение графика эмпирической функции распределения в Excel

Будем предполагать, что в выборке нет повторяющихся значений.

Сначала необходимо сформировать столбцы данных для построения графика. Поскольку в каждой точке вариационного ряда эмпирическая функция распределения имеет скачок, то на графике каждому аргументу (выборочному значению) должно соответствовать два значения эмпирической функции, отличающиеся на величину этого скачка. Чтобы иметь возможность каждому выборочному значению поставить в соответствие два числа, необходимо продублировать выборочные значения в порядке возрастания. Для этого надо дважды скопировать вариационный ряд, построенный в задании 1, в один столбец с помощью функции Меню/Правка/Копировать, а затем упорядочить полученные данные с помощью Меню/Данные/Сортировка/По возрастанию. При этом будет занято *2n* ячеек. Столбец соответствующих значений эмпирической функции распределения должен состоять из чисел, принадлежащих отрезку *[0;1]* и образующих арифметическую прогрессию с шагом $\frac{1}{n}$. Этот столбец можно построить следующим образом: в две верхние ячейки записать соответственно значения 0 и $\frac{1}{n}$, затем, выделив их и удерживая курсор в правом нижнем углу выделенного диапазона (там должен появиться крестик) левой кнопкой мыши, «протянуть» значения до 1 – это займет *(n+1)* ячеек, а в оставшиеся *(n-1)* ячеек скопировать значения от $\frac{1}{n}$ до $\frac{n-1}{n}$; при этом



также будет заполнено *2n* ячеек. Теперь для того, чтобы полученные данные соответствовали скачкам эмпирической функции распределения, их надо упорядочить по возрастанию с помощью Меню/Данные/Сортировка/По возрастанию.

*Замечание 4.* Если в выборке есть повторяющиеся значения, то соответствующие значения эмпирической функции распределения рекомендуется вычислять и вводить вручную.

Для построения графика используется МАСТЕР ДИАГРАММ в Меню/Вставка/Диаграмма. На первом шаге выбирается тип диаграммы: Меню/Вставка/Диаграмма/Точечная/Точечная диаграмма со значениями, соединенными отрезками без маркеров. Далее, на втором шаге, вводятся данные, по которым будет строиться диаграмма: Источник данных диаграммы/Ряд/Добавить; в поле *Значения X* копируется столбец, содержащий значения вариационного ряда, а в поле *Значения Y* – столбец значений эмпирической функции распределения. Нажав кнопку Готово, получаем график эмпирической функции распределения.

*Замечание 5.* Полученный график – это график эмпирической функции распределения на отрезке $[X_1^*;X_n^*]$; $X_1^*$ и $X_n^*$ - соответственно минимальное и максимальное выборочные значения. При $x \le X_1^*$ $\hat{F}_n(x)=0$, а при $x > X_n^*$ $\hat{F}_n(x)=1$.

## Построение графика теоретической функции распределения

Для сравнения получившейся эмпирической функции распределения с теоретической рекомендуется на той же координатной плоскости построить график теоретической функции распределения на некотором отрезке [*a; b*], таком, что $a \le X_1^*$, $b \ge X_n^*$.

Сформируем столбцы данных для построения этого графика. Для этого разобьем отрезок [*a; b*] на интервалы фиксированной длины *d*, величина которой выбирается в зависимости от длины этого отрезка таким образом, чтобы получилось достаточное количество точек разбиения. Это делается тем же способом, который был описан для построения значений эмпирической функции распределения: в две верхние ячейки записываются соответственно значения *a* и *a+d*, а затем значения «протягиваются» до *b*. Во второй столбец вносятся значения теоретической функции распределения в точках разбиения $u_j$, которые вычисляются для распределения Вейбулла с помощью функции ВЕЙБУЛЛ($u_j, r, \lambda, 1$) в Меню/Вставка/Функция/Статистические, а для логнормального распределения - с помощью функции ЛОГНОРМРАСП($u_j, \mu, \sigma$) в Меню/Вставка/Функция/Статистические.

Для построения графика нужно, щелкнув правой кнопкой мыши в Области построения диаграммы, выбрать Исходные данные/Ряд/Добавить и скопировать в поле *Значения X* столбец, содержащий точки разбиения, а в поле

*Значения Y* – столбец значений теоретической функции распределения. Нажав кнопку ОК, получаем график теоретической функции распределения.

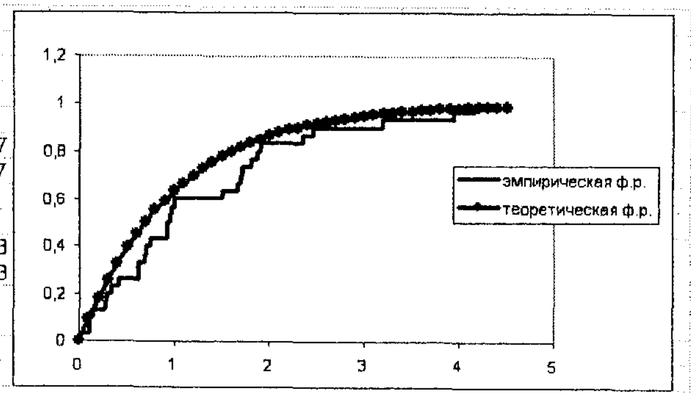
В *таблице 3* приведены столбцы данных для построения графиков эмпирической и теоретической функций распределения для примера из задания 1 – распределения Вейбулла $W(1,1)$, а также построенные по этим данным графики.

*Таблица 3*

| Эмпирическая функция распределения. | | Теоретическое распределение | |
|---|---|---|---|
| 0,019239271 | 0 | 0 | 0 |
| 0,019239271 | 1/30 | 0,1 | 0,095162582 |
| 0,10697888 | 1/30 | 0,2 | 0,181269247 |
| 0,10697888 | 0,066666667 | 0,3 | 0,259181779 |
| 0,110407836 | 0,066666667 | 0,4 | 0,329679954 |
| 0,110407836 | 1/10 | 0,5 | 0,39346934 |
| 0,148217048 | 1/10 | 0,6 | 0,451188364 |
| 0,148217048 | 0,133333333 | 0,7 | 0,503414696 |
| 0,263454902 | 0,133333333 | 0,8 | 0,550671036 |
| 0,263454902 | 1/6 | 0,9 | 0,59343034 |
| 0,284450738 | 1/6 | 1 | 0,632120559 |
| 0,284450738 | 0,2 | 1,1 | 0,667128916 |
| 0,3279382 | 0,2 | 1,2 | 0,698805788 |
| 0,3279382 | 7/30 | 1,3 | 0,727468207 |
| 0,417126872 | 7/30 | 1,4 | 0,753403036 |
| 0,417126872 | 0,266666667 | 1,5 | 0,77686984 |
| 0,615129532 | 0,266666667 | 1,6 | 0,798103482 |
| 0,615129532 | 3/10 | 1,7 | 0,817316476 |
| 0,617270619 | 3/10 | 1,8 | 0,834701112 |
| 0,617270619 | 0,333333333 | 1,9 | 0,850431381 |
| 0,683307635 | 0,333333333 | 2 | 0,864664717 |
| 0,683307635 | 11/30 | 2,1 | 0,877543572 |
| 0,690556227 | 11/30 | 2,2 | 0,889196842 |
| 0,690556227 | 0,4 | 2,3 | 0,899741156 |
| 0,745317271 | 0,4 | 2,4 | 0,909282047 |
| 0,745317271 | 13/30 | 2,5 | 0,917915001 |
| 0,921818008 | 13/30 | 2,6 | 0,925726422 |
| 0,921818008 | 0,466666667 | 2,7 | 0,932794487 |
| 0,927980224 | 0,466666667 | 2,8 | 0,939189937 |
| 0,927980224 | 1/2 | 2,9 | 0,94497678 |
| 0,950930092 | 1/2 | 3 | 0,950212932 |
| 0,950930092 | 0,533333333 | 3,1 | 0,954950798 |
| 0,970320123 | 0,533333333 | 3,2 | 0,959237796 |
| 0,970320123 | 17/30 | 3,3 | 0,963116833 |
| 0,994894594 | 17/30 | 3,4 | 0,96662673 |
| 0,994894594 | 0,6 | 3,5 | 0,969802617 |
| 1,50861118 | 0,6 | 3,6 | 0,972676278 |
| 1,50861118 | 19/30 | 3,7 | 0,975276474 |
| 1,662138304 | 19/30 | 3,8 | 0,977629228 |
| 1,662138304 | 0,666666667 | 3,9 | 0,979758089 |
| 1,700465321 | 0,666666667 | 4 | 0,981684361 |
| 1,700465321 | 7/10 | 4,1 | 0,983427325 |
| 1,717636578 | 7/10 | 4,2 | 0,985004423 |
| 1,717636578 | 0,733333333 | 4,3 | 0,986431441 |
| 1,802090764 | 0,733333333 | 4,4 | 0,98772266 |
| 1,802090764 | 23/30 | 4,5 | 0,988891003 |



| | |
|---|---|
| 1,879811142 | 23/30 |
| 1,879811142 | 0,8 |
| 1,897124181 | 0,8 |
| 1,897124181 | 5/6 |
| 2,342313608 | 5/6 |
| 2,342313608 | 0,866666667 |
| 2,460582537 | 0,866666667 |
| 2,460582537 | 9/10 |
| 3,186887282 | 9/10 |
| 3,186887282 | 0,933333333 |
| 3,931873824 | 0,933333333 |
| 3,931873824 | 29/30 |
| 4,160674435 | 29/30 |
| 4,160674435 | 1 |




Гистограмма и полигон частот

Для наглядности представления результатов наблюдений иногда бывает удобно построить другие виды графиков статистического распределения, в частности, гистограмму и полигон частот.

