Министерство образования и науки Российской Федерации

Федеральное государственное автономное
образовательное учреждение высшего образования
«ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

А. И. Луценко

Математическая статистика

Ростов-на-Дону
Издательство Южного федерального университета
2014

УДК 519.2
ББК 22.172
Л

Печатается по решению Редакционно-издательского совета
Южного федерального университета

Р е ц е н з е н т ы :
Доктор технических наук, доцент кафедры «Строительство и техносферная безопасность» Институт сферы обслуживания и предпринимательства (филиал ФГБОУ ВПО ДГТУ, г. Шахты) Елисеев И.Н.

Кандидат физико-математических наук, доцент кафедры «Прикладной математики и вычислительной техники» ФГБОУ ВПО РГСУ Бычков А.Б.

Издание осуществлено в рамках выполнения проекта «Интернационализация учебных планов на уровне магистра в российских вузах Южного региона» программы Tempus-IV

Луценко А. И.
Математическая статистика.---- Ростов н/Д: Изд-во ЮФУ, 157с., ил.
ISBN

Книга написана на основе лекций, которые в течение многих лет читались автором в VII семестре студентам специальности «прикладная математика и информатика» института математики, механики и компьютерных наук ЮФУ.

При написании книги ставились две цели. Первая цель: дать возможность математику, интересующемуся статистикой, овладеть основными методами математической статистики необходимыми при экспериментальной работе. Вторая цель: дать возможность статистику, интересующемуся математикой, ознакомиться с теоретической основой построения математической статистики.

Рассмотрены математические основы выборочного метода, теории точечных и интервальных оценок, проверки статистических гипотез, корреляционного и дисперсионного анализов.

ISBN

УДК 519.2
ББК 22.172

Предисловие

Одним из недостатков нашей системы образования является недостаточная мотивация образования. Начиная изучать новый курс, студент нередко задаёт себе вопросы: «Зачем мне это надо?», « Пригодятся ли мне эти знания в моей будущей научной или практической работе?». Ясно, что получить однозначный ответ на эти вопросы – нельзя. Но, желая возбудить интерес к излагаемому курсу, преподаватель стремится реализовать две цели. Первая – расширить теоретический кругозор будущего специалиста, повысить его уровень научного мышления, что в будущем позволит ему легче вникнуть в суть возникающей проблемы и выбрать теоретическую модель аналогичную по содержанию с возникшей практической задачей, которая поможет решить эту задачу. Вторая – познакомить будущего специалиста с методологией последовательного движения мыслительного процесса от изучении простых моделей до решения сложных теоретических и практических задач, с правилами выбора пути решения возникшей задачи. Выбор пути решения возникшей задачи, основывается на умении увидеть аналогию в постановке этой задачи с формулировками и путями решения других, казалось бы, на первый взгляд непохожих, теоретических моделей и задач.

У молодого специалиста, аспиранта, научного работника, занимающегося научным исследованием, проведением эксперимента, наблюдений за интересующим его объектом исследования через некоторое время скапливается более или менее значительный объём данных. Часто возникает вопрос: что делать с полученными данными, как их обрабатывать, как формулировать и истолковывать получающиеся результаты? Тем более что курс высшей математики с основами теории вероятностей и математической статистики прослушивался и сдавался в форме экзамена два-три года назад.

Надо ясно представлять себе, что квалифицированное применение методов математической статистики требует привлечения специалистов в этой области. Но, в то же время, мы считаем, что лучший способ для создания представления о математической статистике за ограниченное время – подробно разъяснить теоретические основы построения математической статистики и проиллюстрировать некоторые частные приёмы решения задач методами математической статистики.

В пособии большое внимание уделяется вопросам теоретической постановки задач математической статистики и вопросам формулирования выводов по результатам статистической обработки экспериментального материала.

Столь подробное изложение вопросов теоретической постановки задачи, хода её выполнения и, особенно, формулирования результата статистической проверки основной гипотезы вызвано тем, что

практический работник, в виду недостаточной теоретической подготовки по теории вероятностей и математической статистики, решая какую-то проблему, задачу по своей специальности, не знает возможностей, которые дают методы математической статистики.

Большинству существующих учебных пособий по математической статистике для студентов и практических работников нематематических специальностей присущ один недостаток – это рецептурное изложение материала. Теоретический материал излагается в слишком сжатой форме, недостаточное внимание уделяется вопросам теоретического содержания задач статистической обработки материалов исследований, и, что особенно важно для практического работника или специалиста, приёмам правильного и убедительного для оппонентов формулирования выводов из результатов статистической обработки. Вместе с этим в этих пособиях много внимания уделяется изложению технической стороны вопроса и излишне подробно описывается вычислительная процедура статистической обработки экспериментального материала.

В других учебных пособиях посвящённых математической статистике в основном уделяется внимание теоретической стороне построения курса и мало отводится места практической иллюстрации теоретических построений, что затрудняет чтение материала практику-статистику.

Всё это является одним из основных препятствий к более широкому практическому применению методов математической статистики при обработке результатов экспериментальных или статистических наблюдений.

В данной книге выбраны для изучения приёмы обработки результатов наблюдений и обоснование этих приёмов для наиболее часто встречающихся в практике задач. Закон больших чисел, по существу, утверждает, что различие между средними значениями величин, наблюдаемыми в выборке, и истинным средним значением по всей совокупности будет уменьшаться при увеличении объёма выборки. Центральная предельная теорема объясняет использование нормального закона при построении параметрических критериев статистической проверки гипотез.

Трудности вычислительного характера отходят на второй план, и становится вполне доступным использование весьма трудоёмких методов статистического анализа. Однако останутся трудности понимания существа используемых методов, выбора наиболее подходящей математической модели и правильной интерпретации полученных результатов.

В предлагаемом пособии изложение материала ведется тремя параллельными путями.

В пунктах «**теория**» излагается по возможности строгое теоретическое построение курса математической статистики. Здесь преследуется цель

повышения уровня теоретической подготовки будущего специалиста, что позволит ему более глубоко, на более высоком уровне осмысливать возникающие в практической деятельности задачи.

В пункте *«**комментарий**»* обсуждается методология постановки цели исследования методами математической статистики. Описывается варианты правильной, чёткой формулировки задачи и объяснения получившегося ответа.

В пункте «**практика**» приводятся конкретные примеры постановки проблемы, пути её решения и формулирования получающегося результата. Рассматриваемые модели преследуют в основном одну цель: увеличить «библиотеку» вариантов задач и способов их решения, что в будущем позволит специалисту легче и быстрее найти аналогию с возникшей перед ним задачей.

Первый пункт «**теория»** присутствует в любом конспекте лекций. Студент во время лекции практически не успевает записать идеи, связывающие отдельные теоретические построения, обсуждения результатов, которые приводятся в «***комментарии***». А спустя некоторое время, при подготовке к экзамену, забываются логические связи записанных в «**теории»** теоретических положений и построений. Всё это затрудняет изучение курса. Процесс подготовки к экзамену сводится к механическому запоминанию определений, теорем и формул. Всё это механически, без логического обоснования, заученное быстро забывается. И порой через год студент вспоминает лишь название курса, но не может чётко описать цель и методологию этого курса. Как тогда такими знаниями он будет пользоваться в практической работе?

В пункте «**практика**», на который обычно при подготовке к экзамену студент не обращает внимания, приводятся типичные практические задачи. Они могут помочь специалисту-статистику увидеть аналогию между этими задачами и с возникшей перед ним задачей и найти способ её решения.

Данная книга рассчитана на математически образованного естествоиспытателя, который не собирается стать специалистом по математической статистике. Однако ему необходимо чётко представлять себе, в каких случаях эта наука может принести существенную пользу, и уметь воспользоваться помощью специалиста. Лучший способ для создания такого представления за ограниченное время – подробное изложение некоторых частных приёмов математической статистики, доведя их до возможности приложений, но оставив в стороне другие приёмы за недостатком времени. Сама структура математической статистики позволяет такое исчерпывающее изучение одних приёмов за счёт полного игнорирования других, так как в этой науке сравнительно небольшую роль играют общие математические теоремы, а отдельные приёмы часто логически независимы друг от друга.

При написании книги ставились две цели. Первая цель: дать возможность математику, интересующемуся статистикой, овладеть основными методами математической статистики необходимыми при экспериментальной работе. Вторая цель: дать возможность статистику, интересующемуся математикой, ознакомиться с теоретической основой построения математической статистики.

Рассмотрены математические основы выборочного метода, теории точечных и интервальных оценок, проверки статистических гипотез, корреляционного и дисперсионного анализов.

Хочу выразить искреннюю благодарность и глубокую признательность директору института математики, механики и компьютерных наук им. И.И.Воровича, доктору физико-математических наук Карякину Михаилу Игоревичу и заведующему кафедрой дифференциальных и интегральных уравнений, доктору физико-математических наук Авсянкину Олегу Геннадьевичу за стимул к написанию данного учебника.

«Измеряй всё доступное измерению
и делай недоступное измерению доступным»
Галилео Галилей

Модуль первый
Первичная обработка статистических данных
§1.1 Основные понятия и определения математической статистики
Комментарий

Основными понятиями, с которых начинается знакомство с математической статистикой, являются понятия: «генеральная совокупность» и «выборка». Практическая направленность учебного пособия предполагает более упрощённое введение этих понятий. Однако, при достаточном уровне математической подготовки, естествоиспытателю при рассмотрении практических задач необходимо понимать суть постановки задач и способов их решения, которые решаются методами математической статистики. Упрощённо описав эти понятия, дадим теоретическое обоснование их, базирующееся на основных понятиях теории вероятностей.

Теория

Рассматриваются совокупность однородных объектов. Изучению подлежит некоторый количественный или качественный признак, характеризующий эти объекты и объединяющий их в эту рассматриваемую совокупность. Эта совокупность может состоять из бесконечного числа элементов. Результатом изучения этой совокупности будет набор данных, то есть числовых значений признака у каждого объекта, являющегося элементом этой совокупности. Всё множество этих числовых значений будем называть ***генеральной совокупностью***. Практически получение и изучение данных измерений количественной характеристики у всех элементов совокупности иногда просто невозможно, если их множество бесконечно, а иногда технически или экономически сильно затруднено. Поэтому приходится прибегать к ограниченному числу наблюдений. Случайным образом отбирается n объектов и у них измеряется значение интересующего признака. Полученное множество, состоящее из n числовых значений, называется ***выборкой из генеральной совокупности.***

Ясно, что оба типа совокупностей – и генеральная, и выборочная будут характеризоваться одинаковыми общими закономерностями проявления наиболее характерных особенностей количественного признака. Дадим объяснение сказанному, опираясь на теорию вероятностей. Это поможет нам понять точный смысл основных понятий и получить теоретическое обоснование приёмов и методов математической статистики.

Рассматривается вероятностное пространство $\langle \Omega, \mathcal{A}, P \rangle$, определение которого даётся в аксиоматическом построении А.Н.Колмогорова. Элементы множества Ω - изучаемые однородные объекты. На этом вероятностном пространстве определяется измеримая функция ξ, называемая случайной величиной, которая отображает множество Ω в некоторое подмножество Ω_R множества действительных чисел R. Это множество действительных чисел Ω_R, то есть множество возможных значений ξ, будем называть ***генеральной совокупностью***. Множество Ω_R может быть бесконечным, может совпадать с множеством R. Вероятностная функция P есть распределение вероятностей случайной величины ξ.

Пусть $\mathcal{G}$ случайный эксперимент, заключающийся в том, что мы выбираем случайно, наугад какой-нибудь объект - элемент из множества Ω. Регистрируем значение его количественной характеристики, то есть, получаем число x – одно из возможных значений случайной величины ξ, $x \in \Omega_R$. Значит $\{\xi = x\}$ это случайное событие, одноточечный элемент алгебры событий $\mathcal{B}$ – σ -алгебры борелевских множеств действительных чисел. Если этот эксперимент повторить n раз, то можем сказать, что мы рассматриваем заданные на одном и том же вероятностном пространстве $\langle \Omega, \mathcal{A}, P \rangle$ случайные величины $\xi_1, \xi_2, ..., \xi_n$ и при этом наблюдаем n штук событий: $\{\xi_1 = x_1\}$, $\{\xi_2 = x_2\}$,..., $\{\xi_n = x_n\}$.

Случайные величины $\xi_1, \xi_2, ..., \xi_n$ назовём *выборочными случайными величинами*, а множество их значений $(x_1, x_2, ..., x_n)$ - ***выборкой*** из генеральной совокупности Ω. Ясно, что любой элемент x_i выборки $(x_1, x_2, ..., x_n)$ будет элементом множества Ω_R. Число элементов выборки n называется ***объёмом выборки***.

Комментарий

Мы не будем касаться технических вопросов связанных с процессом получения или процессом формирования выборки $(x_1, x_2, ..., x_n)$. Ясно, что все элементы Ω_R должны иметь равные возможности попасть в выборку $(x_1, x_2, ..., x_n)$. Не должно быть оснований предполагать, что среди элементов выборки есть «привилегированные» элементы. Для этого мы должны требовать, чтобы:

а) все эксперименты $\mathcal{G}$, а, следовательно, и их результаты, были независимыми друг от друга;

б) во всех n экспериментах $\mathcal{G}$ условия их проведения оставались одинаковыми;

в) во всех n экспериментах G измерялся один и тот же количественный признак по постоянной, одной и той же методике.

Всё это подразумевает, чтобы, по возможности при минимальных финансовых и трудовых затратах, выборка была по возможности наиболее информативной. Если мы называем набор чисел $(x_1, x_2, ..., x_n)$ *выборкой*, то мы подразумеваем, что все перечисленные требования соблюдались при организации выборки, при сборе экспериментальных данных. Такую выборку мы будем называть – ***репрезентативной*** (representative – представительная). И впредь, строя теоретические модели математической статистики, рассматривая примеры, иллюстрирующие теоретические модели, мы всегда будем считать, что выборка, с которой мы работаем,- *репрезентативная*.

Необходимо заметить, что не существует исчерпывающего способа проверки того, что данные измерений $x_1, x_2, ..., x_n$ можно считать репрезентативной выборкой $(x_1, x_2, ..., x_n)$. То есть, нет надёжного способа оценки репрезентативности выборки. При этом мы должны иметь в виду, что человек – удивительно плохой инструмент для сознательного осуществления случайной, представительной выборки. На проблемы организации случайной представительной выборки стал обращать внимание ещё в 20-е годы прошлого столетия английский математик и биолог Р.А.Фишер.

Мы должны знать, что уверенность в правильности соответствующих методов работы при организации выборки приобретается только в процессе накопления опыта.

Практика

Пример 1.1 [13]

На стол положили 1200 камней различных размеров и массы. Было предложено 12 лицам (каждому трижды) отобрать 20 камней «типичных» для этой совокупности. После каждого испытания камни возвращались обратно и перемешивались, так что все наблюдатели, и каждый из них лично, при повторном отборе находились в одинаковых условиях. Результаты эксперимента приведены в таблице.

повторность	экспериментатор											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	1,9	2,4	2,4	1,9	2,2	2,8	2,4	1,6	2,2	2,6	2,4	2,4
2	1,8	3,0	2,4	2,0	2,7	2,6	2,6	2,0	2,2	2,2	2,4	3,0
3	1,7	2,4	2,1	2,0	3,1	2,8	2,5	2,0	2,2	3,1	1,8	2,4
среднее, $\overline{x}_i$	1,8	2,6	2,3	2,0	2,7	2,7	2,5	1,9	2,2	2,6	2,2	2,6

Рассмотрим этот эксперимент с позиций сформулированных выше понятий генеральной совокупности и выборки.

Исследуемым количественным признаком будет масса отбираемых в каждом повторном опыте 20 камней. Отбирая 20 «типичных» камней из генеральной совокупности, состоящей из 1200 камней, каждый из 12 участников эксперимента стремится, чтобы пропорции количеств мелких, средних и крупных камней в его выборке были примерно равны пропорциям количеств мелких, средних и крупных камней в генеральной совокупности. Каждый участник эксперимента делает три отбора по 20 камней. Значит, в результате мы будем иметь 36 выборок, каждая из которых, по мнению создавшего её экспериментатора, обладает характерными особенностями генеральной совокупности.

В качестве меры «предпочтений» экспериментаторов при формировании каждой выборки возьмём (выберем) средний вес камня в выборке.

Средний вес одного камня в трёх выборках по 20 штук у i – того экспериментатора обозначим $\bar{x}_i$, $i = 1,2,...,12$. Последняя строка таблицы показывает диапазон изменения усреднённых «личных предпочтений» экспериментаторов: от 1,8 до 2,7. Средний вес одного камня в 36 выборках оказался равным $\bar{x} = 2,34$. В то время как средний вес одного камня в генеральной совокупности был равен $\bar{\xi} = 1,91$. Разница между реальным средним весом камня и средним весом камня, на который оказали субъективные предпочтения экспериментаторов, составила более 22,5%. Рассмотренный пример наглядно иллюстрирует возможность нарушения репрезентативности выборки, вызванную субъективизмом экспериментатора.

Наиболее надёжным инструментом при отборе из генеральной совокупности в выборку n элементов являются датчик или таблица случайных чисел.

§2.1 Эмпирическая случайная величина. Вариационный ряд. Гистограмма

Теория

Выборка $(x_1, x_2, ..., x_n)$ будет рассматриваться нами как множество возможных значений случайной величины дискретного типа, которую будем называть *эмпирической случайной величиной* и обозначать ξ^*. Случайный выбор элементов $x_1, x_2, ..., x_n$ из множества Ω_R производится со стремлением обеспечить наибольшую репрезентативность. Это означает, что все элементы генеральной совокупности имеют равные возможности попасть в выборку. То есть, для любого элемента x_i выборки, состоящей из n элементов, его вероятность $P(\xi^* = x_i) = \frac{1}{n}$. Следовательно, вероятность

любого борелевского множества B, которое является случайным событием, согласно классическому определению вероятности, будет равна $P^*(B)=\frac{\mu(B)}{n}$, где $\mu(B)$ - число (количество) элементов выборки $(x_1, x_2, ..., x_n)$, попавших во множество B.

Ясно, что функция распределения эмпирической случайной величины ξ^*, определяемой выборкой объёма n, будет иметь вид $F_n^*(x)=\frac{\mu(S_x)}{n}$, где $S_x=(-\infty; x)$ - борелевское множество.

Комментарий

Интуитивно мы понимаем, что с увеличением объёма выборки n различие между функцией распределения $F_n^*(x)$ случайной величины ξ^* и функцией распределения $F(x)$ случайной величины ξ будет уменьшаться. Но так как мы осуществляем случайный отбор элементов в выборку и состав каждой выборки случаен, то нельзя утверждать, что если $n_2 > n_1$, то различие между функциями $F_{n_2}^*(x)$ и $F(x)$ обязательно будет меньше различия между функциями $F_{n_1}^*(x)$ и $F(x)$. При этом остаётся неясным: как оценивать величину различия между функциями распределения? Полный ответ на этот вопрос и чёткое объяснение процесса «сближения» эмпирической функции распределения $F_n^*(x)$ с функцией распределения $F(x)$ даёт теорема Гливенко.

Теория

Теорема Гливенко:

Пусть числа $x_1, x_2, ..., x_n$ образуют выборку из распределения, имеющего теоретическую функцию распределения $F(x)$. Тогда при любом x $-\infty < x < +\infty$, будет справедливо $\lim\limits_{n\to\infty} P\left(\left|F_n^*(x)-F(x)\right|<\varepsilon\right)=1$.

То есть, при неограниченном увеличении объёма выборки n, для любого фиксированного значения x, эмпирическая функция распределения сходится по вероятности к теоретической функции распределения.

Пусть при выбранном фиксированном значении аргумента x функция $F(x)$ принимает значение p, $0 \le p \le 1$, то есть p – вероятность события $\{\xi \in S_x\}$, коротко: $p = P(\xi \in S_x)$. При этом значении x функция $F_n^*(x)$ принимает значение $\frac{m}{n}$, где m – число элементов выборки $(x_1, x_2, ..., x_n)$, которые оказались меньше чем x, то есть $\mu(S_x)=m$. Но тогда мы можем

сказать, что если p – вероятность события $\{\xi \in S_x\}$, то $\frac{m}{n}$ - относительная частота этого события. Согласно теореме Я.Бернулли из Закона больших чисел относительная частота наступлений некоторого события, при неограниченно возрастающем числе проведения испытаний, сходится по вероятности к вероятности наступления этого события в единичном испытании, то есть $\lim\limits_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1$. Заменяя значение относительной частоты события $\{\xi \in S_x\}$ и значение вероятности этого события на значения функций распределения $F_n^*(x)$ и $F(x)$, получим утверждение теоремы.

Комментарий

Рассмотренная теорема имеет принципиальное значение для всего курса математической статистики. Мы теперь понимаем, что неизвестное распределение вероятностей наблюдаемого или исследуемого количественного признака, то есть – распределение случайной величины ξ, с большой степенью уверенности может быть сколько угодно точно восстановлено по выборке достаточно большого объёма. Значит, по выборке мы можем составить мнение о распределении вероятностей случайной величины ξ, о значениях параметров функции распределения, о значениях числовых характеристик этой случайной величины.

Теория

Так как выборка имеет конечное число элементов, то мы всегда можем найти минимальный x_{min} и максимальный x_{max} элементы этой выборки. Величина разности $x_{max} - x_{min}$, то есть диапазон изменения значений элементов выборки, называется *размах выборки*. Эту величину делим на выбранное число k, то есть диапазон изменения значений элементов выборки делим на k интервалов и обозначим $\delta = \frac{x_{max} - x_{min}}{k}$ - длину каждого интервала. Записываем границы интервалов: $a_1 \le x_{min}$, $a_2 = a_1 + \delta$, $a_3 = a_2 + \delta, \ldots$, $a_{j+1} = a_j + \delta, \ldots$, $a_{k+1} = a_k + \delta$. Получаем k интервалов: $(a_1; a_2], (a_2; a_3], \ldots, (a_j; a_{j+1}], \ldots, (a_k; a_{k+1}]$. Эти полузакрытые справа интервалы будут случайными событиями $\{B_j\}, j = 1, \ldots, k$. Подсчитываем значения частот $\{m_j\}$, где $m_j = \mu(B_j)$ - число элементов выборки $(x_1, x_2, \ldots, x_n)$, которые попали в интервал $(a_j; a_{j+1}]$. Результаты записываем в таблицу, которая называется *частотный вариационный ряд*.

$(a_1;a_2]$	$(a_2;a_3]$	…	$(a_j;a_{j+1}]$	…	$(a_k;a_{k+1}]$
m_1	m_2	…	m_j	…	m_k

Ясно, что всегда будет справедливо $\sum_{j=1}^{k} m_j = n$. Если во второй строке таблицы записать значения относительных частот, то получаемая таблица:

$(a_1;a_2]$	$(a_2;a_3]$	…	$(a_j;a_{j+1}]$	…	$(a_k;a_{k+1}]$
$\frac{m_1}{n}$	$\frac{m_2}{n}$	…	$\frac{m_j}{n}$	…	$\frac{m_k}{n}$

будет называться *вариационным рядом относительных частот.*

Практика

Рассмотрим два набора статистических данных, записанные в разных формах. Статистические данные, которые мы понимаем как *репрезентативные выборки*, являются значениями одного и того же количественного признака, случайной величины ξ, - рост мужчин в *см.* Эти данные взяты из разных источников и были получены разными исследователями, поэтому мы не будем здесь обсуждать вопросы о генеральных совокупностях, с которыми работали исследователи. Хотя кратко, в довольно общей форме, мы можем сказать, что здесь генеральная совокупность – «мужчины, проживающие в некотором регионе, в некоторой области, некотором государстве, и т.п.».

Пример 2*а*.1 [11] В результате обследования некоторой группы мужчин (115 чел.) получены следующие данные о росте (*см*):

162 151 161 170 167 164 166 164 173 172 165 153 164 169 170 154 163 159
161167 168 164 170 166 176 157 159 158 160 161 167 155 166 167 173 165
175 165 174 167 170 169 159 159 160 156 161 162 161 181 159 169 160 169
161 161 166 164 170 180 158 167 169 165 166 172 168 171 178 178 171 165
161 162 182 164 171 169 176 177 170 169 171 160 165 165 179 161 178 173
168 171 163 165 166 166 166 169 167 166 167 172 169 171 168 162 165 168
171 174 165 168 167 170 170

Пример 2*б*.1 [13] В результате медицинского обследования 906 призывников получены следующие данные их роста (*см*):

x_i *(см)*	частота (m_i)	x_i *(см)*	частота (m_i)	x_i *(см)*	частота (m_i)	x_i *(см)*	частота (m_i)	x_i *(см)*	частота (m_i)
147	1	155	6	163	48	170	47	177	13

148	0	156	12	164	47	171	48	178	9
149	0	157	14	165	60	172	36	179	9
150	2	158	25	166	63	173	31	180	3
151	4	159	22	167	74	174	36	181	3
152	3	160	30	168	60	175	21	182	4
153	4	161	35	169	64	176	24	183	1
154	7	162	43						

Разбивая диапазоны изменений значений на одинаковое число интервалов $k=6$, в каждом интервале возьмём по пять последовательных возможных значений элементов этих выборок. При этом частоты наиболее редких минимальных и максимальных значений x_i включим, соответственно в первый и последний интервалы.

Подсчитав частоты m_j попаданий элементов выборок в каждый интервал $(a_j ; a_{j+1}]$, $j = 1,2,3,4,5,6$, построим, соответственно, два вариационных ряда относительных частот.

Пример.2*а*.1

$(150;155]$	$(155;160]$	$(160;165]$	$(165;170]$	$(170;175]$	$(175;180]$
$\frac{4}{115} = 0,035$	$\frac{13}{115} = 0,113$	$\frac{31}{115} = 0,270$	$\frac{41}{115} = 0,356$	$\frac{16}{115} = 0,139$	$\frac{10}{115} = 0,087$

Пример.2*б*.1

$(145;155]$	$(155;160]$	$(160;165]$	$(165;170]$	$(170;175]$	$(175;185]$
$\frac{27}{906} = 0,030$	$\frac{103}{906} = 0,114$	$\frac{233}{906} = 0,257$	$\frac{308}{906} = 0,340$	$\frac{169}{906} = 0,186$	$\frac{66}{906} = 0,073$

Даже беглый, без анализа, взгляд на значения относительных частот, вычисленных по разным выборкам, для одних и тех же интервалов приводит нас к заведомо ожидаемому выводу, что теоретические распределения случайных величин, из генеральных совокупностей которых взяты выборки, - одинаковые. Хотя объявление подобных выводов без математически обоснованного анализа получаемых результатов неубедительно и может привести к ошибкам.

Теория

Построенному вариационному ряду относительных частот можно построить геометрическую иллюстрацию, которая наглядно будет показывать характер изменения значений относительных частот в области значений элементов выборки. Для этого на оси абсцисс отмечаются

интервалы разбиения $(a_j ; a_{j+1}]$ ($j=1,2,\ldots,k$) множества возможных значений выборки. Эти интервалы будут основаниями прямоугольников, высоты которых равны $\frac{m_j}{n \cdot \delta}$, соответственно. Получающаяся при этом ступенчатая фигура называется ***гистограммой*** выборки.

Комментарий

Гистограммы, построенные по вариационным рядам примеров 2*а* и 2*б*, приведены на рис.1 и рис.2. Площадь S_j каждого прямоугольника гистограммы равна значению относительной частоты $S_j = \frac{m_j}{n \cdot \delta} \cdot \delta = \frac{m_j}{n}$ попадания элементов выборки в интервал $(a_j ; a_{j+1}]$. Следовательно, общая площадь всех k прямоугольников всегда будет равна единице. Если $F(x)$ – функция распределения вероятностей случайной величины ξ, то вероятность события $\{\xi \in (a_j ; a_{j+1}]\}$ будет равна: $P(a_j < \xi \le a_{j+1}) = F(a_{j+1}) - F(a_j) = p_j$. В частности, если случайная величина ξ непрерывного типа, то вероятность p_j мы обычно интерпретируем как площадь криволинейной трапеции, границы которой определяются интервалом $(a_j ; a_{j+1}]$ и графиком плотности вероятности $p(x)$, то есть: $p_j = \int_{a_j}^{a_{j+1}} p(x)dx$.

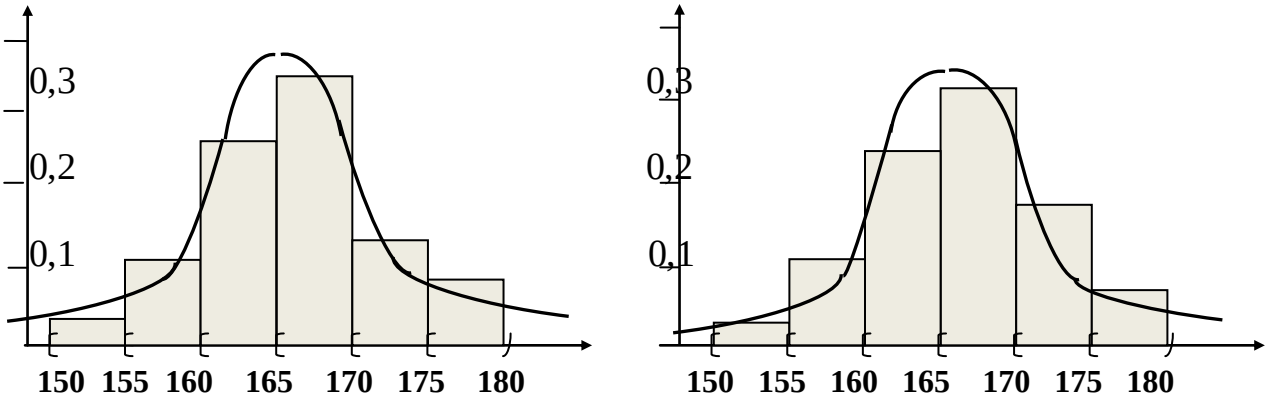


Рис.1.1

Относительная частота $\frac{m_j}{n}$ при неограниченном увеличении объёма выборки n сходится по вероятности к вероятности p_j. То есть мы можем

сказать, гистограмма является *статистической моделью* графика плотности вероятности изучаемой (исследуемой) случайной величины ξ.

Число k - число интервалов вариационного ряда, практически определяется из условия, чтобы при малых объёмах выборки ($n<100$) минимальное значение частот наблюдений m_j элементов выборки, попавших в интервал $(a_j; a_{j+1}]$, было примерно равным $3\sim5$, то есть должно быть: $\min m_j=3\sim5$. Ясно, что при увеличении объёма выборки n минимальное x_{min} и максимальное x_{max} значения элементов выборки могут не измениться, а если они и изменятся, то эти изменения будут незначительными. Это означает, что при увеличении объёма выборки размах выборки $x_{max} - x_{min}$ или не изменяется, или изменяется незначительно. Это, исходя из условия $\min m_j=3\sim5$, позволяет увеличить число интервалов k, а тогда длины δ интервалов $(a_j; a_{j+1}]$ будут уменьшаться. Одновременно длины и вертикальных, и горизонтальных отрезков верхней части контура гистограммы будут уменьшаться. Этот верхний контур будет статистическим аналогом (статистической моделью) графика теоретической плотности вероятности. Значит, по виду гистограммы мы сможем сделать предварительное суждение о законе распределения вероятностей случайной величины ξ.

§3.1. Статистические оценки числовых характеристик случайных величин

Комментарий

Зная закон распределения вероятностей случайной величины, например, в виде $F(x)$ - её функции распределения, можем получить любую информацию о случайной величине ξ. Интегрированную информацию, нивелирующую незначительные характерные особенности, о случайной величине ξ дают нам числовые характеристики случайной величины. Числовыми характеристиками случайной величины являются: математическое ожидание $M\xi$, дисперсия $D\xi$, начальные $\alpha_k = M\xi^k$ и центральные моменты $\mu_k = M[\xi - M\xi]^k$, медиана распределения, среднее квадратическое отклонение, коэффициенты асимметрии и эксцесса и т.д.

Вариационный ряд и гистограмма позволяют нам составить представление о теоретическом законе распределения вероятностей случайной величины ξ. Поэтому мы называем их статистическими моделями теоретического закона.

Точных значений числовых характеристик мы не знаем. Наша задача, используя элементы выборки $(x_1, x_2, ..., x_n)$, получить представление о

значениях числовых характеристик случайной величины ξ и оценить точность этих представлений.

Теория

Определим значения числовых характеристик эмпирической случайной величины ξ^*. Так как ξ^* - случайная величина дискретного типа, у которой множеством возможных значений является выборка $(x_1, x_2, ..., x_n)$ и все её значения равновозможные - $P(\xi^* = x_i) = \frac{1}{n}$, то математическое ожидание ξ^* будет равно: $M\xi^* = \sum_{i=1}^{n} x_i \cdot P(\xi^* = x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$. То есть, математическое ожидание ξ^* равно среднему арифметическому элементов выборки $\bar{x}$.

Рассуждая аналогично, получим, что дисперсия ξ^* будет равна: $D\xi^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S^2$. Начальные и центральные моменты k-того порядка случайной величины ξ^* обозначим их: a_k и m_k, будут равны:

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad \text{и} \quad m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k .$$

Любую числовую характеристику случайной величины ξ будем обозначать ϑ (или θ). Аналогично, любую числовую характеристику эмпирической случайной величины ξ^* будем обозначать ϑ^* (или θ^*). Если распределение вероятностей эмпирической случайной величины ξ^* мы можем принимать как статистическую модель неизвестного теоретического распределения случайной величины ξ, то мы предлагаем принять числовые характеристики эмпирической случайной величины ξ^* в качестве оценок неизвестных значений соответствующих числовых характеристик случайной величины ξ. Например, неизвестное значение m математического ожидания $M\xi$ мы будем оценивать значением $\bar{x}$ - значением среднего арифметического элементов выборки, неизвестное значение σ^2 дисперсии $D\xi$ мы будем оценивать значением S^2 - значением дисперсии элементов выборки.

В общем случае, если ϑ - числовая характеристика случайной величины ξ, то соответствующую эмпирическую числовую характеристику ϑ^* будем называть ***точечной оценкой*** числовой характеристики ϑ.

«Если Эксперимент – Король наук,
то Статистические методы – его Телохранитель»
М. Трайбус

Модуль второй

Точечные и интервальные оценки числовых характеристик случайных величин

§1.2. Требования к точечным оценкам. Теория

Формируя выборку, мы проводим n раз G случайный эксперимент. Это означает, что мы наблюдаем n случайных величин $\xi_1,\xi_2,...,\xi_n$, которые мы назвали *выборочными случайными величинами.* Распределением каждой случайной величины ξ_i, i=1,2,…,n, является распределение исследуемой случайной величины ξ. Можно сказать, что выборочные случайные величины это независимые, одинаково распределённые компоненты n-мерного вектора. Каждая реализация проведённого n раз случайного эксперимента G, есть набор чисел $(x_1,x_2,...,x_n)$ - называемый выборкой. Вычисляя значение среднего арифметического $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, мы определяем по имеющейся выборке значение $\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$ - функции случайных величин $\xi_1,\xi_2,...,\xi_n$. То есть $\bar{\xi}$ является случайной величиной.

Определение. Любую функцию $\vartheta^*=g(\xi_1,\xi_2,...,\xi_n)$ выборочных случайных величин будем называть ***статистикой***.

Следовательно, приведённые выше точечные оценки ϑ^* числовых характеристик ϑ являются значениями статистик – значениями функций выборочных случайных величин $\xi_1,\xi_2,...,\xi_n$.

Комментарий

Наша задача состоит в следующем: желая оценить неизвестное значение числовой характеристики ϑ, мы по выборочным случайным величинам $\xi_1,\xi_2,..\xi_n$ задаём некоторую статистику $\vartheta^*=g(\xi_1,\xi_2,...,\xi_n)$. Значение этой статистики, получаемое по числовым элементам имеющейся у нас выборки $(x_1,x_2,...,x_n)$, принимается нами как оценка значения теоретической числовой характеристики ϑ.

Но статистик - функций выборочных случайных величин для рассматриваемой числовой характеристики ϑ можно придумать много. Каждый автор может придумать свою статистику – функцию $\vartheta^*=g(\xi_1,\xi_2,...,\xi_n)$ и объявить её точечной оценкой числовой характеристики ϑ.

Возникает вопрос: какая из предлагаемых статистик $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ будет наилучшей для числовой характеристики ϑ?

Очевидно, что для выбора наилучшей статистики мы должны сформулировать исходящие из здравого смысла требования к статистикам и проверять выполнение этих требований к предлагаемой статистике. Если проверяемая статистика будет не удовлетворять хотя бы одному из этих требований, то будем эту статистику отклонять.

Теория

Сформулируем три требования к точечным оценкам числовых характеристик в виде определений.

Определение 1. Точечная оценка $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ числовой характеристики называется ***состоятельной,*** если она сходится по вероятности к оцениваемой числовой характеристике ϑ, то есть, если выполняется: $\lim\limits_{n \to \infty} P\left(\left|\vartheta^* - \vartheta\right| < \varepsilon\right) = 1.$

Это требование означает, что, если оценка *состоятельная*, то с увеличением объёма выборки $(n \to \infty)$ наша уверенность в том, что оценка ϑ^* будет мало отличаться (ε - мало) от оцениваемой характеристики ϑ, - должна возрастать. Уверенность в точности получаемой оценки ϑ^* мы оцениваем величиной вероятности P случайного события $\left\{\left|\vartheta^* - \vartheta\right| < \varepsilon\right\}$, которая должна мало от 1.

Определение 2. Точечная оценка $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ числовой характеристики называется ***несмещённой,*** если её математическое ожидание $M\vartheta^*$ равно величине оцениваемой характеристики ϑ, то есть, если $M\vartheta^* = \vartheta$.

Любая статистика $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ - случайная величина. Её значения зависят от элементов, составляющих выборку $(x_1, x_2, ..., x_n)$. У случайной величины ϑ^* есть математическое ожидание $M\vartheta^*$, называемое нами средним значением случайной величины. Требование *несмещённости* рассматриваемой оценки означает, что *среднее значение* получаемых для разных выборок значений оценки ϑ^* должны быть равны значению ϑ.

Перед формулированием третьего определения предположим, что мы рассматриваем две разные статистики $\vartheta_1^* = g_1(\xi_1, \xi_2, ..., \xi_n)$ и $\vartheta_2^* = g_2(\xi_1, \xi_2, ..., \xi_n)$. Обе статистики, как случайные величины, имеют дисперсии $D\vartheta_1^*$ и $D\vartheta_2^*$, значения которых для одной и той же выборки, – различные.

Определение 3. Точечная оценка $\vartheta_1^* = g_1(\xi_1, \xi_2, ..., \xi_n)$ числовой характеристики называется ***более эффективной***, чем точная оценка

$\vartheta_2^* = g_2(\xi_1, \xi_2, ..., \xi_n)$, если её дисперсия $D\vartheta_1^*$ меньше дисперсии $D\vartheta_2^*$, то есть, если $D\vartheta_1^* < D\vartheta_2^*$.

Дисперсия случайной величины это мера разброса значений случайной величины около её математического ожидания. Для каждой новой выборки мы получим некоторое новое значение оценки числовой характеристики. Эти значения имеют разброс около математического ожидания, величина которого оценивается дисперсией. Ясно, что из двух статистик мы должны отдать предпочтение той статистике, у которой разброс значений для одних и тех же выборок будет меньше.

Комментарий

Анализ сформулированных требований к точечным оценкам приводит нас к выводу, что эти требования действительно исходят из здравого смысла и, если применяемая оценка удовлетворяет этим требованиям, то наше доверие к получаемым значениям точечной оценки у нас будет наибольшим.

Рассмотрим примеры проверки выполнения (соответствия) этих требований точечными оценками наиболее часто используемых при исследованиях случайных величин числовых характеристик.

Пусть математическое ожидание $\vartheta = M\xi$ - числовая характеристика случайной величины, неизвестное значение которого мы хотим оценить статистикой $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$. Для оценки значения математического ожидания могут быть предложены такие статистики:

1) $\vartheta_1^* = \frac{1}{n}\sum_{i=1}^{n}\xi_i = \bar{\xi}$ - среднее арифметическое выборочных случайных величин.

2) $\vartheta_2^* = \frac{1}{n}\sum_{i=1}^{n}\beta_i \cdot \xi_i = \bar{\xi}_{взв.}$ - среднее взвешенное выборочных случайных величин.

Создавая эту статистику, мы считаем, что элементы выборочной совокупности ξ_i не являются для нас равнозначными. Мы, руководствуясь личными соображениями, присваиваем каждому элементу ξ_i «вес» β_i, называемый весовым коэффициентом, $i = 1,2,...,n$. Все весовые коэффициенты – неотрицательные числа: $\beta_i \geq 0$ и сумма их значений равна единице: $\sum_{i=1}^{n}\beta_i = 1$.

3) $\vartheta_3^* = \sqrt[n]{\prod_{i=1}^{n}\xi_i} = \bar{\xi}_{геом.}$ - среднее геометрическое выборочных случайных величин. (В практических задачах, где рекомендуется применять $\bar{\xi}_{геом.}$,

исследуемая случайная величина ξ принимает только положительные значения.)

4) $\vartheta_4^* = n : \sum_{i=1}^{n} \frac{1}{x_i} = \bar{\xi}_{гарм.}$ - среднее гармоническое выборочных случайных величин.

5) $\vartheta_5^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \bar{\xi}_{кв.}$ - среднее квадратическое выборочных случайных величин.

Ясно, что при наличии желания и некоторой доли фантазии придумать статистик для оценки математического ожидания можно много. Мы, основываясь на знании основных положений теории вероятностей, проверим выполнение сформулированных требований к точечным оценкам для некоторых из этих предлагаемых статистик.

Теория

Проверим состоятельность первых трёх предлагаемых оценок математического ожидания.

1) Статистика $\vartheta_1^* = \frac{1}{n}\sum_{i=1}^{n} \xi_i = \bar{\xi}$ согласно нашим правилам организации выборки, является средним арифметическим n независимых, одинаково распределённых случайных величин $\xi_1, \xi_2, ..., \xi_n$. У всех этих случайных величин одинаковое математическое ожидание, равное математическому ожиданию исследуемой случайной величины $M\xi_i = M\xi = \vartheta$. Все условия теоремы Хинчина выполнены. Согласно которой выборочная последовательность $\{\xi_1, \xi_2, ..., \xi_m\}$ подчиняется закону больших чисел, то есть $\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i - M\xi\right| < \varepsilon\right) = 1$ или $\lim_{n\to\infty} P\left(\left|\vartheta^* - \vartheta\right| < \varepsilon\right) = 1$. Значит статистика $\vartheta^* = \bar{\xi}$ - состоятельная оценка математического ожидания $\vartheta = M\xi$.