Построение гистограммы и полигона основано на группировке статистических данных. Для этого отрезок $[a;\ b]$, содержащий все выборочные значения, т.е. $a \le X_1^*$ и $b \ge X_n^*$, разбивается на $m$ непересекающихся интервалов длины $\Delta$: $(z_0; z_1], (z_1; z_2], \ldots, (z_{m-1}; z_m]$, где $z_0 = a$, $z_m = b$, $z_k - z_{k-1} = \Delta$, $k = 1, \ldots, m$; рекомендуемое количество интервалов вычисляется по формуле Стерджесса $m = 1 + 1{,}4 \ln n$, а длина каждого интервала $\Delta = \dfrac{b-a}{m}$; в частности, для выборки объема $n$=30 рекомендуемое количество интервалов $m$=6. Затем по вариационному ряду подсчитывается число выборочных значений, попавших в каждый интервал; *число наблюдений, попавших в k-й интервал* $(z_{k-1}; z_k]$, *называется эмпирической частотой и обозначается* $n_k$; *величина* $W_k = \dfrac{n_k}{n}$ *называется относительной эмпирической частотой.*

*Определение 4. Гистограммой относительных частот называется функция* $\hat{f}_n(x) = \begin{cases} 0, & x \notin [a;b] \\ \dfrac{W_k}{\Delta}, & x \in (z_{k-1}; z_k],\ k = 1, \ldots, m \end{cases}$.

Геометрически гистограмму можно представить как фигуру на плоскости, состоящую из прямоугольников, основаниями которых (т.е. горизонтальными сторонами) являются интервалы $(z_0; z_1], (z_1; z_2], \ldots, (z_{m-1}; z_m]$, а высоты (вертикальные стороны) равны соответственно $\dfrac{W_k}{\Delta}$. Площадь $k$-го прямоугольника равна $\dfrac{W_k}{\Delta}\Delta = W_k$, а сумма площадей всех прямоугольников,

составляющих гистограмму, равна $\sum_{k=1}^{m} W_k = \frac{1}{n}\sum_{k=1}^{m} n_k = \frac{n}{n} = 1$. Отметим, что это свойство гистограммы аналогично свойству плотности распределения $\int_{-\infty}^{\infty} f(x)dx = 1$, т.к. геометрический смысл этого интеграла – площадь фигуры, ограниченной осью абсцисс и графиком функции $f(x)$. Поэтому гистограмму разумно считать оценкой для теоретической плотности распределения.

Для построения полигона вычисляются середины интервалов группировки: $z_k^* = \frac{z_{k-1} + z_k}{2} = z_{k-1} + \frac{\Delta}{2}$, $k = 1,...,m$.

*Определение 5. Полигоном относительных частот называется ломаная, соединяющая точки* $\left(z_1^*; \frac{W_1}{\Delta}\right), \left(z_2^*; \frac{W_2}{\Delta}\right), ..., \left(z_m^*; \frac{W_m}{\Delta}\right)$.

## Построение гистограммы и полигона

Сформируем данные для построения гистограммы и полигона. Сначала разобьем выбранный отрезок [*a; b*] на интервалы. Для этого по приведенной формуле найдем величину $\Delta$ и вычислим границы интервалов (точки разбиения) $z_k$, $k = 1,...,m$. Это делается описанным ранее способом: в две первые ячейки записываются соответственно значения $a$ и $a+\Delta$, а затем значения «протягиваются» до $b$. Далее по вариационному ряду подсчитываются эмпирические частоты $n_k$ и по соответствующей формуле вычисляются значения $\frac{W_k}{\Delta}$: для удобства их рекомендуется сохранить как значения с помощью Меню/Правка/Специальная вставка/Значения. Затем формируются столбцы данных для построения гистограммы. Заметим, что точкам $z_0$ и $z_m$ в гистограмме соответствует по два значения: 0 и соответственно $\frac{W_1}{\Delta}$ или $\frac{W_m}{\Delta}$, а остальным $z_k$, $k = 1,...,m-1$ - по три: 0 , $\frac{W_k}{\Delta}$ и $\frac{W_{k+1}}{\Delta}$. Поэтому в столбце аргументов значения $z_0$ и $z_m$ должны повторяться по 2 раза, а остальные $z_k$ - по 3. Чтобы получить столбец аргументов, два раза копируем все $z_k$ и один раз $z_k$ для $k = 1,...,m-1$ в один столбец с помощью Меню/Правка/Копировать, а затем упорядочиваем их, используя Меню/Данные/Сортировка/По возрастанию. В столбец значений записываем следующие величины: $0; \frac{W_1}{\Delta}; \frac{W_1}{\Delta}; 0; \frac{W_2}{\Delta}; \frac{W_2}{\Delta}; 0; \ldots; 0; \frac{W_m}{\Delta}; \frac{W_m}{\Delta}; 0$. Сформированные столбцы данных для построения гистограммы можно интерпретировать как таблицу, в которой перечислены координаты вершин всех прямоугольников, составляющих гистограмму, в том порядке, в котором надо соединять эти вершины, чтобы получить гистограмму.



Теперь можно построить гистограмму. Выберем тип диаграммы в Меню/Вставка/Диаграмма/Точечная/Точечная диаграмма со значениями, соединенными отрезками без маркеров и введем данные, по которым будет строиться диаграмма, в Источник данных диаграммы/Ряд/Добавить; в поле *Значения X* копируется столбец аргументов, а в поле *Значения Y* – столбец значений. Нажав кнопку Готово, получаем изображение гистограммы.

Чтобы построить полигон, необходимо вычислить середины интервалов $z_k^*$, затем, щелкнув правой кнопкой мыши в Области построения диаграммы, выбрать Исходные данные/Ряд/Добавить и скопировать полученные значения в поле *Значения X*; в поле *Значения Y* копируется столбец значений $\frac{W_k}{\Delta}$. Нажав кнопку ОК, получаем полигон: это будет ломаная линия, соединяющая середины горизонтальных сторон прямоугольников, образующих гистограмму.

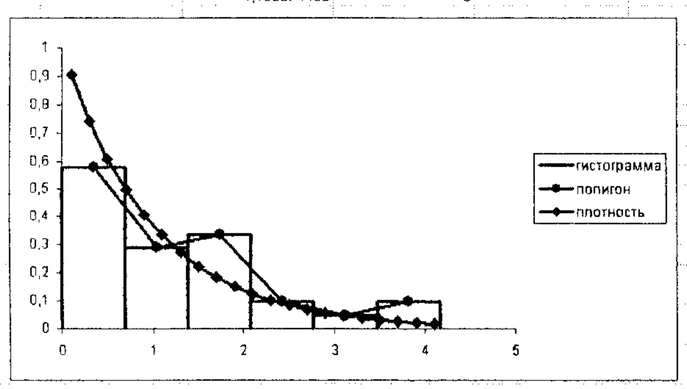
Для сравнения гистограммы и полигона с графиком теоретической плотности распределения рекомендуется на той же координатной плоскости построить этот график на выбранном отрезке [*a; b*]. Это делается точно так же, как и для теоретической функции распределения; при этом значения теоретической плотности в точках разбиения $u_j$ вычисляются для распределения Вейбулла с помощью функции ВЕЙБУЛЛ($u_j, r, \lambda, 0$) в Меню/Вставка/Функция/Статистические, а для логнормального распределения рекомендуется использовать функцию НОРМРАСП($ln(u_j), \mu, \sigma, 0$) в Меню/Вставка/Функция/Статистические: значение логнормальной плотности распределения $LogN(\mu, \sigma)$ в точке $u_j$ равно $\frac{1}{u_j}$НОРМРАСП($ln(u_j), \mu, \sigma, 0$) (см. *таблицу 1*).