2) Статистика $\vartheta_2^* = \frac{1}{n}\sum_{i=1}^{n} b_i \cdot \xi_i = \bar{\xi}_{взв.}$ является средним арифметическим n независимых случайных величин $b_1\xi_1, b_2\xi_2, ..., b_n\xi_n$. Так как весовые коэффициенты b_i в общем случае – различные числа, то случайные величины $b_1\xi_1, b_2\xi_2, ..., b_n\xi_n$ - имеют разные распределения вероятностей. Числа b_i удовлетворяют лишь двум условиям: каждое b_i - неотрицательное число и $\sum_{i=1}^{n} b_i = n$. Значит, может быть не более одного b_i, которое равно n. Ясно, что случай когда $b_{i_0} = n$, а остальные b_i =0 - не

представляет никакого интереса, следовательно, найдётся такая константа C, что для любого номера i будет выполняться $D[b_i\xi_i]=b_i^2\cdot D\xi_i=b_i^2\sigma^2\le C$.

Вычислим среднее арифметическое математических ожиданий случайных величин-слагаемых $b_i\xi_i$. Если обозначить $M\xi=m$ и так как для любого номера i справедливо $M\xi_i=M\xi$, то:

$$\frac{1}{n}\sum_{i=1}^{n}M[b_i\xi_i]=\frac{1}{n}\sum_{i=1}^{n}b_i\cdot M\xi_i=\frac{1}{n}m\sum_{i=1}^{n}b_i=m.$$

Все условия теоремы Чебышёва из закона больших чисел – выполнены, следовательно: $\lim\limits_{n\to\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}b_i\xi_i-M\xi\right|<\varepsilon\right)=1$, то есть статистика $\vartheta_2^*=\bar{\xi}_{взв.}$ сходится по вероятности к числовой характеристике $\vartheta=M\xi$. Значит она состоятельная оценка математического ожидания.

3) Преобразуем статистику $\vartheta_3^*=\sqrt[n]{\prod_{i=1}^{n}\xi_i}=\bar{\xi}_{геом.}$ - среднее геометрическое выборочных случайных величин и рассмотрим: $\ln\vartheta_3^*=\ln\bar{\xi}_{геом.}=\frac{1}{n}\sum_{i=1}^{n}\ln\xi_i$. Так как случайные величины $\xi_1,\xi_2,...,\xi_n$ - независимые и одинаково распределены, то и случайные величины $\ln\xi_1,\ln\xi_2,...,\ln\xi_n$ будут независимыми и одинаково распределёнными. Если $y=f(x)=\ln x$, то запишем $f(\vartheta_3^*)=\ln\vartheta_3^*$ и $f(\vartheta)=M[\ln\vartheta]$. При рассмотрении первой точечной оценки ϑ_1^* мы показали, что среднее арифметическое элементов выборки – состоятельная оценка математического ожидания. Значит, последовательность средних арифметических $\ln\vartheta_3^*$ при увеличении объёма выборки n сходится по вероятности к $M[\ln\vartheta]=\ln\vartheta$. Так как $y=f(x)=\ln x$ - непрерывная функция для любого положительного x, то случайные события $\{|f(\vartheta_3^*)-f(\vartheta)|<\varepsilon\}$ и $\{|\vartheta_3^n-\vartheta|<\delta\}$ для достаточно малых ε и δ имеют одинаковые вероятности наступления.

Отсюда получаем: $\lim\limits_{n\to\infty}P(|\ln\vartheta_3^n-\ln\vartheta|<\varepsilon)=\lim\limits_{n\to\infty}P(|\vartheta_3^*-\theta|<\delta)=1$. То есть, среднее геометрическое $\bar{\xi}_{геом.}=\sqrt[n]{\prod_{i=1}^{n}\xi_i}$ является состоятельной оценкой математического ожидания $M\xi$.

Теорема Хинчина А.Я. из Закона больших чисел позволяет установить состоятельность оценок $a_k=\frac{1}{n}\sum_{i=1}^{n}\xi_i^k$ всех существующих у случайной величины ξ начальных моментов $\alpha_k=M[\xi^k]$. Так как из того, что случайные величины $\xi_1,\xi_2,...,\xi_n$ - независимые и имеют одинаковое

распределение вероятностей следует, что будут независимыми и будут иметь одинаковое распределение вероятностей случайные величины $\xi_1^k, \xi_2^k, ..., \xi_n^k$. Для всех i и любого возможного k будет справедливо $M\left[\xi_i^k\right] = \alpha_k$. Все условия теоремы Хинчина выполнены. Значит: $\lim\limits_{n\to\infty} P\left(\left|a_k - \alpha_k\right| < \varepsilon\right) = 1$, то есть каждая точечная оценка a_k при увеличении объёма выборки сходится по вероятности к начальному моменту α_k. Приходим к выводу, что каждая a_k - состоятельная оценка начального момента α_k, если этот момент существует.

Комментарий

Проверка состоятельности точечных оценок $\vartheta^* = m_k = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^k$ центральных моментов $\vartheta = \mu_k = M\left[\xi - M\xi\right]^k$ исследуемой случайной величины ξ кроме применения теорем Закона больших чисел использует следующую лемму.

Лемма. Пусть $\{\xi_n\}$ и $\{\eta_n\}$, $n = 1,2,3,...$, две последовательности случайных величин, которые сходятся по вероятности к постоянным величинам a и b, соответственно. Если функция $f(x,y)$ непрерывна в точке (a,b), то последовательность случайных величин $\{f(\xi_n,\eta_n)\}$, $n = 1,2,3,...$, сходится по вероятности к постоянной величине $f(a,b)$, то есть:

$$\lim_{n\to\infty} P\left(\left|f(\xi_n,\eta_n) - f(a,b)\right| < \varepsilon\right) = 1.$$

Доказательство. Ясно, что $P\left(\left|f(\xi_n,\eta_n) - f(a,b)\right| < \varepsilon\right) = 1 - P\left(\left|f(\xi_n,\eta_n) - f(a,b)\right| \ge \varepsilon\right)$. Так как функция $f(x,y)$ непрерывна в точке (a,b), то случайное событие $\left\{\left|f(\xi_n,\eta_n) - f(a,b)\right| \ge \varepsilon\right\}$ наступит, если, согласно определению непрерывности функции двух переменных, наступит событие $\left\{\left|\xi_n - a\right| \ge \delta\right\}$ или событие $\left\{\left|\eta_n - b\right| \ge \delta\right\}$. Значит, вероятность события $\left\{\left|f(\xi_n,\eta_n) - f(a,b)\right| \ge \varepsilon\right\}$ равна вероятности суммы событий $\left\{\left|\xi_n - a\right| \ge \delta\right\}$ и $\left\{\left|\eta_n - b\right| \ge \delta\right\}$.

Тогда будет справедливо:

$$1 - P\left(\left|f(\xi_n,\eta_n) - f(a,b)\right| \ge \varepsilon\right) = 1 - P\left(\left|\xi_n - a\right| \ge \delta \cup \left|\eta_n - b\right| \ge \delta\right).$$

Вероятность суммы событий не превосходит суммы вероятностей этих событий-слагаемых, то есть:

$$P\left(\left|\xi_n - a\right| \ge \delta \cup \left|\eta_n - b\right| \ge \delta\right) \le P\left(\left|\xi_n - a\right| \ge \delta\right) + P\left(\left|\eta_n - b\right| \ge \delta\right).$$

Значит: $P\left(\left|f(\xi_n,\eta_n)-f(a,b)\right|<\varepsilon\right)\geq 1-\left[P\left(\left|\xi_n-a\right|\geq\delta\right)+P\left(\left|\eta_n-b\right|\geq\delta\right)\right]$. Перейдём в обеих частях этого неравенства к пределу при $n\to\infty$:

$$\lim_{n\to\infty}P\left(\left|f(\xi_n,\eta_n)-f(a,b)\right|<\varepsilon\right)\geq 1-\left[\lim_{n\to\infty}P\left(\left|\xi_n-a\right|\geq\delta\right)+\lim_{n\to\infty}P\left(\left|\eta_n-b\right|\geq\delta\right)\right].$$

Последовательности $\{\xi_n\}$ и $\{\eta_n\}$ сходятся по вероятности к постоянным числам a и b, следовательно, при $n\to\infty$ пределы в правой части неравенства равны нулю. А так как вероятность любого события не может быть больше единицы, то получаем: $\lim_{n\to\infty}P\left(\left|f(\xi_n,\eta_n)-f(a,b)\right|<\varepsilon\right)=1$. Что и требовалось доказать.

Теория

Проверим состоятельность статистической дисперсии $\vartheta^*=S^2=\frac{1}{n}\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2$ - точечной оценки дисперсии $\vartheta=D\xi$ случайной величины ξ.

В теории вероятностей, применяя свойства математического ожидания, мы получаем, что дисперсию можно представить так: $D\xi=M\xi^2-M^2\xi$. То есть дисперсия выражается через начальные моменты второго и первого порядков $D\xi=\alpha_2-\alpha_1^2$. Используя формулу разности квадратов и проведя суммирование в выражении для оценки S^2, мы получим: $\vartheta^*=S^2=a_2-\overline{x}^2$. Точечная оценка дисперсии выражается через точечные оценки начальных моментов. Эти точечные оценки a_2 и $\overline{x}$, как было выше показано, сходятся по вероятности к соответствующим числовым характеристикам $\alpha_2=M\xi^2$ и $\alpha_1=M\xi$. Для того, чтобы применить лемму, рассмотрим функцию $f(x,y)=x-y^2$. Эта функция непрерывная в любой точке плоскости R^2 и, в частности, в точке с координатами $(\alpha_2;\alpha_1)$, в которой $f(\alpha_2,\alpha_1)=\alpha_2-\alpha_1^2=D\xi$. Но тогда, согласно лемме, можем записать: $\lim_{n\to\infty}P\left(\left|f(a_2,a_1)-f(\alpha_2,\alpha_1)\right|<\varepsilon\right)=1$.

Или, так как $f(a_2,a_1)=S^2$, $\lim_{n\to\infty}P\left(\left|S^2-D\xi\right|<\varepsilon\right)=1$. Значит точечная оценка $\vartheta^*=S^2$ - состоятельная оценка дисперсии $\vartheta=D\xi$.

Используя формулу бинома Ньютона и свойства математического ожидания случайной величины, мы можем выразить центральный момент μ_k любого порядка k через начальные моменты α_j порядка не выше, чем k, то есть для $j=1,2,...,k$. А тогда, учитывая состоятельность оценок a_j начальных моментов a_j , подбирая соответствующую непрерывную функцию и применяя лемму, мы убеждаемся, что точечная оценка m_k соответствующей числовой характеристики μ_k является состоятельной.

Комментарий

Второе требование к точечным оценкам ϑ^* числовых характеристик ϑ, выполнение которого мы должны проверить, выбирая наилучшую статистику, предусматривает наличие несмещённости точечной оценки. Это означает, что среднее значение используемой статистики $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ должно совпадать со значением оцениваемой числовой характеристики, то есть должно выполняться: $M\vartheta^* = \vartheta$.

1) Статистика $\vartheta_1^* = \frac{1}{n}\sum_{i=1}^{n}\xi_i = \bar{\xi}$ является средним арифметическим n независимых, одинаково распределённых случайных величин $\xi_1, \xi_2, ..., \xi_n$. Каждая ξ_i, i=1,2,…,n, имеет математическое ожидание равное математическому ожиданию случайной величины ξ, то есть $M\xi_i = M\xi = m$. Тогда получаем: $M\vartheta_1^* = M\bar{\xi} = M\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\right] = \frac{1}{n}\sum_{i=1}^{n}M\xi_i = \frac{1}{n}nm = m$. Значит, среднее арифметическое $\bar{\xi}$ выборочных случайных величин является несмещённой оценкой математического ожидания случайной величины $M\xi$.

2) Математическое ожидание статистики $\vartheta_2^* = \frac{1}{n}\sum_{i=1}^{n}b_i \cdot \xi_i = \bar{\xi}_{взв.}$ было нами вычислено при проверке состоятельности этой статистики. Получили $M\vartheta_2^* = M\bar{\xi}_{взв.} = m$. Значит, среднее взвешенное выборочных случайных величин $\bar{\xi}_{взв.}$ является несмещённой оценкой математического ожидания случайной величины $M\xi$.

3) Статистика $\vartheta_3^* = \sqrt[n]{\prod_{i=1}^{n}\xi_i} = \bar{\xi}_{геом.}$ - среднее геометрическое выборочных случайных величин, используется при изучении случайных величин, значения которой на числовой оси распределяются подобно значениям членов геометрической прогрессии со знаменателем $1+q$, где $q>0$. Такие случайные величины описывают, например, рост численности некоторой популяции, увеличение размера колонии бактерий, прибавку веса при росте некоторого вида животных, увеличение внутреннего валового продукта государства или отрасли промышленности.

Практика

Допустим, что мы получили выборку n значений такой случайной величины: $\{1, 1+q, (1+q)^2, ..., (1+q)^n\}$. Так как все элементы выборки равновозможные, то для любого k полагаем $P(\xi_k = (1+q)^k) = \frac{1}{n}$. Значения

статистик $\vartheta_1^* = \bar{\xi}$ и $\vartheta_3^* = \bar{\xi}_{геом.}$, как точечных оценок математического ожидания $M\xi$, получаются следующими: $\bar{\xi} = \frac{(1+q)^n - 1}{nq}$ и $\bar{\xi}_{геом.} = (1+q)^{\frac{n-1}{2}}$. При фиксированном значении q для всех значений n, $n \geq 2$, получаем, что $\bar{\xi} > \bar{\xi}_{геом.}$.

Отсюда следует, что в среднем значение точечной оценки $\vartheta_3^* = \bar{\xi}_{геом.}$ для одних и тех же выборок будет всегда меньше значения точечной оценки $\vartheta_1^* = \bar{\xi}$. Этот вывод противоречит нашей обычной интерпретации значения математического ожидания как координаты центра тяжести единичной массы, распределённой на числовой оси. Поэтому будем считать, что статистика $\vartheta_3^* = \sqrt[n]{\prod_{i=1}^{n} \xi_i} = \bar{\xi}_{геом.}$ не удовлетворяет требованию несмещённости точечной оценки.

Теория

Проверим выполнение требования несмещённости точечной оценки дисперсии $\vartheta^* = S^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$. Для этого дисперсию $D\xi$ и её оценку S^2 - статистическую дисперсию, представим, соответственно, в виде: $D\xi = \sigma^2 = \alpha_2 - m^2$ и $S^2 = a_2 - \bar{\xi}^2$. Вычислим математическое ожидание статистической дисперсии: $MS^2 = Ma_2 - M\bar{\xi}^2$.

Так как мы проверили, что точечные оценки начальных моментов – несмещённые, то $Ma_2 = \alpha_2$. Вычислим значение математического ожидания квадрата среднего арифметического случайных величин $\xi_1, \xi_2, ..., \xi_n$:

$$M\bar{\xi}^2 = M\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\right]^2 = \frac{1}{n^2}M\left[\sum_{i=1}^{n}\xi_i^2 + 2\sum_{1\leq i<j\leq n}\xi_i \cdot \xi_j\right] = \frac{1}{n^2}\left[n \cdot M\xi_i^2 + 2C_n^2 \cdot M[\xi_i \cdot \xi_j]\right]$$

. Вспомнив, что случайные величины $\xi_1, \xi_2, ..., \xi_n$ - независимые и одинаково распределены, получаем:

$$M\bar{\xi}^2 = \frac{1}{n}\alpha_2 + \frac{n(n-1)}{n^2}M\xi_i \cdot M\xi_j = \frac{1}{n}\alpha_2 + \frac{n-1}{n}m^2.$$

Математическое ожидание статистической дисперсии S^2 будет равно:

$$MS^2 = \alpha^2 - \frac{1}{n}\alpha^2 - \frac{n-1}{n}m^2 = \frac{n-1}{n}(\alpha_2 - m^2) = \frac{n-1}{n}\sigma^2.$$

Так как мы получили, что: $MS^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2 = D\xi$, то можем сказать, что точечная оценка $\vartheta^* = S^2$ - является смещённой оценкой

дисперсии. Значит, она не удовлетворяет второму требованию к точечным оценкам и должна быть отклонена.

Комментарий

Желая всё же иметь хорошую точечную оценку дисперсии $D\xi$, попробуем ликвидировать смещение оценки S^2, взяв статистику $\vartheta^* = s^2 = \frac{n-1}{n} S^2$. Так как $Ms^2 = \frac{n-1}{n} MS^2 = \sigma^2$, то предлагаемая статистика $\vartheta^* = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^2$ будет несмещённой оценкой дисперсии.

Но теперь нам надо проверить: не потеряли ли мы, ликвидируя смещение, состоятельность точечной оценки дисперсии?

Воспользуемся леммой. При $n \to \infty$ статистика S^2 сходится по вероятности к дисперсии $D\xi$. Последовательность $\left\{\frac{n-1}{n}\right\}$ сходится в обычном смысле к единице. Тем более она сходится к единице по вероятности. Функция $f(x,y) = x \cdot y$ - непрерывная на всей области определения R^2. Все условия леммы выполнены. Значит, последовательность $\left\{\frac{n-1}{n} \cdot S^2\right\}$ сходится по вероятности к значению функции $f\left(1,\sigma^2\right) = 1 \cdot \sigma^2 = \sigma^2$. Отсюда заключаем, что точечная оценка $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^2$ является состоятельной и несмещённой оценкой дисперсии $D\xi$.

Замечание. Требование несмещённости точечной оценки числовой характеристики объясняет нам необходимость использования оценки s^2 вместо оценки S^2. Хотя ясно, что при достаточно больших объёмах выборки деление суммы квадратов $\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^2$ на число n, или на число $n-1$ не приведёт к значительной разнице в значениях S^2 и s^2. Тем не менее, мы будем всегда использовать оценку s^2 и будем не обращать внимания на встречающиеся рекомендации в каких случаях какую оценку нужно применять.

Комментарий

Третье требование к точечным оценкам ϑ^* числовых характеристик ϑ, выполнение которого мы должны проверить, выбирая наилучшую статистику, предусматривает эффективность точечной оценки. Каждая

статистика $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ - случайная величина. Её значения зависят от элементов генеральной совокупности, попавших в выборку. Мы никак не можем заранее предсказать точное значение оценки в каждом конкретном случае. Поэтому о качестве оценки $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ мы должны судить не по её индивидуальным значениям для каждой выборки, а по распределению получающихся значений оценки для достаточно большой совокупности выборок. Значения несмещённой оценки сосредотачиваются в окрестности истинного значения оцениваемой числовой характеристики ϑ. Естественно, что мы будем лучшей называть ту оценку, у которой будет меньше рассеяние значений около значения характеристики ϑ.

Замечание. Подчеркнём ещё раз, что истинного значения числовой характеристики ϑ мы не знаем и, даже после обработки большой совокупности выборок, не будем его знать. Мы можем лишь сказать, что значение этой характеристики ϑ «близко» к полученному значению его точечной оценки ϑ^*. Оценить эту «близость» мы не можем, но мы понимаем, что она зависит от элементов генеральной совокупности, попавших в выборку, то есть - от репрезентативности выборки.

Сравним эффективность двух оценок математического ожидания: статистики $\vartheta_1^* = \frac{1}{n}\sum_{i=1}^{n}\xi_i = \bar{\xi}$ и статистики $\vartheta_2^* = \frac{1}{n}\sum_{i=1}^{n}b_i \cdot \xi_i = \bar{\xi}_{взв.}$. Обе статистики, как мы проверили, являются состоятельными и несмещёнными оценками математического ожидания $M\xi$ случайной величины ξ.

Чтобы выбрать лучшую из них, вычислим дисперсии этих статистик.

Учитывая условия формирования последовательности $\xi_1, \xi_2, ..., \xi_n$ и свойства дисперсии случайной величины, получим следующие результаты.

1) $D\vartheta_1^* = D\bar{\xi} = D\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}D\xi_i = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$.

2) $D\vartheta_2^* = D\bar{\xi}_{взв.} = D\left[\frac{1}{n}\sum_{i=1}^{n}b_i \cdot \xi_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}b_i^2 \cdot D\xi_i = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}b_i^2$. Представим каждое b_i в виде суммы: $b_i = 1 + \delta_i$, где δ_i или ≥ 0, или ≤ 0. Так как $\sum_{i=1}^{n}b_i = n$, то всегда $\sum_{i=1}^{n}\delta_i = 0$. Получаем:

$$\sum_{i=1}^{n}b_i^2 = \sum_{i=1}^{n}(1+\delta_i)^2 = n + 2\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\delta_i^2 = n + \sum_{i=1}^{n}\delta_i^2$$

Следовательно: $D\vartheta_2^* = \frac{\sigma^2}{n} + \frac{\sum_{i=1}^{n}\delta_i^2}{n^2}$, то есть для рассматриваемых точечных оценок получаем $D\vartheta_1^* \leq D\vartheta_2^*$, причём равенство этих дисперсий оценок будет лишь в том случае, когда все $b_i = 1$.

Приходим к выводу: только статистика $\vartheta_1^* = \overline{\xi}$ более предпочтительна в смысле предъявляемых нами требований к точечным оценкам. Значит, для оценки значения математического ожидания случайной величины лучше применять оценку $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ - среднее арифметическое значений элементов выборки.

Практика

В первом модуле рассматривается пример.2.1, в котором приведены два набора статистических данных – значений случайной величины ξ - рост мужчин в *см.* Определим значения точечных оценок математического ожидания и дисперсии случайной величины ξ.

Результаты вычислений по элементам первой выборки, объём которой равен n=115, получились такими:

среднее арифметическое $\overline{x} = \frac{1}{115}\sum_{i=1}^{115} x_i = 166{,}61;$

оценка дисперсии $s^2 = \frac{1}{115-1}\sum_{i=1}^{115}(x_i - \overline{x})^2 = 36{,}28;$

оценка среднего квадратического отклонения $s = \sqrt{s^2} = 6{,}02.$

Вычисления по элементам второй выборки, объём которой равен $n = 906$, проведены с учётом частоты m_j каждого из 37 наблюдаемых одинаковых значений элементов выборки $\tilde{x}_j$ (например: значение $\tilde{x}_1 = 147$ имеет в выборке частоту $m_1 = 1$; значение $\tilde{x}_9 = 155$ имеет в выборке частоту $m_9 = 6$; значение $\tilde{x}_{17} = 163$ имеет в выборке частоту $m_{17} = 48$; а значение $\tilde{x}_{24} = 170$ имеет в выборке частоту $m_{24} = 47$ и т.д.). Ясно, что $\sum_{j=1}^{37} m_j = 906$.

Результаты получились такими:

среднее арифметическое $\overline{x} = \frac{1}{906}\sum_{j=1}^{37} m_j x_j = 166{,}77;$

оценка дисперсии $s^2 = \frac{1}{906-1}\sum_{j=1}^{37} m_j(x_j - \overline{x})^2 = 34{,}70;$

оценка среднего квадратического отклонения $s = \sqrt{s^2} = 5{,}89.$

Анализируя полученные результаты, мы можем предположить, что средний рост мужчин равен 166,75*см*, то есть $M\xi = 166{,}75$. Но можно предположить, что средний рост равен 166,69*см*, то есть $M\xi = 166{,}69$. Позже мы познакомимся с правилами оценки надёжности наших выводов.

Полученные значения точечных оценок математического ожидания и дисперсии, естественно, различны, но разница полученных значений – мала и не имеет принципиального значения.. Мы должны познакомиться с

методами оценки значимости разности полученных значений оценок средних арифметических и статистических дисперсий.

Если анализируемые выборки взяты из разных генеральных совокупностей, мы сможем проверить предположение, о том, что случайные величины, определённые на этих совокупностях, подчиняются одинаковому закону распределения вероятностей.

И, наконец, вид гистограмм, построенных по вариационным рядам относительных частот, позволяет сделать предположение о том, что случайная величина ξ - рост мужчин подчиняется нормальному закону распределения вероятностей. Мы должны будем оценить вероятность правильности этого предположения.

§2.2. Неравенство Крамера-Рао

Комментарий

Выбирая из двух несмещённых и состоятельных точечных оценок математического ожидания более эффективную оценку, мы вычислили дисперсии оценок $\vartheta_1^* = \overline{\xi}$ и $\vartheta_2^* = \overline{\xi}_{взв.}$. Дисперсия первой оценки оказалась меньше дисперсии второй оценки, поэтому мы выбрали первую оценку – среднее арифметическое элементов выборки, как более лучшую оценку, удовлетворяющую трём требованиям к точечным оценкам.

Но при этом может возникнуть вопрос: «А является ли среднее арифметическое наилучшей оценкой математического ожидания? Нет ли другой несмещённой и состоятельной статистики $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$, которая имеет меньшую дисперсию, чем дисперсия оценки $\overline{\xi}$, а потому будет лучшей оценкой $\vartheta = M\xi$? Этот вопрос можно сформулировать так: Как следует использовать статистические данные, чтобы получить точечные оценки с минимальной дисперсией? Ответ на этот вопрос даёт применение неравенства Крамера-Рао.

Теория

Пусть исследуемая случайная величина имеет плотность вероятности $p(x, \vartheta)$, зависящую от параметра ϑ, числовое значение которого неизвестно. Повторяя n раз эксперимент G, мы формируем выборку, то есть, получаем выборочные случайные величины $\xi_1, \xi_2, ..., \xi_n$, которые будут независимыми и имеют одинаковую плотность вероятности $p(x, \vartheta)$.

Мы будем рассматривать выборочные случайные величины $\xi_1, \xi_2, ..., \xi_n$ как независимые компоненты n-мерного вектора, плотность вероятности которого равна: $p(x_1, x_2, ..., x_n, \vartheta) = \prod_{i=1}^{n} p(x_i, \vartheta)$. Нас будет интересовать экстремум этой функции. Но исследование на экстремум произведения большого числа функций затруднительно с технической стороны. Функция

$y = \ln x$ - монотонная функция, и точки экстремума функции $p(x_1, x_2, ..., x_n, \vartheta)$ совпадают с точками экстремума функции $\ln p(x_1, x_2, ..., x_n, \vartheta)$. Поэтому нам будет удобней рассматривать функцию:

$$L_n(\vartheta) = \ln p(x_1, x_2, ..., x_n, \vartheta) = \sum_{i=1}^{n} p(x_i).$$

Возьмём производную от этой функции по параметру ϑ:

$$\frac{\partial L_n(\vartheta)}{\partial \vartheta} = \frac{\partial \ln p(x_1, x_2, ..., x_n, \vartheta)}{\partial \vartheta} = \sum_{i=1}^{n} \frac{1}{p(x_i, \vartheta)} \cdot \frac{\partial p(x_i, \vartheta)}{\partial \vartheta}.$$

Так как все функции $p(x_i, \vartheta)$ - одинаковые, то, обозначив для любого номера i: $\frac{1}{p(x_i, \vartheta)} \cdot \frac{\partial p(x_i, \vartheta)}{\partial \vartheta} = \frac{\partial L_0(\vartheta)}{\partial \vartheta}$, запишем: $\frac{\partial L_n(\vartheta)}{\partial \vartheta} = n \cdot \frac{\partial L_0(\vartheta)}{\partial \vartheta}$.

Аргументы функции, как значения элементов выборки, – случайные величины. Поэтому логарифмическая производная $\frac{\partial L_n(\vartheta)}{\partial \vartheta}$ функции $p(x_1, x_2, ..., x_n, \vartheta)$ будет статистикой, то есть – случайной величиной. Обозначим: $M\left[\frac{\partial L_n(\vartheta)}{\partial \vartheta}\right]^2 = J_n(\vartheta)$ и будем называть $J_n(\vartheta)$ - информацией Фишера.

Если Θ множество возможных значений числовой характеристики ϑ, то возможные значения статистик $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ будут значениями некоторой функции $\gamma(\vartheta)$, определённой на этом множестве Θ. Значит, в общем случае, $M\vartheta^* = \gamma(\vartheta)$. Если рассматриваемая статистика $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ - несмещённая оценка числовой характеристики ϑ, то $M\vartheta^* = \vartheta$.

Ясно, что для плотности вероятности $p(x_1, x_2, ..., x_n, \vartheta)$ n-мерного вектора $(\xi_1, \xi_2, ..., \xi_n)$ всегда выполняется:

$$\int_{R^n} p(x_1, x_2, ..., x_n, \vartheta) dx = 1 \quad (*).$$

Согласно общему определению, математическое ожидание $M[f(\xi)]$ непрерывной функции $f(\xi)$ случайной величины ξ есть значение интеграла Римана-Стилтьеса: $M[f(\xi)] = \int_R f(x) p(x) dx$, где $p(x)$ - плотность вероятности ξ.

Значит: $M[\vartheta^*] = \int_{R^n} g(x_1, x_2, ..., x_n) \cdot p(x_1, x_2, ..., x_n, \vartheta) dx = \gamma(\vartheta) \quad (**)$.

Справедлива <u>теорема</u>:

Если $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ *состоятельная точечная оценка числовой характеристики* ϑ *и кратные интегралы* $(*)$ *и* $(**)$ *дифференцируемы по числовой характеристике* ϑ, *то будет справедливо неравенство:*

$$D\vartheta^* \geq \frac{[\gamma'(\vartheta)]^2}{J_n(\vartheta)} \quad . \qquad \begin{pmatrix} ** \\ * \end{pmatrix}$$

В частности, если точечная оценка $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ числовой характеристики ϑ - несмещённая, то есть $\gamma(\vartheta) = \vartheta$, то последнее неравенство будет иметь более простой вид: $D\vartheta^* \geq \frac{1}{J_n(\vartheta)}$.

Неравенство $\begin{pmatrix} ** \\ * \end{pmatrix}$ называется *неравенством Крамера-Рао*. Оно показывает, что существует точная нижняя граница для дисперсий возможных состоятельных точечных оценок $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ числовой характеристики ϑ. И, если нам удастся найти такую несмещённую точечную оценку $\vartheta_0^* = g(\xi_1, \xi_2, ..., \xi_n)$, что будет выполняться $D\vartheta_0^* = \frac{1}{J_n(\vartheta)}$, то лучше этой оценки, в смысле, эффективности - нет.

Комментарий

Доказательство теоремы.

Условимся записывать n-мерную переменную $(x_1, x_2, ..., x_n)$ в n-кратном интеграле одной буквой x. Возьмём от интегралов $(*)$ и $(**)$ производные по переменной ϑ:

$$\int_{R^n} \frac{\partial p(x, \vartheta)}{\partial \theta} dx = 0 \quad (*') \quad \text{и} \quad \int_{K^m} g(x) \cdot \frac{\partial p(x, \vartheta)}{\partial \vartheta} dx = \gamma'(\vartheta). \ (**')$$

Первое равенство умножим на $\gamma(\vartheta)$ и вычтем полученное из второго интеграла:

$$\int_{K^m} (g(x) - \gamma(\vartheta)) \cdot \frac{\partial p(x, \vartheta)}{\partial \vartheta} dx = \gamma'(\vartheta).$$

Это равенство не изменится, если подинтегральное выражение умножим и разделим на одну и ту же функцию:

$$\int_{K^m} (g(x) - \gamma(\vartheta)) \cdot \frac{1}{p(x, \vartheta)} \cdot \frac{\partial p(x, \vartheta)}{\partial \vartheta} \cdot p(x, \vartheta) dx = \gamma'(x).$$

Так как $g(x) = \vartheta^*$, $\gamma(\vartheta) = M\vartheta^*$, а $\frac{1}{p(x, \vartheta)} \cdot \frac{\partial p(x, \vartheta)}{\partial \vartheta} = \frac{\partial L_n(\vartheta)}{\partial (x)}$, то запишем:

$$\int_{R^n}\left[\left(\vartheta^* - M\vartheta^*\right)\cdot\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right]\cdot p(x,\vartheta)dx = \gamma'(\vartheta).$$

Выражение в квадратной скобке есть произведение двух функций случайных величин, то есть – это функция случайных величин. Тогда, по общему определению математического ожидания функции случайных величин, получаем:

$M\left[\left(\vartheta^* - M\vartheta^*\right)\cdot\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right] = \gamma'(\vartheta)$. Возведём обе части этого равенства в квадрат: $M^2\left[\left(\vartheta^* - M\vartheta^*\right)\cdot\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right] = \left[\gamma'(\vartheta)\right]^2$ и воспользуемся неравенством Коши-Буняковского ($M^2[\xi\cdot\eta]\le M\xi^2\cdot M\eta^2$):

$M\left[\vartheta^* - M\vartheta^*\right]^2\cdot M\left[\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right]^2 \ge \left[\gamma'(\vartheta)\right]^2$. То есть: $D\vartheta^*\cdot J_n(\vartheta)\ge\left[\gamma'(\vartheta)\right]^2$.

Окончательно получаем: $D\vartheta^* \ge \frac{\left[\gamma'(\vartheta)\right]^2}{J_n(\vartheta)}$, что и требовалось доказать.

Замечание. Неравенство Коши-Буняковского доказывается и используется в курсе математического анализа. Проведём доказательство этого неравенства в терминах теории вероятностей.

Определим случайную величину: $\zeta = (a\xi + b\eta)^2$. Ясно, что эта случайная величина принимает неотрицательные значения, то есть, всегда $\zeta \ge 0$. Но тогда и её математическое ожидание будет неотрицательным числом: $M\zeta \ge 0$. Значит всегда $M[a\xi + b\eta]^2 \ge 0$, или $a^2M\xi^2 + 2abM[\xi\cdot\eta] + b^2M\eta^2 \ge 0$. Получилась квадратичная форма, которая принимает неотрицательные значения. Квадратичная форма будет принимать неотрицательные значения, если её дискриминант принимает неположительные значения, то есть, если $M^2[\xi\cdot\eta] - M\xi^2\cdot M\eta^2 \le 0$. Отсюда получаем неравенство, которое использовалось при доказательстве теоремы: $M^2[\xi\cdot\eta]\le M\xi^2\cdot M\eta^2$.

Комментарий

Прежде, чем использовать в примерах неравенство Крамера-Рао для выяснения наиболее эффективной точечной оценки, выясним вероятностный смысл информации Фишера $J_n(\vartheta)$.

Если $\xi_1,\xi_2,...,\xi_n$ - выборочные случайные величины, то логарифмическая производная $\frac{\partial L_n(\vartheta)}{\partial\vartheta}$ функции

$p(\xi_1,\xi_2,...,\xi_n,\vartheta)=\prod_{i=1}^{n} p(\xi_i,\vartheta)$ будет статистикой этих выборочных случайных величин, то есть $\frac{\partial L_n(\vartheta)}{\partial\vartheta}=g(\xi_1,\xi_2,...,\xi_n)$. Вычислим математическое ожидание и дисперсию этой статистики.

$$M\left[\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right]=\int_{R^n}\frac{\partial \ln p(x_1,x_2,...,x_n,\vartheta)}{\partial\vartheta}\cdot p(x_1,x_2,...,x_n)dx.$$

Как и при доказательстве теоремы, будем записывать n-мерную переменную $(x_1,x_2,...,x_n)$ в n-кратном интеграле буквой x. Следующие преобразования подинтегральной функции не требуют пояснений. В результате получаем интеграл $(*')$, который равен нулю:

$$M\left[\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right]=\int_{R^n}\left(\frac{1}{p(x,\vartheta)}\cdot\frac{\partial p(x,\vartheta)}{\partial\vartheta}\right)\cdot p(x,\vartheta)dx=\int_{R^n}\frac{\partial p(x,\vartheta)}{\partial\vartheta}dx=0.$$

Вычисляя дисперсию, запишем:

$D\left[\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right]=M\left[\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right]^2-M^2\left[\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right]$. Здесь уменьшаемое является информацией Фишера, а вычитаемое равно нулю. Получаем: $D\left[\frac{\partial L_n(\vartheta)}{\partial\vartheta}\right]=J_n(\vartheta)$. Значит, информация Фишера это *дисперсия статистики* – логарифмической производной плотности вероятности вектора $(\xi_1,\xi_2,...,\xi_n)$. А так как компоненты этого вектора независимые и одинаково распределённые случайные величины, то $J_n(\vartheta)=nJ_0(\vartheta)$, где $J_0(\vartheta)$ - дисперсия логарифмической производной по ϑ функции $p(\xi,\vartheta)$. Отсюда вытекает, что для того, чтобы вычислить значение информации Фишера надо **знать** плотность вероятности $p(x,\vartheta)$, если исследуемая случайная величина непрерывного типа, или распределение вероятностей $\{p_k\}$, $k=1,2,\ldots$, если исследуемая случайная величина дискретного типа.

Предположим, что для числовой характеристики ϑ имеется некоторое множество точечных оценок: $\vartheta_1^*,\vartheta_2^*,...,\vartheta_l^*$. У каждой точечной оценки вычислили её дисперсию и получили числовое множество: $\{D\vartheta_1^*,D\vartheta_2^*,...,D\vartheta_l^*\}$. Это числовое множество имеет точную нижнюю границу, которая может не принадлежать этому числовому множеству. Если среди этих оценок $\vartheta_1^*,\vartheta_2^*,...,\vartheta_l^*$ существует такая оценка ϑ_0^* , что $D\vartheta_0^*=\inf\{D\vartheta_1^*,D\vartheta_2^*,...,D\vartheta_l^*\}$, то такую оценку будем называть *наиболее эффективной,* или просто – *эффективной.*

Практика

Рассмотрим примеры применения неравенства Крамера-Рао при определении эффективности точечной оценки числовой характеристики.

Требуется оценить эффективность среднего арифметического элементов выборки $\vartheta_1^* = \overline{\xi}$, принимаемого нами в качестве оценки математического ожидания $\vartheta = M\xi = m$. Чтобы применить неравенство Крамера-Рао для несмещённых оценок: $D\vartheta^* \geq \frac{1}{J_n(\vartheta)}$, необходимо определить значение информации Фишера $J_n(m)$ для математического ожидания. Мы установили, что информация Фишера является дисперсией логарифмической производной плотности вероятности $p(x_1, x_2, ..., x_n, \vartheta) = \prod_{i=1}^{n} p(x_i, \vartheta)$ вектора с независимыми компонентами $\xi_1, \xi_2, ..., \xi_n$. Так как все компоненты имеют одинаковое распределение, то $J_n(m) = nJ_0(m)$, где $J_0(m)$ - информация Фишера любой компоненты вектора. Вычисление $J_0(m)$ можно провести, если будет известна плотность вероятности $p(x, \vartheta)$ случайной величины ξ.

Предположим, что случайная величина ξ подчиняется нормальному распределению с параметрами m и σ^2, где оцениваемая характеристика m – математическое ожидание, то есть: $p(x, m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$. Вычислим $J_0(m)$ для этой плотности вероятности, имея в виду, что $J_0(m) = M\left[\frac{\partial L_0(m)}{\partial m}\right]^2$.

Так как $L_0(m) = ln\, p(x, m) = ln\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-m)^2}{2\sigma^2}$, то $\frac{\partial L_0(m)}{\partial m} = \frac{x-m}{\sigma^2}$.

Значит: $J_0(m) = M\left[\frac{\partial L_0(m)}{\partial m}\right]^2 = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}\left(\frac{x-m}{\sigma^2}\right)^2 \cdot e^{-\frac{(x-m)^2}{2\sigma^2}} dx$.

Учитывая, что подинтегральная функция – чётная и, сделав замену переменной интегрирования по формуле: $t = \frac{(x-m)^2}{2\sigma^2}$, получаем:

$$J_0(m) = \frac{2}{\sqrt{\pi}\sigma^2}\int_0^{\infty} t^{1/2} e^{-t} dt = \frac{1}{\sqrt{\pi}} \cdot \frac{2}{\sigma^2}\Gamma\left(3/2\right) = \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sigma^2}\Gamma\left(1/2\right) = \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sigma^2} \cdot \sqrt{\pi} = \frac{1}{\sigma^2}$$

Окончательно получаем, что $J_n(m)=n\cdot J_0(m)=\frac{n}{\sigma^2}$. Ранее мы показали, что дисперсия среднего арифметического в n раз меньше дисперсии каждого отдельного слагаемого, то есть: $D[\vartheta^*]=D[\bar{\xi}]=\frac{\sigma^2}{n}$. Подставляя полученные значения информации Фишера и дисперсии среднего арифметического элементов выборки в неравенство Крамера-Рао, мы видим, что это неравенство становится равенством: $\frac{\sigma^2}{n}=\frac{1}{\frac{n}{\sigma^2}}$. Делаем вывод: Если исследуемая случайная величина ξ подчиняется нормальному распределению, кратко мы записываем $\xi\in\mathcal{N}(m,\sigma)$, то, в смысле наших требований к точечным оценкам, лучшей, чем среднее арифметическое $\vartheta_1^*=\bar{\xi}$, оценки для математического ожидания $\vartheta=m$ не существует.

2) Кратко, не углубляясь в вычислительные подробности, проверим эффективность оценки $\vartheta_1^*=\bar{\xi}$ для математического ожидания $\vartheta=m$ в том случае, когда случайная величина ξ подчиняется экспоненциальному распределению. Плотность вероятности экспоненциального распределения имеет вид: $p(x,\mu)=\begin{cases}0, & x<0\\ \frac{1}{\mu}e^{-\frac{x}{\mu}}, & x\ge 0\end{cases}$.

Значит: $L_0(\mu)=-\ln\mu-\frac{x}{\mu}$.

Тогда $J_0(\mu)=M\left[\frac{\partial L_0(\mu)}{\partial\mu}\right]^2=\int_0^\infty\left(\frac{1}{\mu}-\frac{t}{\mu}\right)^2\cdot e^{-\frac{x}{\mu}}dx$. После упрощений и замены переменной интегрирования $\frac{x}{\mu}=t$. Получаем:

$$J_0(\mu)=\frac{1}{\mu^2}\int_0\left(t^2-2t+1\right)\cdot e^{-t}dt=\frac{1}{\mu^2}\left[\Gamma(3)-2\cdot\Gamma(2)+\Gamma(1)\right]=\frac{1}{\mu^2}\left[2-2\cdot1+1\right]=\frac{1}{\mu^2}.$$

Информация Фишера для вектора ($\xi_1,\xi_2,...,\xi_n$) будет равна: $J_n(m)=n\cdot J_0(m)=\frac{n}{\mu^2}$.

Учитывая, что дисперсия случайной величины ξ равна $D\xi=\mu^2$, получаем, что дисперсия точечной оценки математического равна $D[\vartheta^*]=D[\bar{\xi}]=\frac{\mu^2}{n}$. Замечаем, что значения $[J_n\mu]^{-1}=\left(\frac{n}{\mu^2}\right)^{-1}$ и $D[\vartheta^*]=\frac{\mu^2}{n}$ -

равны. Следовательно, неравенство Крамера-Рао превратится в равенство. Значит, среднее арифметическое элементов выборки $\vartheta_1^* = \overline{\xi}$ будет наиболее эффективной оценкой математического ожидания $M\xi = \mu$. А так как среднее арифметическое $\overline{\xi}$, как было ранее установлено, является состоятельной и несмещённой оценкой математического ожидания $M\xi$, то опять мы приходим к выводу, что и при экспоненциальном распределении вероятностей лучшей оценки для $M\xi$, чем $\overline{\xi}$, - не существует.

3) Проверим теперь эффективность среднего арифметического $\overline{\xi}$ для оценки значения математического ожидания $M\xi$ в случае, когда случайная величина ξ будет случайной величиной дискретного типа.