В *таблице 4* приведены данные для построения гистограммы, полигона и графика теоретической плотности распределения для примера из задания 1 – распределения Вейбулла $W(1,1)$, а также построенные по этим данным графики.

*Таблица 4*

| Длина интервала | | 0,693445739 | Гистограмма | | Середины интервалов | Теоретическая плотность | |
|---|---|---|---|---|---|---|---|
| Границы интервалов | | | 0 | 0 | 0,34672287 | 0,1 | 0,904837418 |
| | z0 | 0 | 0 | 0,576829559 | 1,040168609 | 0,3 | 0,740818221 |
| | z1 | 0,693445739 | 0,693445739 | 0,576829559 | 1,733614348 | 0,5 | 0,60653066 |
| | z2 | 1,386891478 | 0,693445739 | 0 | 2,427060087 | 0,7 | 0,496585304 |
| | z3 | 2,080337218 | 0,693445739 | 0,28841478 | 3,120505826 | 0,9 | 0,40656966 |
| | z4 | 2,773782957 | 1,386891478 | 0,28841478 | 3,813951566 | 1,1 | 0,332871084 |
| | z5 | 3,467228696 | 1,386891478 | 0 | | 1,3 | 0,272531793 |
| | z6 | 4,160674435 | 1,386891478 | 0,336483909 | | 1,5 | 0,22313016 |
| Частоты | | Wk/Δ | 2,080337218 | 0,336483909 | | 1,7 | 0,182683524 |
| n1 | 12 | 0,576829559 | 2,080337218 | 0 | | 1,9 | 0,149568619 |
| n2 | 6 | 0,28841478 | 2,080337218 | 0,09613826 | | 2,1 | 0,122456428 |
| n3 | 7 | 0,336483909 | 2,773782957 | 0,09613826 | | 2,3 | 0,100258844 |
| n4 | 2 | 0,09613826 | 2,773782957 | 0 | | 2,5 | 0,082084999 |
| n5 | 1 | 0,04806913 | 2,773782957 | 0,04806913 | | 2,7 | 0,067205513 |
| n6 | 2 | 0,09613826 | 3,467228696 | 0,04806913 | | 2,9 | 0,05502322 |
| | | | 3,467228696 | 0 | | 3,1 | 0,045049202 |
| | | | 3,467228696 | 0,09613826 | | 3,3 | 0,036883167 |

| | | | |
|---|---|---|---|
| 4,160674435 | 0,09613826 | 3,5 | 0,030197383 |
| 4,160674435 | 0 | 3,7 | 0,024723526 |
| | | 3,9 | 0,020241911 |
| | | 4,1 | 0,016572675 |




## Задание 3. Точечные оценки неизвестных параметров

Постановка задачи

Пусть в результате наблюдений получена выборка $(X_1,...,X_n)$ из генеральной совокупности с теоретическим распределением *F(x)*, и относительно функции *F(x)* известно только, что она принадлежит определенному параметрическому семейству $F(x,\theta)$, где $\theta=(\theta_1,...,\theta_k)$ - вектор параметров, т.е. вид ее известен, но неизвестны параметры, определяющие это распределение. Естественно, возникает задача об оценке этих неизвестных параметров по выборке.

Назовем оценкой (статистической оценкой) $\hat{\theta}$ некоторую функцию от выборочных значений $\hat{\theta}=\hat{\theta}_n(X_1,...,X_n)$, предназначенную для статистического оценивания неизвестных параметров распределения. Если вычисление этой функции при подстановке в нее конкретных результатов наблюдений приводит к одному значению параметра (число при *k=1* или вектор при *k>1*), то такая оценка называется точечной.

Задача состоит в выборе такой функции $\hat{\theta}$, которая в некотором смысле мало отличалась бы от истинного значения оцениваемого параметра. Желательными свойствами такой оценки являются:

- несмещенность
- состоятельность
- эффективность.

*Определение 6. Оценка $\hat{\theta}$ называется несмещенной, если ее математическое ожидание равно оцениваемому параметру:* $E\hat{\theta}=\theta$.

Выполнение этого свойства гарантирует, что оценка не будет давать систематического отклонения результата.



*Определение 7. Оценка $\hat{\theta}$ называется состоятельной, если она сходится по вероятности к оцениваемому параметру:* $E\hat{\theta}\xrightarrow{P}\theta$, *т.е. для любого* $\varepsilon>0$ $P\left(\left|\hat{\theta}_n-\theta\right|<\varepsilon\right)\to 1$ *при* $n\to\infty$.

Это свойство означает, что при достаточном объеме выборки оценка с вероятностью, близкой к 1, будет мало отличаться от истинного значения оцениваемого параметра.

*Определение 8. Оценка $\hat{\theta}$ называется эффективной, если она имеет минимальную дисперсию в определенном классе оценок.*

Поскольку дисперсия характеризует разброс значений случайной величины вокруг математического ожидания, то для несмещенной эффективной оценки разброс значений $\hat{\theta}$ вокруг оцениваемого параметра $\theta$ будет наименьшим.

Существуют различные методы получения точечных оценок; рассмотрим два из них – метод моментов и метод максимального правдоподобия.

Метод моментов

Метод моментов основан на том, что с ростом числа наблюдений эмпирическая функция распределения мало отличается от теоретической, поэтому мало отличаются и соответствующие числовые характеристики, что позволяет, в частности, считать приблизительно равными теоретические и эмпирические моменты одинаковых порядков.

Напомним, что моментом (начальным моментом) порядка *N* случайной величины *X* называется математическое ожидание *N*-й степени этой случайной величины: $m_N=EX^N$. Поскольку теоретическое распределение зависит от вектора параметров $\theta=(\theta_1,...,\theta_k)$, то и теоретические моменты также являются функциями этих параметров: для дискретной теоретической случайной величины *X* $m_N=m_N(\theta)=\sum_j x_j^N P_j(\theta)$, где $P_j(\theta)$- закон распределения случайной величины *X*; для непрерывной теоретической случайной величины *X* $m_N=m_N(\theta)=\int_{-\infty}^{+\infty}x^N f(x,\theta)dx$, где $f(x,\theta)$ - плотность распределения случайной величины *X*.

Из *определения 3* следует, что эмпирическая функция распределения соответствует распределению дискретной случайной величины, которая принимает каждое из выборочных значений $X_1,...,X_n$ с вероятностью $\frac{1}{n}$, поэтому эмпирический момент порядка *N* вычисляется по формуле: $M_N=\frac{1}{n}\sum_{i=1}^{n}X_i^N$.

Метод моментов состоит в приравнивании теоретических и эмпирических моментов соответствующих порядков. Полученные при этом равенства

образуют систему уравнений относительно параметров $\theta_1,...,\theta_k$; число уравнений в этой системе должно совпадать с числом неизвестных параметров:

$$\begin{cases} m_1(\theta_1,...,\theta_k) = M_1 \\ m_2(\theta_1,...,\theta_k) = M_2 \\ \quad ... \\ m_k(\theta_1,...,\theta_k) = M_k \end{cases}$$

. Решая эту систему, получаем точечную оценку для вектора неизвестных параметров $\hat{\theta} = (\hat{\theta}_1,...,\hat{\theta}_k)$.

В частности, если $k$=1, т.е. требуется оценить один параметр, достаточно выписать и решить одно уравнение относительно этого параметра. Приравнивая теоретический и эмпирический моменты первого порядка, получаем: $m_1(\theta) = M_1$. Заметим, что $m_1(\theta) = EX$ - это теоретическое математическое ожидание, а $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_в$ - выборочное среднее, т.е. среднее арифметическое выборочных значений.

Построим с помощью метода моментов точечные оценки неизвестных параметров распределения Вейбулла и логнормального распределения.

Распределение Вейбулла $W(r,\lambda)$. В задании 3 требуется построить точечную оценку параметра $\lambda$, считая параметр $r$ известным. Приравняем теоретическое математическое ожидание и выборочное среднее: $\lambda\Gamma\left(1+\frac{1}{r}\right) = \overline{X}_в$. Решая это уравнение, получаем оценку параметра $\lambda$: $\hat{\lambda} = \frac{\overline{X}_в}{\Gamma\left(1+\frac{1}{r}\right)}$. Для вычисления $\hat{\lambda}$ воспользуемся Excel: значение $\overline{X}_в$ определяется с помощью функции СРЗНАЧ($X_1,...,X_n$) в Меню/Вставка/ /Функция/Статистические, где в поле аргументов копируется выборка, а гамма-функция $\Gamma\left(1+\frac{1}{r}\right)$ вычисляется с помощью функций ГАММАНЛОГ$\left(1+\frac{1}{r}\right)$ в Меню/Вставка/Функция/Статистические, значением которой является $\ln\Gamma\left(1+\frac{1}{r}\right)$, и EXP*(x)* в Меню/Вставка/Функция/Математические, которая возвращает экспоненту заданного числа *x*; тогда $\Gamma\left(1+\frac{1}{r}\right)$=EXP(ГАММАНЛОГ$\left(1+\frac{1}{r}\right)$).