Пусть распределением вероятностей случайной величины ξ будет распределение вероятностей Пуассона, то есть $P(\xi = k) = \dfrac{\lambda^k \cdot e^{-\lambda}}{k!}$, где $k = 0,1,2,\ldots$. Мы знаем, что числовой параметр распределения λ является математическим ожиданием и дисперсией случайной величины ξ. Мы уже привыкаем к тому, что в качестве оценки $M\xi$ прежде других оценок проверяется среднее арифметическое $\overline{\xi}$ элементов выборки.

Как и в первых двух примерах, мы должны вычислить значения информации Фишера $J_n(\vartheta) = n \cdot J_0(\vartheta)$, где $\vartheta = \lambda$. Информация Фишера $J_0(\lambda)$ - это значение математического ожидания функции $\left[\dfrac{\partial L_0(\lambda)}{\partial \lambda}\right]^2$.

Сначала вычисляем: $L_0(\lambda) = \ln p_k = k \ln \lambda - \lambda - \ln(k!)$ и $\dfrac{\partial L_0(\lambda)}{\partial \lambda} = k \cdot \dfrac{1}{\lambda} - 1.$

Так как случайная величина ξ - случайная величина дискретного типа, то в рассматриваемом примере $M\left[\dfrac{\partial L_0(\lambda)}{\partial \lambda}\right]^2$ будет значением суммы ряда, а не значением несобственного интеграла, как было в первых двух примерах.

$$M\left[\frac{\partial L_0(\lambda)}{\partial \lambda}\right]^2 = \sum_{k=0}^{\infty}\left(\frac{k}{\lambda} - 1\right)^2 \cdot p_k = \sum_{k=0}^{\infty}\left(\frac{k}{\lambda} - 1\right)^2 \cdot \frac{\lambda^k \cdot e^{-\lambda}}{k!} =$$
$$= \sum_{k=1}^{\infty}\frac{k^2}{\lambda^2} \cdot \frac{\lambda^k \cdot e^{-\lambda}}{k!} - \frac{2}{\lambda}\sum_{k=1}\frac{k \cdot \lambda^k \cdot e^{-\lambda}}{k!} + \sum_{k=0}\frac{\lambda^k \cdot e^{-\lambda}}{k!}.$$

Вычисления первой и второй суммы сводятся к сокращениям первого множителя в факториалах, представлению в первой сумме переменной

суммирования в виде $k=(k-1)+1$ и к заменам индекса суммирования $k=j+1$. Третья сумма будет равна единице.

В результате вычислений получаем: $M\left[\frac{\partial L_0(\lambda)}{\partial\lambda}\right]^2=1+\frac{1}{\lambda}-2+1=\frac{1}{\lambda}$.

Значит $J_n(\vartheta)=n\cdot J_0(\vartheta)=\frac{n}{\lambda}$. А так как $D(\bar{\xi})=\frac{\lambda}{n}$, то снова приходим к выводу, что если случайная величина ξ распределена по закону Пуассона, то наилучшей оценкой параметра λ, в смысле требований к точечным оценкам числовых характеристик, будет среднее арифметическое элементов выборки $\bar{\xi}$.

Комментарий

Мы трижды для трёх различных законов распределения вероятностей ξ убедились в том, что лучшей, чем среднее арифметическое элементов выборки $\bar{\xi}$, оценки математического ожидания $M\xi$ - нет.

Так как второй важнейшей числовой характеристикой случайной величины является дисперсия $D\xi$, то возникает вопрос о существовании наилучшей оценки для дисперсии.

Рассмотрим три статистики $\vartheta^*=g(\xi_1,\xi_2,...,\xi_n)$, которые можно предложить в качестве точечной оценки дисперсии:

1) $\vartheta_1^*=S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$, 2) $\vartheta_2^*=s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$, 3) $\vartheta_3^*=s_0^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2$. Третья статистика предусматривает, что нам известно значение $m=M\xi$ математического ожидания случайной величины. Что, вообще-то говоря, довольно редко может встретиться в реальных условиях.

Первая оценка $\vartheta_1^*=S^2$ - состоятельная, но смещённая. Она не удовлетворяет второму требованию к точечным оценкам, и мы её рассматривать не будем.

Вторая оценка $\vartheta_2^*=s^2$- состоятельная и несмещённая. Поэтому стоит проверить её на эффективность. Для этого нам надо сначала, зная закон распределения случайной величины ξ, вычислить значение информации Фишера $J_n(\vartheta)=n\cdot J_0(\vartheta)$, где $\vartheta=D\xi=\sigma^2$, и вычислить значение дисперсии оценки s^2.

Опуская техническую сторону вычислений, сразу запишем результаты: $J_n(\sigma^2)=n\cdot J_0(\sigma^2)=\frac{n}{2\sigma^4}$ и $D[s^2]=\frac{2\sigma^4}{n-1}$. Подставляя эти результаты в

неравенство Крамера-Рао для несмещённой оценки $D\vartheta^* \geq \frac{1}{J_n(\vartheta^*)}$, получаем: $\frac{2\sigma^4}{n-1} \geq \frac{1}{\frac{n}{2\sigma^4}}$.

Наш первый вывод: оценка $\vartheta_2^* = s^2$ не будет эффективной. Её эффективность можно оценить числом $\frac{n-1}{n}$, которое при любом n будет меньше единицы.

Второй, более важный вывод: точная нижняя граница числового множества оценок дисперсий, подсчитанных для последовательности $\{n_j\}$, где $n_1 < n_2 < n_3 < .....$, не принадлежит этому множеству. А это означает, что наиболее эффективной оценки ϑ^* для дисперсии $D\xi = \sigma^2$ - не существует.

Третья оценка $\vartheta_3^* = s_0^2$ - состоятельная и несмещённая. Для проверки её на эффективность вычисляем значение дисперсии этой оценки: $D[s_0^2] = \frac{2\sigma^4}{n}$. Так как информация Фишера равна: $J_n(\sigma^2) = \frac{n}{2\sigma^4}$, то неравенство Крамера-Рао $D\vartheta^* \geq \frac{1}{J_n(\vartheta^*)}$, в этом случае становится равенством $\frac{2\sigma^4}{n} = \frac{1}{\frac{n}{2\sigma^4}}$.

Оценка $\vartheta_3^* = s_0^2$ - эффективная оценка, следовательно, она удовлетворяет всем трём требованиям, предъявляемым к точечным оценкам. Значит, она является наилучшей из трёх рассмотренных оценок для дисперсии. Но возможности её применения в практических ситуациях будут очень редкими, так как вычисление её значения предусматривает знание значения математического ожидания $M\xi = m$ исследуемой случайной величины ξ. А это бывает редко, так как в своём распоряжении исследователь, как правило, имеет только набор чисел $(x_1, x_2, ..., x_n)$, называемых репрезентативной выборкой и являющимися значениями *выборочных случайных величин* $\xi_1, \xi_2, ..., \xi_n$.

§3.2. Методы получения точечных оценок.

Комментарий

Мы сказали, что формул, выражающих значения точечных оценок числовых параметров случайной величины, можно придумать много. В основном числовыми параметрами распределений вероятностей являются

числовые характеристики случайной величины. Это - начальные и центральные моменты случайной величины или функции от них. Любая предлагаемая точечная оценка $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ перед её применением в практических задачах должна быть проверена на состоятельность, несмещённость и эффективность.

Рассмотрим два метода получения точечных оценок, использующих наше представление выборки как n-мерный вектор $(\xi_1, \xi_2, ..., \xi_n)$ с независимыми и одинаково распределёнными компонентами.

Теория

Метод моментов

Суть метода, предложенного К.Пирсоном, заключается в том, что вычисляемые моменты эмпирической случайной величины ξ^* принимаются в качестве оценок соответствующих теоретических моментов исследуемой по выборке случайной величины ξ.

То есть, если, например, нам надо оценить значение числовой характеристики $\theta : M\xi, D\xi, \alpha_k, \mu_k$ и т.д., то мы берём значения соответствующих эмпирических числовых характеристик $\vartheta^* : \bar{x}, S^2, a_k, m_k$ и т.д., которые получаем, используя элементы выборки $(x_1, x_2, ..., x_n)$.

Здесь:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_t - \bar{x})^2, \quad a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \quad m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Применяя теоремы Бернулли, Хинчина, Чебышёва из Закона больших чисел, мы показываем, что эти оценки – состоятельные. Но их надо проверять на несмещённость и эффективность.

Аналогично поступаем и в том случае, когда исследуемая случайная величина ξ - многомерная.

Метод максимального правдоподобия

Этот метод, предложенный Фишером, исходит из предположения, что имеющаяся у нас выборка $(x_1, x_2, ..., x_n)$ самая правдоподобная из всех возможных выборок в смысле представления наиболее характерных особенностей исследуемой случайной величины ξ.

Из этого предположения следует, что значение вероятности нашей выборки $P(x_1, x_2, ..., x_n, \vartheta) = \prod_{i=1}^{n} P(x_i, \vartheta)$ максимально по сравнению со значениями вероятностей других выборок. Так как значение числовой характеристики ϑ мы не знаем, то естественно поставить задачу найти оценку $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$, при которой для элементов нашей выборки будет выполняться:

$P(x_1, x_2, ..., x_n, \vartheta^*) = max\, P(x_1, x_2, ..., x_n, \vartheta)$ для всех возможных значений ϑ.

Мы пришли к обычной задаче на определение аргумента ϑ^*, при котором функция $P(x_1, x_2, ..., x_n, \vartheta)$ принимает максимальное значение.

Числовая характеристика ϑ может быть многомерной $\vartheta = (\vartheta_1, ..., \vartheta_l)$. Например, если ξ имеет нормальное распределение вероятностей, то l будет равно двум. В этом случае: $\vartheta_1 = m$ и $\vartheta_2 = \sigma^2$.

Для решения получившейся задачи мы логарифмическую производную по параметру ϑ функции $P(x_1, x_2, ..., x_n, \vartheta)$ приравниваем к нулю и решаем получившееся уравнение относительно неизвестной характеристики ϑ. Решением этого уравнения и будет оценка ϑ^*, которую мы будем называть оценкой максимального правдоподобия.

Если числовая характеристика ϑ многомерная: $\vartheta = (\vartheta_1, ..., \vartheta_l)$, то мы составляем систему l уравнений, каждое из которых есть частная логарифмическая производная по параметру ϑ_j, где $j = 1, ..., l$. Решением этой системы будет многомерная оценка максимального правдоподобия $\vartheta^* = (\vartheta_1^*, ..., \vartheta_l^*)$. Получающиеся таким методом оценки ϑ^* будут состоятельными, если у параметра ϑ существует эффективная оценка ϑ^*, то она обязательно получится при применении метода максимального правдоподобия. Получающиеся оценки надо проверять на несмещённость.

При определении оценок числовых параметров теоретического распределения методом максимального правдоподобия может получиться сложное для решения уравнение, или система уравнений. Для решения уравнения, или системы уравнений, приходится применять различные методы решений, вплоть до применения метода касательных приближённого нахождения корней уравнения (метод Ньютона).

Практика

Рассмотрим примеры применения метода максимального правдоподобия, если случайная величина ξ дискретного или непрерывного типа и нам известен её закон распределения вероятностей.

Простейший пример 1.2 Пусть случайная величина ξ - бернуллиевская, кратко записываем: $\xi \in B_1(p)$. Такая случайная величина встречается при однократном проведении опыта. Если событие A наступило, то $\xi = 1$, если наступило событие $\overline{A}$, то $\xi = 0$. Ясно, что $P(\xi = 1) = P(A) = p$ и $P(\xi = 0) = P(\overline{A}) = q$. Всегда будет выполняться: $p + q = 1$. Вероятность p – числовая характеристика ϑ бернуллиевского распределения. Точное значение вероятности нам неизвестно и мы хотим с

помощью метода максимального правдоподобия найти ей точечную оценку ϑ^*.

Мы проводим n раз одно и то же испытание и получаем числовую выборку $(x_1, x_2, ..., x_n)$, где любой её элемент x_i равен 1, если в i–том испытании наступило событие A, или равен 0, если наступило событие $\overline{A}$. Пусть в этой выборке 1 встретилась m раз, а 0 встретился n-m раз.

Записываем вероятность получения такой выборки:

$$P(x_1, x_2, ..., x_n, \vartheta) = \prod_{i=1}^{n} P(x_i, \vartheta) = p^m \cdot (1-p)^{n-m},$$

то есть $L_n(\vartheta) = p^m \cdot (1-p)^{n-m}$. Логарифмируем эту функцию: $\ln L_n(p) = m \ln p + (n-m)\ln(1-p)$ и дифференцируем её по параметру p: $\frac{1}{L_n(p)} \cdot \frac{\partial L_n(p)}{\partial p} = m \cdot \frac{1}{p} - (n-m) \cdot \frac{1}{1-p}$. Согласно методу максимального правдоподобия, логарифмическую производную функции $L_n(p)$ приравниваем к нулю и получаем уравнение: $m \cdot \frac{1}{p} - (n-m) \cdot \frac{1}{1-p} = 0$.

Решением этого уравнения будет $p^* = \frac{m}{n}$ - относительная частота наступлений события A при проведении n испытаний. Значение относительной частоты $\frac{m}{n} = \vartheta^*$ мы принимаем как точечную оценку неизвестного значения вероятности $p = \vartheta$ – числовой характеристики бернуллиевского распределения.

Пример 2.2 Пусть случайная величина ξ подчиняется закону распределения вероятностей Пуассона с параметром λ, кратко записываем: $\xi \in \Pi(\lambda)$. То есть любое своё возможное значение k случайная величина ξ принимает с вероятностью $p_k = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$, где $k = 0,1,2,...$. Параметр λ - числовая характеристика закона Пуассона, значение которой мы не знаем. Мы хотим, применяя метод максимального правдоподобия, найти точечную оценку λ^* этой числовой характеристики.

Из генеральной совокупности мы отбираем выборку $(x_1, x_2, ..., x_n)$, объём которой равен n. Элементы выборки – это неотрицательные целые числа $k = 0,1,2,...$. Обозначим m_k число элементов выборки $(x_1, x_2, ..., x_n)$, которые равны числу k, то есть m_k - это частота значений k. Ясно, что множество различных значений k, которые встречаются в выборке, будет конечным. То есть, только конечное число частот m_k не будет равно нулю.

Вероятность получения выборки $(x_1, x_2, ..., x_n)$ будет равна:

$P(x_1, x_2, ..., x_n, \lambda) = \prod_{k=0}^{\infty} \left(\frac{\lambda^k \cdot e^{-\lambda}}{k!} \right)^{m_k}$. Эта вероятность будет функцией неизвестного параметра λ, множество возможных значений которого мы обозначили Θ. Как и в предыдущем примере, прологарифмируем эту функцию:

$$L_n(\lambda) = \ln P(x_1, x_2, ..., x_n, \lambda) = \sum_{k=0}^{\infty} m_k \cdot \ln \left(\frac{\lambda^k \cdot e^{-k}}{k!} \right) =$$

$$= \ln \lambda \sum_{k=0}^{\infty} k \cdot m_k - \lambda \sum_{k=0}^{\infty} m_k - \ln k! .$$

Так как числа k – возможные значения элементов выборки $(x_1, x_2, ..., x_n)$, а m_k - частоты этих возможных значений, то: $\sum_{k=0}^{\infty} k \cdot m_k = \sum_{i=1}^{n} x_i$. Сумма значений частот m_k равна объёму выборки: $\sum_{k=0}^{\infty} m_k = n$. Упростим вид функции $L_n(\lambda)$ и запишем производную этой функции по параметру λ :

$$L_n(\lambda) = \left(\sum_{i=1}^{n} x_i \right) \ln \lambda - n \cdot \lambda - \ln k! ; \qquad \frac{\partial L_n(\lambda)}{\partial \lambda} = \frac{1}{\lambda} \cdot \sum_{i=1}^{n} x_i - n .$$

Согласно обязательной последовательности действий определяемой методом максимального правдоподобия, приравниваем логарифмическую производную к нулю: $\frac{\partial L_n(\lambda)}{\partial \lambda} = 0$ и решаем получившееся уравнение. Решение $\lambda^* = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ мы называем значением точечной оценки $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ числовой характеристики $\vartheta = \lambda$ распределения Пуассона, определённой методом максимального правдоподобия.

Замечание. Числовой параметр λ распределения Пуассона – это математическое ожидание $M\xi$ случайной величины ξ, то есть λ является числовой характеристикой случайной величины ξ.

Вводя определение точечной оценки $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ числовой характеристики ϑ, и разбирая простейшие примеры точечных оценок числовых характеристик, мы интуитивно, без теоретического обоснования назвали среднее арифметическое элементов выборки оценкой математического ожидания. Затем мы проверили, что среднее арифметическое элементов выборки удовлетворяет всем трём требованиям к точечным оценкам, а поэтому назвали его наилучшей оценкой математического ожидания.

То есть, мы теоретически обосновали и проверили справедливость такого выбора оценки математического ожидания. Воистину прав был китайский мыслитель Конфуций, который сказал: «*Когда бросаешь камешек в воду, всегда попадаешь точно в центр круга*».

И, тем не менее, рассмотрим пример 3.2 с целью проследить ещё раз применение метода максимального правдоподобия в случае, когда случайная величина является случайной величиной непрерывного типа, а числовая характеристика ϑ - многомерная $\vartheta=(\vartheta_1,...,\vartheta_l)$

Пример 3.2 Пусть случайная величина ξ подчиняется нормальному закону распределения вероятностей, кратко записываем: $\xi\in\mathcal{N}(m,\sigma)$. Плотность вероятности нормального закона $p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$ зависит от значений числовых параметров m и σ^2. Здесь числовая характеристика ϑ - двумерная, то есть $\vartheta=(\vartheta_1,\vartheta_2)$, где $\vartheta_1=m$ и $\vartheta_2=\sigma^2$. Вероятностный смысл этих числовых характеристик нам известен. Пусть $(x_1,x_2,...,x_n)$ - имеющаяся у нас числовая выборка. Вероятность получения этой выборки $(x_1,x_2,...,x_n)$ будет равна:

$$P(x_1,x_2,...,x_n,m,\sigma^2)=\prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-m)^2}{2\sigma^2}}\right]=\left(\frac{1}{\sqrt{2\pi}}\right)^n\cdot\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2}$$

.

Логарифмируем эту функцию:

$$L_n(m,\sigma^2)=\ln P(x_1,x_2,...,x_n,m,\sigma^2)=n\cdot\ln\frac{1}{\sqrt{2\pi}}-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2$$

Берём частные производные по параметрам m и σ^2, приравниваем их к нулю и получаем систему двух уравнений с двумя неизвестными:

$$\begin{cases}\frac{\partial L_n(m,\sigma^2)}{\partial m}=0\\ \frac{\partial L_n(m,\sigma^2)}{\partial\sigma^2}=0\end{cases};\qquad \begin{cases}0-0+\frac{1}{2\sigma^2}\cdot 2\sum_{i=1}^{n}(x_i-m)=0\\ 0-\frac{n}{2}\cdot\frac{1}{\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-m)^2=0\end{cases}.$$

Решением этой системы, как и ожидалось, будут известные нам точечные оценки $m^*=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ и $(\sigma^2)^*=S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Качество оценки $\bar{x}$ нами уже обсуждались. Оценка S^2 – состоятельная, но смещённая оценка. Смещение оценки ликвидируется заменой S^2 на оценку s^2, то есть, считаем, что оценкой дисперсии σ^2 будет статистика $(\sigma^2)^*=s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

Комментарий

В заключение отметим, что составление статистик для применения их в качестве точечных оценок $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ значений числовых параметров ϑ законов распределения случайной величины и проверка выполнения требований к точечным оценкам предполагает хорошее знание основ теории вероятностей и должна выполняться специалистами. Только после положительного вердикта специалистов можно применять точечные оценки в практической работе.

§4.2. Распределения, используемые в математической статистике

Комментарий

Напомним, исследуется случайная величина ξ. Организация выборки понимается нами как проведение n раз эксперимента G. Результатом этих экспериментов будут выборочные случайные величины $\xi_1, \xi_2, ..., \xi_n$. Значения $(x_1, x_2, ..., x_n)$ этих случайных величин, которые мы наблюдаем, мы называем числовой выборкой. Желая оценить значения ϑ числовых параметров, которыми в основном являются числовыми характеристиками случайной величины, или функциями этих характеристик, мы составляем статистики $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$, которые назвали точечными оценками числовых параметров.

Любая статистика $\vartheta^* = g(\xi_1, \xi_2, ..., \xi_n)$ является случайной величиной. А поэтому для дальнейшего изучения методов и приёмом математической статистики надо знать законы распределения этих статистик – случайных величин. Рассмотрим три, наиболее часто используемых, распределения вероятностей. Эти распределения были получены в разное время основателями математической статистики К. Пирсоном, Стьюдентом (В.Госсетом) и Р.Фишером. Эти распределения предполагают, что генеральная совокупность, из которой организуется выборка, подчиняется нормальному распределению вероятностей.

Предположение о нормальности распределения генеральной совокупности базируется на теоремах Муавра-Лапласа, Леви, Линдеберга, Ляпунова, объединяемых общим термином, предложенным Ляпуновым, - Центральная предельная теорема. Суть этих теорем состоит в том, что любое возможное значение случайной величины ξ является результатом влияния большого числа различных факторов. Каждый фактор, независимо от других, вносит свою «лепту» в формирование общей суммы результатов. Эта «лепта», доля влияния каждого фактора, очень мала по сравнению с общей суммой результатов. Возможные значения сумм результатов являются возможными значениями случайной величины ξ. Если эти ограничения независимости и малости влияния большого числа факторов

выполняются, то случайная величина ξ имеет или нормальное, или «близкое» к нормальному, распределение вероятностей.

Эти довольно общие ограничения, при которых возникает нормальный закон, приводят нас к мысли об универсальности нормального закона. Однако стоит напомнить замечание Липмана, цитируемое Пуанкаре, которое мы, не меняя смысла, немного изменим. «…каждый уверен в универсальности нормального закона. Экспериментаторы – потому, что они думают, что это математическая теорема, математики – потому, что они думают, что это экспериментальный факт». Статистические исследования показывают, что при некоторых ограничениях мы вправе чаще всего ожидать приближённо нормальное распределение.

Теория

1) Распределение Пирсона ("χ^2 – ***распределение***").

Пусть случайные величины $\xi_1,\xi_2,...,\xi_n$ - независимые, одинаково распределённые по нормальному закону $\mathcal{N}(0;1)$. Образуем из этих случайных величин функцию $g(\xi_1,\xi_2,...,\xi_n)=\sum_{k=1}^{n}\xi_k^2$. Эта функция будет случайной величиной, которую обозначим χ_n^2, то есть: $\chi_n^2=\sum_{k=1}^{n}\xi_n^2$.

Будем говорить: случайная величина χ_n^2 имеет распределение вероятностей Пирсона или «распределение хи-квадрат» с параметром n. Плотность вероятности этого распределения имеет вид:

$$k_n(x)=\begin{cases}\dfrac{1}{\Gamma\left(\dfrac{n}{2}\right)\cdot 2^{\frac{n}{2}}}\cdot x^{\frac{n}{2}-1}\cdot e^{-\frac{x}{2}}, & \text{при} \quad x>0\\ 0, & \text{при} \quad x\le 0\end{cases}.$$

Число независимых слагаемых n, образующих случайную величину χ_n^2 называется числом степеней свободы.

Для положительных значений аргумента при $n\le 2$ функция плотности вероятности $k_n(x)$ будет убывающей функцией, а при $n>2$ функция $k_n(x)$ будет иметь максимум при $x=n-2$.

Во многих практических задачах необходимо знать α - вероятность P события $\{\chi_n^2\ge\chi_\alpha^2\}$, или, наоборот, по заданной вероятности α необходимо определить квантиль χ_α^2 такой, что будет выполняться $P(\chi_n^2\ge\chi_\alpha^2)=\alpha$. Для решения этих вопросов практически во всех монографиях, учебниках и пособиях посвящённых математической статистике протабулирована случайная величина χ_α^2 как функция вероятности α и параметра n (чаще всего встречаются таблицы для $1\le n\le 30$).

Определим основные числовые характеристики случайной величины χ_n^2 . Так как для любого номера k дисперсия $D\xi_k = M\xi_k^2 = 1$, то математическое ожидание $M\left[\chi_n^2\right] = M\left[\sum_{k=1}^{n}\xi_k^2\right] = \sum_{k=1}^{n} M\xi_k^2 = n$.

Для вычисления дисперсии случайной величины χ_n^2 воспользуемся формулой $D\left[\chi_n^2\right] = M\left[\chi_n^k\right]^2 - M^2\left[\chi_n^k\right]$.

Надо вычислить значение уменьшаемого:

$$M\left[\chi_n^2\right]^2 = M\left[\sum_{k=1}^{n}\xi_k^2\right]^2 = M\left[\sum_{k=1}^{n}\xi_k^4 + 2\cdot C_n^2\cdot\xi_k^2\cdot\xi_l^2\right],$$

здесь индексы k и l удовлетворяют условию: $1 \le k \ne l \le n$. Ясно, что в силу независимости случайных величин ξ_k^2 и ξ_l^2 , будет выполняться $M\left[\xi_k^2\cdot\xi_l^2\right] = M\xi_k^2\cdot M\xi_l^2 = 1$. Вычислим $M\xi_k^4 = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^4\cdot e^{-\frac{x^2}{2}}dx$. Учтём, что подинтегральная функция - чётная и сделаем замену переменной интегрирования $\frac{x^2}{2} = t$, тогда $x^3 = 2^{3/2}\cdot t^{3/2}$ и $xdx = dt$. Дальнейшая цепочка преобразований не нуждается в объяснениях:

$$M\xi_k^4 = \frac{2}{\sqrt{2\pi}}\int_0 2^{3/2}\cdot t^{3/2}e^{-t}dt = \frac{4}{\sqrt{\pi}}\Gamma\left(5/2\right) = \frac{3}{\sqrt{\pi}}\Gamma\left(1/2\right) = 3.$$

Теперь можно закончить вычисление дисперсии случайной величины χ_n^2:

$$D\left[\chi_n^2\right] = \left[3n + n(n-1)\cdot 1\right] - (n)^2 = 2n.$$

Так как случайная величина χ_n^2 есть сумма независимых одинаково распределённых случайных величин $\left\{\xi_\kappa^2\right\}, k = 1,...,n,$ имеющих конечное математическое ожидание $M\xi_k^2 = 1$, то, согласно теореме Хинчина из Закона больших чисел, приходим к выводу, что $\lim_{n\to\infty} P\left(\left|\frac{1}{n}\chi_n^2 - 1\right| < \varepsilon\right) = 1$, то есть, случайная величина $\frac{1}{n}\chi_n^2$ при $n\to\infty$ сходится по вероятности к единице. Случайная величина $\frac{\eta - M\eta}{\sqrt{D\eta}} = \overset{0}{\eta}$ называется центрированной и нормированной случайной величиной, так как $M\overset{0}{\eta} = 0$ и $D\overset{0}{\eta} = 1$. Если к условиям, обеспечившим обращение к теореме Хинчина, добавить существование у каждой из случайных величин $\left\{\xi_\kappa^2\right\}, k = 1,...,n,$ конечной дисперсии $D\xi_k^2 = 3$, то, согласно теореме Леви из Центральной предельной

теоремы, заключаем что центрированная и нормированная случайная величина $\frac{\chi_n^2 - n}{\sqrt{2n}} = \overset{0}{\chi}{}_n^2$ - асимптотически нормальна. Это означает, что при больших значениях параметра n (n>50) распределение вероятностей случайной величины $\overset{0}{\chi}{}_n^2$ мало отличается от нормального распределения $\mathcal{N}(0;1)$. Значит, для вычисления вероятности наступления события $\left\{a \le \overset{0}{\chi}{}_n^2 < b\right\}$ можно при больших значениях n использовать функцию Лапласа, то есть $P\left(a \le \overset{0}{\chi}{}_n^2 < b\right) = \Phi(b) - \Phi(a)$.

И, в конце, отметим ещё одно полезное свойство "χ_n^2 - распределения", которое в теории вероятностей называется свойством "устойчивости распределения". Случайная величина, являющаяся суммой двух случайных величин, подчиняющихся "χ^2 – распределению" с параметрами m и n, подчиняется "χ^2 – распределению" с параметром $n+m$. Кратко это можно записать так: $\chi_m^2 + \chi_n^2 = \chi_{n+m}^2$.

2) Распределение Стьюдента ("t^2 – ***распределение***").

Пусть случайные величины $\xi_0, \xi_1, \xi_2, ..., \xi_n$ - независимые, одинаково распределённые по нормальному закону $\mathcal{N}(0;1)$. Образуем из этих случайных величин функцию $g(\xi_0, \xi_1, \xi_2, ..., \xi_n) = \frac{\xi_0}{\sqrt{\frac{1}{n}\sum_{k=1}^{n} \xi_k^2}}$. Эта функция будет случайной величиной, которую обозначим t_n, то есть: $t_n = \frac{\xi_0}{\sqrt{\frac{1}{n}\sum_{k=1}^{n} \xi_k^2}}$.

Будем говорить: случайная величина t_n имеет распределение вероятностей Стьюдента или "t-распределение" с параметром n. (Стьюдент – псевдоним английского математика и статистика У. Госсета).

Плотность вероятности этого распределения имеет вид:

$$s_n(x) = \frac{1}{\sqrt{\pi \cdot n}} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Случайную величину t_n можно записать так: $t_n = \dfrac{\xi_0}{\sqrt{\dfrac{1}{n}\chi_n^2}}$. А так как случайная величина $\dfrac{1}{n}\chi_n^2$ при $n \to \infty$ сходится по вероятности к единице, то приходим к выводу, что при больших значениях n плотность вероятности случайной величины t_n мало отличается от плотности вероятности случайной величины ξ_0, то есть от плотности нормального закона $\mathcal{N}(0;1)$.

Так как плотность вероятности $s_n(x)$ - чётная функция, то все существующие начальные моменты нечётного порядка $\alpha_{2l+1} = M[t_n]^{2l+1}$, включая и математическое ожидание Mt_n, случайной величины t_n равны нулю. Из центральных моментов чётного порядка отметим только дисперсию $Dt_n = \dfrac{n}{n-2}$.

При решении практических задач с использованием распределения Стьюдента необходимо по заданному значению α определять значения t_α из условий $P(|t_n| > t_\alpha) = \alpha$ или $P(t_n > t_\alpha) = \alpha$. Так как плотность вероятности распределения Стьюдента чётная функция, то вероятность случайного события $\{|t_n| > t_\alpha\}$ будет вдвое больше вероятности события $\{t_n > t_\alpha\}$. Поэтому при фиксированном n в таблице распределения Стьюдента значение t_α будет одинаковым для условий $P(|t_n| > t_\alpha) = \alpha$ и $P(t_n > t_\alpha) = 2\alpha$. Это необходимо учитывать при пользовании таблицами.

3) Распределение Фишера-Снедекора ("F – ***распределение***").

Пусть случайные величины $\xi_1, \xi_2, ..., \xi_m$ и $\eta_1, \eta_2, ..., \eta_n$ - независимые, одинаково распределённые по нормальному закону $\mathcal{N}(0;1)$. Образуем из этих $n+m$ случайных величин функцию $g(\xi_1, ..., \xi_m, \eta_1, ..., \eta_n) = \dfrac{\dfrac{1}{m}\sum_{k=1}^{m}\xi_k^2}{\dfrac{1}{n}\sum_{k=1}^{n}\eta_k^2}$.

Вспомнив определение случайной величины χ_n^2, упростим вид этой функции и будем говорить, что случайная величина $f_{m,n} = \dfrac{\dfrac{1}{m}\chi_m^2}{\dfrac{1}{n}\chi_n^2}$ имеет распределение вероятностей Фишера-Снедекора или "F-распределение" с параметрами m и n.

Плотность вероятности этого распределения имеет вид:

$$p_{m,n}(x)=\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\left(\frac{m}{n}\right)^{\frac{m}{2}}\cdot x^{\frac{m}{2}-1}\cdot\left(1+\frac{mx}{n}\right)^{-\frac{m+n}{2}}$$, если $x>0$ и, как это видно из определения, $p_{m,n}(x)=0$, если $x\le 0$.

Ясно, что знание вида плотности вероятности распределения Фишера-Снедекора, не может помочь быстро определить f_α такое, при котором вероятность события $\{f_{m,n} > f_\alpha\}$ будет равна α.

Имеются таблицы вероятностей $P(f_{m,n} > f_\alpha)=\alpha$, в которых для выбранной вероятности α и конкретных значений параметров m и n выбираем значение f_α. Так как для определения любого значения f_α нужно иметь три «входных» параметра, то для каждого конкретного значения α составляется отдельная таблица, в которой приводятся различные значения параметров m и n. Обычно встречаются таблицы для значений $\alpha=0{,}05$ и $\alpha=0{,}01$.

§5.2. Интервальные оценки числовых характеристик

Комментарий

Точечная оценка $\vartheta^*=g(\xi_1,\xi_2,...,\xi_n)$ числовой характеристики ϑ случайной величины ξ используется в тех случаях, когда мы не знаем истинного значения характеристики ϑ. По элементам выборки $(x_1,x_2,...,x_n)$ мы вычисляем значение статистики $\vartheta^*=g(x_1,x_2,...,x_n)$ и в дальнейшем пользуемся этим значением вместо неизвестного значения ϑ. Но при этом мы не можем сказать насколько сильно отличается это используемое значение точечной оценки от значения числовой характеристики. Даже, если вычисленное значение точечной оценки совпадёт со значением числовой характеристики, мы этого не будем знать, так как числовая характеристика - это теоретическое понятие и её точное значение мы не знаем.

Возникает предложение: найти такой интервал (α,β), про который мы могли бы сказать с достаточно высокой степенью уверенности γ, что он накрывает неизвестное значение числовой характеристики ϑ.

Например, станок-автомат выпускает шарики для шарикоподшипников. Диаметр шарика это случайная величина ξ и значение математического ожидания $M\xi$ - числовой характеристики, характеризующей качество работы станка, нам неизвестно. О соответствии продукции станка требованиям технологии мы можем судить по величине значения среднего арифметического $\bar{x}$ некоторой выборки шариков из

всей продукции станка. Однако мы понимаем, что для суждения о качестве продукции станка, нам гораздо важнее и ценнее было бы знать с некоторой степенью уверенности некоторый интервал, в котором находится значение m математического ожидания $M\xi$, то есть среднего значения диаметров изготовленных шариков.

Оценивание неизвестных значений числовых характеристик с помощью интервалов называется *интервальным оцениванием.*

Теория

Определение. Пусть ϑ - некоторая числовая характеристика случайной величины ξ. Интервал $I_\gamma = (\alpha, \beta)$, про который можно сказать, что с уверенностью γ выполняется $P(\alpha < \vartheta < \beta) = \gamma$, называется *доверительным интервалом* для числовой характеристики ϑ.

Вероятность γ называется *доверительной вероятностью.* Она выбирается или назначается исследователем или экспериментатором. Из смысла определения доверительного интервала видно, что значения доверительной вероятности γ , как степени уверенности, должны выбираться довольно большими: $(\gamma = 0{,}9; 0{,}95; 0{,}99; ...)$. Ясно, что границы интервала α и β будут случайными – статистиками $\alpha = g_1(\xi_1, \xi_2, ..., \xi_n)$ и $\beta = g_2(\xi_1, \xi_2, ..., \xi_n)$, то есть - функциями выборочной совокупности $\xi_1, \xi_2, ..., \xi_n$.

Рассмотрим решения трёх задач построения доверительных интервалов.

Задача 1.2. Построить доверительный интервал для неизвестного значения m математического ожидания $M\xi$ случайной величины ξ, если значение дисперсии ξ известно, то есть: $D\xi = \sigma^2$.

Замечание. Вряд ли на практике может встретиться задача построения доверительного интервала при таком условии. Но здесь, не отвлекаясь на решение других проблем, возникающих при построении интервала, мы можем сначала познакомиться с самой сутью процесса построения доверительного интервала. А затем мы проанализируем возможные изменения длины интервала, вызываемые увеличением объёма выборки или значения доверительной вероятности.

Пусть в результате проведённого эксперимента мы получили выборочную совокупность $\xi_1, \xi_2, ..., \xi_n$. Составим статистику $\bar{\xi} = g(\xi_1, \xi_2, ..., \xi_n) = \frac{1}{n}\sum_{i=1}^{n} \xi_i$ - среднее арифметическое элементов выборочной совокупности.

Случайная величина $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ является средним значением суммы независимых, одинаково распределённых случайных величин $\xi_1, \xi_2, ..., \xi_n$, имеющих конечную дисперсию.

Знакомясь с Центральной предельной теоремой в курсе "Теория вероятностей", мы из теоремы Леви получили следствие: «Если случайные величины $\xi_1, \xi_2, ..., \xi_n$ - независимы, одинаково распределены и имеют конечную дисперсию, то случайная величина $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$, являющаяся их средним арифметическим, имеет нормальное распределение $\mathcal{N}\left(m; \frac{\sigma}{\sqrt{n}}\right)$». Заметим, что при этом нас не интересует вид закона распределения случайных величин - слагаемых $\xi_1, \xi_2, ..., \xi_n$.

Построим, используя среднее арифметическое $\vec{\xi}$, статистику: $u = \frac{\bar{\xi} - m}{\sigma}\sqrt{n}$. Так как $M\bar{\xi} = M\xi = m$ и $D\bar{\xi} = \frac{D\xi}{n} = \frac{\sigma^2}{n}$, то у этой статистики математическое ожидание $Mu = 0$ и дисперсия $Du = 1$. Значит статистика $u = \frac{\bar{\xi} - m}{\sigma}\sqrt{n}$ - будет центрированной и нормированной суммой случайных величин, удовлетворяющей всем условиям теоремы Леви. Следовательно, статистика $u = \frac{\bar{\xi} - m}{\sigma}\sqrt{n}$ имеет нормальное распределение $\mathcal{N}(0;1)$. Этот вывод позволяет нам закончить построение доверительного интервала для математического ожидания.

По выбранной доверительной вероятности γ определим величину u_γ (квантиль u_γ) такую, что будет выполняться $P(|u| < u_\gamma) = \gamma$. Так как случайная величина u распределена по нормальному закону, $u \in \mathcal{N}(0;1)$, то значение u_γ находим из равенства: $u_\gamma = \Phi^{-1}\left(\frac{\gamma}{2}\right)$, где $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}dz$ - функция Лапласа.

Итак, с уверенностью γ выполняется неравенство $\left|\frac{\vec{\xi} - m}{\sigma}\sqrt{n}\right| < u_\gamma$. Решим это неравенство относительно значения математического ожидания m и получим:

$\bar{\xi} - \frac{u_\gamma \cdot \sigma}{\sqrt{n}} < m < \bar{\xi} + \frac{u_\gamma \cdot \sigma}{\sqrt{n}}$. Если, в соответствии с определением, обозначить α и β статистики $\bar{\xi} \pm \frac{u_\gamma \cdot \sigma}{\sqrt{n}}$ этого двойного неравенства, то мы можем сказать, что с уверенностью γ выполняется $\alpha < m < \beta$, то есть интервал $I_\gamma = (\alpha, \beta)$ с уверенностью γ накрывает оцениваемое нами неизвестное значение m математического ожидания $M\xi$.

По элементам выборки $(x_1, x_2, ..., x_n)$ определим числовые значения границ α и β интервала: $\bar{x} - \frac{u_\gamma \cdot \sigma}{\sqrt{n}} < m < \bar{x} + \frac{u_\gamma \cdot \sigma}{\sqrt{n}}$ и можем сказать, что доверительный интервал для математического ожидания построен.

Комментарий

Значение среднего арифметического $\bar{x}$ элементов выборки всегда будет центром доверительного интервала. Разность $\beta - \alpha = 2\frac{u_\gamma \cdot \sigma}{\sqrt{n}}$ равна длине этого интервала.

Так как функция Лапласа $\Phi(x)$ - возрастающая функция, то увеличение доверительной вероятности γ повлечёт за собой увеличение длины интервала. Желая повысить надёжность результата, мы будем увеличивать величину доверительной вероятности, но тогда может получиться интервал такой большой длины, что он не будет иметь какую-либо практическую ценность.

С увеличением объёма выборки n длина интервала – уменьшается. Но желание увеличить объём выборки может повлечь за собой увеличение материальных затрат и времени на организацию выборки.

Поэтому при построении доверительного интервала, исследователь, исходя из имеющихся у него возможностей, стремится подобрать такие значения γ и n, при которых построенный доверительный интервал имел бы практическую ценность.

При построении доверительных интервалов надо знать закон распределения статистики, которая используется при определении границ интервала. В рассмотренном примере условие того, что нам известно значение σ^2 дисперсии случайной величины ξ позволило применить теорему Леви и, на основании её, для определения значений α и β воспользоваться функцией Лапласа.

Если мы хотим построить доверительный интервал для математического ожидания, а значение дисперсии нам неизвестно, то появляется желание при построении статистики заменить дисперсию её

точечной оценкой s^2 и рассмотреть статистику $\frac{\bar{\xi}-m}{s}\sqrt{n}$. Но нам надо знать закон распределения вероятностей этой статистики. Здесь нам придётся использовать распределения, используемые в математической статистике, которые были определены в §4.2 при предположении, что генеральная совокупность имеет нормальное распределение вероятностей.

Теория

Определим распределения некоторых статистик, которые будем применять не только при построении доверительных интервалов, но и в дальнейшем, при рассмотрении других методов математической статистики.

Согласно предложению Фишера, основанному на выводах из теорем, входящих в Центральную предельную теорему, считаем, что исследуемая случайная величина ξ имеет нормальное распределение вероятностей, то есть $\xi \in \mathcal{N}(m;\sigma)$. Мы имеем выборочную совокупность $\xi_1,\xi_2,...,\xi_n$, члены которой - независимые случайные величины и имеют то же распределение вероятностей, что и ξ. То есть, для любого номера i, $i=1,2,...,n$, справедливо $\xi_i \in \mathcal{N}(m;\sigma)$.

Рассмотрим статистику $\frac{s^2}{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}\frac{\left(\xi_i-\bar{\xi}\right)^2}{\sigma^2}$, которая, как мы видим, состоит из суммы квадратов центрированных и нормированных случайных величин.

Сумма квадратов независимых, одинаково распределённых по закону $\mathcal{N}(0;1)$, случайных величин использовалась при построении случайной величины χ^2. Но в выписанной статистике любые две случайные величины $\frac{\xi_i-\bar{\xi}}{\sigma}$ и $\frac{\xi_j-\bar{\xi}}{\sigma}$ не будут независимыми, так как ξ_j входит в число слагаемых среднего арифметического $\bar{\xi}$ первой дроби $\frac{\xi_i-\bar{\xi}}{\sigma}$, а ξ_i входит в число слагаемых среднего арифметического $\bar{\xi}$ второй дроби $\frac{\xi_j-\bar{\xi}}{\sigma}$.

Преобразуем статистику $\frac{s^2}{\sigma^2}$ так, чтобы она стала суммой квадратов независимых случайных величин. Проводимые преобразования не нуждаются в пояснениях.

$$\frac{s^2}{\sigma^2}=\frac{1}{n-1}\sum_{i=1}^{n}\frac{\left((\xi_i-m)-(\bar{\xi}-m)\right)^2}{\sigma^2}=$$

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}\frac{(\xi_i-m)^2}{\sigma^2}-2\cdot\frac{\bar{\xi}-m}{\sigma^2}\cdot\sum_{i=1}^{n}(\xi_i-m)+n\cdot\frac{(\bar{\xi}-m)^2}{\sigma^2}\right]=$$

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}\frac{(\xi_i-m)^2}{\sigma^2}-2n\cdot\frac{\bar{\xi}-m}{\sigma^2}\cdot(\bar{\xi}-m)+n\cdot\frac{(\bar{\xi}-m)^2}{\sigma^2}\right].$$

Окончательно запишем: $\frac{s^2}{\sigma^2}=\frac{1}{n-1}\left[\sum_{i=1}^{n}\frac{(\xi_i-m)^2}{\sigma^2}-\frac{(\bar{\xi}-m)^2}{\frac{\sigma^2}{n}}\right]$.

Так как $\xi_i\in\mathcal{N}(m;\sigma)$, то для любого номера i , $i=1,2,...,n$, будет справедливо $\frac{\xi_i-m}{\sigma}\in\mathcal{N}(0;1)$. Следовательно, в последнем равенстве уменьшаемое, как сумма квадратов случайных величин, имеет распределение χ_n^2. Вычитаемое в этом равенстве есть квадрат случайной величины, имеющей так же нормальное распределение $\mathcal{N}(0;1)$, значит, оно имеет распределение χ_1^2. Таким образом, если исследуемая случайная величина имеет нормальное распределение, то статистика $\frac{s^2}{\sigma^2}$ будет иметь распределение $\frac{1}{n-1}\left(\chi_n^2-\chi_1^2\right)$. Используя свойство устойчивости распределения χ^2, окончательно запишем: $\frac{s^2}{\sigma^2}=\frac{1}{n-1}\chi_{n-1}^2$.

Рассмотрим теперь статистику $\frac{\bar{\xi}-m}{s}\sqrt{n}$, которую мы предлагаем использовать для построения доверительного интервала.

Ясно, что $\frac{\bar{\xi}-m}{s}\sqrt{n}=\frac{\frac{\bar{\xi}-m}{\sigma}\sqrt{n}}{\frac{s}{\sigma}}=\frac{\frac{\bar{\xi}-m}{\sigma}\sqrt{n}}{\sqrt{\frac{s^2}{\sigma^2}}}$. Числитель этой дроби имеет близкое к нормальному распределению $\mathcal{N}(0;1)$, а подкоренное выражение знаменателя, как мы только что показали, имеет распределение $\frac{1}{n-1}\chi_{n-1}^2$. Но тогда мы получаем, статистика $\frac{\bar{\xi}-m}{s}\sqrt{n}$ имеет распределение Стьюдента с параметром $n-1$, то есть: $\frac{\bar{\xi}-m}{s}\sqrt{n}=t_{n-1}$.

Теория

Задача 1.2*а*. Построить доверительный интервал для неизвестного значения m математического ожидания $M\xi$ случайной величины ξ.

В такой общей постановке задача построения доверительного интервала предполагается только одно ограничение: мы считаем, что генеральная совокупность, из которой отбирается выборка, подчиняется нормальному распределению вероятностей.

Для построения доверительного интервала будем использовать статистику $\frac{\bar{\xi}-m}{s}\sqrt{n}$, распределением которой будет распределение Стьюдента (”*t*-распределение”) с параметром $n-1$.

Для выбранной доверительной вероятности γ определим величину t_γ такую, что будет выполняться $P\left(|t_{n-1}|<t_\gamma\right)=\gamma$. Квантиль t_γ находим в таблицах распределения Стьюдента (таблицах ”*t*-распределения”). Как и при решении предыдущей задачи, мы говорим, что с уверенностью γ выполняется неравенство $\left|\frac{\bar{\xi}-m}{\sigma}\sqrt{n}\right|<t_\gamma$. Решаем это неравенство относительно значения математического ожидания m и получаем аналогичный результат:

$$\bar{\xi}-\frac{t_\gamma\cdot\sigma}{\sqrt{n}}<m<\bar{\xi}+\frac{t_\gamma\cdot\sigma}{\sqrt{n}}.$$

Левая и правая части этого двойного неравенства - это статистики, числовые значения которых мы определяем по элементам выборки $(x_1, x_2, ..., x_n)$.

То есть мы с уверенностью γ можем сказать, что значение m математического ожидания случайной величины ξ находится в интервале $(\alpha;\beta)$, где $\alpha=\bar{x}-\frac{t_\gamma\cdot\sigma}{\sqrt{n}}$ и $\beta=\bar{x}+\frac{t_\gamma\cdot\sigma}{\sqrt{n}}$.

Замечание. Мы всегда должны помнить, что неизвестное нам значение m математического ожидания $M\xi$ – это число, которое мы оцениваем или значением $\bar{x}$ точечной оценки $\bar{\xi}$, или доверительным интервалом $(\alpha;\beta)$. Поэтому мы говорим, что «интервал $(\alpha;\beta)$ *накрывает* неизвестное значение m», а не - «значение m *попадает* в интервал $(\alpha;\beta)$».

Практика

Пример 2*а*.1 и 2*б*.1 Построим доверительные интервалы для математического ожидания $M\xi$ - среднего роста мужчин в *см*, используя элементы двух выборок, объёмы которых равны n=115 и n=906.

При первичной обработке статистических данных были вычислены значения точечных оценок математических ожиданий, дисперсий и средних квадратических отклонений:

а) $n_1 = 115, \quad \bar{x} = 166{,}61, \quad s^2 = 36{,}28, \quad s = 6{,}02;$

б) $n_2 = 906, \quad \bar{x} = 166{,}77, \quad s^2 = 34{,}70, \quad s = 5{,}89;$

Выберем значение доверительной вероятности - $\gamma = 0{,}95$. По таблицам распределения Стьюдента (“***t***-распределения”) мы должны, учитывая объёмы выборок $n_1 = 115$ и $n_2 = 906$, определить значения квантиля t_γ. Так как при больших значениях n $(n \geq 100)$ значения t_n изменяются незначительно, то значения t_γ определим из условий *а)* $P(|t_{100}| < t_\gamma) = 0{,}95$ и *б)* $P(|t_\infty| < t_\gamma) = 0{,}95$. Выписываем из таблиц: *а)* $t_\gamma = 1{,}98$ и *б)* $t_\gamma = 1{,}96$, вычисляем значения дроби $\frac{t_\gamma \cdot s}{\sqrt{n}}$ для соответствующих значений точечных оценок рассматриваемых выборок

а) $\frac{t_\gamma \cdot s}{\sqrt{n}} = 1{,}11$ и *б)* $\frac{t_\gamma \cdot s}{\sqrt{n}} = 0{,}38$ и записываем доверительные интервалы:

а) $(166{,}61 - 1{,}11; 166{,}61 + 1{,}11)$ или $165{,}50 < m < 167{,}72;$

б) $(166{,}77 - 0{,}38; 166{,}77 + 0{,}38)$ или $166{,}39 < m < 167{,}15.$

Краткий анализ получившихся доверительных интервалов иллюстрирует наш вывод о том, что увеличение объёма выборки приводит к уменьшению длины доверительного интервала.

Увеличим значение доверительной вероятности - $\gamma = 0{,}99$ и снова проведём вычисления границ интервалов рассматриваемых выборок.

В соответствии с условиями *а)* $P(|t_{100}| < t_\gamma) = 0{,}99$ и *б)* $P(|t_\infty| < t_\gamma) = 0{,}99$ выписываем из таблиц: *а)* $t_\gamma = 2{,}617$ и *б)* $t_\gamma = 2{,}576$ и вычисляем значения дроби $\frac{t_\gamma \cdot s}{\sqrt{n}}$ для соответствующих значений точечных оценок *а)* $\frac{t_\gamma \cdot s}{\sqrt{n}} = 1{,}47$ и *б)* $\frac{t_\gamma \cdot s}{\sqrt{n}} = 0{,}50$. Записываем доверительные интервалы:

а) $(166{,}61 - 1{,}47; 166{,}61 + 1{,}47)$ или $165{,}14 < m < 168{,}08;$

б) $(166{,}77 - 0{,}50; 166{,}77 + 0{,}50)$ или $166{,}27 < m < 167{,}27.$

Сравнивая для каждой выборки доверительные интервалы, построенные для доверительных вероятностей $\gamma = 0{,}95$ и $\gamma = 0{,}99$, мы видим, что увеличение доверительной вероятности приводит к увеличению длины доверительного интервала. Естественно нам хочется получить как можно более надёжный результат, для этого надо выбирать большое значение доверительной вероятности, а это влечёт увеличение длины интервала. Но

мы должны иметь в виду, что может получиться доверительный интервал такой большой длины, что он не будет иметь никакой практической ценности.

Теория

Задача 2.2. Построить доверительный интервал для неизвестного значения σ^2 дисперсии $D\xi$ случайной величины ξ.

При предположении о нормальном распределении генеральной совокупности мы показали, что статистика $\frac{(n-1)\cdot s^2}{\sigma^2}$ будет подчиняться распределению Пирсона (“χ^2-распределению”) с параметром $(n\text{-}1)$.

Определение доверительного интервала требует, чтобы по выбранной доверительной вероятности γ мы определили величины χ_1^2 и χ_2^2 такие, при которых будет выполняться $P\left(\chi_1^2 < \chi_{n-1}^2 < \chi_2^2\right) = \gamma$.

При построении доверительного интервала для математического ожидания в виду симметричности распределения Стьюдента условия $P\left(|t_{n-1}| < t_\gamma\right) = \gamma$ было достаточно для определения границ интервала. Распределение Пирсона (“χ^2-распределение”) – не симметрично, а поэтому можно указать много пар чисел χ_1^2 и χ_2^2, при которых будет выполняться условие $P\left(\chi_1^2 < \chi_{n-1}^2 < \chi_2^2\right) = \gamma$. То есть для определения двух неизвестных χ_1^2 и χ_2^2 мы имеем одно уравнение. Нужны дополнительные условия.

Наиболее часто применяются следующее условие.

Мы считаем, что вероятности $P\left(0 < \chi_{n-1}^2 < \chi_1^2\right)$ и $P\left(\chi_2^2 < \chi_{n-1}^2 < \infty\right)$ равны между собой и равны $\frac{1-\gamma}{2}$. То есть величина χ_1^2 по таблицам “χ^2-распределения”, в виду их особенностей, определяется из условия $P\left(\chi_1^2 \le \chi_{n-1}^2\right) = \gamma + \frac{1-\gamma}{2} = \frac{1+\gamma}{2}$, а величина χ_2^2 определяется из условия $P\left(\chi_2^2 \le \chi_{n-1}^2\right) = \frac{1-\gamma}{2}$. Определив значения χ_1^2 и χ_2^2, закончим построение доверительного интервала. Двойное неравенство $\chi_1^2 < \frac{(n-1)\cdot s^2}{\sigma^2} < \chi_2^2$ выполняется с вероятностью γ. Решив его относительно σ^2, получим $\frac{(n-1)\cdot s^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)\cdot s^2}{\chi_1^2}$. Считая, что $\frac{(n-1)\cdot s^2}{\chi_2^2} = \alpha$ и $\frac{(n-1)\cdot s^2}{\chi_1^2} = \beta$ мы можем сказать, что доверительный интервал для неизвестного значения σ^2

дисперсии, при дополнительном условии равенства вероятностей событий $\{0 < \chi^2_{n-1} < \chi^2_1\}$ и $\{\chi^2_2 < \chi^2_{n-1} < \infty\}$, построен.

Практика

Пример 2*а*.1 и 2*б*.1 Построим доверительные интервалы для дисперсии $D\xi$ случайной величины ξ - среднего роста мужчин, используя элементы первой и второй выборок, объёмы которых равны n=115 и n=906 . При первичной обработке статистических данных были вычислены значения точечных оценок дисперсий $s_1^2 = 36{,}28$ и $s_2^2 = 34{,}70$.

Пусть выбрана доверительная вероятность γ =0,95. Для первой выборки мы можем «округлить» число степеней свободы до n=101, и считать параметр распределения n-1 равным 100. По таблицам "χ^2 - распределения" определяем χ^2_1 и χ^2_2 , считая, что $P\left(\chi^2_1 \geq \chi^2_{n-1}\right) = \frac{1+\gamma}{2} = 0{,}975$ и $P\left(\chi^2_2 \geq \chi^2_{n-1}\right) = \frac{1-\gamma}{2} = 0{,}025$. Выписав из таблиц: $\chi^2_1 = 74{,}22$ и $\chi^2_2 = 129{,}60$, вычисляем границы интервала: $\alpha = \frac{114 \cdot 36{,}28}{129{,}60}$ и $\beta = \frac{114 \cdot 36{,}28}{74{,}22}$.

Окончательно получаем, что с уверенностью γ =0,95 значение σ^2 дисперсии $D\xi$ случайной величины ξ находится в интервале: $\sigma^2 \in (31{,}91; 55{,}72)$.

Так как объём второй выборки n=906 – большой, то для построения доверительного интервала в этом случае используем асимптотическую нормальность случайной величины χ^2_n. Это означает, что при больших объёмах выборки центрированная и нормированная случайная величина $\tau_n = \frac{\chi^2_n - n}{\sqrt{2n}}$ распределена по закону близкому к нормальному закону $\mathcal{N}(0;1$). При доверительной вероятности γ =0,95 найдём значение τ_γ из условия: $P\left(|\tau_n| < \tau_\gamma\right) = \gamma$. Так как $\gamma = 2\Phi(\tau_m)$, то определяем по таблицам значений функции Лапласа: $\tau_n = \Phi^{-1}\left(\frac{\gamma}{2}\right) = \Phi^{-1}(0{,}475) = 1{,}96$. Значит, с уверенностью γ =0,95 выполняется двойное неравенство: $-1{,}96 < \frac{\chi^2_{n-1} - (n-1)}{\sqrt{2(n-1)}} < 1{,}96$. Решаем это неравенство относительно χ^2_{n-1} и, имея в виду, что $\chi^2_{n-1} = \frac{(n-1) \cdot s^2}{\sigma^2}$, получаем доверительный интервал для дисперсии:

$\frac{(n-1)\cdot s^2}{(n-1)-1{,}96\cdot\sqrt{2(n-1)}}<\sigma^2<\frac{(n-1)\cdot s^2}{(n-1)+1{,}96\cdot\sqrt{2(n-1)}}$. Подставим в это неравенство значения $n-1=905$, $s_2^2=34{,}70$ и, проведя вычисления, получим доверительный интервал $(31{,}77;38{,}22)$ рассматриваемого примера, который с уверенностью $\gamma=0{,}95$ накрывает значение σ^2 дисперсии $D\xi$.

«… большей частью важнейшие жизненные вопросы являются на самом деле лишь задачами из теории вероятностей»
П. Лаплас

Модуль третий

Статистическая проверка гипотез

§1.3 Общая постановка задачи статистической проверки гипотез

Теория

Предположим, что наблюдая некоторый количественный признак, которым обладают объекты, составляющие генеральную совокупность, мы формулируем некоторое предположение, справедливость которого надо проверить. Это предположение мы назовём *основной гипотезой* и будем обозначать H_0. Естественно, сформулировав наше предположение в форме гипотезы H_0, мы можем сформулировать и противоположное предположение, которое мы будем называть *альтернативной гипотезой* и обозначать H_1. Альтернативных гипотез может быть больше чем одна, и мы их будем обозначать H_1, H_2,… Ясно, что в реальности справедливо только одно предположение, только одна гипотеза.

Наше предположение, то есть основную гипотезу H_0, мы должны проверить и, если оно подтвердится, то мы сможем в дальнейшем пользоваться этим предположением как действительным, реальным фактом.

Но мы должны помнить, что у нас всегда имеет место случайность, статистически устойчивая, но всё же – случайность и поэтому при проверке справедливости гипотезы H_0 мы должны пользоваться понятиями и методами теории вероятностей, а именно – основными положениями, которые следуют из Закона больших чисел и Центральной предельной теоремы.

Для проверки справедливости основной гипотезы H_0 мы подбираем *критерий проверки гипотезы* T, который является статистикой $T=g(\xi_1,\xi_2,...,\xi_n)$. Пусть множество действительных чисел $Q\cup S$ является множеством возможных значений статистики $T=g(\xi_1,\xi_2,...,\xi_n)$. Проводится эксперимент, в результате которого может наступить либо

событие $\{T \in Q\}$, либо событие $\{T \in S\}$. Но мы заранее договариваемся, что если наступит событие $\{T \in S\}$, то гипотеза H_0 – неверна, её надо отклонить и принять альтернативную гипотезу H_1, а если наступит событие $\{T \in Q\}$, то наша гипотеза H_0 верна и её надо принять.

Мы считаем, что при верной гипотезе H_0 событие S может наступить, но вероятность α этого события $P(S \mid H_0)$ - мала. То есть, так как у нас имеет место случайность, то при верной в реальности гипотезе H_0 мы будем изредка ошибаться. Это будет тогда, когда будет наступать событие $\{T \in S\}$, и вероятность $P(T \in S \mid H_0) = \alpha$.

Сформулируем чётко последовательность действий при статистической проверке справедливости основной гипотезы H_0.

После того как гипотеза H_0 сформулирована подбирается критерий T проверки её справедливости. Так как критерий является статистикой $T = g(\xi_1, \xi_2, ..., \xi_n)$, то у него есть распределение вероятностей его значений. Пусть это плотность вероятности $p(x \mid H_0)$. Мы считаем, что при верной гипотезе H_0 с маленькой вероятностью α значение критерия T попадает в область S, то есть наступит случайное событие $\{T \in S\}$. Область S будем называть *критической областью* и обозначать $S_{кр.}$. Все остальную область возможных значений критерия T будем называть *областью допустимых значений* и обозначать $Q_{доп.}$. Определив области значений критерия $S_{кр.}$ и $Q_{доп.}$, проводим эксперимент и получаем числовую выборку $(x_1, x_2, ..., x_n)$. По этой выборке подсчитываем наблюдаемое значение критерия $T_{набл.} = g(x_1, x_2, ..., x_n)$ и смотрим: в какую область попало это наблюдаемое значение критерия $T_{набл.}$.

Правило принятия решений:

1) Если значение критерия $T_{набл.}$ попало в область $S_{кр.}$, то есть, если наступило событие $\{T_{набл.} \in S_{кр.}\}$, то, так как при справедливой гипотезе H_0 это событие маловероятное (α - мало), мы говорим: «*Гипотеза H_0 отклоняется с уровнем значимости α*».

2) Если значение критерия $T_{набл.}$ попало в область $Q_{доп.}$, то есть, если наступило событие $\{T_{набл.} \in Q_{доп.}\}$, то мы говорим: «*У нас нет оснований отклонять гипотезу H_0. Гипотеза H_0 - принимается*».

Так “осторожно” во втором случае мы говорим потому, что наступление события $\{T_{набл.} \in Q_{доп.}\}$ - результат однократного проведения эксперимента и получившейся при этом выборки $(x_1, x_2, ..., x_n)$. И не исключено, что при повторных проведениях эксперимента, и для новых выборок мы это событие больше наблюдать не будем.

Для сформулированных основной H_0 и альтернативной H_1 гипотез критерий $T = g(\xi_1, \xi_2, ..., \xi_n)$, который мы применяем для проверки справедливости этих гипотез, имеет, соответственно, плотности вероятности $p(x \mid H_0)$ и $p(x \mid H_1)$. В результате проведения эксперимента, вычислив значение $T_{набл.} = g(x_1, x_2, ..., x_n)$, мы можем наблюдать наступление одного из двух событий $\{T_{набл.} \in Q_{доп.}\}$ или $\{T_{набл.} \in S_{кр.}\}$.

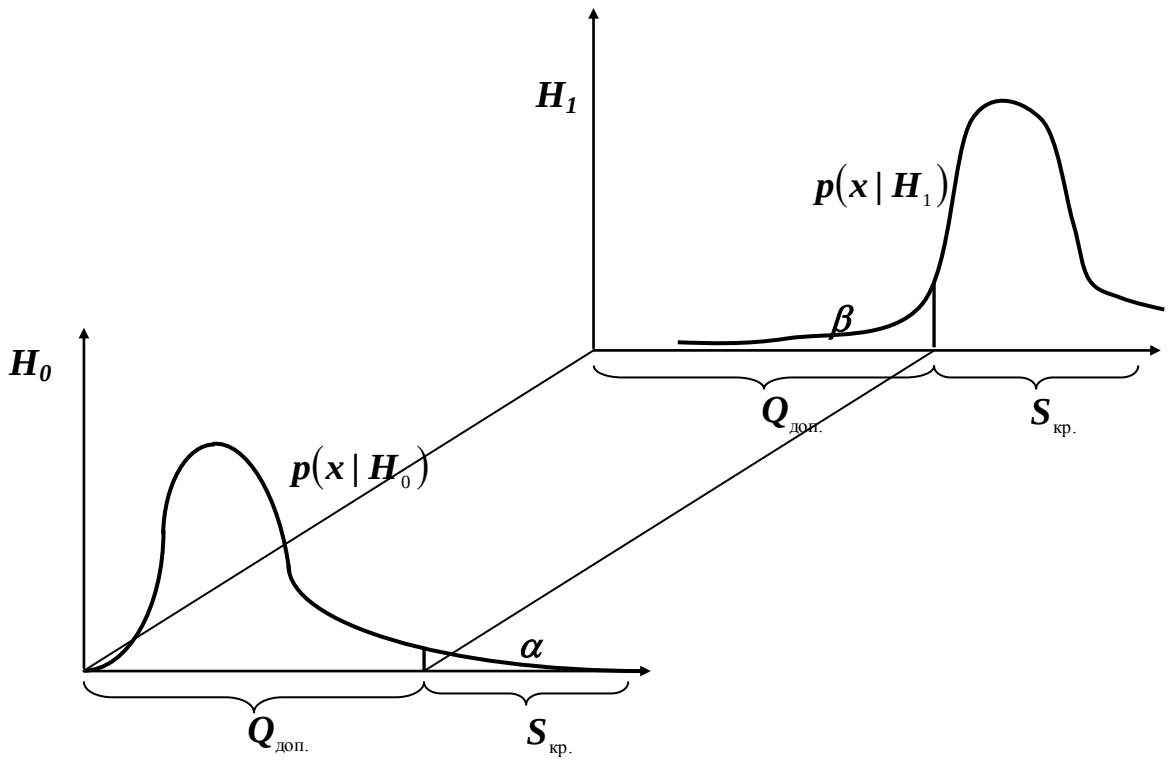


Рис.3.1

Из рис. 3.1 видно, что, проводя процедуру статистической проверки гипотез, мы можем встретить одну из четырёх ситуаций, которые приведены в таблице:

	эксперимент	
	$T \in Q_{доп.}$	$T \in S_{кр.}$
верна H_0	1) всё хорошо	2) ошибка I рода
верна H_1	4) ошибка II рода	3) всё хорошо

Рассмотрим каждую из этих возможных ситуаций отдельно.

1) В реальности справедлива гипотеза H_0, проведя эксперимент и вычислив значение критерия, мы наблюдаем наступление события

$\{T_{набл.} \in Q_{доп.}\}$. По правилу принятия решений мы принимаем гипотезу H_0, Здесь всё хорошо. Но надо помнить, что принятие гипотезы свидетельствует лишь о том, что она не находится в противоречии с имеющимися статистическими данными. Мы не доказываем справедливость, или верность гипотезы H_0, мы лишь констатируем, что наша гипотеза согласуется с имеющимися данными опыта, эксперимента. То есть, результаты эксперимента не противоречат нашей гипотезе, согласуются с нашей гипотезой.

2) В реальности справедлива гипотеза H_0, проведя эксперимент и вычислив значение критерия, мы наблюдаем наступление события $\{T_{набл.} \in S_{кр.}\}$. По правилу принятия решений мы должны отклонить верную гипотезу H_0. При этом мы совершаем ошибку, которая называется *ошибкой первого рода*. Вероятность ошибки первого рода равна вероятности события $\{T_{набл.} \in S_{кр.}\}$, то есть она равна α.

3) В реальности справедлива гипотеза H_1, проведя эксперимент и вычислив значение критерия, мы наблюдаем наступление события $\{T_{набл.} \in S_{кр.}\}$. По правилу принятия решений мы отклоняем основную гипотезу H_0 как ошибочную и принимаем альтернативную гипотезу H_1, которая имеет место быть в реальности. Здесь всё хорошо.

4) В реальности справедлива гипотеза H_1, проведя эксперимент и вычислив значение критерия, мы наблюдаем наступление события $\{T_{набл.} \in Q_{доп.}\}$. По правилу принятия решений мы должны принять гипотезу H_0, которая в реальности неверна. При этом мы совершаем ошибку, которая называется *ошибкой второго рода*. Вероятность ошибки второго рода равна $P(T_{набл.} \in Q_{доп.} \mid H_1) = \beta$.

Пример

Рассмотрим практическую иллюстрацию ситуаций, которые могут произойти во взаимоотношениях «производителя» и «покупателя».

Производитель желает реализовать партию произведённого им товара. Покупатель желает приобрести эту партию товара. В зависимости от доли плохого товара в партии, партию можно признать «хорошей» или признать «плохой». Значит, у нас есть две гипотезы: H_0 –партия товара «хорошая», и H_1 - партия товара «плохая».

Для проверки качества партии товара производитель и покупатель решают с помощью контрольного осмотра некоторой выборки товара из общей партии сделать оценку качества всей партии. Они договариваются о критерии проверки справедливости сформулированных гипотез. Наудачу выбираются десять предметов, и оценивается их качество. Если среди этих предметов окажется не более одного «плохого», то вся партия признаётся «хорошей». И покупатель забирает всю партию. Если среди отобранных

предметов окажется два, или более двух, «плохих» то вся партия признаётся «плохой» и покупатель отказывается от приобретения её.

С позиции производителя гипотеза H_0 будет основной гипотезой, так как он стремится реализовать свой товар и признание партии «плохой», хотя она в действительности «хорошая», для него чревато большими материальными и моральными потерями. С позиции покупателя гипотеза H_1 будет основной гипотезой, так как ему лучше забраковать «хорошую» партию, чем приобрести «плохую», и в этом случае он не пострадает ни материально, ни морально.

Допустим, что мы симпатизируем производителю, то есть гипотеза H_0 для нас основная и она верна в реальности. Провели контроль отобранных десяти предметов и все они оказались «хорошими». По правилу принятия решений вся партия объявляется «хорошей» и покупатель обязан её забрать. В этом случае и производитель, и покупатель будут довольны.

Если среди отобранных десяти предметов три оказались «плохими», то вся партия объявляется «плохой». Это будет ошибка первого рода. Покупатель не покупает партию. Производитель проигрывает, он имеет материальный ущерб и страдает морально.

Пусть в реальности справедлива гипотеза H_1. По результатам контроля отобранных десяти предметов три оказались «плохими». По правилу принятия решений вся партия объявляется «плохой». Покупатель уходит без покупки, а производитель должен проверить всю партию и подумать о причинах большого числа «плохих» предметов в партии.

Если же все десять предметов оказались «хорошими», то вся партия объявляется «хорошей» и покупатель обязан её приобрести. Совершается ошибка второго рода. Покупатель проигрывает, он страдает и материально, и морально. Производитель вроде бы должен быть довольным, но вряд ли этот покупатель обратится за товаром к этому производителю в следующий раз.

Комментарий

Естественно, подбирая критерий проверки основной гипотезы H_0, мы должны стремиться, чтобы для выбранного критерия вероятности ошибок первого и второго рода были по возможности минимальными. Для выбранного критерия $T = g(\xi_1, \xi_2, ..., \xi_n)$ уменьшение одной из вероятностей α или β, как видно из рис. 3.1, приводит к увеличению другой. Значит, если выбранный критерий нас не удовлетворяет по причине большого значения вероятности или ошибки первого рода α, или ошибки второго рода β, то надо подбирать новый критерий.

Возникает проблема: как сравнивать различные критерии для проверки справедливости гипотезы H_0 и как выбирать из них наиболее

«хороший»? Решению этого вопроса посвящены работы Неймана и Е.Пирсона. Суть теории Неймана-Пирсона состоит в следующем.

Для того чтобы мы назвали критерий «хорошим» надо потребовать, чтобы при использовании этого критерия вероятность α *отклонить* гипотезу H_0, когда она *верна* в реальности, была мала, а вероятность $1-\beta$ *отклонить* гипотезу H_0, когда она *неверна* в реальности, была велика.

Вероятность $1-\beta$ называют *функцией мощности* выбираемого критерия. При выборе из двух критериев T_1 и T_2, у которых одна и та же вероятность α отклонить верную в реальности гипотезу H_0 следует признать более «хорошим» тот, который даёт большую вероятность $1-\beta$ отклонить H_0, когда эта гипотеза неверна.

Выбор наилучшего критерия проводится при наличии нескольких альтернативных гипотез. Если при всех альтернативных гипотезах критическое множество $S_{кр.}$, при котором достигается максимальное значение $1-\beta$, одинаково для всех этих гипотез, то такой критерий называется *наиболее мощным* среди всех критериев выбранного уровня значимости α. К сожалению, такое положение, то есть существование наиболее мощного критерия, встречается редко.

Поэтому мы должны иметь в виду, что чаще всего встречающиеся статистические гипотезы можно систематизировать, то есть разбить на несколько групп. Формулировки гипотез, принадлежащих одной группе, звучат практически одинаково, а потому перечень применяющихся критериев будет не очень большим. Если для всех одинаковых по смыслу гипотез существует наилучший критерий, то он описан во всех книгах посвящённых математической статистике. С другой стороны, к такому критерию обычно нетрудно прийти, исходя из здравого смысла, основанного на знании теории вероятностей и, прежде всего, её предельных теорем. Это мы попытаемся сделать в рассматриваемых примерах.

Теория

Все практические задачи статистической проверки гипотез, с которыми может встретиться экспериментатор, можем разбить на четыре группы.

К первой группе относятся задачи статистической проверки гипотезы о равенстве неизвестного значения числового параметра исследуемого качественного признака, то есть случайной величины ξ, некоторому фиксированному числу.

Пусть закон распределения вероятностей $P(x,\vartheta)$ случайной величины ξ зависит от числового параметра ϑ, однозначно определяющего это распределение. Множество Θ - множество возможных значений параметра

ϑ. Это множество представим в виде $\Theta = \Theta_0 \cup \Theta_1$, где $\Theta_0 \cap \Theta_1 = \varnothing$. Формулируем основную гипотезу: $H_0 : \vartheta \in \Theta_0$ и альтернативную гипотезу: $H_1 : \vartheta \in \Theta_1$. Гипотеза $H_0 : \vartheta \in \Theta_0$ называется *простой* гипотезой, если множество Θ_0 состоит из одного элемента ϑ_0, то есть $H_0 : \vartheta = \vartheta_0$. Гипотезу, не являющуюся простой, будем называть *сложной*. Альтернативная гипотеза может быть простой $H_1 : \vartheta = \vartheta_1$ или сложной, например, $H_1 : \vartheta \neq \vartheta_0$.

В задачах этой группы основные гипотезы – простые. Принятие простой гипотезы однозначно определяет распределение $P(x, \vartheta)$.

Ко второй группе относятся задачи статистической проверки гипотезы о равенстве неизвестных значений одноимённых числовых характеристик исследуемого качественного признака двух различных случайных величин ξ и η.

Если законы распределения вероятностей случайных величин ξ и η зависят от числового параметра, значения которого равны, соответственно, ϑ_1 и ϑ_2, мы формулируем основную гипотезу $H_0 : \vartheta_1 = \vartheta_2$. Альтернативная гипотеза может быть такой: $H_1 : \vartheta_1 \neq \vartheta_2$, или $H_1 : \vartheta_1 > \vartheta_2$.

К третьей группе относятся задачи статистической проверки гипотезы о виде закона распределения вероятностей исследуемой случайной величины.

Анализируя вид гистограммы, построенной по выборке $(x_1, x_2, ..., x_n)$, мы формулируем гипотезу о конкретном виде закона распределения случайной величины: H_0 : $P(x)$ - *закон распределения вероятностей случайной величины* ξ. Или, что одно и то же, H_0 : *Случайная величина* ξ *имеет функцию распределения* $F(x)$.

К четвёртой группе относятся задачи статистической проверки гипотезы *а)* о совпадении законов распределения вероятностей исследуемых одноимённых случайных величин ξ и η, и *б)* о независимости законов распределения вероятностей исследуемых одноимённых случайных величин ξ и η.

В случае *а)* мы, анализируя вид гистограмм, построенных по выборкам $(x_1, x_2, ..., x_{n_1})$ и $(y_1, y_2, ..., y_{n_2})$, формулируем гипотезу H_0 : *Законы распределения вероятностей* $P_1(x)$ и $P_2(y)$ *случайных величин* ξ и η - *совпадают.* Или, что одно и то же, H_0 : *Функции распределения* $F_1(x)$ *и* $F_2(y)$ *случайных величин* ξ и η - *одинаковые.*

В случае *б)* мы, анализируя вид корреляционной таблицы, построенной по двумерной выборке $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, формулируем гипотезу H_0 : *Случайные величины* ξ *и* η - *независимые, то*

есть закон распределения $P(x,y)$ *случайного вектора* (ξ,η) *равен произведению частных законов распределения компонент:* $P(x,y)=P_1(x)\cdot P_2(y)$. Или, что одно и то же, H_0 **:** *Функция распределения вектора* (ξ,η) *равна произведению частных функций распределения компонент* ξ *и* η*, то есть:* $F(x,y)=F_1(x)\cdot F_2(y)$.

Рассмотрим конкретные примеры решения задач статистической проверки гипотез каждого из четырёх типов. Процесс решения задачи состоит в формулировании основной и альтернативной гипотез и построении критерия проверки справедливости основной гипотезы.

§2.3 Задачи первого типа статистической проверки гипотез

Теория

Задача 1.3 Проверка гипотезы о равенстве неизвестного значения вероятности наступления случайного события $P(A)$ числу p_0.

Эту задачу сформулируем так: проводится испытание, в результате которого с вероятностью p_0 может наступить или случайное событие A, или с вероятностью $q_0=1-p_0$ случайное событие $\overline{A}$. Определим случайную величину ξ, как измеримое отображение множества Ω в R_1, следующим образом: $\xi=\begin{cases}0, & \text{если наступило событие } \overline{A}\\ 1, & \text{если наступило событие } A\end{cases}$.

Случайная величина ξ - результат однократного проведения опыта, называемая нами «бернуллиевской», $\xi\in B_1(p_0)$, имеет ряд распределения

ξ	0	1
	q_0	p_0

. Здесь вероятность p_0 - числовая характеристика бернуллиевского распределения, значение которой нам неизвестно.

Формулируем основную гипотезу $H_0: P(A)=p_0$. Альтернативную гипотезу запишем так: $H_1: P(A)\neq p_0$. Гипотеза H_0 – простая, гипотеза H_1 - сложная.

Построим критерий $T=g(\xi_1,\xi_2,...,\xi_n)$ проверки справедливости H_0. Если эксперимент, в котором наступают или случайное событие A, или случайное событие $\overline{A}$, проведём n раз, то получим выборочную последовательность $\xi_1,\xi_2,...,\xi_n$ независимых бернуллиевских случайных величин. Известно, что статистика $\sum_{i=1}^{n}\xi_i$ подчиняется биномиальному распределению $B_n(p_0)$ и, согласно интегральной теореме Муавра-Лапласа,

статистика $\frac{\sum_{i=1}^{n}\xi_i - np_0}{\sqrt{np_0q_0}}$ - асимптотически нормальна, то есть при больших значениях n она подчиняется распределению мало отличающемуся от $\mathcal{N}(0;1)$. Поэтому мы можем в роли критерия проверки гипотезы H_0 взять статистику $T = g(\xi_1,\xi_2,...,\xi_n) = \frac{\bar{\xi} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}}$, где статистика $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ - среднее арифметическое элементов выборочной совокупности. Значением этой статистики для любого эксперимента будет дробь $\frac{m}{n}$ - относительная частота наступлений события A при проведении n одинаковых опытов. По выбранному значению уровня значимости основной гипотезы α, при двусторонней альтернативной гипотезе $H_1 : P(A) \neq p_0$ определяем критическое значение $t_{кр.}$ критерия $T = g(\xi_1,\xi_2,...,\xi_n)$ из условия $P(|T| > t_{кр.}) = \alpha$. События $\left\{\frac{\bar{\xi} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}}\right\}$ и $\left\{\frac{\sum_{i=1}^{n}\xi_i - np_0}{\sqrt{np_0q_0}}\right\}$ - равновероятны. Следовательно, применяя интегральную теорему Муавра-Лапласа, при больших значениях n получаем:

$$\alpha = P(|T| > t_{кр.}) = 1 - P(|T| \le t_{кр.}) = 1 - 2\Phi(t_{кр.}),$$ то есть $$t_{кр.} = \Phi^{-1}\left(\frac{1-\alpha}{2}\right).$$

Критическая область $S_{кр.}$, в соответствии с видом альтернативной гипотезы H_1, будет иметь вид: $S_{кр.} = (-\infty;-t_{кр.}) \cup (+t_{кр.};+\infty)$, а областью допустимых значений критерия будет $Q_{доп.} = (-t_{кр.};+t_{кр.})$. Построив области $S_{кр.}$ и $Q_{доп.}$, проводим эксперимент и определяем наблюдаемое значение критерия: $t_{набл.} = \frac{\frac{m}{n} - p_0}{\sqrt{\frac{p_0q_0}{n}}} = \frac{m - np_0}{\sqrt{np_0q_0}}$, где m – число наступлений события A в проведённом эксперименте, то есть в проведённых n раз одинаковых независимых опытах.

Если мы наблюдаем событие $\{t_{набд.} \in S_{кр.}\}$, то, так как при справедливой в реальности гипотезе H_0 это событие маловероятное, по правилу принятия решений гипотезу H_0 отклоняем и принимаем гипотезу H_1. При этом с вероятностью α мы можем совершить ошибку первого рода.

Если мы наблюдаем событие $\{t_{\text{набд.}} \in Q_{\text{доп..}}\}$, то по правилу принятия решений у нас нет оснований отклонять гипотезу H_0. То есть, гипотеза H_0 принимается.

Задача 2.3 Проверка гипотезы о равенстве неизвестного значения математического ожидания $M\xi$ случайной величины ξ некоторому фиксированному числу μ_0 при известном значении дисперсии σ^2.

Формулируем основную гипотезу $H_0 : M\xi = \mu_0$. Альтернативные гипотезы могут иметь разный вид: $H_1 : M\xi \neq \mu_0$; $H_1 : M\xi > \mu_0$ или $H_1 : M\xi < \mu_0$. Это будут сложные гипотезы. Первая гипотеза называется двусторонней альтернативой, вторая и третья гипотезы называются односторонними альтернативами. Альтернативная гипотеза может быть простой: $H_1 : M\xi = \mu_1$.

В случае, когда гипотезы H_0 и H_1 – простые, всегда существует наилучший критерий $T = g(\xi_1, \xi_2, ..., \xi_n)$ проверки справедливости гипотезы. Для построения областей $S_{\text{кр.}}$ и $Q_{\text{доп.}}$ надо знать закон распределения этого критерия T.

Так как по условию задачи нам известно значение дисперсии случайной величины $D\xi = \sigma^2$, то мы будем применять критерий $T = u = \dfrac{\bar{\xi} - \mu_0}{\sigma}\sqrt{n}$, который, согласно теореме Леви из Центральной предельной теоремы, подчиняется распределению близкому к нормальному распределению $\mathcal{N}$ с параметрами 0 и 1, т. е. $\mathcal{N}(0,1)$.

Выбрав критерий $T = g(\xi_1, \xi_2, ..., \xi_n)$, назначаем малое значение вероятности α - уровня значимости основной гипотезы. Обычно выбирается одно из значений 0,05; 0,01; 0,001. Общепринятый выбор из этих трёх значений сокращает объём необходимых таблиц распределений при построении критической области значений $S_{\text{кр.}}$.

Так как среднее арифметическое $\bar{\xi}$ выборочной совокупности является наилучшей оценкой $M\xi$, то при значительных отклонениях значения математического ожидания m от числа μ_0 мы будем иметь и значительные отклонения значения среднего арифметического $\bar{x}$ от числа μ_0. Поэтому, в случае двусторонней альтернативы $H_1 : M\xi \neq \mu_0$, используя таблицу значений функции Лапласа, определим величину $u_{\text{кр.}}$, при которой будет выполняться $P(|u| > u_{кр.}) = \alpha$. Ясно, что $u_{\text{кр.}} = \Phi^{-1}\left(\dfrac{1-\alpha}{2}\right)$. Следовательно, критическая область будет иметь вид: $S_{\text{кр.}} = (-\infty; -u_{\text{кр.}}) \cup (+u_{\text{кр.}}; +\infty)$. Область допустимых значений критерия будет такой: $Q_{\text{доп.}} = (-u_{\text{кр.}}; +u_{\text{кр.}})$.

Построив области $S_{кр.}$ и $Q_{доп.}$, проводим эксперимент и получаем выборку $(x_1, x_2, ..., x_n)$. По элементам этой выборки вычисляем наблюдаемое значение критерия $T_{набл.} = u_{набл.} = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$ и смотрим: в какую область попадает это значение.

Если мы наблюдаем событие $\{u_{набд.} \in S_{кр.}\}$, то, так как при справедливой в реальности гипотезе H_0 это событие маловероятное, по правилу принятия решений гипотезу H_0 отклоняем и принимаем гипотезу H_1. При этом с вероятностью α мы можем совершить ошибку первого рода.

Если мы наблюдаем событие $\{u_{набд.} \in Q_{доп..}\}$, то по правилу принятия решений у нас нет оснований отклонять гипотезу H_0. То есть, гипотеза H_0 принимается.

Если сформулирована односторонняя альтернатива, например, $H_1 : M\xi > \mu_0$, то, при её справедливости, разность $\bar{\xi} - \mu_0$ будет принимать значительные положительные значения. Для построения критической области по выбранному значению α определяем по таблицам значений функции Лапласа величину $u_{кр.}$, при которой будет выполняться $P(u > u_{кр.}) = \alpha$. Ясно, что в этом случае $u_{кр.} = \Phi^{-1}\left(\frac{1}{2} - \alpha\right)$, а критическая область будет иметь вид $S_{кр.} = (u_{кр.}; +\infty)$. Область допустимых значений будет такой: $Q_{доп.} = (-\infty; u_{кр.})$. Дальнейшая последовательность действий будет такой же, как и при проверке справедливости гипотезы H_0 при двусторонней альтернативе.

При простой альтернативной гипотезе $H_1 : M\xi = \mu_1$, построив области $S_{кр.}$ и $Q_{доп.}$, мы можем вычислить вероятность ошибки второго рода β.