В нашем примере $r$=1, тогда $\hat{\lambda} = \frac{\overline{X}_в}{\Gamma(2)}$. Результаты вычислений представлены в *таблице 5.*

*Таблица 5*



| | |
|---|---|
| Гамма-функция | 1 |
| Выборочное среднее | 1,268185 |
| Оценка для λ | 1,268185 |

Логнормальное распределение $LogN(\mu,\sigma)$. В задании 3 требуется построить точечную оценку параметра $\mu$, считая параметр $\sigma$ известным. Приравняем теоретическое математическое ожидание и выборочное среднее: $e^{\mu+\frac{\sigma^2}{2}} = \overline{X}_в$. Решая это уравнение, получаем оценку параметра $\mu$: $\hat{\mu} = \ln(\overline{X}_в) - \frac{\sigma^2}{2}$. Значение $\hat{\mu}$ вычисляется в Excel с помощью функций СРЗНАЧ($X_1,...,X_n$) в Меню/Вставка/Функция/Статистические, где в поле аргументов копируется выборка, значением которой является $\overline{X}_в$, и LN*(x)* в Меню/Вставка/Функция/Математические, которая возвращает натуральный логарифм заданного числа *x*: $\hat{\mu}$ = LN(СРЗНАЧ($X_1,...,X_n$))-$\frac{\sigma^2}{2}$.

## Метод максимального правдоподобия

При нахождении точечной оценки для неизвестных параметров $\theta = (\theta_1,...,\theta_k)$ распределения *F(x,θ)* естественно попытаться выбрать такое значение $\hat{\theta}$, которое бы лучше всего соответствовало полученной выборке $(X_1,...,X_n)$. Это означает, что при $\theta = \hat{\theta}$ вероятность (в дискретном случае) или плотность вероятности (в непрерывном случае) реализации данной выборки $L(\theta) = L(\theta, X_1,...,X_n)$ будет максимальной. Если теоретическая случайная величина $X$ является дискретной, то вероятность реализации выборки $(X_1,...,X_n)$ - это вероятность совместного осуществления событий $\{X = X_i\}, i = 1,...,n$; если $X$ – непрерывна, то соответствующая плотность вероятности - это совместная плотность распределения в точках $X_1,...,X_n$. По *определению 2* случайные величины $X_i$ независимы и одинаково распределены, поэтому их совместное распределение $L(\theta, X_1,...,X_n)$ вычисляется как произведение вероятностей в дискретном случае или как произведение плотностей в непрерывном случае. Функция $L(\theta, X_1,...,X_n)$ была названа Фишером функцией правдоподобия.

*Определение 9. Функция*

$$L(\theta, X_1,...,X_n) = P(X_1,\theta)P(X_2,\theta)...P(X_n,\theta) = \prod_{i=1}^{n} P(X_i,\theta),$$

*где* $P(X_i,\theta) = P\{X = X_i\}$, *в дискретном случае и*

$$L(\theta, X_1,...,X_n) = f(X_1,\theta)f(X_2,\theta)...f(X_n,\theta) = \prod_{i=1}^{n} f(X_i,\theta),$$

*где* $f(x,\theta) = F'(x,\theta)$, *в непрерывном случае называется функцией правдоподобия.*

*Определение 10. Оценкой максимального правдоподобия называется такое значение $\hat{\theta}$, для которого $L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$, где $\Theta$ - замкнутая область допустимых значений параметров.*

Иногда на практике бывает удобнее пользоваться не самой функцией правдоподобия, а ее логарифмом; функция $l(\theta) = lnL(\theta)$ называется логарифмической функцией правдоподобия; очевидно, точки максимума у функций $L(\theta)$ и $l(\theta)$ совпадают.

Если максимум функции $l(\theta)$ достигается внутри области $\Theta$, то в точке максимума выполняются необходимые условия экстремума:

$$\frac{\partial l(\theta)}{\partial \theta_j} = 0, \quad j = 1,...,k \ .$$

Полученные уравнения называются уравнениями правдоподобия. Решения системы уравнений правдоподобия могут быть точками максимума или минимума функции $l(\theta)$, а могут и не являться для нее точками экстремума, поэтому необходимо проверять, что полученное решение действительно является точкой максимума. Если система уравнений правдоподобия не имеет решений внутри области $\Theta$, это означает, что максимум функции $l(\theta)$ достигается на границе области допустимых значений $\Theta$.

Построим с помощью метода максимального правдоподобия точечные оценки неизвестных параметров распределения Вейбулла и логнормального распределения.

Распределение Вейбулла $W(r,\lambda)$. В задании 3 требуется построить точечную оценку параметра $\lambda$, считая параметр $r$ известным; таким образом, вектор неизвестных параметров имеет размерность $k$=1 и $\theta = \lambda$. По *определению 9* функция правдоподобия имеет вид:

$$L(\lambda, X_1,...,X_n) = \prod_{i=1}^{n} f(X_i,\lambda) = \frac{r^n}{\lambda^{rn}} \left( \prod_{i=1}^{n} X_i \right)^{r-1} e^{-\left(\frac{1}{\lambda}\right)^r \sum_{i=1}^{n} X_i}$$ ; тогда логарифмическая функция правдоподобия равна:

$$l(\lambda) = n \ln r - rn \ln \lambda + (r-1) \sum_{i=1}^{n} \ln X_i - \frac{1}{\lambda^r} \sum_{i=1}^{n} X_i \ .$$

Дифференцируя по $\beta$, получаем уравнение правдоподобия:

$$-rn \frac{1}{\lambda} + r \frac{1}{\lambda^{r+1}} \sum_{i=1}^{n} X_i = 0 \ .$$

Легко проверить, что решение этого уравнения $\hat{\lambda} = (\overline{X}_e)^{\frac{1}{r}}$ является точкой максимума для функции $l(\lambda)$, поэтому $\hat{\lambda} = (\overline{X}_e)^{\frac{1}{r}}$ - оценка максимального правдоподобия.

Значение $\hat{\lambda}$ вычисляется в Excel по полученной формуле с помощью функции СРЗНАЧ( $X_1,...,X_n$) в Меню/Вставка/Функция/Статистические.



В нашем примере $r$=1, тогда $\hat{\lambda} = \overline{X}_e$, и значение оценки максимального правдоподобия совпадает со значением оценки, полученной методом моментов (*таблица 5*).

Логнормальное распределение $LogN(\mu,\sigma)$. В задании 3 требуется построить точечную оценку параметра $\mu$, считая параметр $\sigma$ известным. Поэтому вектор неизвестных параметров имеет размерность $k$=1 и $\theta = \mu$. По *определению 9* функция правдоподобия имеет вид:

$$L(\mu, X_1,...,X_n) = \prod_{i=1}^{n} f(X_i,\mu) = \frac{1}{\left(\prod_{i=1}^{n} X_i\right) \sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\ln X_i - \mu)^2}$$ ; тогда

логарифмическая функция правдоподобия равна:

$$l(\mu) = n \ln \sigma - \frac{n}{2} \ln(2\pi) - \sum_{i=1}^{n} \ln X_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\ln X_i - \mu)^2 \ .$$

Дифференцируя по $\mu$, получаем уравнение правдоподобия:

$$\frac{1}{\sigma^2} \left( \sum_{i=1}^{n} \ln X_i - n\mu \right) = 0 \ .$$

Легко проверить, что решение этого уравнения $\hat{\mu} = \frac{\sum_{i=1}^{n} \ln X_i}{n}$ является точкой максимума для функции $l(\mu)$, поэтому $\hat{\mu} = \frac{\sum_{i=1}^{n} \ln X_i}{n}$ - оценка максимального правдоподобия.

Значение $\hat{\mu}$ вычисляется в Excel по полученной формуле с помощью функций LN*(x)*, которая возвращает натуральный логарифм заданного числа *x*, и СУММ(), которая суммирует аргументы, в Меню/Вставка/Функция/ /Математические: сначала вычисляется столбец значений $\ln X_i$, затем он копируется в поле аргументов функции СУММ(), и полученное значение этой функции делится на *n*; можно воспользоваться также функцией СРЗНАЧ() в Меню/Вставка/Функция/Статистические, где в поле аргументов копируется столбец значений $\ln X_i$.