При справедливости основной гипотезы H_0 случайная величина $\bar{\xi}$ имеет распределение близкое к нормальному распределению $\mathcal{N}\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$. Так как $Q_{доп.} = (-u_{кр.}; +u_{кр.})$, то с вероятностью $1 - \alpha$ случайная величина $\bar{\xi}$ примет значение принадлежащее интервалу $\left(\mu_0 - \frac{u_{кр.} \cdot \sigma}{\sqrt{n}}; \mu_0 + \frac{u_{кр.} \cdot \sigma}{\sqrt{n}}\right)$.

При справедливости альтернативной гипотезы $H_1 : M\xi = \mu_1$ случайная величина $\bar{\xi}$ будет иметь распределение близкое к нормальному распределению $\mathcal{N}\left(\mu_1, \frac{\sigma}{\sqrt{n}}\right)$. А тогда вероятность ошибки второго рода β будет равна вероятности этой случайной величины принять значение в

интервале $\left(\mu_0 - \frac{u_{кр.}\cdot\sigma}{\sqrt{n}}; \mu_0 + \frac{u_{кр.}\cdot\sigma}{\sqrt{n}}\right)$.Следовательно, вероятность ошибки второго рода будет равна: $\beta = \Phi\left(\frac{\mu_0-\mu_1}{\sigma}\sqrt{n} + u_{кр.}\right) - \Phi\left(\frac{\mu_0-\mu_1}{\sigma}\sqrt{n} - u_{кр.}\right)$.

Замечание. Если обозначить $1-\alpha=\gamma$, то записанный интервал будет доверительным интервалом для математического ожидания $M\xi$ случайной величины ξ, имеющей нормальное распределение $\mathcal{N}(\mu_0,\sigma)$.

В этом состоит двойственность задач интервального оценивания числовых характеристик случайных величин и статистической проверки гипотез о значениях числовых характеристик случайных величин.

Теория

Задача 3.3 Проверка гипотезы о равенстве неизвестного значения математического ожидания $M\xi$ случайной величины ξ некоторому фиксированному числу μ_0 при неизвестном значении дисперсии σ^2.

Как и в предыдущей задаче формулируем основную гипотезу $H_0: M\xi = \mu_0$. Альтернативные гипотезы могут иметь разный вид: $H_1: M\xi \neq \mu_0$; $H_1: M\xi > \mu_0$ или $H_1: M\xi < \mu_0$. Это будут сложные гипотезы. Простая альтернативная гипотеза записывается так же: $H_1: M\xi = \mu_1$.

Для построения областей $S_{кр.}$ и $Q_{доп.}$ надо подобрать критерий $T = g(\xi_1,\xi_2,...,\xi_n)$, закон распределения которого мы должны определить, зная некоторые специальные распределения, используемые в математической статистике.

Мы воспользуемся распределением Стьюдента, которому подчиняется среднее арифметическое $\bar{\xi}$ выборочных случайных величин $(\xi_1,\xi_2,...,\xi_n)$ в том случае, когда неизвестное значение дисперсии заменяется его точечной оценкой s^2, но при условии, что исследуемая случайная величина ξ подчиняется нормальному распределению вероятностей.

Если у нас есть уверенность в нормальном распределении случайной величины ξ, то мы будем применять критерий $T = t_{n-1} = \frac{\bar{\xi}-\mu_0}{s}\sqrt{n}$, подчиняющийся распределению Стьюдента с параметром *n*-1.

Выбираем значение α - уровень значимости основной гипотезы. При двусторонней альтернативе $H_1: M\xi \neq \mu_0$, используя таблицу распределения Стьюдента («*t*-распределения»), определяем величину $t_{кр.}$, из условия $P(|t_{n-1}| > t_{кр.}) = \alpha$. Критическая область и область допустимых

значений критерия t_{n-1} будут иметь вид, аналогичный виду этих областей в предыдущей задаче: $S_{кр.} = (-\infty; -t_{кр.}) \cup (+t_{кр.}; +\infty)$ и $Q_{доп.} = (-t_{кр.}; +t_{кр.})$.

Построив области $S_{кр.}$ и $Q_{доп.}$, проводим эксперимент и получаем выборку $(x_1, x_2, ..., x_n)$. По элементам этой выборки вычисляем наблюдаемое значение критерия $T_{набл.} = t_{набл.} = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$, определяем область, в которую попадает это наблюдаемое значение и, согласно правилу принятия решений, делаем вывод об отклонении или принятии основной гипотезы H_0.

При других альтернативах проводим действия аналогичные действиям в предыдущей задаче.

Задача 4.3 Проверка гипотезы о равенстве неизвестного значения дисперсии $D\xi$ случайной величины ξ фиксированному числу σ_0^2.

Основная гипотеза в этой задаче имеет вид $H_0 : D\xi = \sigma_0^2$. Альтернативная гипотеза чаще всего берётся в виде: $H_1 : D\xi > \sigma_0^2$. Так как мы увидели двойственность задач интервального оценивания числовых характеристик и проверки гипотез о значениях числовых характеристик, то ясно, что в этой задаче будет применяться критерий подчиняющийся распределению Пирсона («распределению χ^2»), которое мы использовали при построении доверительного интервала для дисперсии. Но это распределение вероятностей построено на основе предположения, что исследуемая случайная величина подчиняется нормальному распределению.

Если у нас есть уверенность в нормальном распределении случайной величины ξ, то для проверки справедливости основной гипотезы будем применять критерий $T = \chi^2_{n-1} = \frac{(n-1)\cdot s^2}{\sigma_0^2}$, подчиняющийся распределению Пирсона (" χ^2 - распределению") с параметром n-1. Для выбранного значения α, по таблицам распределения Пирсона, исходя из условия $P(\chi^2_{n-1} > \chi^2_{кр.}) = \alpha$, определяем величину $\chi^2_{кр.}$. Области критических и допустимых значений критерия будут иметь вид: $S_{кр.} = (\chi^2_{кр.}; +\infty)$ и $Q_{доп.} = (0; \chi^2_{кр.})$. Далее процедура – обычная.

Проводим эксперимент, получаем выборку $(x_1, x_2, ..., x_n)$, по элементам выборки подсчитываем наблюдаемое значение критерия $\chi^2_{набл.} = \frac{(n-1)\cdot s^2}{\sigma_0^2}$.

Определяем область, в которую попало наблюдаемое значение критерия. В соответствии с правилом принятия решений отклоняем, или принимаем основную гипотезу $\boldsymbol{H_0}$.

Практика

Рассмотрим следующую задачу, аналоги которой могут встретиться в практической жизни.

При выпечке сладких булочек по государственному стандарту на 1000 булочек полагается 10000 изюмин. Продавец магазина, продающий эти булочки, подозревает, что при выпечке булочек изюм частично уходит по непредусмотренным законом каналам, и предлагает хозяину магазина проверить соответствие среднего числа изюмин в булочке требованиям стандарта.

С этой целью из привезённых 100 булочек для проверки были случайным образом выбраны 25 штук. В каждой из выбранных булочек подсчитали количество изюмин. Результаты проверки приведены в таблице.

Количество изюмин в булочке	4	5	6	7	8	9	10	11	12
Число булочек	1	3	2	2	4	5	3	3	2

Возникает вопрос: Как должен обработать эти статистические данные хозяин магазина, чтобы проверить, справедливы или нет подозрения продавца? Может ли хозяин магазина предъявить производителю булочек претензии, состоящие в нарушении норм стандарта в его продукции?

Решение.

Определим название исследуемой случайной величины $\boldsymbol{\xi}$ – число изюмин в одной булочке. Так как согласно стандарту в 1000 булочек должно быть 10000 изюмин, а число изюмин в каждой булочке – случайно, то мы можем сказать, что математическое ожидание числа изюмин в одной булочке должно быть равным десяти. Значит решением вопроса о соответствии, или не соответствии среднего числа изюмин в булочке стандарту будет проверка гипотезы о равенстве значения математического ожидания случайного числа изюмин – десяти, то есть $\boldsymbol{H_0}$: $\boldsymbol{M\xi} = 10$. Так как мы подозреваем, что не весь, положенный по стандарту, изюм попадает в булочки, то альтернативной гипотезой будет гипотеза $\boldsymbol{H_1}$: $\boldsymbol{M\xi} < 10$.

Основная гипотеза будет простой гипотезой, а альтернативная гипотеза - сложной.

Для проверки справедливости основной гипотезы будем использовать критерий $\boldsymbol{t_{n-1}} = \dfrac{\bar{\xi} - \mu_0}{s}\sqrt{n}$, где $\mu_0 = 10$, распределением вероятностей которого является распределение Стьюдента с параметром $\boldsymbol{n}$-1.

По виду альтернативной гипотезы H_1 определяется критическое значение $t_{кр}$ критерия проверки гипотезы H_0 и строятся области критических и допустимых значений этого критерия. Затем хозяин магазина должен, используя статистические данные, вычислить среднее арифметическое числа изюмин в одной булочке и статистическую оценку среднего квадратического отклонения. По полученным оценкам вычисляется наблюдаемое значение критерия $t_{набл.} = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$. Это значение сравнивается с критическим $t_{кр}$. По правилу принятия решений будет сделан вывод о принятии или об отклонении гипотезы H_0. Если эта гипотеза будет принята, то подозрения продавца надо отвергнуть. Если эта гипотеза будет отклонена, то можно потребовать от производителя соблюдения норм стандарта и условий договора.

Основной вопрос задачи: «Действительно ли среднее число изюмин в булочке меньше положенного по стандарту числа?» однозначно определяет ил вид альтернативной гипотезы H_1: $M\xi < \mu_0$. Значит, в случае справедливости предположения H_1 о том, что изюм частично уходит по непредусмотренным законом каналам, критерий будет принимать отрицательные значения, так как возможные значения среднего арифметического $\bar{x}$ будут значительно меньше положенного по стандарту числа μ_0. Следовательно, для построения областей $S_{кр}$ и $Q_{доп}$ возможных значений критерия будем использовать условие $P(t_{n-1} < -t_{кр}) = \alpha$.

В виду симметричности распределения вероятностей Стьюдента будет справедливо равенство $P(t_{n-1} < -t_{кр}) = P(t_{n-1} > t_{кр}) = \alpha$. Если мы определим уровень значимости α основной гипотезы H_0 равным $\alpha = 0{,}01$, то по таблицам распределения Стьюдента для односторонних критериев находим $t_{кр} = 2{,}492$. Значит, области $S_{кр}$ и $Q_{доп}$ будут иметь вид: $S_{кр} = (-\infty; -2{,}492)$ и $Q_{доп} = (-2{,}492; +\infty)$.

Вычислим значения точечных оценок математического ожидания, дисперсии и среднего квадратического отклонения случайной величины ξ данным проверки 25 булочек:

$n = 25;\quad \bar{x} = 8{,}36;\quad s^2 = 5{,}24;\quad s = 2{,}289.$

Вычисляем наблюдаемое значение критерия, округляя результат до третьего знака за запятой: $t_{набл} = \frac{8{,}36 - 10}{2{,}289} \cdot \sqrt{25} \approx -3{,}582.$

Наблюдаемое значение критерия попадает в область $S_{кр} = (-\infty; -2{,}492)$, то есть мы наблюдаем случайное событие $\{t_{набл} \in S_{кр}\}$. Вероятность наступления такого события в случае справедливости гипотезы H_0:

$M\xi = 10$ очень мала: $P\left(t_{n-1} \in S_{кр} \middle| H_0\right) = \alpha = 0{,}01$. То есть, это событие мы в среднем можем наблюдать только в одной из ста выборок булочек по 25 штук.

Принимаем решение. Так как в случае справедливости гипотезы H_0, состоящей в том, что среднее число изюмин в булочке равно десяти, событие $\left\{t_{набл} \in S_{кр}\right\}$ является практически невозможным, то гипотезу H_0: $M\xi = 10$ отклоняем и принимаем альтернативную гипотезу H_1: $M\xi < 10$, то есть, при выпечке сладких булочек часть изюма уходит по непредусмотренным законом каналам.

Вероятность ошибки первого рода равна $\alpha = 0{,}01$. Это означает, что если действительно весь положенный по стандарту изюм попадает в тесто, то полученный нами результат $(\bar{x} = 8{,}36)$ эксперимента с n=25 булочками возможен лишь в одном случае из ста, то есть он очень маловероятен.

Допустим, что производитель булочек продолжает настаивать на том, что гипотеза H_0 – верна, то есть при выпечке 1000 булочек в тесто закладываются положенные по стандарту 10000 изюмин, а мы, проведя эксперимент, делаем ошибку первого рода – отклоняем верную в реальности гипотезу H_0: $M\xi = 10$. Формулируя гипотезу H_0: $M\xi = \mu_0 = 10{,}0$, мы руководствовались принципом «презумпции невиновности». Теперь же, учитывая результат проведения эксперимента, мы сформулируем более чётко альтернативную гипотезу H_1: $M\xi = \mu_1 = 8{,}0$. Это будет простая гипотеза.

Определим вероятность ошибки второго рода. Ошибка второго рода – это принятие неверной в реальности гипотезы. Это произойдёт в том случае, если действительно изюм воруется, то есть в реальности справедлива гипотеза H_1: $M\xi = \mu_1 = 8{,}0$, но мы, поддаваясь «нажиму» производителя, принимаем неверную гипотезу H_0 – «изюм не воруется». Вероятность ошибки второго рода β определяется как вероятность того, что при верной гипотезе H_1 наблюдаемое значение критерия попадёт в построенную область допустимых значений $Q_{доп}$.

Если гипотеза H_1 - верна, то область допустимых значений критерия будет иметь вид: $Q_{\text{доп.}} = \left(-t_{\text{кр}} - \dfrac{\mu_1 - \mu_0}{s} \cdot \sqrt{n}; +\infty\right)$, здесь: $\mu_0 = 10$; $\mu_1 = 8$, а значение $-t_{\text{кр}} = -2{,}492$ определили по таблицам при проверке основной гипотезы. Обозначим $t_\beta = -t_{кр.} - \dfrac{\mu_1 - \mu_0}{s} \cdot \sqrt{n}$ и вычислим $P\left(t_{n-1} \in Q_{доп.} \middle| H_1\right) = P\left(t_\beta < t_{n-1} < \infty\right) = \beta$. Так как $t_\beta = -t_{кр} - \dfrac{\mu_1 - \mu_0}{s} \cdot \sqrt{n} \approx 1{,}877$, то по таблицам распределения Стьюдента определяем, что вероятность

ошибки второго рода β (при n-1=24) будет равна: $\beta = P(t_{n-1} > 1{,}877) = 0{,}038 \approx 0{,}04$.

Признавая справедливость гипотезы H_0: $M\xi$ =10, когда в реальности справедлива H_1 мы совершим ошибку второго рода. Если в реальности среднее число изюминок в булочке равно восьми, а мы принимаем гипотезу, что среднее число изюминок равно десяти, то вероятность такого ошибочного решения равна $\beta = 0{,}04$.

Задачи подобного типа приходится почти каждый раз решать организациям Роспотребнадзора, которые должны оценивать качество продуктов питания, лекарств и других видов предметов массового потребления.

§3.3 Задачи второго типа статистической проверки гипотез

Теория

Статистическая проверка гипотез о равенстве значений одноимённых числовых характеристик двух различных случайных величин.

При решении этой группы задач статистической проверки гипотез рассматриваются две случайные величины, которые могут быть и независимыми, и зависимыми, и сравниваются значения их одноимённых числовых характеристик. Но при этом мы всегда должны иметь в виду, что критерии, которые мы будем использовать, построены на предположении, что генеральные совокупности, на которых определены случайные величины и из которых мы формируем выборки, имеют нормальное распределение вероятностей.

Задача 5.3 Статистическая проверка гипотезы о совпадении значений дисперсий двух случайных величин.

Постановка задачи. Мы изучаем две случайные величины ξ и η, имеющих дисперсии $D\xi$ и $D\eta$. Выдвигается гипотеза $H_0: D\xi = D\eta$, - *дисперсии рассматриваемых случайных величин – равны*. Эта гипотеза будет основной и простой, так как в ней сравниваются два числовых значения σ_1^2 и σ_2^2. Альтернативная гипотеза $H_1: D\xi > D\eta$ будет сложной гипотезой.

Комментарий

Задача такого типа может возникнуть, например, при сравнении и оценке уровней технологии производства некоторого продукта двух различных производителей.

Альтернативная гипотеза будет односторонней, а так как нам всё равно: какую из случайных величин считать первой, а какую – второй, то обычно в роли первой случайной величины выбирают, у которой будет больше значение оценки дисперсии.

Теория

Основную и альтернативную гипотезы можно переписать так:

$H_0: \frac{D\xi}{D\eta} = 1$ и $H_1: \frac{D\xi}{D\eta} > 1$. Представление гипотез в такой форме объясняется выбором критерия проверки справедливости основной гипотезы.

В роли критерия $T = g(\xi_1, \xi_2, ..., \xi_n)$ проверки гипотезы выберем случайную величину $T = f_{n_1-1, n_2-1} = \frac{\frac{1}{n_1 - 1} \cdot \chi^2_{n_1-1}}{\frac{1}{n_2 - 1} \cdot \chi^2_{n_2-1}}$, которая подчиняется распределению Фишера - Снедекора ("F-распределению"). При выбранном уровне значимости α по таблицам F-распределения определяем $f_{кр.}$, которое удовлетворяет условию: $P(f_{n_1-1, n_2-1} > f_{кр.}) = \alpha$. Критическая область имеет вид: $S_{кр.} = (f_{кр.}; +\infty)$, а область допустимых значений: $Q_{доп.} = (1; f_{кр.})$.

Построив области $S_{кр}$ и $Q_{доп}$, проводим эксперимент и получаем две выборки: $(x_1, x_2, ..., x_{n_1})$ и $(y_1, y_2, ..., y_{n_2})$. По элементам этих выборок вычисляем значения точечных оценок дисперсий s_1^2 и s_2^2.

Ранее мы показали, что статистика $\frac{s_i^2}{\sigma_i^2}$ подчиняется распределению $\frac{1}{n_i - 1} \cdot \chi^2_{n_i-1}$, где $i = 1,2$. Если гипотеза H_0 верна, то значения дисперсий случайных величин ξ и η равны, то есть: $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Следовательно, наблюдаемое значения критерия проверки гипотезы будет равно: $f_{набл} = \frac{s_1^2}{s_2^2}$. Дальнейшие действия при проверке справедливости гипотезы H_0 такие же, как и при решении предыдущих задач. То есть, определяем в какую область попадает наблюдаемое значение критерия и принимаем решение о принятии или об отклонении основной гипотезы.

Замечание. Условие $s_1^2 > s_2^2$, которое мы используем для выбора первой и второй случайной величины, объясняется сложностью составления и представления в справочниках таблиц F-распределения. Для определения $f_{кр.}$ надо иметь три «входа» в таблицу: вероятность α, и степени свободы $n_1 - 1$ и $n_2 - 1$. Эта таблица должна быть размещена на двумерной странице книги. Поэтому для каждого значения α (обычно это α =0,05 или α =0,01) таблицы печатаются на отдельной странице.

Отсутствие условия $s_1^2 > s_2^2$ привело бы к увеличению объёма таблиц, но, чтобы это избежать, можно использовать справедливость отношения: $f_{m,n} = \frac{1}{f_{n,m}}$.

Пример

Пример 2*а*.1 и *2б*.1 При анализе выборок, получившихся в результате обследования двух групп мужчин были вычислены значения точечных оценок дисперсий $s_1^2 = 36{,}28$ для $n_1 = 115$ и $s_2^2 = 34{,}70$ для $n_2 = 906$.

Формулируем основную гипотезу H_0: Дисперсии случайных величин ξ и η - дисперсии роста мужчин двух генеральных совокупностей, - одинаковые. Вычисляем наблюдаемое значение критерия $f_{набл.} = \frac{s_1^2}{s_2^2} = 1{,}045$. Для двух значений уровня значимости по таблицам "F- распределения" критические значения критерия: $\alpha = 0{,}05$; $f_{кр.} = 1{,}29$; $\alpha = 0{,}01$; $f_{кр.} = 1{,}44$. Так как в обоих случаях выполняется неравенство $f_{набл.} < f_{кр.}$, то у нас нет оснований отклонять гипотезу H_0, то есть эта гипотеза – принимается.

Теория

Задача 6.3 Статистическая проверка гипотезы о совпадении значений математических ожиданий двух случайных величин.

Постановка задачи. Мы изучаем две случайные величины ξ и η, имеющих математические ожидания $M\xi$ и $M\eta$. Выдвигается основная гипотеза $H_0 : M\xi = M\eta$ - математические ожидания рассматриваемых случайных величин – равны.

Комментарий

Задача такого типа может возникнуть, например, при выборе из двух поставщиков продукции наиболее лучшего; при оценке предпочтительности акций или каких-либо ценных бумаг одной из двух компаний; при сравнении двух методик лечения болезни; при сравнении качества обучения по двум отличающимся по содержанию программам или методикам.

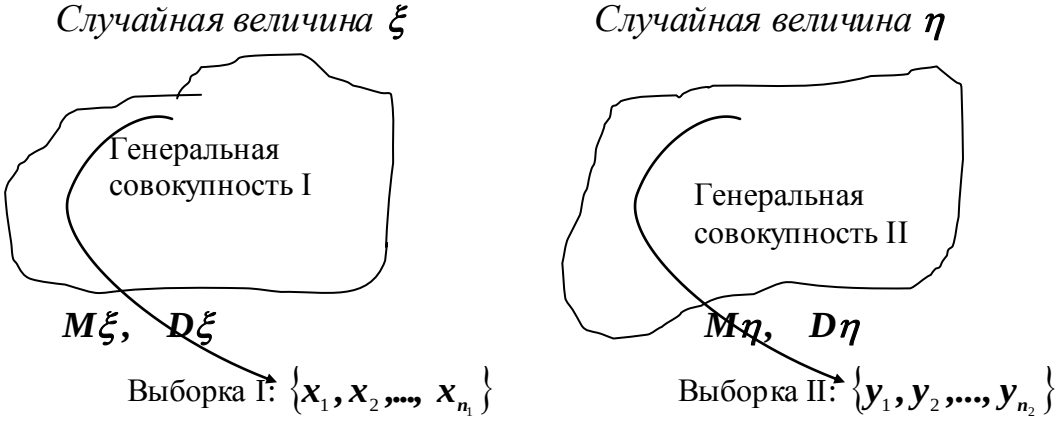
К этому же типу задач, решаемых путём проверки справедливости сформулированной гипотезы H_0 относятся задачи, которые могут возникнуть при оценке эффективности проведённого мероприятия по изменению значения некоторого параметра, характеризующего работу фирмы или предприятия, при оценке проведённых мероприятий по повышению уровня квалификации работников, выполняющих некоторые производственные операции.

Сравнивая значения математических ожиданий количественных показателей деятельности работников «до» и «после» проведения обучения или переподготовки, мы сможем сделать вывод о заметном повышении квалификации и производительности труда персонала и тем самым оценить эффективность проведённых мероприятий. Сравнивая показатели состояния здоровья группы пациентов «до» и «после» проведения курса лечения, мы оцениваем результативность этого курса лечения.

Комментарий

Для проверки справедливости основной гипотезы мы будем проводить эксперимент, в результате которого получим две выборки возможных значений случайных величин ξ и η : $(x_1, x_2, ..., x_{n_1})$ и $(y_1, y_2, ..., y_{n_2})$ объёмами n_1 и n_2. Очевидно, что возможны два варианта выбора критерия проверки гипотезы, зависящие от решаемой задачи. Эти варианты определяются видом выборок.

Вариант I.



Независимые выборки

Рис. 3.2.

Выборки, в зависимости от цели решаемой задачи, могут быть независимыми или зависимыми. Если мы сравниваем качество продукции двух различных производителей, то выборки измерений значений качественного признака будут независимыми (Рис.3.2). Если мы сравниваем, например, производительность труда персонала «до» (исследуемый качественный признак – случайная величина ξ) и - «после» (исследуемый качественный признак – случайная величина η) проведения обучения или переподготовки персонала, то, так как x_i и y_i - значения исследуемого качественного признака у одного и того же i-того объекта генеральной совокупности, мы не можем утверждать, что выборки будут независимыми (Рис.3.3).

Например, пусть случайная величина ξ - данные анализов у больных каким-то определённым заболеванием. Имеется n пациентов, для которых проводится курс лечения по некоторой методике. С целью оценки эффективности проведённого курса лечения, этими же пациентами снова сдаются те же анализы. Случайная величина η - данные анализов после проведённого курса лечения. Мы имеем две выборки $\{x_1, x_2, ..., x_n\}$ и $\{y_1, y_2, ..., y_n\}$. Ясно, что величина разности $x_i - y_i$ позволит судить об эффективности проведённого курса лечения для i-того пациента. Мы не можем сказать, что значения x_i и y_i не зависят друг от друга. Выборки $\{x_1, x_2, ..., x_n\}$ и $\{y_1, y_2, ..., y_n\}$ - зависимые и называются «сообразными».

Вариант II

Рис.3.3

Если выборки зависимые, или «сообразные», то значения x_i и y_i - это значения контролируемого качественного признака у одного и того же объекта генеральной совокупности, попавшего в выборку. А поэтому по этим двум выборкам мы можем записать новую выборку: $\{x_1 - y_1, x_2 - y_2, ..., x_n - y_n\}$ и изучать эффективность проводимых мероприятий по этой выборке.

Теория

Вариант **I.** Рассмотрим сначала задачу проверки справедливости гипотезы H_0 в том случае, когда, выборки $(x_1, x_2, ..., x_{n_1})$ и $(y_1, y_2, ..., y_{n_2})$ полученные из двух генеральных совокупностей, - независимые.

Записываем основную гипотезу $H_0: M\xi = M\eta$. Альтернативная гипотеза может быть двусторонней $H_1: M\xi \neq M\eta$, или односторонней $H_1: M\xi > M\eta$. Построим сами критерий проверки справедливости основной гипотезы. Основную гипотезу можно записать так $H_0: M\xi - M\eta = 0$. В соответствии с этой формой записи основной гипотезы рассмотрим статистику: $\overline{\xi} - \overline{\eta} = \frac{1}{n_1}\sum_{i=1}^{n_1}\xi_i - \frac{1}{n_2}\sum_{i=1}^{n_2}\eta_i$.

Ясно, что: $M(\bar{\xi}-\bar{\eta})=M\xi-M\eta=m_1-m_2$.

Средние арифметические элементов выборочных совокупностей при больших объёмах выборок имеют распределение, мало отличающееся от нормального (будем впредь говорить: средние арифметические асимптотически нормальны). Надо оценить дисперсию этой новой статистики.

а) Если нам известны значения дисперсий исследуемых случайных величин $D\xi=\sigma_1^2$ и $D\eta=\sigma_2^2$, то, используя независимость выборок, запишем: $D(\bar{\xi}-\bar{\eta})=D\bar{\xi}+D\bar{\eta}=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$.

Центрируя и нормируя статистику $\bar{\xi}-\bar{\eta}$, придём к статистике $T=g(\xi_1,\xi_2,...,\xi_n)=\frac{(\bar{\xi}-\bar{\eta})-(m_1-m_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$, которая, согласно теореме Леви, асимптотически нормальна. Если основная гипотеза верна, то здесь будет $m_1-m_2=0$.

Для проверки справедливости гипотезы $H_0: M\xi-M\eta=0$ в качестве критерия T мы получаем статистику $u=\frac{(\bar{\xi}-\bar{\eta})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$, которая подчиняется нормальному распределению $\mathcal{N}(0,1)$.

По выбранному (или назначенному) уровню значимости α, при альтернативной гипотезе $H_1: M\xi\neq M\eta$, определим величину $u_{кр}$ такую, что будет справедливо $P(|u|>u_{кр})=\alpha$. Величину $u_{кр}$, используя таблицу значений функции Лапласа, находим из условия $u_{кр}=\Phi^{-1}\left(\frac{1-\alpha}{2}\right)$.

Область $S_{кр}$ критических значений критерия u для гипотезы H_0 будет иметь вид: $(-\infty;-u_{кр})\cup(u_{кр};+\infty)$. Область $Q_{доп}$ допустимых значений критерия u будет иметь вид: $(-u_{кр};+u_{кр})$.

Построив области $S_{кр}$ и $Q_{доп}$, проводим эксперимент и получаем две выборки $(x_1,x_2,...,x_{n_1})$ и $(y_1,y_2,...,y_{n_2})$. Вычисляем наблюдаемое значение

критерия $u_{набл} = \dfrac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ и смотрим: в какую из областей оно попало. Принимаем решение об отклонении или принятии гипотезы H_0.

б) Если дисперсии случайных величин ξ и η - неизвестны, то мы, чтобы использовать построенные ранее распределения статистик, должны предположить, что эти случайные величины подчиняются нормальному распределению.

Построим критерий проверки справедливости основной гипотезы H_0 при условии, что предварительно была проверена и принята гипотеза $H_0 : D\xi = D\eta$ (Задача 5.3).

Рассматривая объединённую выборку, содержащую $n_1 + n_2$ элементов, мы строим статистику $T = g(\xi_1, \xi_2, ..., \xi_n) = \dfrac{(\bar{\xi} - \bar{\eta}) - (m_1 - m_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$. Учитывая, что дисперсии случайных величин ξ и η равны $D\xi = D\eta = \sigma^2$ и что при верной гипотезе H_0 значения m_1 и m_2 математических ожиданий этих случайных величин тоже - равны, статистика T примет вид $u = \dfrac{\bar{x} - \bar{y}}{\sigma} \cdot \sqrt{\dfrac{n_1 \cdot n_2}{n_1 + n_2}}$.

Но мы не знаем величину значения σ^2 равных дисперсий $D\xi$ и $D\eta$, а поэтому определим её оценку s^2 по элементам объединённой выборки:

$$s^2 = \frac{1}{n_1 + n_2 - 2} \cdot \left[\sum_{i=1}^{n_1} (\xi_i - \bar{\xi})^2 - \sum_{i=1}^{n_2} (\eta_i - \bar{\eta})^2 \right].$$

Или $s^2 = \dfrac{1}{n_1 + n_2 - 2} \cdot \left[(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 \right]$. Желая определить закон распределения оценки дисперсии s^2, полученной по объединённой выборке, вычислим значение математического ожидания этой оценки дисперсии:

$$Ms^2 = M\left[\frac{1}{n_1 + n_2 - 2} \left[(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 \right] \right] = \frac{1}{n_1 + n_2 - 2} \left[(n_1 - 1) Ms_1^2 + (n_2 - 1) Ms_2^2 \right]$$

То есть: $Ms^2 = \dfrac{1}{n_1 + n_2 - 2} \left[(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2 \right] = \sigma^2$.

Так как $\left[(n_1-1)s_1^2+(n_2-1)s_2^2\right]=\sigma^2\cdot\chi^2_{n_1-1}+\sigma^2\cdot\chi^2_{n_2-1}=\sigma^2\cdot\chi^2_{n_1+n_2-2}$, то распределением вероятностей статистики $\frac{s^2}{\sigma^2}$ будет распределение вероятностей случайной величины $\frac{1}{n_1+n_2-2}\chi^2_{n_1+n_2-2}$.

Значит отношение статистик $\frac{\frac{\bar{\xi}-\bar{\eta}}{\sigma}\cdot\sqrt{\frac{n_1\cdot n_2}{n_1+n_2}}}{\sqrt{\frac{s^2}{\sigma^2}}}=\frac{\frac{\bar{\xi}-\bar{\eta}}{\sigma}\cdot\sqrt{\frac{n_1\cdot n_2}{n_1+n_2}}}{\sqrt{\frac{1}{n_1+n_2-2}\chi^2_{n_1+n_2-2}}}$, являясь случайной величиной, подчиняется распределению Стьюдента с параметром n_1+n_2-2. Так как все «составные элементы» последней формулы нам ясны и их значения мы всегда можем вычислить по элементам двух выборок, то эту случайную величину мы возьмём в качестве критерия проверки справедливости гипотезы $H_0: M\xi=M\eta$:

$$t_{n_1+n_2-2}=\frac{\bar{\xi}-\bar{\eta}}{s}\cdot\sqrt{\frac{n_1\cdot n_2}{n_1+n_2}},\quad \text{где } s^2=\frac{1}{n_1+n_2-2}\left[(n_1-1)s_1^2+(n_2-1)s_2^2\right].$$

Далее процедура обычная – строим критическую и допустимую области, проводим эксперимент, получаем две выборки, подсчитываем наблюдаемое значение критерия $t_{\text{набл.}}=\frac{\bar{x}-\bar{y}}{s}\cdot\sqrt{\frac{n_1\cdot n_2}{n_1+n_2}}$ и смотрим: в какую область попадает это наблюдаемое значение критерия. Принимаем решение о принятии или об отклонении гипотезы $H_0: M\xi=M\eta$.

Сообразуясь со здравым смыслом и применяя свои знания теории вероятностей, мы построили критерий проверки справедливости гипотезы H_0 в случае независимых выборок $(x_1, x_2, ..., x_{n_1})$ и $(y_1, y_2, ..., y_{n_2})$.

Практика

Пример 2*а*.1 и *2б*.1 При анализе выборок, получившихся в результате обследования двух групп мужчин были вычислены значения точечных оценок математических ожиданий $\bar{x}=166{,}61$ для $n_1=115$ и $\bar{y}=166{,}77$ для $n_2=906$.

Формулируем основную гипотезу H_0: Математические ожидания случайных величин ξ и η - средний рост мужчин двух генеральных совокупностей, - одинаковые.

Ранее мы приняли гипотезу о том, что дисперсии двух анализируемых генеральных совокупностей равны. Поэтому будем использовать критерий

$t_{\text{набл.}} = \dfrac{\bar{x} - \bar{y}}{s} \cdot \sqrt{\dfrac{n_1 \cdot n_2}{n_1 + n_2}}$. Вычисляем значение точечной оценки дисперсии по объединённой выборке $s^2 = \dfrac{1}{n_1 + n_2 - 2}\left[(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2\right] = 34{,}88$ и наблюдаемое значение критерия $t_{набл.} = -0{,}274$. Выберем уровень значимости основной гипотезы $\alpha = 0{,}05$. По таблицам "t-распределения" при альтернативной гипотезе H_1: $M\xi \neq M\eta$ определяем критическое значение критерия $t_{кр.} = 124{,}3$. Так как $|t_{набл.}| < t_{кр.}$ формулируем решение: «У нас нет оснований отклонять гипотезу H_0. Гипотеза H_0: $M\xi = M\eta$ - принимается».

Вариант **II**. Рассмотрим теперь задачу проверки справедливости основной гипотезы H_0: $M\xi = M\eta$ в том случае, когда выборочные последовательности $(\xi_1, \xi_2, ..., \xi_n)$ и $(\eta_1, \eta_2, ..., \eta_n)$ - сообразные.

Комментарий

Мы можем сказать, что перед проведением мероприятий на генеральной совокупности была определена случайная величина ξ. После проведения мероприятий с целью изменения значений количественного признака на этой же генеральной совокупности определена новая случайная величина η. Значения η при сравнении со значениями ξ должны показать эффективность проведённых мероприятий. Элементы выборочных последовательностей $\{\xi_i\}, i = 1,2,...,n$, и $\{\eta_i\}, i = 1,2,...,n$, принимают значения являющиеся значениями качественного признака одного и того же объекта генеральной совокупности попавшего в выборку. Эти значения получены, соответственно, «до» и «после» проведения некоторых мероприятий с целью изменения качественного признака элементов генеральной совокупности. Очевидно, что для оценки эффективности проведённых мероприятий мы будем, сравнивать значения математических ожиданий этих случайных величин ξ и η.

а) Рассмотрим случайную величину $\zeta = \xi - \eta$. Определим её основные числовые характеристики – математическое ожидание и дисперсию.

Математическое ожидание разности ζ случайных величин всегда равно разности их математических ожиданий $M\zeta = M[\xi - \eta] = M\xi - M\eta = m_1 - m_2$.

Но случайные величины ξ и η зависимые, а поэтому дисперсия разности ζ не будет равна сумме дисперсий величин ξ и η.

Согласно определению дисперсии случайной величины, дисперсию разности ζ запишем так:

$$D\zeta = D[\xi - \eta] = M[(\xi - \eta) - M(\xi - \eta)]^2 = M[(\xi - M\xi) - (\eta - M\eta)]^2 .$$

Используя свойства математического ожидания, получаем: $D\zeta = M[\xi - M\xi]^2 - 2M[(\xi - M\xi)(\eta - M\eta)] + M[\eta - M\eta]^2 = D\xi - 2\mu_{11} + D\eta$, где $\mu_{11} = cov(\xi,\eta) = M[\xi \cdot \eta] - M\xi \cdot M\eta$ - ковариационный момент случайных величин ξ и η.

Теория

Основная гипотеза формулируется также как и в предыдущем случае: $H_0 : M\xi = M\eta$, или $H_0 : m_1 = m_2$.

Построим критерий проверки основной гипотезы, учитывающий зависимость случайных величин ξ и η.

Если обозначить $\zeta = \xi - \eta$, то основная гипотеза H_0 может быть записана так $H_0 : M\zeta = 0$, то есть $H_0 : m_1 - m_2 = 0$.

Рассмотрим статистику: $\overline{\zeta} = \overline{\xi} - \overline{\eta}$.

Ясно, что при справедливости основной гипотезы будет верно равенство: $M[\overline{\xi} - \overline{\eta}] = M\overline{\xi} - M\overline{\eta} = m_1 - m_2 = 0$. А так как дисперсия среднего арифметического $\overline{\zeta} = \frac{1}{n}\sum_{i=1}^{n}\zeta_i$ одинаково распределённых, независимых слагаемых $z_i = x_i - y_i$, $i = 1,2,...,n$, в n раз меньше дисперсии каждого слагаемого, то: $D\overline{\zeta} = D[\overline{\xi} - \overline{\eta}] = \frac{D\zeta}{n} = \frac{D\xi - 2\mu_{11} + D\eta}{n}$.

Случайные элементы выборки случайной величины ζ: $\{\xi_1 - \eta_1, \xi_2 - \eta_2, ..., \xi_n - \eta_n\}$, - независимы и имеют то же распределение вероятностей, что и случайная величина ζ, следовательно, согласно теореме Леви, статистика $\frac{(\overline{\xi} - \overline{\eta}) - (m_1 - m_2)}{\sqrt{\sigma_1^2 - 2\mu_{11} + \sigma_2^2}}\sqrt{n} = u$ - асимптотически нормальна. Эту статистику можно взять в качестве критерия проверки справедливости гипотезы H_0 и при построении областей $S_{кр}$ и $Q_{доп}$ будем использовать таблицы значений функции Лапласа.

Этот критерий построен при предположении, что нам известны значения дисперсий случайных величин ξ и η и ковариационного момента $cov(\xi,\eta) = \mu_{11} = M[(\xi - m_1)(\eta - m_2)]$.

б) Если значения дисперсий и ковариационного момента неизвестны, то, как и в варианте **I*б***), заменив их значения значениями точечных оценок, построим критерий, который подчиняется распределению Стьюдента.

Для этого сначала определим точечную оценку s_ζ^2 дисперсии $D\zeta$:

$$s_\zeta^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left((\xi_i - \eta_i) - (\bar{\xi} - \bar{\eta})\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left((\xi_i - \bar{\xi}) - (\eta_i - \bar{\eta})\right)^2 =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2 - \frac{2}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})(\eta_i - \bar{\eta}) + \frac{1}{n-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2.$$

Вычислим отдельно среднюю сумму:

$$\sum_{i=1}^{n}(\xi_i - \bar{\xi})(\eta_i - \bar{\eta}) = \sum_{i=1}^{n}(\xi_i\eta_i - \bar{\xi}\eta_i - \xi_i\bar{\eta} + \bar{\xi}\bar{\eta}) =$$

$$= n\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\eta_i - \bar{\xi}\cdot\bar{\eta} - \bar{\xi}\cdot\bar{\eta} + \bar{\xi}\cdot\bar{\eta}\right) = n\cdot(a_{11} - \bar{\xi}\cdot\bar{\eta}) = n\cdot m_{11}.$$

Здесь $m_{11} = \vartheta^* = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})(\eta_i - \bar{\eta})$ - точечная оценка ковариационного момента $\mu_{11} = \vartheta$ - теоретической характеристики двумерной случайной величины (ξ, η).

Таким образом, точечная оценка дисперсии $D\zeta$ будет равна: $s_\zeta^2 = s_1^2 - 2\frac{n}{n-1}m_{11} + s_2^2$. Определим значение математического ожидания Ms_ζ^2 этой оценки:

$$Ms_\zeta^2 = M\left[s_1^2 - 2\frac{n}{n-1}m_{11} + s_2^2\right] = Ms_1^2 - 2\frac{n}{n-1}Mm_{11} + Ms_2^2.$$

Вычислим значение математического ожидания оценки ковариационного момента: $Mm_{11} = M[a_{11} - \bar{\xi}\cdot\bar{\eta}] = Ma_{11} - M[\bar{\xi}\cdot\bar{\eta}]$.

Ясно что: $Ma_{11} = M\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\eta_i\right] = \frac{1}{n}\sum_{i=1}^{n}M[\xi_i\eta_i] = \frac{1}{n}n\alpha_{11} = \alpha_{11} = M[\xi\eta]$.

Вычисляя значение $M[\bar{\xi}\cdot\bar{\eta}]$, мы должны учитывать, что средние арифметические элементов выборок $\bar{\xi}$ и $\bar{\eta}$ не будут независимыми случайными величинами, так как анализируемые выборки не являются независимыми.

$$M[\bar{\xi}\cdot\bar{\eta}] = M\left[\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right)\cdot\left(\frac{1}{n}\sum_{j=1}^{n}\eta_j\right)\right] = \frac{1}{n^2}M\left[\sum_{i=1}^{n}\left(\xi_i\cdot\sum_{j=1}^{n}\eta_j\right)\right].$$

Проводя дальнейшие преобразования, мы должны иметь в виду, что случайные величины с одинаковыми индексами ξ_i и η_i двумерной выборочной совокупности - зависимые случайные величины. Поэтому $M[\xi_i\cdot\eta_i] = M[\xi\cdot\eta] = \alpha_{11}$. Элементы выборочной совокупности ξ_i и η_j с разными индексами – независимые случайные величины, имеющие те же распределения вероятностей, что и случайные величины ξ и η, поэтому $M[\xi_i\cdot\eta_j] = M\xi_i\cdot M\eta_j = M\xi\cdot M\eta = m_1\cdot m_2$.

Следовательно: $M\left[\bar{\xi}\cdot\bar{\eta}\right]=\frac{1}{n^2}M\left[\xi_1\left(\sum_{j=1}^{n}\eta_j\right)+\xi_2\left(\sum_{j=1}^{n}\eta_j\right)+\ldots+\xi_n\left(\sum_{j=1}^{n}\eta_j\right)\right]=$

$$=\frac{1}{n^2}\left(n\cdot\alpha_{11}+n(n-1)\cdot m_1\cdot m_2\right)=\frac{1}{n}\alpha_{11}+\frac{n-1}{n}\cdot m_1\cdot m_2.$$

Таким образом:

$$Mm_{11}=Ma_{11}-M\left[\bar{x}\cdot\bar{y}\right]=\alpha_{11}-\frac{1}{n}\alpha_{11}-\frac{n-1}{n}\cdot m_1\cdot m_2=\frac{n-1}{n}\left(\alpha_{11}-m_1\cdot m_2\right).$$

Комментарий

Как промежуточный результат заметим, что точечная оценка $m_{11}=a_{11}-\bar{x}\cdot\bar{y}$ является смещённой оценкой ковариационного момента $\mu_{11}=\alpha_{11}-m_1\cdot m_2$. Несмещённую оценку ковариационного момента обозначим $\tilde{m}_{11}$. Значит: $\tilde{m}_{11}=\frac{n}{n-1}\cdot m_{11}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$.

Задачи проверки справедливости рассматриваемой основной гипотезы в случае, когда выборки - зависимые, в литературе практически не встречаются. Поэтому мы здесь уделяем повышенное внимание построению критерия для решения таких задач, что может понадобиться в практической работе статистику, интересующемуся математикой.

Теория

Закончим построение критерия проверки справедливости основной гипотезы.

Математическое ожидание Ms_ζ^2 - оценки дисперсии случайной величины ζ будет равно:

$$Ms_\zeta^2=Ms_1^2-2\frac{n}{n-1}Mm_{11}+Ms_2^2=\sigma_1^2-2\frac{n}{n-1}\cdot\frac{n-1}{n}\left(\alpha_{11}-m_1\cdot m_2\right)+\sigma_2^2 \quad .$$

То есть: $Ms_\zeta^2=\sigma_1^2-2\mu_{11}+\sigma_2^2=\sigma_\zeta^2$. А так как распределением вероятностей статистики $\frac{s_\zeta^2}{\sigma_\zeta^2}$ будет распределение $\frac{1}{n-1}\chi_{n-1}^2$, то статистика $\frac{\frac{\bar{\zeta}-m_z}{\sigma_\zeta}\sqrt{n}}{\sqrt{\frac{s_\zeta^2}{\sigma_\zeta^2}}}=\frac{\bar{\zeta}-m_z}{s_\zeta}\sqrt{n}$ подчиняется распределению Стьюдента с параметром n-1.

Случайную величину $t_{n-1} = \dfrac{\bar{\xi} - m_\zeta}{s_\xi}\sqrt{n} = \dfrac{(\bar{\xi} - \bar{\eta}) - (m_1 - m_2)}{\sqrt{s_1^2 - 2\tilde{m} + s_2^2}}\sqrt{n}$ мы будем брать в качестве критерия проверки гипотезы $H_0 : m_1 - m_2 = 0$, когда случайные величины ξ и η не являются независимыми.

Критерий построен. Далее последовательность действий при построении областей возможных значений критерия и принятии решения - обычная.

§4.3 Задача статистической проверки гипотезы о виде закона распределения вероятностей случайной величины. Критерий согласия К. Пирсона

Теория

Задача 7.3 Статистическая проверка гипотезы о виде закона распределения вероятностей исследуемой случайной величины.

Знание закона распределения случайной величины X даёт нам всю информацию об области возможных значений исследуемого количественного признака и о распределении вероятностей этих возможных значений, даёт возможность определить значения числовых характеристик и вычислить вероятность наступления любого случайного события.

Комментарий

Предположим, что по имеющейся в нашем распоряжении некоторой выборке мы построили вариационный ряд (таблица 1.4.3) и сделали его геометрическую иллюстрацию, то есть построили гистограмму (рис.1.4.3)

Таблица 1.4.3

$[a_j ; a_{j+1})$	$[1;2)$	$[2;3)$	$[3;4)$	$[4;5)$	$[5;6)$	$[6;7)$	$[7;8)$	$[8;9)$	$[9;10)$	$[10;11)$
$\dfrac{m_j}{n}$	$\dfrac{3}{250}$	$\dfrac{5}{250}$	$\dfrac{21}{250}$	$\dfrac{60}{250}$	$\dfrac{70}{250}$	$\dfrac{50}{250}$	$\dfrac{27}{250}$	$\dfrac{8}{250}$	$\dfrac{4}{250}$	$\dfrac{2}{250}$

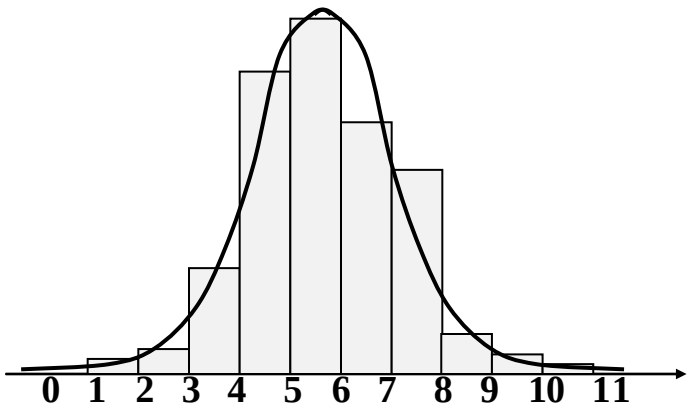


Рис.1.4.3

Анализ вида гистограммы позволяет сделать предположение, что анализируемые статистические данные являются выборкой значений случайной величины ξ , подчиняющейся нормальному закону распределения вероятностей. Это предположение мы формулируем на основании ранее сказанного утверждения, базирующегося на теореме Гливенко, о том, что гистограмма является статистической моделью теоретической плотности вероятности.

Ясно, что анализируя вид гистограммы с целью сформулировать гипотезу о виде закона распределения вероятностей случайной величины ξ , мы должны знать типы и виды законов распределения, хотя бы те, которые наиболее часто встречаются в практике. Если мы не знаем законов распределения вероятностей, не знаем, как выглядит график плотности вероятности того, или иного закона, то нам не с чем будет мысленно сравнивать вид построенной гистограммы. Значит, мы не сможем на основе анализа построенной статистической модели закона сформулировать гипотезу о виде и названии, если оно существует, теоретического закона распределения вероятностей случайной величины ξ .

Теория

На основании анализа результатов первичной обработки статистических данных формулируется основная простая гипотеза:

H_0: *Случайная величина ξ имеет распределение вероятностей, описываемое вероятностной функцией $P(x)$ (кратко: $\xi \in P$).*

Или, что одно и то же:

H_0: *Функция $F(x)$ является функцией распределения исследуемой случайной величины ξ .*

Альтернативная гипотеза будет сложной $H_1: \xi \in \mathcal{P} \neq P$. То есть, распределение ξ принадлежит некоторому семейству распределений $\mathcal{P}$, не содержащему P.

В основе построения критериев проверки простой гипотезы H_0 лежит оценка «различия» эмпирического распределения вероятностей $P_n^*(x)$ эмпирической случайной величины ξ^* и предполагаемого теоретического распределения $P(x)$. Это «различие» оценивается введением «меры отклонения» $\mathcal{D}(P_n^*, P)$ эмпирического распределения $P_n^*(x)$ и предполагаемого теоретического распределения $P(x)$. Строится критерий, который основывается на свойствах выборочного распределения этой «меры отклонения».

В основном, чаще всего применяются три вида «меры отклонения» и, соответственно, строятся три критерия проверки гипотезы H_0. Это: “критерий Колмогорова”, “критерий Мизеса” и “критерий К.Пирсона”.

Рассмотрим построение “критерия К.Пирсона”, который называют ”критерием согласия”.

Комментарий

Поясним последовательность действий при построении этого критерия в предположении, что исследуется случайная величина непрерывного типа. В этом случае мы основную гипотезу запишем так:

H_0: *Плотность вероятности* $p(x)=F'(x)$ *является плотностью вероятности исследуемой случайной величины* ξ.

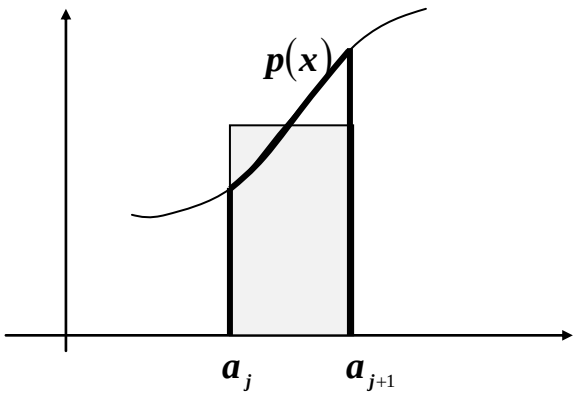
Имеется выборочная совокупность $(\xi_1,\xi_2,...,\xi_n)$, по которой при первичной обработке статистических данных строится вариационный ряд.

Таблица 2.4.3

интервалы	$[a_1,a_2)$	$[a_2,a_3)$	…	$[a_j,a_{j+1})$	…	$[a_k,a_{k+1})$
относительные частоты	$\frac{m_1}{n}$	$\frac{m_2}{n}$	…	$\frac{m_j}{n}$	…	$\frac{m_k}{n}$

Предполагаемая функция $p(x)$ - это такая функция, график которой наилучшим образом, на наш взгляд, высказанный в форме гипотезы H_0, представляет построенная нами гистограмма.

Если функция $p(x)$ является плотностью вероятностей предполагаемого теоретического закона распределения вероятностей, то вероятность того, что случайная величина ξ примет значение, принадлежащее полуинтервалу $[a_j;a_{j+1})$,то есть $p_j=P(\xi\in[a_j,a_{j+1}))$, будет равна: $p_j=\int_{a_j}^{a_{j+1}}p(x)dx$. Значение этого интеграла равно площади криволинейной трапеции, ограниченной сверху частью графика функции $p(x)$, находящейся над полуинтервалом $[a_j;a_{j+1})$.



Каждый случайный элемент выборки ξ_i с вероятностью p_j может принять значение, принадлежащее полуинтервалу $[a_j; a_{j+1})$, и может с вероятностью $q_j = 1 - p_j$ не принять значение, принадлежащее этому полуинтервалу.

Число $\frac{m_j}{n}$ - относительная частота попаданий случайных элементов выборки $(\xi_1, \xi_2, ..., \xi_n)$ в полуинтервал $[a_j; a_{j+1})$. Относительная частота $\frac{m_j}{n}$ является статистической оценкой теоретической вероятности p_j.

Площадь прямоугольника гистограммы, имеющего основанием полуинтервал $[a_j; a_{j+1})$, равна значению относительной частоты $\frac{m_j}{n}$. А так как относительная частота наступлений события является статической оценкой вероятности наступления этого события, то, при верном предположении о виде закона распределения случайной величины ξ, величины площадей прямоугольников гистограммы будут мало отличаться от величин площадей соответствующих криволинейных трапеций.

Теория

Для выбранного числа k вычислим вероятности p_j. Если $j = 2,3,...,k-1$, то $p_j = \int_{a_j}^{a_{j+1}} p(x)dx$. Если j=1, то $p_1 = \int_{-\infty}^{a_2} p(x)dx$, а если j=k, то $p_k = \int_{a_k}^{+\infty} p(x)dx$. Ясно, что $\sum_{j=1}^{k} p_j = 1$.

Определим случайную величину ν_{ij}, которая принимает значение ν_{ij} =1, если случайная величина ξ_i принимает значение, принадлежащее полуинтервалу $[a_j; a_{j+1})$, и ν_{ij} =0, если значение ξ_i не принадлежит этому полуинтервалу. Случайная величина ν_{ij} имеет ряд распределения

ν_{ij}	0	1
	$1-p_j$	p_j

, то есть - это случайная величина $\nu_{ij} \in B_1(p_j)$, которую мы назвали бернуллиевской. Сумма $\sum_{i=1}^{n} \nu_{ij} = m_j$ таких случайных величин для каждого фиксированного значения j является случайной величиной и принимает значение равное количеству элементов числовой выборки $\{x_1, x_2, ..., x_n\}$, попавших в полуинтервал $[a_j; a_{j+1})$.

Случайная величина $\boldsymbol{m_j}$ подчиняется биномиальному распределению $\boldsymbol{m_j} \in \mathcal{B}_n\left(\boldsymbol{p_j}\right)$ и $\boldsymbol{M m_j = np_j}$, $\boldsymbol{D m_j = np_j(1-p_j)}$. Согласно интегральной теореме Муавра-Лапласа центрированная и нормированная случайная величина $\dfrac{\boldsymbol{m_j - np_j}}{\sqrt{\boldsymbol{np_j(1-p_j)}}}$ при большом объёме выборки $\boldsymbol{n}$ подчиняется распределению близкому к нормальному распределению, то есть мы можем считать, что $\dfrac{\boldsymbol{m_j - np_j}}{\sqrt{\boldsymbol{np_j}}} \in \mathcal{N}(0,1)$.

Здесь мы полагаем, что при большом объёме выборки $\boldsymbol{n}$ мы можем выбрать достаточно большое число полуинтервалов $\boldsymbol{k}$. Это приведёт к уменьшению длин полуинтервалов $\left[\boldsymbol{a_j; a_{j+1}}\right)$, а тогда мы можем считать, что $1 - \boldsymbol{p_j} \approx 1$.

Случайная величина, являющаяся суммой квадратов случайных величин, подчиняющихся нормальному распределению $\mathcal{N}(0,1)$, подчиняется распределению Пирсона. То есть: $\sum\limits_{j=1}^{k} \dfrac{\left(\boldsymbol{m_j - np_j}\right)^2}{\boldsymbol{np_j}} = \boldsymbol{\chi^2_{k-1}}$. Параметр распределения Пирсона здесь равен $\boldsymbol{k}$-1, так как у нас есть одно уравнение «связи»: $\sum\limits_{j=1}^{k} \boldsymbol{m_j = n}$, уменьшающее число независимых случайных величин-слагаемых. К. Пирсон (1900г.) предложил случайную величину $\boldsymbol{\chi^2_{k-1}} = \sum\limits_{j=1}^{k} \dfrac{\left(\boldsymbol{m_j - np_j}\right)^2}{\boldsymbol{np_j}}$ использовать в качестве критерия проверки справедливости основной гипотезы $\boldsymbol{H_0}$.

Критерий проверки справедливости основной гипотезы построен. Но при практических исследованиях чаще всего встречаются случаи, когда предполагаемое теоретическое распределение $\boldsymbol{P\left(x, \vartheta_1, ..., \vartheta_l\right)}$ содержит числовые параметры $\boldsymbol{\vartheta_1, ..., \vartheta_l}$, значения которых мы не знаем. О значениях этих параметров мы можем судить лишь по значениям их точечных оценок $\boldsymbol{\vartheta_1^*, ..., \vartheta_l^*}$, которые мы можем определить по элементам числовой выборки $\left\{\boldsymbol{x_1, x_2, ..., x_n}\right\}$. Поэтому фактически мы вычисляем величину $\sum\limits_{j=1}^{k} \dfrac{\left(\boldsymbol{m_j - np_j\left(\vartheta_1^*, ..., \vartheta_l^*\right)}\right)^2}{\boldsymbol{np_j\left(\vartheta_1^*, ..., \vartheta_l^*\right)}}$, заменяя значения числовых параметров значениями их точечных оценок. Для того чтобы применять эту случайную величину в качестве критерия, Р.Фишер (1922г.) предложил уменьшить число степеней свободы предельного распределения на $\boldsymbol{l}$ единиц, то есть на число параметров, заменяемых их точечными оценками.

Итак, для проверки справедливости основной гипотезы H_0 мы будем использовать критерий $\chi^2_{k-1-l} = \sum_{j=1}^{k} \frac{\left(m_j - np_j\left(\vartheta_1^*,...,\vartheta_l^*\right)\right)^2}{np_j\left(\vartheta_1^*,...,\vartheta_l^*\right)}$.

Например, если мы проверяем справедливость гипотезы H_0: *Случайная величина* ξ *подчиняется нормальному распределению вероятностей, то есть* $\xi \in \mathcal{N}(m,\sigma)$. Аналитическая запись нормального закона зависит от значений числовых параметров m и σ^2, то есть здесь l=2. При вычислении значения критерия χ^2_{k-1-2} мы эти числовые параметры заменим их точечными оценками $\bar{x}$ и s^2. Дальнейшие действия при проверке справедливости основной гипотезы – обычные.

Практика

Рассмотрим примеры 2*а*.1 и 2*б*.1, в которых приводятся выборки объёмами n=115 и n=906 значений случайной величины ξ - рост мужчин некоторой генеральной совокупности. Вид гистограмм этих примеров позволяет нам сформулировать основную гипотезу H_0: *Случайная величина* ξ *подчиняется нормальному распределению вероятностей*. Или коротко H_0: $\xi \in \mathcal{N}(m,\sigma)$.

Для проверки справедливости этой гипотезы мы, согласно К.Пирсону, должны использовать критерий $\chi^2_{k-1} = \sum_{j=1}^{k} \frac{\left(m_j - np_j(m,\sigma)\right)^2}{np_j(m,\sigma)}$,

где $p(x,m,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$ - плотность нормального закона $\mathcal{N}(m,\sigma)$. Аналитическая запись плотности вероятности содержит две числовые характеристики – математическое ожидание m и дисперсию σ^2, то есть здесь, как было ранее подчёркнуто, $l = 2$. Точного значения их мы не знаем, поэтому при вычислении вероятностей $p_j = P\left(\xi \in \left[a_j, a_{j+1}\right)\right)$ мы заменим эти числовые характеристики значениями их точечных оценок – средним арифметическим $\bar{x}$ и статистической дисперсией s^2. Значит, следуя Р.Фишеру, мы будем использовать критерий $\chi^2_{k-1-2} = \sum_{j=1}^{k} \frac{\left(m_j - np_j(\bar{x},s)\right)^2}{np_j(\bar{x},s)}$.

Результаты вычислений необходимых для определения наблюдаемого значения критерия $\chi^2_{набл.}$ удобно записывать в виде таблицы. Для выборки, объём которой равен n=906, получены результаты, приведённые в таблице.

№	Интервалы $[a_j, a_{j+1})$	Вероятности p_j	Теоретич. частоты np_j	Эксперимент. частоты m_j	$\frac{(m_j - np_j)^2}{np_j}$
1	$(-\infty;155)$	0,0269	24,3714	27	0,2835
2	[155;160)	0,1105	100,113	103	0,0825
3	[160;165)	0,2594	235,0164	233	0,0173
4	[165;170)	0,3165	286,7490	308	1,5749
5	[170;175)	0,2054	186,0924	169	1,5699
6	$[175; +\infty)$	0,0813	73,6578	66	0,1040
	суммы	1,000	906	906	$\chi^2_{набл.} = 3{,}6321$

По таблицам распределения «хи-квадрат» для уровня значимости $\alpha = 0{,}05$ при значении параметра $k = 6\text{-}1\text{-}2=3$ определяем критическое значение критерия $\chi^2_{кр.} = 7{,}815$. Критическая область будет иметь вид $S_{кр.} = (7{,}815; +\infty)$, а область допустимых значений критерия: $Q_{доп.} = (0; 7{,}815)$. Наблюдаемое значение критерия равно $\chi^2_{набл.} = 3{,}632$. Так как мы получили, что $\chi^2_{набл.} = 3{,}632 < \chi^2_{кр.} = 7{,}815$ или $\{\chi^2_{набл.} \in Q_{доп.}\}$, то, согласно правилу принятия решений, у нас нет оснований отклонять гипотезу H_0. Формулируем вывод: «Случайная величина ξ - рост мужчин анализируемой генеральной совокупности, - подчиняется нормальному распределению вероятностей с параметрами $M\xi = 166{,}61$ и $D\xi = 36{,}24$.

Аналогичный результат и вывод получаются при обработке статистических данных другой выборки, объём которой равен n=115. По элементам этой выборки получаем $\chi^2_{набл.} = 3{,}672$, то есть, при таком же уровне значимости $\alpha = 0{,}05$ основной гипотезы H_0 мы наблюдаем наступление события $\{\chi^2_{набл.} \in Q_{доп.}\}$. Значит, у нас нет оснований отклонять гипотезу H_0. Для этой генеральной совокупности мы можем считать, что $M\xi = 166{,}77$ и $D\xi = 34{,}69$. Хотя принципиальной разницы в принятых значениях числовых характеристик этих генеральных совокупностей – нет.

Комментарий

Построенный критерий Пирсона называется *параметрическим* критерием, так как в нём «участвуют» значения числовых параметров предполагаемого теоретического распределения вероятностей. Значения числовых параметров заменены значениями их точечных оценок. Это привело к уменьшению значения числа степеней свободы, то есть к уменьшению параметра "χ^2-распределения".

Критерии Колмогорова и Мизеса называются *непараметрическими* критериями, так как в них мера отличия $\mathcal{D}(P_n^*, P)$ эмпирического

распределения $P_n^*(x)$ от предполагаемого теоретического распределения $P(x)$ не предполагает знания значений числовых параметров теоретического распределения.

Теория

Если вероятностным распределениям $P(x)$ и $P_n^*(x)$ соответствуют функции распределения $F(x)$ и $F_n^*(x)$, а случайная величина ξ - непрерывного типа, то, согласно А.Н.Колмогорову (1933г.), рассматривается мера отличия $D(P_n^*, P)$ эмпирического распределения от предполагаемого распределения следующего вида: $D(P_n^*, P) = \sup_{x \in R} |F_n^*(x) - F(x)|$. Эта мера - статистика, Она не зависит от вида предполагаемой функции распределения $F(x)$. У этой статистики должна быть функция распределения. Справедлива теорема А.Н. Колмогорова:

Если $F(x)$ - функция распределения непрерывной случайной величины ξ, то существует $\lim_{n \to \infty} P(\sqrt{n} \cdot D(P_n^*, P) < x) = K(x)$.

Здесь $K(x) = \sum_{k=-\infty}^{+\infty} (-1)^k \cdot e^{-2k^2x^2}$ - функция распределения Колмогорова, которая табулирована во многих руководствах по математической статистике. Для выбранного уровня значимости α по таблицам находим величину $k_{кр.}$, для которой $K(k_{кр.}) = 1 - \alpha$. Значит с уверенностью 1-α для любого $x \in R$ должно выполняться неравенство $|F_n^*(x) - F(x)| < k_{кр.}$. Если хотя бы для одного значения аргумента это неравенство не выполняется, то гипотезу H_0 отклоняем.

Критерий Мизеса-Смирнова (1931г.,1936г.) (называется «критерий ω^2 ») в качестве меры отличия $D(P_n^*, P)$ рассматривает статистику $\omega^2 = n \cdot \int_R (F(x) - F_n^*(x))^2 dF(x)$.

Имеются таблицы предельного распределения этой статистики $\lim_{n \to \infty} P(\omega^2 < x) = \Omega(x)$. То есть, при $n \to \infty$ распределение статистики ω^2 не зависит от проверяемой функции распределения $F(x)$. Критерий ω^2 применяется, как и критерий Колмогорова, при исследовании случайных величин непрерывного типа.

Для определения наблюдаемого значения критерия сначала ранжируем числовую выборку $\{x_1, x_2, ..., x_n\}$, то есть, переписываем её в виде $\{x_{(1)}, x_{(2)}, ..., x_{(m)}\}$, где для любого номера i выполняется $x_{(i)} < x_{(i+1)}$. Так как случайная величина непрерывного типа, то мы считаем, что среди элементов выборки $\{x_1, x_2, ..., x_n\}$ нет одинаковых.

Наблюдаемое значение критерия вычисляется по формуле:

$$\omega^2_{набл.} = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left(F\left(x_{(i)}\right) - \frac{2i-1}{2n}\right)^2 .$$

Дальнейшая процедура проверки справедливости гипотезы H_0 – обычная.

Комментарий

Сделаем несколько замечаний об особенностях рассмотренных критериев.

Если применяется критерий К.Пирсона, то, при вычислении вероятностей p_j для определения наблюдаемого значения критерия, желательно объединять рядом находящиеся группы данных, чтобы каждая получающаяся группа содержала не менее 10 значений элементов числовой выборки $\{x_1, x_2, ..., x_n\}$. Естественно, объединяя данные этих групп в одну группу, мы увеличиваем длину полуинтервала интегрирования, по которому вычисляется p_j. При этом приходится жертвовать некоторой долей информации, содержащейся в выборке. Значение $k-1-l$ параметра "χ^2 - распределения" соответственно уменьшается.

Асимптотическим уровням значимости α можно доверять при больших значениях объёма выборки n.

Критерии Колмогорова и Мизеса-Смирнова не зависят от конкретного вида проверяемого вероятностного распределения $P(x)$. Поэтому таблицы функций распределения этих критериев могут быть использованы при любых $P(x)$ и любых объёмах выборок n.

Как было сказано, применение критерия К.Пирсона предусматривает группировку статистических данных по интервалам разбиения для вычисления значений частот m_j. На практике, когда объём выборки не очень большой (например, $n \approx 50 \div 120$) объединение нескольких полуинтервалов в один осуществляется до тех пор, пока не будет $\min m_j \approx 3 \div 5$.

При сравнительно небольшой выборке, особенно когда при малом числе полуинтервалов трудно высказать предположение о виде закона распределения вероятностей случайной величины, желательно применение критериев, учитывающих индивидуальные особенности выборочных данных. Такими критериями являются непараметрические критерии и, в частности, критерий Мизеса или, критерий Манна-Уитни [4].

§5.3 Задачи статистической проверки гипотезы о совпадении законов распределения вероятностей случайных величин

Комментарий

Задачи такого типа могут возникнуть при сравнении законов распределения вероятностей случайных величин – количественных признаков, определённых на разных одноимённых генеральных совокупностях.

Например, возникает необходимость оценить: одинаковы ли распределения растений нивяника по числу язычковых цветков в соцветии, растущих в разных местах? Или - необходимо сравнить распределение количественных показателей качества знаний по одному и тому же предмету двух групп студентов или двух потоков абитуриентов.

Мы изучаем случайные величины ξ и η - два одноимённых количественных признака и нам надо проверить предположение о том, что эти случайные величины имеют одинаковый закон распределения вероятностей. Прийти к такому, или противоположному предположению мы можем, анализируя гистограммы, построенные по выборкам возможных значений изучаемых случайных величин.

Теория

Основная гипотеза о совпадении законов распределения вероятностей $P_1(x)$ и $P_2(y)$ двух случайных величин ξ и η записывается так:

H_0: *Случайные величины ξ и η имеют одинаковый закон распределения вероятностей*.

Заметим, что при этом нас не интересует вид этого закона распределения вероятностей.

Производится эксперимент, то есть - отбор экспериментальных данных из двух рассматриваемых генеральных совокупностей. Мы получаем две выборочные совокупности: $\{\xi_1, \xi_2, ..., \xi_{n_1}\}$ и $\{\eta_1, \eta_2, ..., \eta_{n_2}\}$, объёмы которых равны n_1 и n_2. Как обычно, при первичной обработке статистических данных сначала определяем количество полуинтервалов вариационного ряда k и границы этих полуинтервалов $\Delta_j \equiv [a_j; a_{j+1})$, где $j = 1,2,...,k$. И количество интервалов, и границы этих интервалов для рассматриваемых выборок должны быть одинаковыми.

Составляем два вариационных ряда частот:

1	2	...	j	...	k
Δ_1	Δ_2	...	Δ_j	...	Δ_k
m'_1	m'_2	...	m'_j	...	m'_k

1	2	...	j	...	k
Δ_1	Δ_2	...	Δ_j	...	Δ_k
m''_1	m''_2	...	m''_j	...	m''_k

.

Здесь m'_j – количество элементов первой числовой выборки $\{x_1, x_2, ..., x_{n_1}\}$, а m''_j – количество элементов второй числовой выборки $\{y_1, y_2, ..., y_{n_2}\}$, которые попали в интервал $\Delta_j \equiv [a_j; a_{j+1})$. Ясно, что $\sum_{j=1}^{k} m'_j = n_1$ и $\sum_{j=1}^{k} m''_j = n_2$.

Обозначим p'_j и p''_j - вероятности того, что значения случайных величин ξ и η, соответственно, принадлежит интервалу Δ_j, то есть: $p'_j = P(\xi \in \Delta_j)$ и $p''_j = P(\eta \in \Delta_j)$.

Элементы выборки – это независимые случайные величины, имеющие то же распределение, что и исследуемая случайная величина. Как и при построении критерия Пирсона, «пересмотрим» все элементы совокупности: $\{\xi_1, \xi_2, ..., \xi_{n_1}\}$. Вероятность того, что значение элемента ξ_i первой выборки будет принадлежать интервалу Δ_j, равна p'_j. Перебирая все элементы этой выборки и подсчитывая количество элементов выборки, значения которых принадлежат интервалу Δ_j, мы получаем значение случайной величины m'_j. Следовательно, случайная величина m'_j – количество элементов первой выборки, попавших в интервал Δ_j, имеет биномиальное распределение вероятностей и $Mm'_j = n_1 p'_j$, $Dm'_j = n_1 p'_j q'_j$. Аналогично рассуждая при анализе элементов второй совокупности $\{\eta_1, \eta_2, ..., \eta_{n_2}\}$, приходим к выводу, что случайная величина m''_j так же имеет биномиальное распределение вероятностей и $Mm''_j = n_2 p''_j$, $Dm''_j = n_2 p''_j q''_j$.

Числа p'_j и p''_j - это значения теоретических вероятностей. Если гипотеза H_0 - верна, то для всех номеров j эти вероятности будут равны, то есть будет справедливо равенство $p'_j = p''_j$. Обозначим общее значение этих вероятностей p_j.

Относительные частоты $\frac{m'_j}{n_1}$ и $\frac{m''_j}{n_2}$ являются состоятельными и несмещёнными оценками вероятностей p'_j и p''_j. Если гипотеза H_0 - верна, то для всех номеров j значения относительных частот будут мало отличаться друг от друга.

Объединим выборочные совокупности: $\{\xi_1, \xi_2, ..., \xi_{n_1}\}$ и $\{\eta_1, \eta_2, ..., \eta_{n_2}\}$ в одну совокупность объёмом $n_1 + n_2$. Для каждого номера j рассмотрим

статистику $\mu_j = \frac{m'_j}{n_1} - \frac{m''_j}{n_2}$. По величине модулей значений этих k штук статистик μ_j, можно судить о справедливости, или – нет, гипотезы H_0.

Если гипотеза H_0 верна, то для всех номеров j математическое ожидание $M\mu_j$ статистики $\mu_j = \frac{m'_j}{n_1} - \frac{m''_j}{n_2}$ должно быть равно нулю:

$$M\mu_j = M\left[\frac{m'_j}{n_1} - \frac{m''_j}{n_2}\right] = p'_j - p''_j = 0.$$

Вычислим значение дисперсии этой статистики, учитывая, что выборки - независимые:

$$D\mu_j = D\left[\frac{m'_j}{n_1} - \frac{m''_j}{n_2}\right] = D\left[\frac{m'_j}{n_1}\right] + D\left[\frac{m''_j}{n_2}\right] = \frac{1}{n_1^2} D\left[m'_j\right] + \frac{1}{n_2^2} D\left[m''_j\right] = \frac{p'_j q'_j}{n_1} + \frac{p''_j q''_j}{n_2}.$$

Как и при построении критерия Пирсона, считаем, что значения вероятностей противоположных событий q'_j и q''_j будут мало отличаться от единицы, так как при больших объёмах выборки, когда длина каждого интервала Δ_j мала. Мы уже сказали, что при справедливости гипотезы H_0 вероятности p'_j и p''_j будут равны. Заменяя значения вероятностей q'_j и q''_j единицей и считая, что $p'_j = p''_j = p_j$, упростим выражение значения дисперсии рассматриваемой статистики μ_j, построенной по объединённой выборке: $D\mu_j = D\left[\frac{m'_j}{n_1} - \frac{m''_j}{n_2}\right] = p_j\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = p_j \cdot \frac{n_1 + n_2}{n_1 n_2}$.

Используя эту объединённую выборку, определим точечную оценку ϑ_j^* теоретической вероятности $\vartheta_j = p_j$. Количество наблюдений элементов объединённой выборки, принадлежащих интервалу Δ_j, будет равно $m'_j + m''_j$. Следовательно, оценкой теоретической вероятности p_j будет относительная частота $\vartheta_j^* = \frac{m'_j + m''_j}{n_1 + n_2}$. Заменим в выражении для дисперсии значение неизвестной теоретической вероятности p_j её точечной оценкой, то есть – относительной частотой:

$$D\mu_j = D\left[\frac{m'_j}{n_1} - \frac{m''_j}{n_2}\right] = \frac{m'_j + m''_j}{n_1 + n_2} \cdot \frac{n_1 + n_2}{n_1 n_2} = \frac{m'_j + m''_j}{n_1 n_2}.$$

Согласно теореме Муавра-Лапласа, мы можем утверждать, что при больших объёмах выборок n_1 и n_2, закон распределения статистики μ_j

будет мало отличаться от нормального закона. То есть, центрированная и нормированная случайная величина $\tau_j = \dfrac{\mu_j - M[\mu_j]}{\sqrt{D[\mu_j]}}$ имеет распределение вероятностей, мало отличающееся нормального распределения $\mathcal{N}(0;1)$.

А тогда сумма квадратов $\sum_{j=1}^{k} \tau_j^2$ этих случайных величин будет подчиняться “ χ^2 - распределению”. Так как статистики $m'_j + m''_j$, определяемые по элементам объединённой выборки, «связаны» уравнением $\sum_{j=1}^{k} \left(m'_j + m''_j\right) = n_1 + n_2$, то на единицу уменьшается число степеней свободы. Следовательно, параметр “ χ^2 -распределения” будет равен k-1. Упростим вид суммы квадратов $\sum_{j=1}^{k} \tau_j^2$ и запишем в окончательной форме построенный критерий проверки гипотезы H_0:

$$\chi^2_{k-1} = \sum_{j=1}^{k} \tau_j^2 = \sum_{j=1}^{k} \frac{\left(\dfrac{m'_j}{n_1} - \dfrac{m''_j}{n_2}\right)^2}{\dfrac{m'_j + m''_j}{n_1 n_2}} = n_1 n_2 \sum_{j=1}^{k} \frac{1}{m'_j + m''_j} \cdot \left(\frac{m'_j}{n_1} - \frac{m''_j}{n_2}\right)^2 .$$

Вид критерия несколько упрощается, если объёмы выборок будут одинаковыми, то есть если $n_1 = n_2$: $\chi^2_{k-1} = \sum_{j=1}^{k} \dfrac{\left(m'_j - m''_j\right)^2}{m'_j + m''_j}$.

Комментарий

Процесс построения критерия проверки гипотезы о совпадении законов распределения двух случайных величин во многом аналогичен процессу построения критерия Пирсона для проверки гипотезы о виде закона распределения случайной величины. Дальнейшие действия экспериментатора при проверке справедливости основной гипотезы полностью совпадают с теми действиями, которые описаны при проверке гипотез предыдущих групп.

Практика

В примерах 2*а*.1 и 2*б*.1 введём две одноимённые случайные величины ξ и η - рост мужчин, двух различных генеральных совокупностей. Из первой совокупности произведена выборка объёмом n_1=115, а из второй – объёмом n_2=906. При первичной обработке статистических данных выбрано одинаковое число интервалов (k=6) и построены интервалы с одинаковыми границами. Построены вариационные ряды и сделаны гистограммы. Выдвигается основная гипотеза:

$\boldsymbol{H_0}$: *Законы распределения вероятностей случайных величин* ξ *и* η *- совпадают.*

Для проверки справедливости гипотезы $\boldsymbol{H_0}$ вычислим сначала наблюдаемое значение критерия $\chi^2_{набл.} = n_1 n_2 \sum_{j=1}^{6} \frac{1}{m'_j + m''_j} \cdot \left(\frac{m'_j}{n_1} - \frac{m''_j}{n_2} \right)^2$.

Все данные и результаты промежуточных вычислений запишем в таблице.

j	m'_j	m''_j	$m'_j + m''_j$	$\frac{m'_j}{n_1}$	$\frac{m''_j}{n_2}$
1	4	27	31	0,035	0,030
2	13	103	116	0,113	0,140
3	31	233	264	0,270	0,257
4	41	308	349	0,356	0,340
5	16	169	175	0,139	0,186
6	10	66	76	0,087	0,073

Наблюдаемое значение критерия будет равно:

$$\chi^2_{набл.} = n_1 n_2 \sum_{j=1}^{6} \frac{1}{m'_j + m''_j} \cdot \left(\frac{m'_j}{n_1} - \frac{m''_j}{n_2} \right)^2 = 2{,}466 .$$

По таблицам "χ^2 - распределения " при уровне значимости $\alpha = 0{,}05$ для k-1=6-1=5 определяем критическое значение $\chi^2_{кр.} = 11{,}07$. Мы можем не строить области $S_{кр}$ и $Q_{доп}$. Так как для наступления события $\{\chi^2_{набл.} \in Q_{доп.}\}$ достаточно выполнения неравенства $\chi^2_{набл.} < \chi^2_{кр.}$, то заключаем, что у нас нет оснований отклонять гипотезу $\boldsymbol{H_0}$. То есть, законы распределения вероятностей случайных величин ξ и η - совпадают.

Практика

Рассмотрим теперь пример статистической проверки гипотезы о совпадении законов распределения случайных величин ξ и η дискретного типа.

Пример. В любом художественном произведении длина наудачу выбранного предложения, то есть количество слов в предложении, - случайна. Ясно, что каждый автор имеет свой индивидуальный стиль изложения своего произведения: у одного автора предложения в основном короткие, состоящие из малого количества слов, у другого автора более часто встречаются длинные предложения, состоящие из большого количества слов. Если обозначить ξ и η – случайные величины – длины предложений в произведениях двух авторов, то, изучая выборки значений

этих случайных величин, можно составить некоторое суждение и сделать некоторые заключения о способах написания своих произведений этими авторами, делать выводы о совпадении, или различии законов распределения случайных величин ξ и η.

Не меняя постановки задачи: сравнение законов распределения двух случайных величин – количество слов в предложении в произведениях двух авторов, рассмотрим следующий пример.

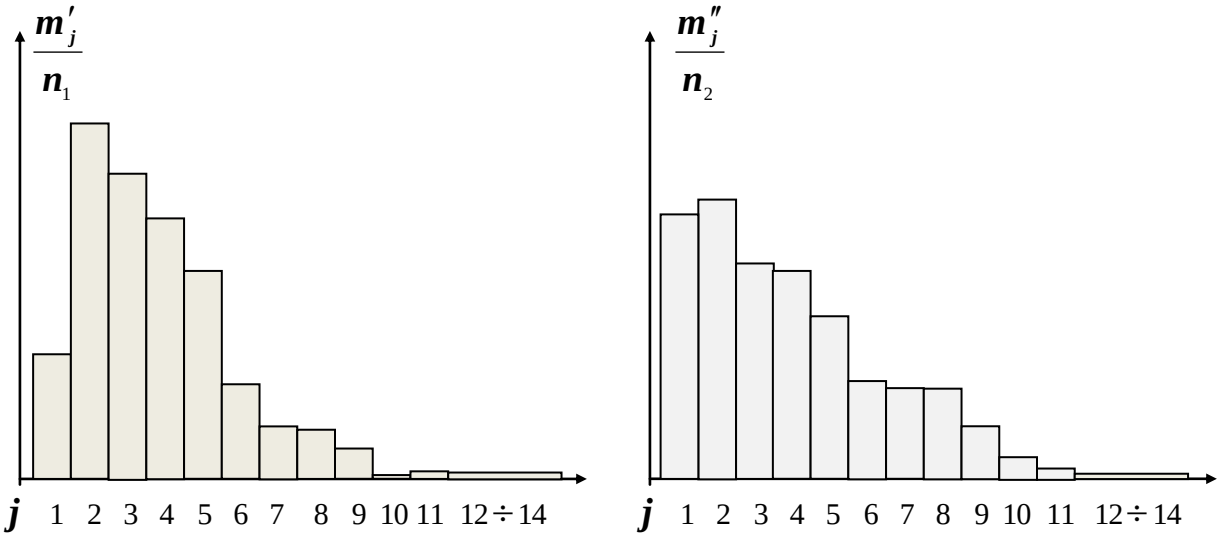
Пусть: ξ – количество слов в предложении в научно- фантастическом романе «Ариэль» А. Беляева; η – количество слов в предложении в романе «Два капитана» В. Каверина.

В этих произведениях были выбраны наудачу по четыреста предложений, в которых были подсчитаны количества слов. В результате были получены две выборки, объёмы которых равны n=400.

Запишем результаты первичной обработки статистических данных и, для наглядного представления о теоретическом распределении этих случайных величин, сделаем гистограммы.

J	1	2	3	4	5	6	7	8	9	10	11	12
Количество слов	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	≥ 34
Выборка ***Б***	33	94	81	69	55	25	14	13	8	1	2	5
Выборка ***К***	70	74	57	55	43	26	24	24	14	6	3	4

Вычисляем значения точечных оценок основных числовых характеристик - математического ожидания и дисперсии случайных величин ξ и η: $\bar{x} = 10{,}485$, $s_1^2 = 43{,}378$; $\bar{y} = 11{,}065$, $s_2^2 = 61{,}169$.



Сформулируем сначала предположение - гипотезу о том, что средние количества слов в произведениях А.Беляева и В.Каверина равны. Проверка гипотезы о равенстве математических ожиданий одноимённых случайных

величин относится к задачам второго типа статистической проверки гипотез. Основная гипотеза записывается так: $H_0: M\xi = M\eta$.

При подборе критерия проверки справедливости гипотезы H_0 примем во внимание два факта. Во-первых, так как оценка дисперсии случайной величины η примерно в 1,4 раза больше оценки дисперсии случайной величины ξ, то мы не можем достаточно уверенно считать, что дисперсии случайных величин ξ и η равны, или их различие - незначимо. Критерий Фишера – Снедекора показывает, что $f_{набл.} = \frac{s_2^2}{s_1^2} = 1{,}410 > f_{кр.} = 1{,}28$, что укрепляет наше предположение, что нельзя считать, что значения дисперсий $D\xi$ и $D\eta$ равны.

Так как объёмы выборок $n_1 = n_2 = 400 = n$ достаточно большие, то мы можем, учитывая теорему Леви (ЦПТ) об асимптотической нормальности среднего арифметического, выбрать в качестве критерия статистику $u = \frac{\bar{x} - \bar{y}}{\sqrt{s_1^2 + s_2^2}}\sqrt{n}$, распределение вероятностей которой в случае справедливости гипотезы H_0 мало отличается от нормального $\mathcal{N}(0;1)$. То есть, при построении областей $S_{кр}$ и $Q_{доп}$ значений критерия u, используются значения функции Лапласа. При двусторонней альтернативной гипотезе $H_1: M\xi \neq M\eta$, с уровнем значимости $\alpha = 0{,}05$ по таблицам определяем $u_{кр}$ такое, при котором будет справедливо: $P(|u| > u_{кр}) = \alpha$, то есть: $u_{кр} = \Phi^{-1}\left(\frac{1-\alpha}{2}\right) = \Phi^{-1}(0{,}475) = 1{,}96$. Значит: $S_{кр} = (-\infty; -1{,}96) \cup (1{,}96; +\infty)$ и $Q_{доп} = (-1{,}96; +1{,}96)$. Вычисляем наблюдаемое значение критерия: $u_{набл} = \frac{10{,}485 - 11{,}065}{\sqrt{43{,}378 + 61{,}169}} \cdot 20 = -1{,}1345.$

Принимаем решение: так как наблюдаемое значение $u_{набл}$ критерия попало в область $Q_{доп}$, то у нас нет оснований отклонять гипотезу H_0, то есть гипотеза H_0 – принимается. Практически мы можем утверждать, что разность значений математических ожиданий числа слов в предложениях этих авторов - незначительная.