## Задание 4. Доверительные интервалы

### Постановка задачи

При рассмотрении методов получения точечных оценок было показано, что точечная оценка не совпадает с оцениваемым параметром; при малом объеме выборки такая оценка может значительно отличаться от истинного значения параметра. Поэтому разумно было бы указывать те допустимые границы, в которых может находиться неизвестный параметр $\theta$ при условии реализации выборки $(X_1,...,X_n)$, т.е. возникает задача интервального

оценивания. Доверительный интервал – это статистическая оценка параметра вероятностного распределения, имеющая вид интервала, границы которого являются функциями от результатов наблюдений и который с заданной вероятностью «накрывает» неизвестное значение параметра.

Итак, пусть в результате наблюдений получена выборка $(X_1,...,X_n)$ из генеральной совокупности с теоретическим распределением $F(x,\theta)$, зависящим от числового параметра $\theta$, $\theta \in \Theta \subseteq R$, значение которого неизвестно; $\alpha$, $0<\alpha<1$ - фиксированное число.

*Определение 11. Интервал* $I(X_1,...,X_n)=(\hat{\theta}_1;\hat{\theta}_2)$ *с границами* $\hat{\theta}_1=\hat{\theta}_1(X_1,...,X_n)$ *и* $\hat{\theta}_2=\hat{\theta}_2(X_1,...,X_n)$, $\hat{\theta}_1<\hat{\theta}_2$, *такой, что* $\inf\limits_{\theta\in\Theta} P(\hat{\theta}_1<\theta<\hat{\theta}_2)=1-\alpha$, *называется доверительным интервалом надежности* $1-\alpha$ *для неизвестного параметра* $\theta$.

Задача состоит в том, чтобы по выборке $(X_1,...,X_n)$ при заданном уровне надежности $1-\alpha$ найти функции $\hat{\theta}_1=\hat{\theta}_1(X_1,...,X_n)$ и $\hat{\theta}_2=\hat{\theta}_2(X_1,...,X_n)$.

Из *определения 11* следует, что интервальные оценки позволяют установить точность и надежность точечных оценок.

Действительно, пусть $\hat{\theta}$ - точечная оценка неизвестного параметра $\theta$. Точность этой оценки определяется отклонением $|\hat{\theta}-\theta|$; если $\varepsilon>0$ и выполняется неравенство $|\hat{\theta}-\theta|<\varepsilon$, то чем меньше $\varepsilon$, тем точнее оценка. Таким образом, положительное число $\varepsilon$ характеризует точность оценки.

Поскольку оценка $\hat{\theta}=\hat{\theta}_n(X_1,...,X_n)$ является функцией от случайных величин, то нельзя наверняка утверждать, что оценка $\hat{\theta}$ удовлетворяет неравенству $|\hat{\theta}-\theta|<\varepsilon$; можно говорить только о вероятности, с которой это неравенство выполняется; эта вероятность называется надежностью (или доверительной вероятностью) оценки $\hat{\theta}$.

Пусть $P(|\hat{\theta}-\theta|<\varepsilon)=1-\alpha$. Неравенство $|\hat{\theta}-\theta|<\varepsilon$ равносильно неравенству $\hat{\theta}-\varepsilon<\theta<\hat{\theta}+\varepsilon$, поэтому $P(\hat{\theta}-\varepsilon<\theta<\hat{\theta}+\varepsilon)=1-\alpha$, т.е. интервал $(\hat{\theta}-\varepsilon;\hat{\theta}+\varepsilon)$ является доверительным интервалом надежности $1-\alpha$ для неизвестного параметра $\theta$. Таким образом, интервальная оценка определяет надежность точечной оценки.

Число $\alpha$ называется уровнем значимости. В математической статистике обычно задаются следующие значения уровня значимости: $\alpha$=0,01; 0,05; 0,1.

### Доверительный интервал для неизвестного математического ожидания при известной дисперсии в случае нормального распределения генеральной совокупности

Пусть $(X_1,...,X_n)$ - выборка из генеральной совокупности с нормальным распределением $N(\theta,\sigma)$, $\theta$ - неизвестное математическое ожидание; дисперсия



$\sigma^2$ предполагается известной. Построим доверительный интервал для $\theta$ при заданном уровне значимости $\alpha$.

Точечную оценку $\hat{\theta}$ для неизвестного математического ожидания $\theta$ несложно получить, например, с помощью метода моментов, приравнивая теоретический и эмпирический моменты первого порядка: $\hat{\theta}=\overline{X}_в$. Теперь, чтобы построить доверительный интервал заданной надежности $1-\alpha$, надо найти такое число $\varepsilon$, чтобы $P(|\overline{X}_в-\theta|<\varepsilon)=1-\alpha$.

По *определению 2* случайные величины $X_1,...,X_n$ независимы, и распределение каждой из них совпадает с теоретическим распределением: $X_i \sim N(\theta,\sigma)$; соответственно, $EX_i=\theta$, $DX_i=\sigma^2$. Напомним, что сумма независимых нормально распределенных случайных величин снова имеет нормальное распределение, а линейное преобразование $aX+b$ при $a>0$ не меняет вид распределения, поэтому случайная величина $\overline{X}_в-\theta=\frac{1}{n}\sum\limits_{i=1}^{n}X_i-\theta$ имеет нормальное распределение. Найдем параметры этого распределения. Используя свойства математического ожидания и дисперсии, получаем:

$$E(\overline{X}_в-\theta)=\frac{1}{n}\sum_{i=1}^{n}EX_i-\theta=\frac{n\theta}{n}-\theta=0;$$

$$D(\overline{X}_в-\theta)=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}DX_i=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}.$$

Таким образом, $\overline{X}_в-\theta\sim N\left(0,\frac{\sigma}{\sqrt{n}}\right)$. Разделим эту случайную величину на корень из дисперсии (напомним, что такая процедура называется нормированием случайной величины; при этом вид распределения не меняется, а дисперсия нормированной случайной величины равна 1); тогда случайная величина $Y_0=\frac{\sqrt{n}(\overline{X}_в-\theta)}{\sigma}$ имеет стандартное нормальное распределение: $Y_0\sim N(0,1)$.

Следовательно,

$$P(|\overline{X}_в-\theta|<\varepsilon)=P\left(\frac{\sqrt{n}|\overline{X}_в-\theta|}{\sigma}<\frac{\sqrt{n}\varepsilon}{\sigma}\right)=P\left(|Y_0|<\frac{\sqrt{n}\varepsilon}{\sigma}\right)=P\left(-\frac{\sqrt{n}\varepsilon}{\sigma}<Y_0<\frac{\sqrt{n}\varepsilon}{\sigma}\right)=$$

$$=\int_{-\infty}^{\frac{\sqrt{n}\varepsilon}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx-\int_{-\infty}^{-\frac{\sqrt{n}\varepsilon}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=2\int_{-\infty}^{\frac{\sqrt{n}\varepsilon}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx-1=2F_0\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)-1=1-\alpha,$$ где

$F_0(t)=\int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$ - функция распределения стандартного нормального распределения $N(0,1)$.

Получаем, что $F_0\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right)=1-\frac{\alpha}{2}$.

Напомним, что решение $x_p$ уравнения $F(x_p)=p$, где $F(x)$ - некоторое заданное распределение, называется квантилью уровня вероятности $p$ распределения $F$. Тогда $\frac{\sqrt{n}\varepsilon}{\sigma}=d_{1-\alpha/2}$, где $d_{1-\alpha/2}$ - квантиль уровня вероятности $1-\frac{\alpha}{2}$ стандартного нормального распределения, т.е. решение уравнения $\int_{-\infty}^{d_{1-\alpha/2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=1-\frac{\alpha}{2}$. Отсюда следует, что $\varepsilon=\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}$, а интервал $\left(\overline{X}_в-\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}};\overline{X}_в+\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}\right)$ является доверительным интервалом заданной надежности $1-\alpha$ для неизвестного математического ожидания $\theta$.

Таким образом, с заданной вероятностью $1-\alpha$ выполняется неравенство:

$$\overline{X}_в-\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}<\theta<\overline{X}_в+\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}.$$

Полученная интервальная оценка называется классической; ее точность равна $\varepsilon=\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}$, а границы имеют вид: $\hat{\theta}_1=\overline{X}_в-\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}$; $\hat{\theta}_2=\overline{X}_в+\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}$.

### Доверительный интервал для неизвестного математического ожидания в случае произвольного теоретического распределения генеральной совокупности

Пусть теперь $(X_1,...,X_n)$ - выборка из генеральной совокупности с произвольным теоретическим распределением $F(x,\theta)$, отличным от нормального; $\theta$ - неизвестное математическое ожидание; дисперсия $\sigma^2$ предполагается известной.

В этом случае точечной оценкой $\hat{\theta}$ для неизвестного математического ожидания $\theta$, как и при нормальном распределении генеральной совокупности, является выборочное среднее: $\hat{\theta}=\overline{X}_в$. Это следует, например, из метода моментов: независимо от вида теоретического распределения первый теоретический момент – это всегда математическое ожидание, а эмпирический – выборочное среднее. По *определению 2* случайные величины $X_1,...,X_n$ независимы и одинаково распределены, поэтому и $EX_i=\theta$, $DX_i=\sigma^2$, $i=1,...,n$. Согласно центральной предельной теореме, если независимые одинаково распределенные случайные величины $X_1,...,X_n$ имеют конечную дисперсию, то каков бы ни был их закон распределения, сумма $\sum_{i=1}^{n}X_i$ при достаточно больших $n$ ($n\geq 30$) имеет распределение, близкое к нормальному. Поскольку линейное преобразование случайной величины $X$ $aX+b$ при $a>0$ не меняет вид распределения, то при достаточно большом объеме выборки можно считать,



что случайная величина $\overline{X}_в-\theta=\frac{1}{n}\sum_{i=1}^{n}X_i-\theta$ имеет нормальное распределение с теми же параметрами, что и в случае нормального распределения генеральной совокупности, так как общие свойства математического ожидания и дисперсии не зависят от вида функции распределения, и эти числовые характеристики вычисляются аналогично.

Таким образом, $\overline{X}_в-\theta\sim N\left(0,\frac{\sigma}{\sqrt{n}}\right)$. Поэтому все рассуждения, проведенные при построении доверительного интервала для неизвестного математического ожидания в случае нормального распределения генеральной совокупности, остаются справедливыми и для генеральной совокупности с произвольным теоретическим распределением. Соответственно, и доверительный интервал заданной надежности $1-\alpha$ для неизвестного математического ожидания имеет тот же самый вид: $\left(\overline{X}_в-\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}};\overline{X}_в+\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}\right)$.

Если дисперсия $\sigma^2$ неизвестна, то можно использовать несмещенную оценку для дисперсии, полученную по выборке; напомним, что такая оценка называется исправленной выборочной дисперсией и вычисляется по формуле: $\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_в)^2$. При достаточно большом объеме выборки оценку $\hat{\sigma}^2$ можно считать приблизительно равной истинному значению $\sigma^2$, т.к. с ростом числа наблюдений эмпирическая функция распределения мало отличается от теоретической, поэтому и соответствующие числовые характеристики отличаются незначительно. Заменяя $\sigma$ на $\hat{\sigma}$, получаем доверительный интервал заданной надежности $1-\alpha$ для неизвестного математического ожидания генеральной совокупности с произвольным теоретическим распределением при неизвестной дисперсии:

$$\left(\overline{X}_в-\frac{d_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}};\overline{X}_в+\frac{d_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}\right),\text{ где }\hat{\sigma}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_в)^2}.$$

Отметим, что замена $\sigma$ на $\hat{\sigma}$ оправдана при достаточно большом объеме выборки; для малых выборок такая замена может привести к существенным ошибкам, в частности, к значительному сужению доверительного интервала.

### Вычисление границ доверительных интервалов в Excel

При известной дисперсии границы доверительного интервала для неизвестного математического ожидания вычисляются следующим образом: $\hat{\theta}_1=\overline{X}_в-\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}$; $\hat{\theta}_2=\overline{X}_в+\frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}$. В условии курсовой работы значение дисперсии не задано, и его необходимо вычислить по формулам, представленным в *таблице 1*. Значение $\overline{X}_в$ можно найти, используя функцию

СРЗНАЧ($X_1,...,X_n$) в Меню/Вставка/Функция/Статистические, для $\sqrt{n}$ используется функция КОРЕНЬ($n$) в Меню/Вставка/Функция/Математические, а квантиль стандартного нормального распределения $d_{1-\alpha/2}$ вычисляется с помощью функции НОРМСТОБР($1-\frac{\alpha}{2}$) в Меню/Вставка/Функция/ /Статистические. Кроме того, для вычисления точности $\varepsilon = \frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}$ можно воспользоваться функцией ДОВЕРИТ($\alpha,\sigma,n$) в Меню/Вставка/Функция/ /Статистические.

Если дисперсия неизвестна, то границы доверительного интервала для неизвестного математического ожидания вычисляются по формулам: $\hat{\theta}_1 = \overline{X}_n - \frac{d_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}$; $\hat{\theta}_2 = \overline{X}_n + \frac{d_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}$. Значение $\hat{\sigma}$ вычисляется в Excel с помощью функции СТАНДОТКЛОН($X_1,...,X_n$) в Меню/Вставка/Функция/ /Статистические, а для вычисления точности $\varepsilon = \frac{d_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}$ снова можно воспользоваться функцией ДОВЕРИТ($\alpha,\hat{\sigma},n$) в Меню/Вставка/Функция/ /Статистические.

Результаты вычислений границ доверительных интервалов надежности 0,99 ($\alpha$=0,01) для неизвестного математического ожидания при известной и неизвестной дисперсии для распределения Вейбулла $W(1,1)$ представлены в *таблице 6.*

*Таблица 6*

| | | | | | |
|---|---|---|---|---|---|
| Выборочное среднее | 1,268185 | | | | |
| Гамма-функция(2) | 1 | | | | |
| Гамма-функция(3) | 2 | | | | |
| σ | 1 | | | | |
| оценка для σ | 1,094862 | | | | |
| корень из n | 5,477226 | | | | |
| квантиль | 2,575829 | дисперсия известна | | дисперсия неизвестна | |
| точность | | | 0,470279938 | | 0,514891843 |
| функция доверит | | | 0,470279938 | | 0,514891843 |
| границы | | 0,797905 | 1,73846485 | 0,7532931 | 1,783076751 |

По представленным в *таблице 6* результатам можно сделать следующие выводы:

1) оценка $\hat{\sigma}$ незначительно отличается от истинного значения $\sigma$;
2) значения точности, найденные по формулам $\varepsilon = \frac{d_{1-\alpha/2}\sigma}{\sqrt{n}}$ для известной дисперсии и $\varepsilon = \frac{d_{1-\alpha/2}\hat{\sigma}}{\sqrt{n}}$ для неизвестной дисперсии, совпадают с соответствующими значениями, вычисленными с помощью функции ДОВЕРИТ();



3) точность доверительного интервала при известной дисперсии выше, чем при неизвестной; это можно объяснить тем, что в случае известной дисперсии имеется больше информации о генеральной совокупности;
4) построенные доверительные интервалы накрывают истинное значение математического ожидания; напомним, что для распределения Вейбулла $W(1,1)$ $EX$=1.

## Задание 5. Проверка статистических гипотез

### Постановка задачи

*Статистической гипотезой* называют предположение о вероятностных закономерностях, которым подчиняется изучаемое случайное явление или процесс. Как правило, статистическая гипотеза – это предположение о виде неизвестного теоретического распределения и его свойствах или о значениях параметров известных распределений.

Гипотеза называется *простой*, если она содержит только одно предположение, т.е. определяет единственное распределение или единственную точку из области возможных значений параметров. *Сложной* называется гипотеза, которая состоит из конечного или бесконечного числа простых гипотез.

Одну из гипотез выделяют в качестве *основной (нулевой)* и обозначают $H_0$, а другую – в качестве *альтернативной (конкурирующей)* и обозначают $H_1$; конкурирующая гипотеза противоречит нулевой.

Выдвинутая основная гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Задача проверки статистических гипотез состоит в том, чтобы на основе выборки $(X_1,...,X_n)$ принять (т.е. считать справедливой) либо основную гипотезу $H_0$, либо конкурирующую $H_1$.

### Общая схема проверки гипотез

Для проверки гипотезы формулируется правило, в соответствии с которым принимается или отклоняется основная гипотеза. Это правило определяется выбором подходящей функции $K = K(X_1,...,X_n)$ от результатов наблюдений, которая служит мерой расхождения между опытными (выборочными) и гипотетическими (теоретическими) значениями. При этом предполагается, что функция $K$, зависящая от выборочных значений, является случайной величиной, распределение которой при правильной нулевой гипотезе точно или приближенно известно. Такая специально подобранная случайная величина $K$ называется *критерием*. Значение критерия, вычисленное по результатам наблюдений, называется *наблюдаемым значением критерия*; обозначим его $K_{набл}$.

Все множество возможных значений критерия $K$ разбивается на две непересекающиеся области: *допустимую* и *критическую.*

*Допустимая область* $D$ – это совокупность значений критерия, при которых нулевая гипотеза принимается. Если наблюдаемое значение критерия $K_{набл}$ попадает в допустимую область, то считается, что результаты наблюдений не противоречат нулевой гипотезе $H_0$, и она принимается.