То есть, средняя длина предложения в произведении А.Беляева практически не отличается от средней длины предложения в произведении В.Каверина.

Однако, даже визуальный анализ вида гистограмм, построенных по выборкам, в каждой из которых подсчитывалась длина 400 наудачу взятых предложений, приводит нас к мысли, что законы распределения вероятностей случайных величин ξ и η не совпадают.

Следуя принципу «презумпции невиновности», формулируем основную гипотезу H_0: «*Случайные величины* ξ и η *имеют одинаковый закон распределения вероятностей*».

Так как объёмы выборок одинаковы, то будем подсчитывать наблюдаемое значение критерия по формуле: $\chi^2_{набл.} = \sum_{j=1}^{k} \frac{(m'_j - m''_j)^2}{m'_j + m''_j}$.

Как и в предыдущем примере, где рассматривались распределения роста мужчин, результаты промежуточных вычислений записываем в таблице. Последние три интервала вариационных рядов, ввиду малых значений относительных частот, объединим в один интервал.

j	m'_j	m''_j	$(m'_j - m''_j)^2$	$m'_j + m''_j$	$\frac{(m'_j - m''_j)^2}{m'_j + m''_j}$
1	33	70	1369	103	13,291
2	94	74	400	168	2,381
3	81	57	576	138	4,174
4	69	55	196	124	1,581
5	55	43	144	98	1,469
6	25	26	1	51	0,020
7	14	24	100	38	2,632
8	13	24	121	37	3,270
9	8	14	36	22	1,636
10	8	13	25	21	1,190

Сумма чисел, записанных в последнем столбце таблицы, будет наблюдаемым значением критерия $\chi^2_{набл}$:

$$\chi^2_{набл} = \sum_{j=1}^{10} \frac{(m'_j - m''_j)^2}{m'_j + m''_j} = 31{,}644.$$

При значении параметра распределения критерия k-1=10-1=9 и уровне значимости $\alpha = 0{,}05$ по таблицам ” χ^2 - распределения” определяем критическое значение критерия: $\chi^2_{кр} = 16{,}92$.

Так как наблюдаемое значение критерия оказалось больше его критического значения: $\chi^2_{набл} = 31{,}644 > \chi^2_{кр} = 16{,}92$, то по правилу принятия решений гипотеза H_0 отклоняется, то есть случайные величины ξ и η подчиняются разным законам распределения вероятностей.

Комментарий

Делая выводы об изучаемых случайных величинах, основанные на результатах статистической обработки экспериментальных данных, мы всегда должны помнить афоризм: *«Статистике часто принадлежит первое слово, но никогда - последнее»*. Чтобы пояснить смысл этого афоризма проанализируем возможные выводы и суждения из возможных ответов о совпадении, или различии законов распределения длин предложений в произведениях двух различных авторов, которые формулируются по результатам статистической проверки гипотез.

Ясно, что взятое наугад предложение из литературного произведения любого писателя не является произвольным набором слов. Построение предложения для передачи мысли автора проводится по чётким синтаксическим правилам, составляющим основу любого литературного языка. Требования синтаксических правил накладывают вполне определённые ограничения, которые всегда учитываются при построении, или написании предложения. Поэтому нельзя рассматривать наудачу выбранное из текста предложение как случайный набор некоторого количества слов. Случайные величины ξ и η - длины предложений в двух различных литературных произведениях, принадлежащим разным авторам, могут иметь или совпадающие, или очень близкие законы распределения вероятностей. При отличительных индивидуальных особенностях авторов, очевидных внимательному и мыслящему читателю, конструкция любого предложения подчиняется синтаксическим правилам, которые оказывают влияние на числовые результаты статистической обработки выборочных данных. В то же время ясно, что статистическая обработка этих данных, её результаты и выводы никогда не смогут количественно учесть эмоциональную составляющую особенности литературного языка того, или иного автора.

Если качественным признакам объектов генеральной совокупности можно присвоить количественные характеристики, то, так как в формуле критерия участвуют только частоты признаков, теорию статистической проверки гипотез можно применить и при сравнении качественных многообразий различных генеральных совокупностей.

Практика

Пример. Изучались качественные составы по породам деревьев двух лесных массивов. Были получены следующие данные числа деревьев каждой из имеющихся пород.

Лесной массив **А**

	Ель	Пихта	Граб	Клён	Бук	Вяз	Ясень	Тисс
i	1	2	3	4	5	6	7	8
m_i'	300	50	200	150	38	137	180	10

Лесной массив **B**

	Ель	Пихта	Граб	Клён	Бук	Вяз	Ясень	Тисс
I	1	2	3	4	5	6	7	8
m_i''	200	42	183	203	34	117	150	12

Можно ли утверждать, что качественный состав пород деревьев в этих лесных массивах одинаков?

Качественным признакам составов лесных массивов – встречающимся породам деревьев, мы можем дать количественные признаки, то есть - просто перенумеровать в одинаковом порядке все породы деревьев этих лесных массивов. Получается, что случайные величины ξ и η принимают значения от **1** до **8**. Основная гипотеза формулируется так H_0: *Видовая структура деревьев в этих лесных массивах – одинаковая*. То есть H_0: Случайные величины ξ и η подчиняются одинаковому закону распределения вероятностей.

Для проверки справедливости основной гипотезы будем применять критерий: $\chi^2_{k-1} = n_1 \cdot n_2 \sum_{i=1}^{k} \frac{1}{m_i' + m_i''} \cdot \left(\frac{m_i'}{n'} - \frac{m_i''}{n''} \right)^2$. Где $n_1 = \sum_{i=1}^{8} m_i'$ и $n_2 = \sum_{i=1}^{8} m_i''$ общие количества деревьев в каждом лесном массиве.

Наблюдаемое значение критерия принимает значение $\chi^2_{набл.} = 26{,}545$. По таблицам “ χ^2 - распределения” для $k-1=8-1=7$, при уровне значимости $\alpha = 0{,}05$ определяем критическое значение критерия $\chi^2_{кр.} = 14{,}07$. Так как мы наблюдаем неравенство $\chi^2_{набл.} > \chi^2_{кр.}$, то гипотезу H_0 – отклоняем, то есть мы не можем утверждать, что видовая структура деревьев этих лесных массивов одинаковая.

§6.3 Задача статистическая проверка гипотезы о независимости законов распределения вероятностей двух случайных величин

Комментарий

Два или несколько количественных признаков, наблюдаемых у элементов генеральной совокупности, можно рассматривать как компоненты двумерной или многомерной случайной величины. У наблюдателя может возникнуть вопрос: зависимы, или нет исследуемые признаки?

Наблюдаемые значения двух количественных признаков у объектов генеральной совокупности трактуются как возможные значения компонент ξ и η двумерной случайной величины $\zeta = (\xi, \eta)$.

Компоненты ξ и η двумерной случайной величины, являясь случайными величинами, имеют законы распределения вероятностей своих возможных значений. Законы распределения вероятностей компонент могут быть или независимыми, или – зависимыми.

Например, отклонение длины (случайная величина ξ) и толщины (случайная величина η) детали от номинала; вес (случайная величина ξ) и количество зёрен (случайная величина η) в одном колосе изучаемого сорта пшеницы; цена одного барреля нефти (случайная величина ξ) и курс доллара (случайная величина η) в течение некоторого анализируемого периода.

Теория

Основная гипотеза при решении задач такого типа записывается так:

$\boldsymbol{H_0}$**:** *Случайные величины* ξ *и* η *- независимы.*

Согласно определению в курсе «Теория вероятностей» независимости случайных величин вероятностная функция вектора должна быть равна произведению частных вероятностных функций компонент. А поэтому мы основную гипотезу можем сформулировать так:

$\boldsymbol{H_0}$**:** *Функция распределения векторной случайной величины* $\zeta=(\xi,\eta)$ *равна произведению частных функций распределения компонент* ξ *и* η *:* $F(x,y)=F_1(x)\cdot F_2(y)$.

То есть, каждая компонента принимает свои возможные значения с вероятностями, которые не зависят от значений принятых другой компонентой.

Если гипотеза $\boldsymbol{H_0}$ верна, то:

а) для любого элемента (x,y) из множества возможных значений двумерной случайной величины $\zeta=(\xi,\eta)$ должны выполняться равенства:

$P(\xi=x_i;\eta=y_j)=P(\xi=x_i)\cdot P(\eta=y_j)$, если случайные величины ξ и η - дискретного типа. Или, более кратко, для любой пары индексов (i,j) должно быть справедливо: $p_{ij}=p_{i*}\cdot p_{*j}$.

б) плотность вероятности двумерной $\zeta=(\xi,\eta)$ должна быть равна произведению частных плотностей вероятностей компонент ξ и η, если случайные величины непрерывного типа. Или, более кратко: $p(x,y)=p_1(x)\cdot p_2(y)$.

Построим критерий проверки справедливости гипотезы $\boldsymbol{H_0}$.

В рассматриваемой задаче элементами выборки будут значения двумерной случайной величины $\zeta=(\xi,\eta)$. То есть, в нашем распоряжении имеется двумерная выборка $\{(\xi;\eta)_k\}$, или $\{(\xi_k;\eta_k)\}$ $k=1,2,\ldots,n$, объём которой равен n.

Как и в случае изучения одномерной случайной величины, проводим первичную обработку статистических данных, то есть, составляем вариационную таблицу – «двумерный вариационный ряд».

Область значений первой случайной компоненты ξ_k, попавших в выборку разобьём на r интервалов: $\Delta_i = [a_i ; a_{i+1})$, $i = 1,2,...,r$. Область значений компоненты η_k, попавших в выборку разобьём на s интервалов: $\varpi_j = [b_j ; b_{j+1})$, $j = 1,2,...,s$. Каждый элемент выборки, являясь двумерной случайной величиной, $(\xi_k ; \eta_k)$с вероятностью $p_{ij} = P(\xi \in \Delta_i ; \eta \in \varpi_j)$ может принять значение в «прямоугольнике» $\Delta_i \times \varpi_j = [a_i ; a_{i+1}) \times [b_j ; b_{j+1})$. Переберём все n элементов выборки, проверяя, принадлежит, или не принадлежит каждый значение случайной величины $(\xi_k ; \eta_k)$ «прямоугольнику» $\Delta_i \times \varpi_j$. Случайное количество элементов выборки, значения которых будут принадлежать этому прямоугольнику, обозначим m_{ij}. Все результаты запишем в таблице 1.6.3.

Таблица 1.6.3

η / ξ	$j=1$ ϖ_1	$j=2$ ϖ_2	...	$j=j$ ϖ_j	...	$j=s$ ϖ_s	
$i=1$ Δ_1	m_{11}	m_{12}		m_{1j}		m_{1s}	m_{1*}
$i=2$ Δ_2	m_{21}	m_{22}		m_{2j}		m_{2s}	m_{2*}
...							...
$i=i$ Δ_i	m_{i1}	m_{i2}		m_{ij}		m_{is}	m_{i*}
...							...
$i=r$ Δ_r	m_{r1}	m_{r2}		m_{rj}		m_{rs}	m_{r*}
	m_{*1}	m_{*2}	...	m_{*j}	...	m_{*s}	n

В последнем столбце таблицы записываем количества m_{i*} элементов двумерной выборки, первая координата которых принимает значение в интервале Δ_i, $i = 1,2,...,r$. Ясно, что $m_{i*} = \sum_{j=1}^{s} m_{ij}$. Аналогично, в последней строке таблицы записываем количества m_{*j} элементов двумерной выборки, вторая координата которых принимает в интервале ϖ_j, $j = 1,2,...,r$. Ясно, что $m_{*j} = \sum_{i=1}^{r} m_{ij}$. Если при определении частот попаданий m_{ij} в «прямоугольники» $\Delta_i \times \varpi_j$ все числовые значения элементов выборки

$\{(x;y)_k\}$, $k=1,2,...,n$, были учтены, то всегда будет справедливо равенство: $\sum_{i=1}^{r} m_{i*} = \sum_{j=1}^{s} m_{*j} = n$.

Случайная величина m_{ij} - количество элементов выборки $\{(\xi_k;\eta_k)\}$, попавших в «прямоугольник» $\Delta_i \times \varpi_j$, подчиняется биномиальному распределению вероятностей $B_n(p_{ij})$.

Математическое ожидание случайной величины m_{ij} будет равно: $Mm_{ij} = np_{ij}$. Для достаточно большого объёма выборки n, можно считать, что дисперсия случайной величины m_{ij} будет мало отличаться от этого же числа, то есть можно считать, что: $Dm_{ij} \approx np_{ij}$.

Если гипотеза H_0 верна, то для любой пары индексов (i,j) должно быть справедливо: $p_{ij} = p_{i*} \cdot p_{*j}$, где $p_{i*} = P(\xi \in \Delta_i)$ и $p_{*j} = P(\eta = \varpi_j)$. То есть, при независимости компонент ξ и η математическое ожидание и дисперсия частоты m_{ij} будут равны: $Mm_{ij} = np_{i*} \cdot p_{*j}$ и $Dm_{ij} \approx np_{i*} \cdot p_{*j}$.

Так как все частоты m_{ij} подчиняются биномиальному распределению, то, согласно теореме Муавра-Лапласа, при больших значениях n центрированная и нормированная случайная величина $\frac{m_{ij} - np_{i*} \cdot p_{*j}}{\sqrt{np_{i*} \cdot p_{*j}}}$ имеет распределение вероятностей, мало отличающееся от нормального распределения $\mathcal{N}(0,1)$.

Используя теорему Бернулли, заменим в этой записи неизвестные значения теоретических вероятностей p_{i*} и p_{*j} их точечными оценками $\frac{m_{i*}}{n}$ и $\frac{m_{*j}}{n}$. Значит, если гипотеза H_0 верна, то для всех пар индексов (i,j) случайные величины $\frac{m_{ij} - \frac{m_{i*} \cdot m_{*j}}{n}}{\sqrt{\frac{m_{i*} \cdot m_{*j}}{n}}}$ имеют распределение вероятностей, мало отличающееся от нормального распределения $\mathcal{N}(0,1)$.

Сумма квадратов независимых, подчиняющихся нормальному распределению $\mathcal{N}(0,1)$, случайных величин подчиняется ” χ^2 – распределению” с параметром равным числу независимых слагаемых. В рассматриваемом случае мы имеем $r \cdot s$ центрированных и нормированных случайных величин m_{ij}. Но, так как r штук случайных величин m_{i*} и s

штук случайных величин m_{*j} «связаны» условиями: $\sum_{i}^{r} m_{i*} = n$ и $\sum_{j=1}^{s} m_{*j} = n$, то случайная величина $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(m_{ij} - \frac{m_{i*}\cdot m_{*j}}{n}\right)^2}{\frac{m_{i*}\cdot m_{*j}}{n}}$ подчиняется "χ^2 – распределению" с параметром $(r-1)\cdot(s-1)$. Если гипотеза H_0 верна, то значения m_{ij} будут мало отличаться от значений дробей $\frac{m_{i*}\cdot m_{*j}}{n}$, и статистика $\chi^2_{(r-1)(s-1)}$ будет принимать «малые» значения. Если гипотеза H_0 неверна, среди слагаемых, формирующих значение статистики χ^2, будут слагаемые, которые принимают большие значения из-за большой разницы между значением m_{ij} и значением дроби $\frac{m_{i*}\cdot m_{*j}}{n}$. В итоге статистика $\chi^2_{(r-1)(s-1)}$ примет «большое» значение, что, при справедливости гипотезы H_0 является практически невозможным событием.

В составленной двойной сумме возведём каждое слагаемое в квадрат и, используя уравнения «связи», упростим вид статистики $\chi^2_{(r-1)(s-1)}$. Получим более простую и удобную для вычислений форму статистики – функции элементов двумерной выборки:

$$\chi^2_{(r-1)(s-1)} = n\cdot\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{m_{ij}^2}{m_{i*}\cdot m_{*j}} - 1\right).$$

Эту статистику будем использовать в качестве критерия проверки справедливости гипотезы H_0.

Практика

Пример. 5.3.2 [9] Изучались длина пластинки листа в *мм*, число зубчиков на всей пластинке и число зубчиков на 1*см* края пластинки у нивяника круглолистного.

Мы можем сказать, что рассматривалась трёхмерная случайная величина $\psi = (\xi, \eta, \zeta)$ с компонентами: ξ – длина пластинки листа в *мм* у нивяника круглолистного; η – число зубчиков на всей пластинке листа и ζ – число зубчиков на 1 *см* края пластинки листа.

Проводим эксперимент, суть которого заключается в следующем. Из генеральной совокупности, её составляют все листья растений нивяника круглолистного, растущих на изучаемом лугу, или на опушке леса, отбираются $n=100$ объектов – по одному листу с каждого растения. У каждого листа была измерена длина пластинки x в *мм*, подсчитано число зубчиков на всей пластинке y и число зубчиков z на 1 *см* края пластинки.

Результаты измерений этих трёх качественных признаков листьев нивяника обыкновенного составляют трёхмерную выборку $\{(x;y;z)_k\}$, где $k=1,2,...,100$.

Таблица 2.6.3

ξ	56	55	66	58	55	54	53	55	64	62	53	52	50	70	53	57	69	51	54	63
η	63	52	56	46	55	33	41	63	55	62	58	43	52	58	40	41	55	52	45	36
ζ	19	12	13	10	13	9	9	14	12	11	14	11	14	12	13	10	10	12	9	9

ξ	35	28	66	20	60	50	39	55	59	55	72	52	57	44	46	45	54	45	48	42
η	38	30	39	38	44	41	41	48	56	42	49	52	47	48	62	39	46	36	42	45
ξ	15	14	9	13	8	10	14	11	8	11	9	13	9	15	18	11	11	11	13	15

ξ	82	48	34	63	60	64	42	56	55	37	50	47	52	74	57	50	35	36	63	44
η	66	48	39	39	40	59	37	50	51	36	50	57	65	55	52	49	37	32	44	57
ζ	10	13	14	9	8	12	11	11	14	13	16	15	13	8	13	13	14	9	9	18

ξ	48	54	34	69	57	44	41	41	49	28	39	31	58	45	44	56	30	33	63	47
η	38	37	47	62	42	51	35	49	52	33	31	35	42	41	47	40	40	46	47	38
ζ	10	9	16	11	10	16	11	15	14	15	9	15	10	10	13	10	14	15	12	12

ξ	41	36	54	43	55	74	36	52	43	60	40	63	30	38	51	55	62	70	51	42
η	58	34	35	47	36	59	36	45	34	47	54	34	33	38	32	57	63	53	42	54
ζ	17	13	10	12	10	11	13	11	10	11	15	6	17	12	9	14	14	11	11	18

Возникает вопрос: «Зависимы, или - независимы между собой изучаемые количественные показатели пластинки листа у нивяника круглолистного?»

То есть, применяя построенный критерий, нам надо проверить гипотезы о зависимости или независимости законов распределения вероятностей трёх пар компонент трёхмерной случайной величины $\psi=(\xi,\eta,\zeta)$.

Случайная величина ξ - длина пластинки листа в *мм*, является случайной величиной непрерывного типа. Случайная величина η - число зубчиков на всей пластинке листа и случайная величина ζ – количество зубчиков на *1см* края пластинки - случайные величины дискретного типа.

Сформулируем основную гипотезу для первой пары компонент:

H_0: *Случайные величины ξ и η – независимые, то есть, число зубчиков на пластинке листа не зависит от её размеров*.

Результаты ста измерений длины пластинки листа записаны первыми координатами ста элементов трёхмерной выборки. Диапазон полученных

значений измерений $(20 \le x \le 82)$ разделим на десять интервалов Δ_i, $i = 1,2,...,10$. Результаты подсчётов числа зубчиков записаны вторыми координатами трёхмерной выборки. Диапазон полученных значений числа зубчиков $(30 \le y \le 66)$ разделим на девять интервалов ϖ_j, $j = 1,2,...,9$. Для каждой пары индексов $(i; j)$ запишем в вариационной таблице число m_{ij} - количество элементов выборки, у которых первые две координаты x и y будут принадлежать прямоугольнику $\Delta_i \times \varpi_j$. Получаемые результаты запишем в таблицу 3.6.3.

Критерий проверки справедливости гипотезы H_0 подчиняется распределению "χ^2 – распределению" с параметром $(r - 1)(s - 1) = 9 \cdot 8 = 72$. По таблицам распределения "χ^2 – распределению" при уровне значимости $\alpha = 0,05$ и значении параметра равном 72 определяем критическое значение критерия $\chi^2_{кр}$=92,76, исходя из условия, что при справедливости гипотезы H_0 событие $\{\chi^2_{72} > \chi^2_{кр}\}$ будет практически невозможным. То есть вероятность наступления этого события равна $\alpha = 0,05 : P(\chi^2_{72} > \chi^2_{кр}) = \alpha$. Критическая область и область допустимых значений критерия будут иметь вид: $S_{кр} = (92,76; +\infty)$ и $Q_{доп} = (0; 92,76)$.

Вычислим наблюдаемое значение критерия:

$$\chi^2_{набл} = n \cdot \left(\sum_{i=1}^{10} \sum_{j=1}^{9} \frac{m_{ij}^2}{m_{i*} \cdot m_{*j}} - 1 \right) = 100 \cdot (1,8883 - 1) = 88,83 .$$

Так как наблюдаемое значение критерия оказалось меньше его критического значения: $\chi^2_{набл} < \chi^2_{кр}$, то есть результатом обработки статистических данных стало наступление случайного события $\{\chi^2_{набл} \in Q_{доп}\}$, то по правилу принятия решений у нас нет оснований отклонять гипотезу H_0.

Казалось бы, что число зубчиков на пластинке листа должно зависеть от размеров пластинки листа. Но на основании проведённого эксперимента, состоящего:

а) в случайном отборе ста листьев из листьев нивяника круглолистного растущего в данном ареале, составляющих генеральную совокупность,

б) измерения длины пластинки листа и подсчёта числа зубчиков на этой пластинке листа,

в) статистической обработки результатов измерений и подсчётов,

мы можем утверждать, что длина пластинки листа и количество зубчиков на пластинке являются независимыми случайными величинами.

Таблица 3.6.3

$\xi \backslash \eta$	ϖ_1 [30;34]	ϖ_2 [35;38]	ϖ_3 [39;42]	ϖ_4 [43;46]	ϖ_5 [47;50]	ϖ_6 [51;54]	ϖ_7 [55;58]	ϖ_8 [59;62]	ϖ_9 [63;66]	m_{i*}
Δ_1 [20;26]	2	1								3
Δ_2 [27;32]	1	1	1							3
Δ_3 [33;38]	2	5	1	1	1					10
Δ_4 [39;44]	2	2	1	1	4	3	2			15
Δ_5 [45;50]	0	3	4	0	3	2	1	1		14
Δ_6 [51;56]	2	3	5	4	2	4	2	0	3	26
Δ_7 [57;62]	0	0	4	2	2	1	1	1	1	12
Δ_8 [63;68]	1	1	2	1	1	0	3	2	0	11
Δ_8 [69;74]					1	1	2	1	0	5
Δ_{10} [75;82]									1	1
m_{*j}	10	16	18	9	14	11	12	5	5	100

То есть, вероятности возможных значений одной из компонент не зависят от значений принимаемых другой компонентой. Количество зубчиков на пластинке листа не зависит от размеров пластинки.

Комментарий

Независимость размеров пластинки листа и количества зубчиков на ней приводит нас к следующему выводу. На пластинке листьев малых размеров может быть столько же зубчиков, сколько и на пластинке листьев больших размеров, но тогда зубчики должны быть мелкими. И, наоборот, на большой по размерам пластинке может быть столько же зубчиков сколько и на пластинке маленьких размеров, но эти зубчики будут крупными.

Анализ расположения в таблице 5.3.2. частот m_{ij}, значения которых не равны нулю, вызывает сомнение в категоричности сделанного вывода о независимости случайных величин ξ и η. Из таблицы видно, что на маленьких пластинках листа (20мм $\le \xi \le$ 30мм) число зубчиков не будет

большим ($30 \le \eta \le 42$). На больших пластинках (69мм $\le \xi \le$ 82мм) число зубчиков будет большим ($47 \le \eta \le 66$). И хотя листьев с такими «крайними» размерами и количествами зубчиков оказалось только 12 из 100, тенденция увеличения числа зубчиков с увеличением размера пластинки просматривается довольно чётко. В следующей главе мы вернёмся к этому примеру, дадим объяснение замеченной тенденции и сделаем вывод окончательный вывод о силе и виде зависимости случайных величин ξ и η.

Здесь мы лишь вспомним афоризм: «Статистике часто принадлежит первое слово, но никогда - последнее».

§7.3 Анализ совокупностей качественных признаков

Комментарий

Статистическую проверку гипотез о независимости случайных величин ξ и η можно проводить и при анализе качественных признаков.

Задачи такого типа чаще всего встречаются в медицине при выборе наиболее эффективного лекарства или курса лечения больных, в биологии при выборе наиболее эффективного уровня концентрации удобрения, или препарата для предпосевной обработки семян, в социологии при анализе распределения семей по годовому доходу и количеству детей.

Ясно, что математик-статистик, постоянно работающий с объектами, различаемыми по количественному признаку, предлагаемые примеры будет стараться интерпретировать в терминах теории вероятностей и математической статистики.

Пример. Исследуется влияние предпосевной обработки семян фунгицидом ТМТД на качество проростков семян. Семена делятся на две группы. Одна группа семян обрабатывается фунгицидом, другая – не обрабатывается. Семена первой группы высеиваются в грунт на одном участке, а семена второй группы высеиваются в грунт на другом участке. Когда семена прорастут, проводится анализ качества проростков. Определяются проростки здоровые, и проростки – больные.

Ознакомившись с целью опыта и правилами его постановки и проведения, мы стараемся построить соответствующую математическую модель подходящего статистического метода. Применение избранного метода будет тем успешнее, чем ближе наша модель к действительности.

Рассматривается генеральная совокупность, элементы которой обладают качественными признаками, а каждый из этих признаков имеет не менее двух уровней различия. Если каждому качественному признаку присвоить «ранг». То мы можем сказать, что рассматривается случайная величина ξ, назначенные «ранги» - её возможные значения. Уровни

различия контролируемого качественного признаки мы можем называть возможными значениями случайной величины η. После этого проблема: влияет, или нет, предпосевная обработка семян фунгицидом на качество проростков в терминах математической статистики будет формулироваться так: зависимы, или нет, случайные величины ξ и η.

Формулируется основная гипотеза H_0: *Качественные признаки* ξ *и* η - *независимые.* Или, как более привычно формулировать математику-статистику, H_0: *Случайные величины* ξ *и* η *– независимые.*

Составляется таблица, в которой вместо возможных значений случайных величин ξ *и* η записываются контролируемые названия качественных признаков. Результаты проведённого эксперимента – наблюдаемые частоты осуществления различных комбинаций качественных признаков. Так как мы видим, что это обычая задача проверки гипотезы о независимости случайных величин, то критерием проверки справедливости основной гипотезы будет критерий $\chi^2_{(r-1)(s-1)} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{m_{ij}^2}{m_{i*}\cdot m_{*j}} - 1\right)$, вид которого значительно упрощается при небольших значениях r и s.

В частности, если продолжить построение статистической модели проверки влияния предпосевной обработки семян фунгицидом на качество проростков, то получим таблицу, называемую таблицей сопряжённости исследуемых признаков.

Таблица 4.6.3

η / ξ	больные проростки	здоровые проростки	
необработанные семена	m_{11}	m_{12}	$m_{1*} = m_{11} + m_{12}$
обработанные семена	m_{21}	m_{22}	$m_{2*} = m_{21} + m_{22}$
	$m_{*1} = m_{11} + m_{21}$	$m_{*2} = m_{12} + m_{22}$	n

Таблица 4.6.3 - это упрощённая таблица 1.6.3, в которой случайные величины ξ и η принимают только по два значения. То есть, здесь r=2 и s=2.

Мы можем считать, что ξ =1, если наудачу выбранное семя обработано фунгицидом, и ξ=0, если семя – не обработано.

Аналогично: η=1, если наудачу выбранный проросток – здоров, и η =0, если проросток – больной.

Общее количество семян, с которыми проводился эксперимент равно n - это объём выборки,

m_{11} - количество необработанных фунгицидом семян, давших больные проростки, то есть m_{11} - количество наблюдений события $\{\xi = 0; \eta = 0\}$;

m_{12} - количество необработанных фунгицидом семян, давших здоровые проростки, то есть m_{12} - количество наблюдений события $\{\xi = 0; \eta = 1\}$;

m_{21} - количество обработанных фунгицидом семян, давших больные проростки; то есть m_{21} - количество наблюдений события $\{\xi = 1; \eta = 0\}$;

m_{22} - количество обработанных фунгицидом семян, давших здоровые проростки; то есть m_{22} - количество наблюдений события $\{\xi = 1; \eta = 1\}$;

Ясно, что m_{1*} и m_{2*}, соответственно, количества значений 0 и 1 случайной величины ξ - количеств необработанных и обработанных семян, а m_{*1} и m_{*2} количества значений 0 и 1 случайной величины η – количеств больных и здоровых проростков.

Так как в рассматриваемом примере r=2 и s=2, то вид критерия проверки справедливости упрощается:

$$\chi_1^2 = \frac{(m_{11} \cdot m_{22} - m_{12} \cdot m_{21})^2}{m_{1*} \cdot m_{2*} \cdot m_{*1} \cdot m_{*2}} \cdot n.$$

При выбранном уровне значимости α по таблицам распределения "χ^2-распределения" из условия $P(\chi_1^2 \geq \chi_{кр}^2) = \alpha$ определяем значение $\chi_{кр}^2$ и сравниваем его с наблюдаемым значением критерия. Принимаем решение об отклонении или принятии основной гипотезы H_0.

Запись формулы критерия проверки справедливости гипотезы H_0, в том случае когда число «рангов» или «уровней различия» качественных признаков ξ или η будет более чем два, не вызовет особых затруднений. Надо только перед применением таблиц при определении числа степеней свободы критерия $\chi^2_{(r-1)(s-1)}$ вычислить значение произведения $(r-1)(s-1)$.

Практика

Пример. Сто учащихся средней школы сдавали экзамены по биологии и математике. Надо проверить, есть ли статистическая зависимость между альтернативными признаками: результатами экзаменов по биологии и по математике, которые приведены в следующей таблице:

		Экзамен по биологии	
		сдали	провалили
Экзамен по математике	сдали	20	18
	провалили	23	39

Формулируем, исходя из принципа «презумпции невиновности», основную гипотезу $\boldsymbol{H_0}$: *Статистической зависимости между результатами экзаменов по биологии и по математике нет.*

Ясно, что здесь: $m_{1*} = m_{11} + m_{12}$=20+18=38; $m_{2*} = m_{21} + m_{22}$=23+39=62; $m_{*1} = m_{11} + m_{21}$ =20+23=43 и $m_{*2} = m_{12} + m_{22}$ =18+39=57. Подсчитываем наблюдаемое значение критерия $\chi^2_{набл.} = \dfrac{(20 \cdot 39 - 18 \cdot 23)^2 \cdot 100}{38 \cdot 62 \cdot 43 \cdot 57} = 2{,}320.$

При уровне значимости $\alpha = 0{,}05$ по таблицам "χ^2-распределения" для $(r-1)(s-1)$ =1 определяем, что $\chi^2_{кр.} = 3{,}841$. Так как мы наблюдаем выполнение неравенства $\chi^2_{набл.} < \chi^2_{кр.}$, то, по правилу принятия решений, у нас нет оснований отклонять гипотезу $\boldsymbol{H_0}$, то есть мы будем считать, что статистической зависимости между результатами экзаменов по биологии и по математике нет.

«Опыт – вот учитель жизни»
И.В. Гёте
«… опыт – сын ошибок трудных»
А.С. Пушкин

Модуль четвёртый

Элементы корреляционного и регрессионного анализов

§ 1.4 Статистическая зависимость случайных величин

Комментарий

При изучении объектов генеральной совокупности, обладающих двумя, или более, качественными признаками, интересующими исследователя, мы говорим, что изучается двумерная, или многомерная, случайная величина. Каждая компонента является случайной величиной, значения которой - количественные значения качественного признака. Компоненты многомерной случайной величины могут быть независимыми, могут быть зависимыми. Независимость компонент мы понимаем в том смысле, что каждая компонента многомерной случайной величины принимает свои возможные значения с вероятностями, которые не зависят от значений, принимаемых другой компонентой. Закон распределения вероятностей многомерной случайной величины с независимыми компонентами равен произведению частных законов распределения вероятностей её компонент.

Теория

Пусть элементы генеральной совокупности обладают двумя количественными признаками, которые являются случайными величинами ξ и η. В этом случае мы говорим, что изучается двумерная случайная величина $\zeta = (\xi;\eta)$ с компонентами ξ и η. Если компоненты ξ и η – независимые случайные величины, то закон распределения вероятностей двумерной случайной величины ζ равен произведению частных законов распределения вероятностей его компонент. В частности, если двумерная случайная величина ζ – непрерывного типа, то её плотность вероятности равна произведению частных плотностей вероятности компонент: $p(x,y) = p_1(x) \cdot p_2(y)$. Если двумерная случайная величина ζ – дискретного типа, то для любого её возможного значения $(x_i; y_k)$ будет справедливо: $P(\zeta \in (x_i; y_k)) = p_{ik} = p_{i*} \cdot p_{*k}$, где $p_{i*} = P(\xi = x_i)$ и $p_{*k} = P(\eta = y_k)$.

Зависимость ξ и η случайных величин – компонент двумерной случайной величины, являющихся двумя различными количественными признаками объектов исследования, проявляется в том, что значения одной случайной величины – одной из компонент и распределение вероятностей этих значений зависят от значений, принимаемых другой случайной величиной – другой компонентой. Различают два вида зависимостей случайных величин.

Зависимость случайных величин называется ***функциональной***, если возможные значения одной из них являются функциями значений другой. Распределение вероятностей возможных значений случайной величины η можно определить, если известен закон распределения вероятностей случайной величины ξ (плотность вероятности $p(x)$ или набор вероятностей $\{p_k\}, k = 1,2,..n$) и функцию $y = f(x)$, определяющую связь между значениями случайных величин ξ и η. В таких случаях мы будем говорить, что случайные величины ξ и η связаны функциональной зависимостью и записывать: $\eta = f(\xi)$.

Комментарий

Пример 1.1.4 Ясно, что при любом, даже достаточно высоком уровне технологии, значения диаметров шариков подшипников можно рассматривать как значения некоторой случайной величины. Пусть диаметр шарика – случайная величина ξ. Каждый шарик имеет вес и ясно, что значения весов шариков этих подшипников тоже будут случайными. Вес шарика равен произведению удельной плотности стали ρ, из которой он изготовлен, и объёма шарика $v = \frac{4}{3}\pi \cdot R^3$. Если вес шарика – случайная

величина η, то каждое возможное значение η однозначно определяется значением, которое приняла случайная величина ξ, то есть мы можем записать: $\eta = \rho \cdot \frac{4}{3}\pi \cdot \left(\frac{\xi}{2}\right)^3$.

Функция, определяющая связь между случайными величинами ξ и η, имеет вид: $y = f(x) = \frac{\pi}{6}\rho \cdot x^3$. Ясно, что распределение вероятностей возможных значений η определяется распределением вероятностей возможных значений ξ.

Как правило, в реальных наблюдениях или исследованиях мы не встречаем функциональной зависимости между количественными значениями изучаемых качественных признаков. Но, анализируя числовые значения признаков, то есть - значения случайных величин ξ и η, мы можем заметить, что увеличение или уменьшение значений одной из случайных величин вызывает вполне определённую тенденцию изменения значений другой случайной величины.

Пример 2.1.4 Если случайная величина ξ – рост студента первого курса ЮФУ, а случайная величина η – вес этого студента, то мы можем сказать, что увеличение значений ξ влечёт в основном увеличение возможных значений η, то есть, чем больше рост студента, тем больше будет его вес. Однако ясно, что среди студентов одинакового роста встречаются и худощавые, и предрасположенные к полноте, то есть – студенты, имеющие разный вес.

Другими словами, каждому возможному числовому значению случайной величины ξ будут соответствовать значения η из некоторого числового интервала. Увеличение возможных значений ξ вызывает смещение этого интервала. Мы не можем найти функцию $y = f(x)$, с помощью которой мы смогли бы для каждого конкретного значения ξ вычислить точное значение η. Но мы можем сказать, что с увеличением возможных значений ξ, интервалы возможных значений η изменяются и, если их отмечать на числовой оси, будет замета тенденция смещения этих интервалов по оси ординат вправо.

Пример 3.1.4 Цена одного барреля нефти на сырьевой бирже зависит от многих причин и постоянно изменяется, то увеличиваясь, то уменьшаясь. Пусть цена одного барреля нефти - случайная величина ξ. Курс доллара по отношению к рублю тоже постоянно меняется и будем считать, что его значения – значения случайной величины η. Анализируя одновременные изменения значений случайных величин ξ и η в течение некоторого периода, мы можем заметить, что повышение цены одного

барреля нефти сопровождается, как правило, снижением курса доллара, и наоборот. Здесь мы также не сможем найти функцию $y = f(x)$, с помощью которой мы могли бы для каждого конкретного значения ξ вычислить точное значение η, но отрицать наличие статистической зависимости между случайными величинами ξ и η мы не сможем.

Теория

Зависимости между случайными величинами, в которых изменение значений одной из случайных величин вызывает вполне определённую тенденцию изменений значений другой случайной величины, называются ***статистическими зависимостями***. В некоторых источниках эту взаимозависимость между случайными величинами называют ***корреляционной.***

Первое, ориентировочное представление о характере статистической (корреляционной) зависимости между случайными величинами мы можем получить, отобрав из генеральной совокупности выборку и построив по ней геометрическую иллюстрацию, которую называют «скэттер-диаграммой».

Пусть исследуется двумерная случайная величина с компонентами ξ и η. Выбрав из генеральной совокупности n объектов, и измерив у каждого значения исследуемых качественных признаков ξ и η, мы получим двумерную выборку $\{(x_i; y_i)\}$, $i = 1,2,...,n$, объём которой равен n.

При небольших объёмах выборки $(n \approx 20 \div 50)$ можно сделать геометрическую иллюстрацию этой двумерной выборки. В выбранной системе координат отмечаются n точек с координатами $(x_i; y_i)$. Мы получим «скэттер-диаграмму», иллюстрирующую статистическую зависимость между случайными величинами. По виду диаграммы можно сделать предварительный вывод о силе и характере связи между случайными величинами ξ и η.

При больших объёмах выборки возникает затруднение в графическом изображении большого числа точек. В таких случаях предварительное представление о силе и характере связи между случайными величинами ξ и η можно получить, построив корреляционную таблицу частот.

Практика

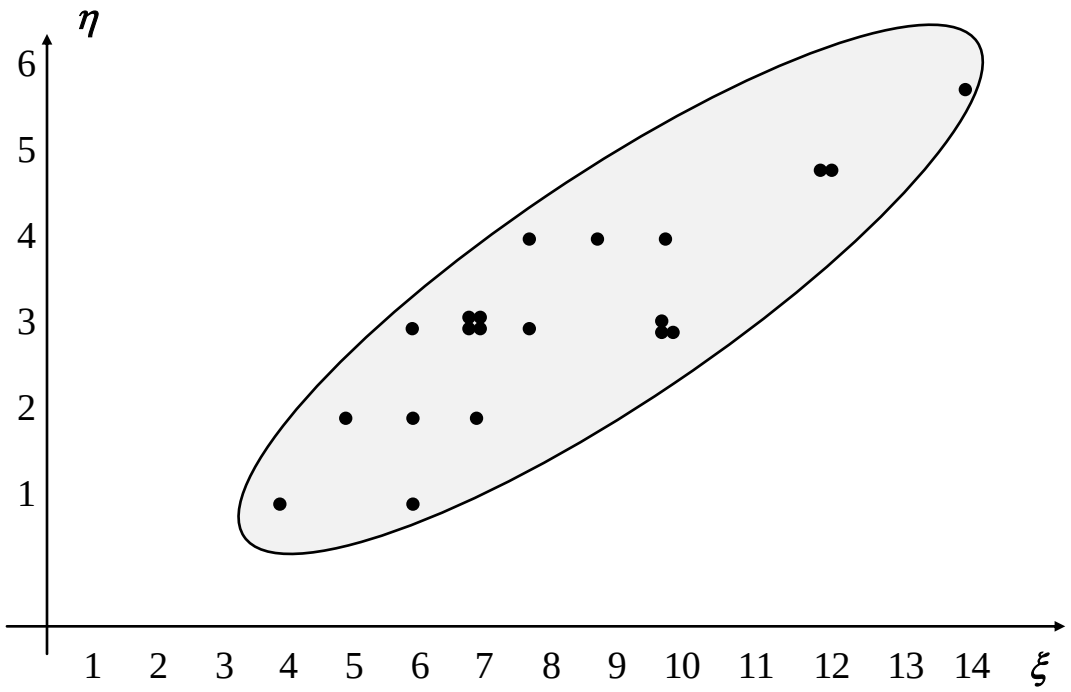
Пример 4.1.4. Из «Словаря русского языка» С.И.Ожегова наудачу выбрано двадцать слов ($n=20$). В каждом слове подсчитано общее количество букв и количество гласных букв. Обозначим: ξ – количество букв в наудачу выбранном слове; η – количество гласных букв в этом слове. Полученную двумерную выборку $\{(x_i; y_i)\}$, $i = 1,2,...,20$ запишем в виде ранжированного по первой координате вариационного ряда.

Таблица 4.1.4

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i	4	5	6	6	6	7	7	7	7	7	8	8	9	10	10	10	10	12	12	14
y_i	1	2	1	2	3	2	3	3	3	3	3	4	4	3	3	3	4	5	5	6

Уже из записи элементов выборки в виде двумерного вариационного ряда, видно, что увеличение значений случайной величины ξ, вызывает заметную тенденцию к увеличению значений случайной величины η.

«Скэттер-диаграмма», являясь геометрической иллюстрацией выборки, подтверждает этот вывод.



Комментарий

И без организации выборки и последующей статистической обработки полученных данных ясно, что, чем больше будет букв в наудачу выбранном слове, тем больше можно будет насчитать в нём гласных букв. Проблема статистической зависимости длины слова и количества гласных букв в нём может быть интересной только для специалиста-филолога. Мы же рассматриваем здесь этот эксперимент только в виду его простоты и наглядности. Нас интересует только количественная картина, то есть возможность получения по числовым данным, которыми являются элементы двумерной выборки, **количественных оценок** характера связи между случайными величинами и формулирование **качественных выводов** из этих оценок.

Практика

Пример 5.1.4. Продолжим эксперимент. Из этого же словаря С.И.Ожегова сделаем выборку, содержащую ***n=1760*** слов.