*Критическая область* $S$ – это совокупность значений критерия, при которых нулевая гипотеза отвергается. Если наблюдаемое значение критерия $K_{набл}$ попадает в критическую область, то расхождение между гипотетическими и опытными данными считается значимым, основная гипотеза $H_0$ отвергается, и принимается конкурирующая гипотеза $H_1$.

*Критическими точками* $k_{кр}$ называются точки, отделяющие критическую область $S$ от допустимой области $D$. Поскольку критерий $K$ - это случайная величина, все его возможные значения принадлежат некоторому интервалу; соответственно, критическая область $S$ и допустимая область $D$ также являются интервалами; границы этих интервалов и называются критическими точками.

Значение критерия зависит от результатов наблюдений, которые являются случайными величинами, поэтому выводы, сделанные при проверке гипотез на основе статистических данных, могут оказаться ошибочными. При этом возможны ошибки двух родов.

*Ошибка первого рода* состоит в том, что отвергается правильная нулевая гипотеза. В этом случае нулевая гипотеза верна, но значение критерия попадет в критическую область, и принимается конкурирующая гипотеза $H_1$; вероятность ошибки первого рода $\alpha$ называется уровнем значимости критерия: $\alpha = P(K \in S|H_0)$.

*Ошибка второго рода* состоит в том, что принимается неправильная нулевая гипотеза. В этом случае нулевая гипотеза неверна, но значение критерия попадет в допустимую область, и нулевая гипотеза $H_0$ принимается; вероятность ошибки второго рода обозначим $\beta$: $\beta = P(K \in D|H_1)$.

*Мощностью критерия* называется вероятность принятия конкурирующей гипотезы $H_1$, если она верна; это происходит в случае, когда при правильной конкурирующей гипотезе значение критерия попадает в критическую область $S$. При этом $P(K \in S|H_1) = 1 - P(K \in D|H_1) = 1 - \beta$ - мощность критерия.

При построении допустимой области желательно было бы выбрать ее границы таким образом, чтобы при проверке гипотез как можно реже происходили ошибки как первого, так и второго рода. Но одновременно минимизировать вероятности $\alpha$ и $\beta$ невозможно, потому что для уменьшения вероятности ошибки первого рода $\alpha$ необходимо расширять границы допустимой области $D$, и, наоборот, для уменьшения вероятности ошибки второго рода $\beta$ (т.е. увеличения мощности критерия) необходимо расширять границы критической области $S$. Поэтому обычно поступают следующим образом: фиксируют уровень значимости $\alpha$, как более важный с практической



точки зрения, а затем при заданном уровне $\alpha$ выбирают критерий, имеющий наибольшую мощность $1-\beta$.

Построение допустимой области заключается в выборе таких критических точек $k_{кр}^1$ и $k_{кр}^2$ (возможно, $k_{кр}^1 = -\infty$ или $k_{кр}^2 = +\infty$), чтобы при заданном уровне значимости $\alpha$ вероятность попадания критерия в допустимую область (т.е. вероятность принять нулевую гипотезу, если она верна) была равна $1-\alpha$: $P\left(k_{кр}^1 < K < k_{кр}^2 | H_0\right) = 1-\alpha$. Очевидно, выбор допустимой области не является однозначным, так как при заданном уровне $\alpha$ и известном распределении критерия $K$ существует сколько угодно интервалов, удовлетворяющих этому условию. Поэтому критические точки по возможности выбираются таким образом, чтобы вероятность ошибки второго рода $\beta = P\left(k_{кр}^1 < K < k_{кр}^2 | H_1\right)$ была минимальной, т.е. выбирается наиболее мощный критерий.

В зависимости от вида конкурирующей гипотезы можно построить правостороннюю, левостороннюю или двустороннюю критическую область.

Если $k_{кр}^1 = -\infty$, то критическая область $S$ – это интервал $[k_{кр};+\infty)$, и его граница находится из уравнения $P(K \geq k_{кр}) = \alpha$; такая критическая область называется правосторонней.

Если $k_{кр}^2 = +\infty$, то критическая область $S$ – это интервал $(-\infty;k_{кр}]$, и его граница находится из уравнения $P(K \leq k_{кр}) = \alpha$; такая критическая область называется левосторонней.

Двусторонняя критическая область – это объединение двух интервалов: $S = (-\infty;k_{кр}^1] \cup [k_{кр}^2;+\infty)$. При этом $P(K \leq k_{кр}^1) + P(K \geq k_{кр}^2) = \alpha$; часто выбирается симметричная критическая область, для которой $P(K \leq k_{кр}^1) = P(K \geq k_{кр}^2) = \dfrac{\alpha}{2}$.

Таким образом, общая схема проверки гипотез сводится к следующим этапам:

1) задается уровень значимости $\alpha$;
2) выбирается критерий для проверки нулевой гипотезы;
3) определяются границы критической области;
4) по выборочным данным вычисляется наблюдаемое значение критерия;
5) если значение $K_{набл}$ попадает в критическую область, то нулевая гипотеза $H_0$ отвергается и принимается конкурирующая гипотеза $H_1$; если $K_{набл}$ попадает в допустимую область, то нулевая гипотеза $H_0$ принимается.

Отметим, что даже если нулевая гипотеза $H_0$ принимается, это не является доказательством того, что она верна, потому что принятие гипотезы происходит на некотором фиксированном уровне надежности и основывается на случайных результатах наблюдений. Принятие гипотезы $H_0$ означает только, что на выбранном уровне надежности эта гипотеза не противоречит полученным выборочным данным.

## Проверка гипотезы о распределении генеральной совокупности. Критерий согласия Пирсона (критерий $\chi^2$)

Иногда под конкурирующей гипотезой подразумевается то, что просто не выполнена основная. В этом случае задача проверки нулевой гипотезы ставится следующим образом: требуется проверить, согласуются ли результаты наблюдений с высказанным предположением. Соответствующие критерии для проверки таких гипотез называются критериями согласия.

Пусть имеется выборка объема $n$ $(X_1,...,X_n)$ из генеральной совокупности с неизвестным теоретическим распределением. По выборочным данным при заданном уровне значимости $\alpha$ требуется проверить предположение, что теоретическое распределение имеет определенный вид $F(x,\theta)$, где $\theta=(\theta_1,...,\theta_k)$ - вектор неизвестных параметров распределения, т.е. гипотеза $H_0$ состоит в том, что функция распределения теоретической случайной величины $X$ равна $F(x,\theta)=P(X<x)$; $H_0: X\sim F(x,\theta)$.

Для проверки данной гипотезы необходимо построить критерий, характеризующий меру расхождения между теоретической и эмпирической функцией распределения. Один из таких критериев был предложен Пирсоном.

Критерий Пирсона основан на сравнении теоретических и эмпирических частот. Для построения критерия множество возможных значений теоретической случайной величины $X$ разбивается на $m$ непересекающихся интервалов: $(-\infty;z_1),[z_1;z_2),...,[z_{m-1};+\infty)$; затем вычисляются теоретические и эмпирические частоты.

Напомним, что эмпирической частотой $n_j$ называется число точек вариационного ряда, попавших в $j$-й интервал $[z_{j-1};z_j)$. Под теоретической частотой понимается математическое ожидание числа наблюдений, которые должны попасть в $j$-й интервал в соответствии с теоретическим распределением $F(x,\theta)$. По *определению 2* случайные величины $X_1,...,X_n$ независимы, и распределение каждой из них совпадает с гипотетическим распределением: $X_i\sim F(x,\theta)$. Обозначим $p_j=P(X_i\in[z_{j-1};z_j))$ - вероятность того, что $i$-е наблюдение попало в $j$-й интервал; поскольку случайные величины $X_1,...,X_n$ независимы и одинаково распределены, эта вероятность не зависит от $i$. Поэтому выборку можно интерпретировать как результат $n$ независимых испытаний, где в каждом испытании успех (попадание наблюдения в $j$-й интервал) происходит с вероятностью $p_j$, т.е. как схему независимых испытаний Бернулли. Напомним, что число успехов в схеме Бернулли имеет биномиальное распределение, а математическое ожидание числа успехов вычисляется как произведение числа испытаний на вероятность успеха в одном испытании. Таким образом, теоретическая частота равна $np_j$. Для вычисления



вероятностей $p_j$ воспользуемся известными свойствами функции распределения:

$p_j=P(X_i\in[z_{j-1};z_j))=F(z_j,\theta)-F(z_{j-1},\theta)$ ; для первого интервала эта вероятность равна $p_1=P(X_i\in(-\infty;z_1))=F(z_1,\theta)$, а для последнего $p_m=P(X_i\in[z_{m-1};+\infty))=1-F(z_{m-1},\theta)$; очевидно, $\sum_{j=1}^{m}p_j=1$.

Поскольку теоретическое распределение зависит от параметров, значения которых неизвестны, то в формулы для вычисления $p_j$ подставляют точечные оценки этих параметров $\hat{\theta}=(\hat{\theta}_1,...,\hat{\theta}_k)$, найденные по результатам наблюдений.

Пирсон показал, что при $n\to\infty$, независимо от закона распределения генеральной совокупности, распределение случайной величины $K=\sum_{j=1}^{m}\frac{(n_j-np_j)^2}{np_j}$ сходится к распределению $\chi^2$ с $m$-$k$-$1$ степенями свободы; $m$ – количество интервалов разбиения, $k$ – число неизвестных параметров гипотетического распределения.

Для проверки гипотезы $H_0$ построим правостороннюю критическую область; критическая точка является решением уравнения $P(K\geq k_{кр})=\alpha$, где случайная величина $K$ имеет $\chi^2$-распределение с $K\sim\chi^2_{m-k-1}$ с $m$-$k$-$1$ степенями свободы.

По выборочным данным найдем наблюдаемое значение критерия $K_{набл}$ и сравним его с $k_{кр}$. Если $K_{набл}$ попадает в критическую область, т.е. $K_{набл}\geq k_{кр}$, то гипотеза $H_0$ отвергается; это означает, что экспериментальные данные не согласуются с выдвинутым предположением о виде распределения генеральной совокупности. Если $K_{набл}$ попадает в допустимую область, т.е. $K_{набл}<k_{кр}$, то гипотеза $H_0$ принимается, т.е. гипотеза о виде теоретического распределения не противоречит результатам наблюдений.

*Замечание 6.* При построении критерия Пирсона рекомендуется выбирать интервалы разбиения таким образом, чтобы для каждого интервала теоретическая частота была не менее 10: $np_j\geq 10$.

## Проверка гипотезы об экспоненциальном распределении генеральной совокупности

В качестве примера проверим гипотезу об экспоненциальном распределении генеральной совокупности на основе выборочных данных, полученных в задании 1.

По *замечанию 2* распределение Вейбулла $W(1,1)$ – это экспоненциальное распределение с параметром 1 (*exp(1)*). Поэтому выборка, смоделированная в задании 1, – это выборка из генеральной совокупности с экспоненциальным теоретическим распределением *exp(1)*. Зададим уровень значимости $\alpha$=0,01 и с

помощью критерия Пирсона проверим, согласуются ли полученные выборочные данные с гипотезой об экспоненциальном распределении генеральной совокупности.

*По замечанию 6* интервалы разбиения при построении критерия следует выбирать таким образом, чтобы $np_j \geq 10$. Поскольку объем выборки $n$=30, то $p_j \geq \frac{1}{3}$; при этом должно выполняться условие $\sum_{j=1}^{m} p_j = 1$, поэтому максимально возможное число интервалов разбиения $m$=3, и для каждого интервала $p_j = \frac{1}{3}$. Найдем точки разбиения $z_1$ и $z_2$, определяющие границы этих интервалов. По формулам для вычисления $p_j$ получаем: $p_1 = F(z_1,\theta) = \frac{1}{3}$; $p_3 = 1 - F(z_2,\theta) = \frac{1}{3}$. Отсюда, учитывая, что гипотетическое распределение $F$ не зависит от неизвестных параметров, получаем, что $F(z_1) = \frac{1}{3}$, $F(z_2) = \frac{2}{3}$. Для вычисления границ интервалов разбиения в Excel можно использовать формулу обратной функции для функции распределения Вейбулла, полученную в задании 1 при моделировании выборки. Если проверяется гипотеза о логнормальном распределении генеральной совокупности, то для вычисления границ интервалов разбиения можно воспользоваться функцией ЛОГНОРМОБР() в Меню/Вставка/Функция/Статистические. Для распределения $W(1,1)$ формула обратной функции имеет вид: $z_i = -\ln(1 - F(z_i))$, отсюда $z_1 = -\ln\left(\frac{2}{3}\right)$, $z_2 = -\ln\left(\frac{1}{3}\right)$. Зная границы интервалов разбиения, можно найти эмпирические частоты, и, учитывая, что $np_j = \frac{1}{3} \cdot 30 = 10$, вычислить наблюдаемое значение критерия $K_{набл} = \sum_{j=1}^{m} \frac{(n_j - np_j)^2}{np_j} = \sum_{j=1}^{3} \frac{(n_j - 10)^2}{10} = \frac{1}{10} \sum_{j=1}^{3} (n_j - 10)^2$.

Для построения правосторонней критической области найдем критическую точку $k_{кр}$. Для этого надо решить уравнение $P(K \geq k_{кр}) = \alpha$, где уровень значимости задан $\alpha$=0,01, а случайная величина $K$ имеет распределение $\chi^2$ с $m$-$k$-$l$=3-0-1=2 степенями свободы: $k$=0, поскольку гипотетическое распределение полностью задано и не зависит от неизвестных параметров. Решение этого уравнения можно найти с помощью Excel, используя функцию ХИ2ОБР($\alpha$,2) в Меню/Вставка/Функция/Статистические.

Результаты вычислений всех перечисленных параметров представлены в *таблице 7*.



*Таблица 7*

| | |
|---|---|
| z1 | 0,4054651 |
| z2 | 1,0986123 |
| n1 | 7 |
| n2 | 11 |
| n3 | 12 |
| Кнабл | 1,4 |
| Ккр | 9,2103404 |

Из представленной таблицы следует, что $K_{набл} < k_{кр}$, т.е. наблюдаемое значение критерия попало в допустимую область, поэтому гипотеза об экспоненциальном распределении генеральной совокупности принимается.

Литература


1. Белько И.В., Свирид Г.П. Теория вероятностей и математическая статистика. Примеры и задачи. – Минск: Новое знание, 2002. - 250 с.
2. Бочаров П.П., Печинкин А.В. Теория вероятностей. Математическая статистика. - М.: Гардарика, 1998. —327 с.
3. Гмурман В.Е. Теория вероятностей и математическая статистика. - М.: Высшая школа, 2000. - 480 с.
4. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. Учебник. - М.: Издательство ЛКИ, 2010.
5. Коваленко И.Н., Филиппова А.А. Теория вероятностей и математическая статистика. - М.: Высшая школа, 1973. - 368 с.
6. Колемаев В.А., Староверов О.В., Турундаевский В.Б. Теория вероятностей и математическая статистика. - М.: Высшая школа, 1991. - 400 с.


Вопросы для подготовки к защите курсовой работы

1. Задачи математической статистики. Генеральная совокупность. Выборка. Вариационный ряд.
2. Эмпирическая функция распределения.
3. Гистограмма и полигон частот.
4. Точечные оценки. Свойства оценок (несмещенность, состоятельность, эффективность).
5. Выборочное среднее. Выборочная дисперсия. Исправленная выборочная дисперсия.
6. Метод моментов.
7. Найти с помощью метода моментов точечные оценки для неизвестного параметра λ распределения Вейбулла $W(r,\lambda)$, считая параметр $r$ известным; неизвестного параметра $\mu$ логнормального распределения $LogN(\mu,\sigma)$, считая параметр $\sigma$ известным.
8. Метод максимального правдоподобия. Функция правдоподобия.

9. Найти с помощью метода максимального правдоподобия точечные оценки для неизвестного параметра $\lambda$ распределения Вейбулла $W(r,\lambda)$, считая параметр $r$ известным; неизвестного параметра $\mu$ логнормального распределения $LogN(\mu,\sigma)$, считая параметр $\sigma$ известным.
10. Доверительные интервалы. Точность и надежность доверительных интервалов.
11. Доверительный интервал для неизвестного математического ожидания при известной дисперсии (нормальное распределение).
12. Доверительный интервал для неизвестного математического ожидания при известной и неизвестной дисперсии (произвольное распределение).
13. Статистические гипотезы. Критерии. Ошибки первого и второго рода. Уровень значимости. Схема проверки статистических гипотез.
14. Критерий согласия Пирсона ($\chi^2$). Проверка гипотезы о виде распределения генеральной совокупности.

## Требования к оформлению курсовой работы

Отчет по курсовой работе сдается на проверку в распечатанном виде на листах формата А4.

Для каждого задания должны быть изложены необходимые теоретические сведения: приведены определения основных понятий и формулировки теорем, сформулированы постановки задач и указаны методы их решения. Формулы для обратной функции (распределение Вейбулла), точечных оценок неизвестных параметров распределений и для вычисления границ интервалов разбиения при построении критерия Пирсона должны быть приведены с выводом и необходимыми комментариями. Результаты вычислений в Excel должны быть представлены в отчете в виде таблиц, аналогичных *таблицам 2-7* в данном учебно-методическом пособии.

Все формулы и результаты вычислений для каждого задания должны быть сохранены на USB-флеш-накопителе в формате Excel (файл с расширением .xls или .xlsx).