Изучаются те же случайные величины: ξ – количество букв в наудачу выбранном слове; η – количество гласных букв в этом слове. Определим значения частот m_{ij} - количества слов, состоящих из i букв, среди которых j – гласных.

Ясно, что построение геометрической иллюстрации выборки такого объёма, во-первых – затруднительно, а во-вторых, бессмысленно, так как по ней нельзя будет увидеть различие величин отдельных частот m_{ij} различных пар (x_i, y_j). Для предварительного представления о силе и характере статистической зависимости составляем корреляционную таблицу.

Таблица 5.1.4.

9																			1
8															4	1	1	1	1
7												2	5	6	9	2	3		
6									1	6	15	17	19	17	9	7	1		
5								7	18	40	49	41	27	14	7	3			
4						4	38	59	71	125	78	27	9	1					
3				2	12	68	118	168	134	50	9	3							
2			2	43	92	121	75	27	6	2									
1	1	7	21	37	15														
η/ξ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Величины частот m_{ij} и характер изменения их в таблице позволяют сделать первые выводы.

Чем больше длина слова, тем, в основном, больше в нём гласных букв. Если слово, например, состоит из восьми букв $(x_i = 8)$, то в подавляющем большинстве случаев $(m_{ij} = 168)$ в этом слове будет три гласных буквы $(y_j = 3)$. Почти в три раза реже $(m_{ij} = 59)$ встречаются слова, состоящие из восьми букв, среди которых гласных букв - четыре $(x_i = 8,\ \ y_j = 4)$. Очень редко $(m_{ij} = 7)$ встречаются слова, состоящие из восьми букв, среди которых пять гласных $(x_i = 8,\ \ y_j = 5)$.

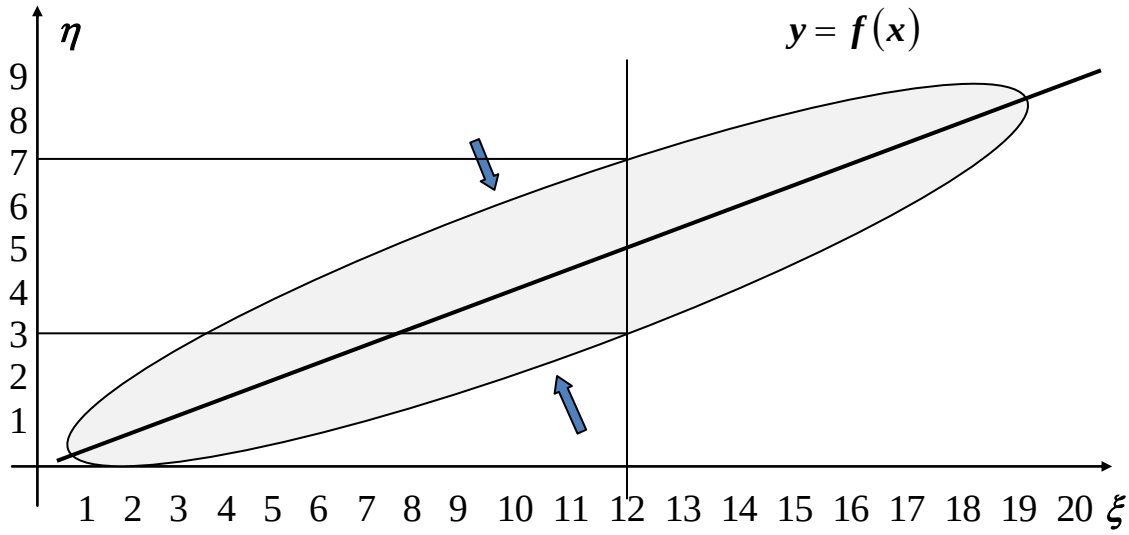
Аналогично можно проследить изменение длины слов с одинаковым количеством гласных букв. Например, хотя в русском алфавите гласные буквы составляют треть от общего числа букв, среди слов, состоящих из

двенадцати букв $(x_i = 12)$, чаще $(m_{ij} = 41)$ встречаются слова, содержащие пять $(y_j = 5)$, а не четыре $(y_j = 4)$, гласных букв $(m_{ij} = 27)$.

Расположение, похожее на «облако», и величины ненулевых частот m_{ij} в «прямоугольнике» корреляционной таблицы 1.4.*а* довольно наглядно подтверждает наличие статистической (корреляционной) связи между случайными величинами ξ и η. Контур «облака» (пунктирная линия) похож на эллипс. В таких случаях будем говорить, что статистическая связь между случайными величинами ξ и η имеет *линейный* характер.

Комментарий

Если мы сделаем геометрическую иллюстрацию корреляционной таблицы 1.4.а. рассматриваемого примера, то получили бы диаграмму, на которой значения двумерной случайной величины $\zeta = (\xi, \eta)$, попавшие в выборку, образуют на плоскости область эллиптической формы.



У исследователя, анализирующего вид «скэттер-диаграммы», возникают два естественных желания.

Первое желание – это желание количественно оценить силу статистической связи случайных величин ξ и η. Сила статистической связи понимается нами как сила влияния значений одной из случайных величин на величину диапазона возможных значений другой случайной величины. Ясно, что чем меньше будет величина «вертикального» диаметра получившегося «эллипса», тем меньше, при каждом конкретном значении случайной величины ξ, будет диапазон возможных значений случайной величины η. То есть, статистическая связь между случайными величинами ξ и η будет сильнее.

Из корреляционной таблицы 5.1.4 мы видим, что при значении $x_i = 12$ случайной величины ξ случайная величина η принимает значения y_j в

диапазоне от 3 до 7. В то время как областью всех возможных значений η является диапазон от 1 до 9. Очевидно, что диапазон от 3 до 7 является областью возможных значений η при данном значении $x_i = 12$ случайной величины ξ.

Второе желание – это желание провести линию, которая отображала бы тенденцию изменения значений случайной величины η при изменении значений случайной величины ξ, и найти уравнение этой линии $y = f(x)$.

Теория

В соответствии со сказанным сформулируем две задачи корреляционного анализа.

Первая задача. Оценить количественно силу статистической связи между случайными величинами ξ и η, и дать ей качественную оценку.

Вторая задача. Найти уравнение линии, то есть – функцию $y = f(x)$, которая будет описывать тенденцию изменения значений одной из случайных величин в зависимости от изменения другой. Вероятностный смысл этой функции позволит прогнозировать наиболее возможные значения случайной величины η при изменении значений случайной величины ξ.

§ 2.4 Коэффициент линейной корреляции

Теория

Силу статистической связи, но только если она имеет линейный характер, оценивает коэффициент линейной корреляции, являющийся математическим ожиданием произведения центрированных и нормированных случайных величин $\frac{\xi - M\xi}{\sigma_\xi}$ и $\frac{\eta - M\eta}{\sigma_\eta}$:

$$\rho = M\left(\frac{\xi - M\xi}{\sigma_\xi} \cdot \frac{\eta - M\eta}{\sigma_\eta}\right).$$

Если воспользоваться свойствами математического ожидания, то формулу коэффициента линейной корреляции, можно упростить: $\rho = \frac{\alpha_{11} - m_\xi \cdot m_\eta}{\sigma_\xi \cdot \sigma_\eta}$, где $\alpha_{11} = M[\xi \cdot \eta]$, $m_\xi = M\xi$, $m_\eta = M\eta$.

Коэффициент линейной корреляции ρ обладает следующими свойствами:

1) Для любых случайных величин ξ и η значения коэффициента линейной корреляции ρ удовлетворяют неравенству $-1 \le \rho \le +1$, или $|\rho| \le 1$;

2) Если случайные величины ξ и η - независимые, то коэффициент линейной корреляции равен нулю: $\rho = 0$.

3) Модуль значения коэффициента линейной корреляции равен единице тогда, и только тогда, когда случайные величины ξ и η связаны линейной функциональной зависимостью. То есть: $|\rho| = 1 \Leftrightarrow \eta = a\xi + b$.

Комментарий

Коэффициент линейной корреляции является числовой характеристикой теоретической двумерной случайной величины $\zeta = (\xi, \eta)$, которую мы изучаем по выборочным данным – по двумерной выборочной совокупности случайных величин $\{(\xi_i ; \eta_i)\}$, $i = 1,2,...,n$.

С одной стороны, элементы числовой выборки $\{(x_i ; y_i)\}$ – это некоторый набор n пар чисел, являющихся возможными значениями изучаемой двумерной случайной величины $\zeta = (\xi, \eta)$. С другой стороны, элементы числовой выборки $\{(x_i ; y_i)\}$ мы можем рассматривать как значения некоторой эмпирической двумерной случайной величины $\zeta^* = (\xi^*, \eta^*)$, которая является случайной величиной дискретного типа. Применяя метод моментов определения точечных оценок числовых характеристик, получаем, что у эмпирической случайной величины ζ^* коэффициент линейной корреляции будет иметь вид: $r = \dfrac{a_{11} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y}$, где

$$a_{11} = M[\xi^* \cdot \eta^*] = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i\,,\ \bar{x} = M\xi^* = \frac{1}{n}\sum_{i=1}^{n} x_i\,,\ \bar{y} = M\eta^* = \frac{1}{n}\sum_{i=1}^{n} y_i\,.$$

Теория

Итак, теоретическую числовую характеристику – коэффициент линейной корреляции $\vartheta = \rho$ двумерной случайной величины $\zeta = (\xi, \eta)$ мы будем оценивать соответствующей числовой характеристикой $\theta^* = r = \dfrac{a_{11} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y}$ эмпирической случайной величины $\zeta^* = (\xi^*, \eta^*)$.

Из свойств коэффициента линейной корреляции следуют возможность его применения при изучении статистической связи между случайными величинами ξ и η, а так же правило формулирования выводов о величине силы этой связи.

Как и теоретический коэффициент корреляции ρ, выборочный коэффициент корреляции r принимает значения между -1 и +1. Граничные значения ($|r| = 1$), он принимает только при наличии линейной связи между наблюдениями x и y, составляющими выборку $\{(x_i ; y_i)\}$, $i = 1,2,...,n$, то

есть, когда все значения эмпирической случайной величины $\zeta^* = (\xi^*, \eta^*)$ удовлетворяют условию: $y = ax + b$. Чем слабее, имеющая линейный характер, статистическая связь, тем будет меньше абсолютное значение выборочного коэффициента корреляции r .

Если теоретический коэффициент линейной корреляции ρ равен нулю, то мы будем говорить, что линейная связь между случайными величинами ξ и η – отсутствует. В этом случае статистическая связь может иметь или нелинейный характер, или совсем отсутствовать, то есть случайные величины ξ и η будут независимыми.

Вычисленное значение выборочного коэффициента линейной корреляции r может получиться близким к нулю. Малое отклонение от нуля значения выборочного коэффициента корреляции может объясняться или слабой статистической связью между случайными величинами ξ и η, или недостаточной репрезентативностью выборки и неизбежными ошибками измерений при организации выборки. Чтобы сделать вывод о наличии, или отсутствии линейной статистической зависимости между случайными величинами ξ и η проверяется гипотеза о значимости коэффициента линейной корреляции. Основная гипотеза записывается так:

$H_0: \rho = 0$, то есть: *Коэффициент линейной корреляции* ρ *равен нулю.*

Критерием проверки справедливости этой простой гипотезы будет статистика $t_{n-2} = r \cdot \sqrt{\dfrac{n-2}{1-r^2}}$, которая подчиняется распределению Стьюдента с параметром n-2. Выбрав уровень значимости основной гипотезы α, по таблицам распределения Стьюдента определяем значение $t_{кр}$ такое, что $P(|t_{n-2}| > t_{кр}) = \alpha$. Критической областью значений критерия будет объединение интервалов $S_{кр} = (-\infty; -t_{кр}) \cup (+t_{кр}; +\infty)$, а областью возможных значений критерия будет интервал $Q_{доп} = (-t_{кр}; +t_{кр})$. Если мы будем наблюдать событие $\{t_{набл} \in S_{кр}\}$, то гипотезу H_0 отклоняем. Это означает, что статистическая связь между исследуемыми случайными величинами существует. Если же мы будем наблюдать событие $\{t_{набл} \in Q_{доп}\}$, то гипотеза H_0 принимается. Это означает, что ненулевое значение выборочного коэффициента корреляции r вызвано недостаточной представительностью выборки и ошибками измерений.

Практика

Анализируя результаты проверки справедливости гипотез о независимости случайных величин в примере 5.3.2, мы отметили ненадёжность сделанных выводов о зависимости, или независимости их во всех трёх случаях.

Возникает предположение о том, что между исследуемыми парами случайных величин имеется статистическая (корреляционная) зависимость, но эта зависимость – слабая.

Продолжим статистическую обработку данных эксперимента, вычислив значения коэффициентов линейной корреляции трёх пар случайных величин.

Рассмотрим первую пару, где ξ – длина пластики листа нивяника круглолистного и η – количество зубчиков на всей пластинке.

Вычислим точечные оценки числовых характеристик этих случайных величин: $n=100;\quad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=50{,}62;\quad s_x=\sqrt{\frac{1}{n-1}\left[\sum x_i^2-n\cdot\bar{x}^2\right]}=11{,}770;$

$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=45{,}98;\quad s_y=\sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-n\cdot\bar{y}^2\right]}=9{,}249;\quad a_{11}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i=2377{,}61.$

Оценка коэффициента линейной корреляции случайных величин ξ и η будет равна: $r_{xy}=\frac{a_{11}-\bar{x}\cdot\bar{y}}{s_x\cdot s_y}=0{,}460$.

Комментарий

При рассмотрении этого примера в предыдущей главе мы приняли гипотезу о независимости случайных величин ξ и η. Это означает, что коэффициент линейной корреляции ρ, оценивающий силу статистической связи между величинами ξ и η, должен быть равным нулю. Расположение ненулевых частот m_{ij} в таблице 5.3.2 и их величины повлияли на нашу уверенность в объявлении независимости случайных величин ξ и η. Вычисленное значение $r_{xy}=0{,}460$ увеличивает наше сомнение в их независимости.

Практика

Гипотеза о значимости коэффициента корреляции величин ξ и η кратко записывается так: $H_0: \rho_{\xi\eta}=0$.

Если принять уровень значимости гипотезы H_0 равным $\alpha=0{,}05$, то по таблицам распределения Стьюдента определяем, что при n-2=98 условие $P\left(\left|t_{n-2}\right|\geq t_{кр}\right)=\alpha$ выполняется, если $t_{кр}=1{,}98$. Критическая область значений критерия будет иметь вид: $S_{кр}=(-\infty;-1{,}98)\cup(+1{,}98;+\infty)$.

Наблюдаемое значение критерия $t_{набл}=r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}}$, где $r_{xy}=0{,}460$, будет равно $t_{набл}=5{,}13$. Сравнивая наблюдаемое и критическое значения критерия, получаем неравенство $t_{набл}>t_{кр}$. Это означает, что мы наблюдаем

событие $\{t_{набл} \in S_{кр}\}$. По правилу принятия решений гипотеза $H_0: \rho_{\xi\eta} = 0$ - отклоняется. Значит, мы не можем утверждать, что ненулевое значение $r_{xy} = 0{,}460$ статистической оценки коэффициента корреляции объясняется случайными причинами или ошибками измерений. Следовательно, между случайными величинами ξ и η существует значимая статистическая связь, то есть такая связь, что пренебрегать её наличием нельзя.

Комментарий

При рассмотрении этого примера в предыдущей главе мы приняли гипотезу о независимости случайных величин ξ и η. Это означает, что коэффициент линейной корреляции $\rho_{\xi\eta}$, оценивающий силу статистической связи между величинами ξ и η, должен, исходя из его свойств, быть равным нулю.

Расположение ненулевых частот m_{ij} в таблице 3.6.3 и их величины повлияли на нашу уверенность в объявлении независимости величин ξ и η. Вычисленное значение $r_{xy} = 0{,}460$ увеличивает наше сомнение в их независимости. Отклонение гипотезы о равенстве нулю коэффициента линейной корреляции $\rho_{\xi\eta}$ привело нас к заключению о том, что статистическая связь между случайными величинами ξ и η существует. То есть у растений, имеющих более крупные пластинки листа, зубчики будут крупнее, и в то же время зубчиков на этих пластинках будет больше, чем на менее крупных пластинках листа.

Практика

Рассмотрим теперь вторую пару случайных величин: ξ – длина пластики листа нивяника круглолистного и ζ – количество зубчиков на *1см* края пластинки. Основная гипотеза записывается аналогично H_0 : $\rho_{\xi\zeta} = 0$.

К записанным выше значениям точечных оценок числовых характеристик случайной величины ξ добавим значения точечных оценок числовых характеристик случайной величины ζ и вычислим статистическую оценку коэффициента линейной корреляции случайных величин ξ и ζ.

$$n = 100; \quad \bar{x} = 50{,}62 \quad s_x = 11{,}770;$$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = 12{,}04; \quad s_z = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} z_i^2 - n \cdot \bar{z}^2\right]} = 2{,}680 \quad \text{и}$$

$$a_{11} = \frac{1}{n}\sum_{i=1}^{n} x_i z_i = 597{,}01.$$

Оценка коэффициента линейной корреляции случайных величин ξ и ζ будет равна: $r_{xz} = \dfrac{a_{11} - \bar{x} \cdot \bar{z}}{s_x \cdot s_z} = -0{,}395$.

Проверим гипотезу о значимости коэффициента линейной корреляции $H_0: \rho_{\xi\zeta} = 0$, применяя тот же критерий t_{n-2}. Если $\alpha = 0{,}05$ и n-2=98, то, как было определено в предыдущем случае, критическая область имеет вид: $S_{кр} = (-\infty; -1{,}98) \cup (+1{,}98; +\infty)$.

Наблюдаемое значение критерия $t_{набл} = r_{xy}\sqrt{\dfrac{n-2}{1-r_{xy}^2}}$, где $r_{xy} = -0{,}395$, будет равно $t_{набл} = -4{,}256$. Мы наблюдаем событие $\{t_{набл} \in S_{кр}\}$. По правилу принятия решений гипотеза $H_0: \rho_{\xi\zeta} = 0$ отклоняется. Значит, между случайными величинами ξ и ζ существует статистическая зависимость. Отрицательное значение оценки коэффициента корреляции свидетельствует о том, что увеличение значений ξ вызывает уменьшение возможных значений ζ. У растений с большими пластинками листьев число зубчиков на 1 *см* края пластинки будет меньше, чем у растений с меньшими пластинками, что вполне естественно, так как у больших листьев будут более крупные зубчики.

Заканчивая анализ статистических данных о качественных свойствах листьев нивяника круглолистного, рассмотрим теперь третью пару случайных величин: η – количество зубчиков на всей пластинке листа и ζ – количество зубчиков на *1см* края пластинки. Основная гипотеза записывается так же, как и при рассмотрении первых двух пар случайных величин, то есть $H_0: \rho_{\eta\zeta} = 0$. Уровень значимости гипотезы оставим тем же $\alpha = 0{,}05$. Следовательно, критическое значение $t_{кр} = 1{,}98$ и критическая область $S_{кр}$ будут такими же, как и в первых двух случаях.

Точечная оценка коэффициента линейной корреляции случайных величин η и ζ равна: $r_{yz} = \dfrac{a_{11} - \bar{y} \cdot \bar{z}}{s_y \cdot s_z} = 0{,}337$.

Вычисляем наблюдаемое значение критерия: $t_{набл} = r_{yz}\sqrt{\dfrac{n-2}{1-r_{yz}^2}} = 3{,}54$.

Сравнивая наблюдаемое и критическое значения критерия, получаем неравенство $t_{набл} > t_{кр}$, то есть мы наблюдаем событие $\{t_{набл} \in S_{кр}\}$. По правилу принятия решений гипотеза $H_0: \rho_{\eta\zeta} = 0$ отклоняется. Значит, мы не можем утверждать, что ненулевое значение статистической оценки коэффициента корреляции $r_{yz} = 0{,}337$ объясняется случайными причинами

или ошибками измерений. Следовательно, между случайными величинами η и ζ существует значимая статистическая связь, то есть такая связь, что пренебрегать её наличием нельзя.

Комментарий

Оценивая значимость коэффициентов линейной корреляции трёх возможных пар случайных величин ξ, η и ζ, мы в первом и в третьем случаях пришли к выводам о наличии статистической зависимости линейного вида между парами ξ, η и ξ, ζ. При этом, вычислив значения оценок коэффициентов линейной корреляции, мы оценили силу статистической зависимости между этими случайными величинами.

Эти выводы противоположны сделанным в предыдущей главе выводам о независимости случайными величин составляющих эти пары.

Возникает вопрос: «В каких границах должно находиться значение оценки коэффициента линейной корреляции **r**, которое, при проверке гипотезы $\mathbf{H_0}: \rho = 0$, мы могли бы объяснить наличием ошибок измерений и недостаточной репрезентативностью выборки»?

Определим границы для значений статистической оценки коэффициента линейной корреляции r, нахождение в которых позволяло бы нам считать, что коэффициент линейной корреляции ρ или равен нулю, или пренебрежимо мало отличается от нуля.

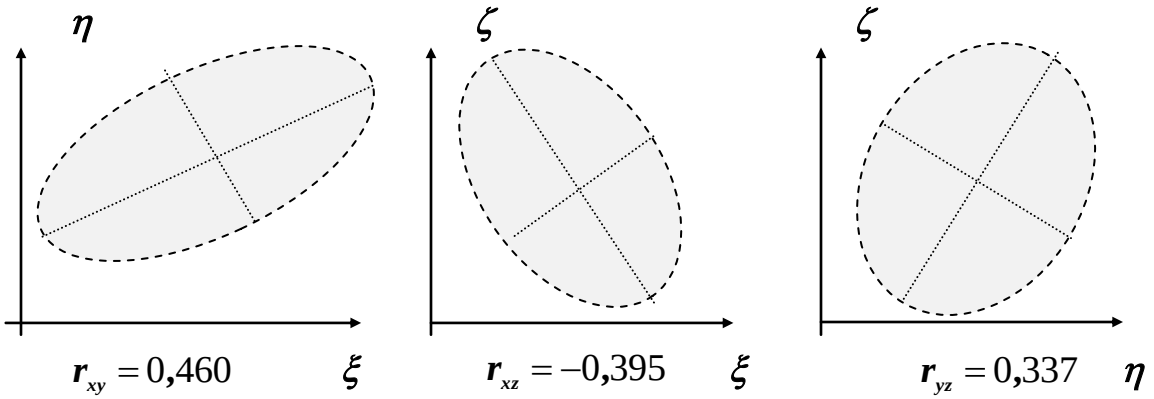
Воспользуемся двойственностью задач интервального оценивания числовых характеристик и статистической проверки простых гипотез о значениях числовых характеристик.

Для проверки справедливости гипотезы H_0 мы применяли критерий: $t_{n-2} = r \cdot \sqrt{\dfrac{n-2}{1-r^2}}$. Решая это уравнение относительно точечной оценки r, получим: $|r| = \dfrac{t_{n-2}}{\sqrt{n-2+t_{n-2}^2}}$.

По таблицам распределения Стьюдента мы определили, что, для уровня значимости основной гипотезы $\alpha = 0{,}05$ и объёма выборки n=100 условие $P\left(|t_{n-2}| \geq t_{кр}\right) = \alpha$ выполняется, если $t_{кр} = 1{,}98$. Подставляя значение $t_{кр} = 1{,}98$ в выражение для $|r|$, получаем: $|r| \approx 0{,}2$. Значит, если значения оценки коэффициента линейной корреляции r будут принадлежать интервалу $(-0{,}202; +0{,}202)$, то с уверенностью равной 1-α=0,95 мы можем утверждать, что теоретическое значение коэффициента корреляции ρ или равно нулю, или близко к нулю. Следовательно, между случайными величинами или нет линейной статистической зависимости, или она настолько слабая, что ею можно пренебречь.

Выше, говоря о задачах математической статистики, решаемых корреляционным анализом, мы говорили, что первой задачей является задача количественной оценки силы статистической связи между исследуемыми случайными величинами.

Мы можем по выборке таблице 2.6.3 построить ещё две корреляционные таблицы, аналогичные таблице 3.6.3, по которым можно увидеть наличие статистической связи между парами случайных величин (ξ,ζ) и (η,ζ). Значения статистических оценок коэффициентов свидетельствуют о наличии статистической связи между случайными величинами. Геометрическая иллюстрация этих корреляционных таблиц позволяет нам чётче увидеть наличие этой связи и её направление.



$r_{xy} = 0{,}460$ $r_{xz} = -0{,}395$ $r_{yz} = 0{,}337$

Во всех трёх случаях мы видим, что отношения длин вертикальных (малых) диаметров «эллипсов» к длинам их горизонтальных (больших) диаметров близки к единице. То есть «эллипсы», характеризующие расположение на плоскости ненулевых значений частот m_{ij}, мало отличаются от «окружностей». Это позволяет нам сделать вывод, что статистическая связь между парами случайных величин – слабая.

В предыдущей главе, проверяя гипотезы о независимости первой пары случайных величин ξ и η и третьей пары η и ζ, мы сделали вывод, что случайные величины, составляющие эти пары, - независимые. Однако, в комментарии полученных результатов была отмечена наша неудовлетворённость сделанными выводами. Эта неудовлетворённость была вызвана относительной близостью вычисленного наблюдаемого значения критерия проверки гипотезы $t_{набл}$ к критическому значению критерия $t_{кр}$.

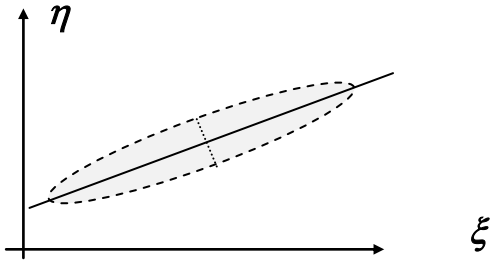
Вычислив точечные оценки коэффициентов линейной корреляции, и проверив гипотезы о значимости этих коэффициентов, мы можем теперь объяснить причину нашей неудовлетворённости: статистическая связь между парами случайных величин – слабая и применяемый в предыдущем

параграфе критерий проверки независимости случайных величин - недостаточно «чувствительный» при наличии слабой статистической связи.

Практика

Вернёмся к примеру 5.1.4, в котором рассматривалась выборка, состоящая из ***n=1760*** слов, и анализировались случайные величины: ξ – количество букв в слове; η – количество гласных букв в слове.

Из геометрической иллюстрации вариационной таблицы 4.1.4 видно, что вертикальный диаметр «эллипса» ненулевых значений относительных частот m_{ij} будет значительно меньше его горизонтального диаметра. Это свидетельствует о наличии сильной статистической связи между случайными величинами.



Оценить силу этой статистической связи мы можем, вычислив значение точечной оценки коэффициента линейной корреляции случайных величин ξ и η. Так как $\bar{x} = 8{,}53$; $s_x = 2{,}776$; $\bar{y} = 3{,}41$; $s_y = 1{,}334$; $a_{11} = 32{,}272$, то

$$r_{xy} = \frac{a_{11} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y} = 0{,}860.$$

Аналогично определим значения точечных оценок числовых характеристик и коэффициента линейной корреляции по ***n=20*** элементам выборки, приведённым в таблице 1.4:

$\bar{x} = 7{,}75$; $s_x = 3{,}878$; $\bar{y} = 3{,}15$; $s_y = 1{,}268$; $a_{11} = 28{,}7$ и $r_{xy} = 0{,}872$.

Вычисленные значения оценок коэффициента линейной корреляции подтверждают наше утверждение о наличии сильной статистической связи между случайными величинами ξ и η – длиной слова и количеством гласных букв в нём.

§ 3.4. Условные распределения и условные математические ожидания. Функция регрессии.

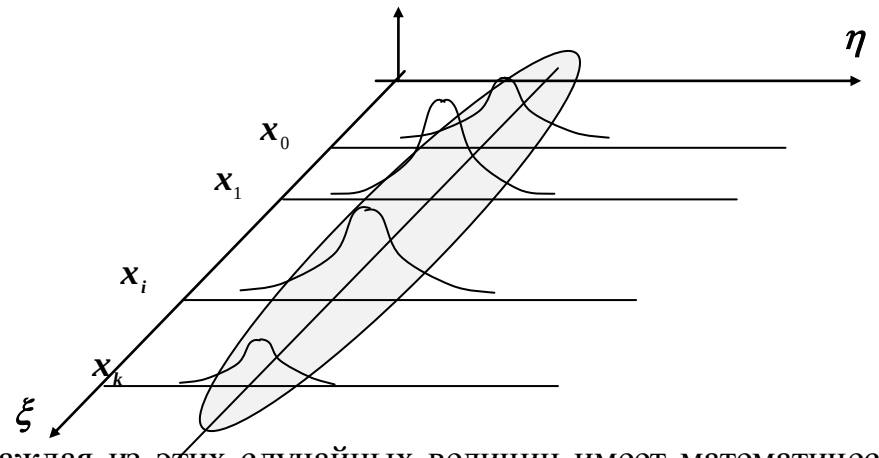
Теория

Рассматривается двумерная случайная величина $\zeta = (\xi, \eta)$ и $F(x, y)$ - её функция распределения. Напомним: частные функции распределения компонент ξ и η определяются, в соответствии со свойством согласованности вероятностных функций, следующим образом:

$F_1(x) = P(\xi \in S_x) = P(\zeta \in S_x \times R_2) = F(x, +\infty)$ и
$F_2(y) = P(\eta \in S_y) = P(\zeta \in R_1 \times S_y) = F(+\infty, y)$.

У случайной величины непрерывного типа $\zeta = (\xi, \eta)$ закон распределения вероятностей задаётся путём определения плотности вероятностей $p(x, y)$. Плотности вероятности компонент определяются так: $p_1(x) = \int_{R_2} p(x, y)dy$, $p_2(y) = \int_{R_1} p(x, y)dx$.

Графиком плотности вероятности $p(x, y)$ в пространстве R^3 будет поверхность, ограничивающая единичный объём. Линией пересечения поверхности $z = p(x, y)$ и плоскости $x = x_0$ будет график функции $z = p(x_0, y)$. Но эта, принимающая неотрицательные значения, функция не будет плотностью вероятности. Разделим эту функцию на значение интеграла $\int_{R_2} p(x_0, y)dy = p_1(x_0)$ и рассмотрим функцию $\frac{p(x_0, y)}{p_1(x_0)} = p_2(y/x_0)$. Эта функция удовлетворяет всем требованиям, предъявляемым к плотности вероятности некоторой непрерывной случайной величины. Эту случайную величину, которую обозначим $\eta/\xi = x_0$, будем называть условной случайной величиной. Аналогично мы можем определить плотность вероятности $\frac{p(x, y_0)}{p_2(y_0)} = p_1(x/y_0)$, которая будет плотностью вероятности условной случайной величины $\xi/\eta = y_0$.



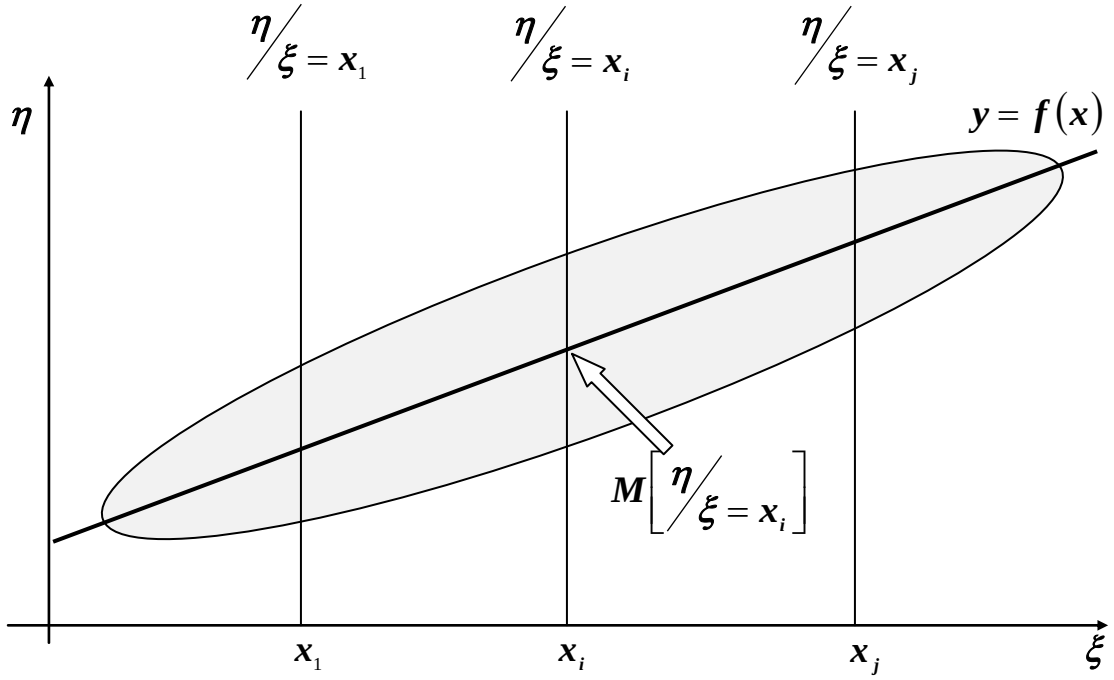
Каждая из этих случайных величин имеет математическое ожидание, которое будем называть условным математическим ожиданием: $M\left[\eta/\xi = x_i\right] = \int_{R_2} y \cdot p(y/x_i)dy$ и $M\left[\xi/\eta = y_j\right] = \int_{R_1} x \cdot p(x/y_j)dx$.

Если случайная величина $\zeta=(\xi,\eta)$ дискретного типа, то её закон распределения вероятностей задаётся определением набора вероятностей $\{p_{ik}\}$, где $i=1,...,n,\infty$; $k=1,...,m,\infty$. Частные распределения её компонент ξ и η - наборы вероятностей: $\{p_{i*}\}$ где $i=1,...,n,\infty$ и $\{p_{*k}\}$, где $k=1,...,m,\infty$, её компонент ξ и η, определяются как суммы для каждого значения i: $p_{i*}=\sum_{k=1}^{m,\infty}p_{ik}$ и для каждого значения k: $p_{*k}=\sum_{i=1}^{n,\infty}p_{ik}$. Распределения вероятностей случайных величин $\eta/\xi=x_{i_0}$ и $\xi/\eta=y_{k_0}$ определяются наборами вероятностей $\left\{p_{k/i_0}\right\}$ и $\left\{p_{i/k_0}\right\}$, где $p_{k/i_0}=P\left(\eta=y_k/\xi=x_{i_0}\right)=\frac{p_{i_0k}}{p_{i_0*}}$ и $p_{i/k_0}=P\left(\xi=x_i/\eta=y_{k_0}\right)=\frac{p_{ik_0}}{p_{*k_0}}$.

Так же, как и для случайных величин непрерывного типа определяем условные математические ожидания условных случайных величин дискретного типа: $M\left[\eta/\xi=x_{i_0}\right]=\sum_{k=1}^{m}y_k\cdot p_{k/i_0}$ и $M\left[\xi/\eta=y_{k_0}\right]=\sum_{i=1}^{n}x_i\cdot p_{i/k_0}$.

Комментарий

Математическое ожидание интерпретируется нами как среднее значение случайной величины.



В примере 5.1.4. по выборке, объём которой равен ***n=1760***, изучаются случайные величины ξ – количество букв в наудачу выбранном слове и η – количество гласных букв в этом слове. Если случайная величина ξ

примет, например, значения x_1 =5, x_i =10 или x_j =15, то выборочные значения случайной величины η равные, соответственно, (1,2,3); (2,3,4,5,6) и (5,6,7,8) будут выборочными значениями условных случайных величин $\eta/_{\xi = x_1}$; $\eta/_{\xi = x_i}$ и $\eta/_{\xi = x_j}$. Вычислив по выборке значения оценок условных математических ожиданий – условные средние арифметические:

$$\bar{y}_{3/_{\xi=5}} = M\left[\eta^*/_{\xi=5}\right] = \frac{1\cdot 15 + 2\cdot 92 + 3\cdot 12}{119} = 1{,}98,$$

$$\bar{y}_{6/_{\xi=10}} = M\left[\eta^*/_{\xi=10}\right] = 3{,}99 \text{ и } \bar{y}_{8/_{\xi=15}} = M\left[\eta^*/_{\xi=15}\right] = 6{,}34,$$

мы можем высказать суждение о средних значениях условных случайных величин $\eta/_{\xi = x_i}$ для указанных конкретных значениях случайной величины ξ. Естественно, что с увеличением возможного значения случайной величины ξ - число букв в наудачу выбранном слове, будет увеличиваться среднее значение η - среднее число гласных букв в этом слове. Исследователя теперь интересует вопрос: как описать динамику изменения средних значений η в зависмости от изменения значений ξ?

Теория

Если случайная величина ξ изменяется в области своих возможных значений, то значения условного математического ожидания условной случайной величины $\eta/_{\xi = x}$ будут функционально изменяться, то есть будут зависеть от этих возможных значений.

Определение. Функция $M\left[\eta/_{\xi = x}\right] = f(x)$, аргументом которой является значение случайной величины ξ, называется *функцией регрессии случайной величины η на случайную величину ξ*.

Замечание. Регрессия не имеет ничего общего с *регрессом.* Слово *регрессия* здесь означает *отклик* контролируемого параметра на изменение независимого фактора. Слово *регресс* означает *движение в обратном направлении* контролируемого параметра при изменении некоторого фактора.

Значит, график функции регрессии случайной величины η на случайную величину ξ показывает, как *откликаются средние значения* случайной величины η на изменение значений случайной величины ξ.

Аналогично определяется функция регрессии $M\left[\xi / \eta = y\right] = \varphi(y)$ случайной величины ξ на случайную величину η.

Комментарий

Во второй задаче корреляционного анализа мы определяем функцию регрессии одной из компонент двумерной случайной величины на другую компоненту. Функция регрессии может быть любого вида: линейной, параболической, гиперболической, экспоненциальной и т.д. Чаще всего изучается линейная регрессия. В учебниках и руководствах по математической статистики рассматривается многомерное, или даже гильбертово пространство значений случайных величин и выводится теоретическое уравнение линейной функции регрессии. Применяя метод моментов, по теоретическому уравнению линейной функции регрессии, выписывается статистическое уравнение линейной регрессии, в котором теоретические коэффициенты уравнения заменяются их точечными оценками.

Так как нас больше интересуют вопросы практической обработки статистических данных, поступим наоборот. Определим статистическое уравнение регрессии, а затем по этому уравнению запишем теоретическое уравнение функции регрессии.

Теория

Рассматривается двумерная случайная величина $\zeta = (\xi, \eta)$, которую мы изучаем по экспериментальным данным – по двумерной выборке $\{(x_i ; y_i)\}$, $i = 1,2,...,n$. По виду «скэттер-диаграммы» - геометрической иллюстрации выборки $\{(x_i ; y_i)\}$ мы определяем тип функции регрессии $M\left[\eta / \xi = x\right] = f(x)$.

Пусть функция регрессии - линейная, то есть $y = f(x) = ax + b$. Надо по элементам выборки найти точечные оценки неизвестных значений коэффициентов a и b. Определим случайную величину $\delta_i = f(\xi_i) - \eta_i$ и для произвольного числового элемента выборки $(x_i ; y_i)$, вычислим разность $f(x_i) - y_i = (ax_i + b) - y_i$. Эта разность равна величине отклонения значения функции регрессии $y = f(x)$ при $\xi = x_i$ от ординаты y_i элемента выборки $(x_i ; y_i)$. Разность $(ax_i - b) - y_i$ может принимать и положительные, и отрицательные значения. Так как элементы выборки мы рассматриваем как выборочную совокупность случайных величин $\{(\xi_i ; \eta_i)\}$, полученное значение разности является значением случайной величины

$\delta_i = f(\xi_i) + \eta_i$. Легко проверить, что $M\left[\sum_{i=1}^{n} \delta_i\right] = 0$, то есть среднее значение отклонений ординат элементов выборки от функции регрессии, в силу её определения, всегда равно нулю.

Каждой паре коэффициентов $(a;b)$ соответствует своя линейная функция $y = f(x)$. Значит выбрать значения $(a;b)$ из условия, чтобы среднее отклонение ординат элементов выборки от значений функции регрессии было минимальным, мы не сможем. Поэтому воспользуемся методом Гаусса – «методом наименьших квадратов».

Построим функцию $\Psi(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$, и найдём оценки коэффициентов a^* и b^* исходя из условия: $\Psi(a^*,b^*) = \min_{(a,b)} \Psi(a,b) = \min_{(a,b)} \sum_{i=1}^{n} (ax_i + b - y_i)^2$.

Получается обычная задача на определение экстремума функции двух переменных. Для решения её мы находим частные производные по переменным a и b, приравниваем их к нулю и решаем систему двух уравнений с двумя неизвестными. Решением этой системы и будет пара коэффициентов a^* и b^*.

$$\begin{cases} \dfrac{\partial \Psi(a,b)}{\partial a} = 0 \\ \dfrac{\partial \Psi(a,b)}{\partial b} = 0 \end{cases} ; \qquad \begin{cases} \sum_{i=1}^{n} (ax_i + b - y_i) \cdot x_i = 0 \\ \sum_{i=1}^{n} (ax_i + b - y_i) \cdot 1 = 0 \end{cases} ;$$

$$\begin{cases} a \cdot \sum_{i=1}^{n} x_i^2 + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a \cdot \sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i \end{cases};$$

Разделив каждое уравнение системы на n – величину объёма выборки, мы получим линейную систему уравнений с привычным для нас видом коэффициентов:

$$\begin{cases} a \cdot \dfrac{1}{n} \sum_{i=1}^{n} x_i^2 + b \cdot \bar{x} = a_{11} \\ a \cdot \bar{x} + b \cdot 1 = \bar{y} \end{cases}.$$ Решением этой системы будут: $a^* = \dfrac{a_{11} - \bar{x} \cdot \bar{y}}{S_x^2}$ и $b^* = \bar{y} - a^* \cdot \bar{x}$. Заменим в получившейся дроби смещённую оценку S_x^2 дисперсии $D\xi$ на несмещённую оценку s_x^2 и умножим эту

дробь на $\frac{s_y}{s_y}$. Так как дробь $\frac{a_{11} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y}$ является статистической оценкой коэффициента линейной корреляции r, то окончательно получаем:

$a^* = r \cdot \frac{s_y}{s_x}$ и $b^* = \bar{y} - r \cdot \frac{s_y}{s_x} \cdot \bar{x}$. Статистическое уравнение линейной регрессии случайной величины η на случайную величину ξ будет иметь вид: $y - \bar{y} = r \cdot \frac{s_y}{s_x}(x - \bar{x})$. Применяя метод моментов «наоборот», запишем теоретическое линейное уравнение регрессии случайной величины η на случайную величину ξ: $y - m_\eta = \rho \cdot \frac{\sigma_\eta}{\sigma_\xi}(x - m_\xi)$.

Комментарий

Аналогично рассуждая, мы можем получить статистическое и теоретическое линейные уравнения регрессии случайной величины ξ на случайную величину η: $x - \bar{x} = r \cdot \frac{s_x}{s_y}(y - \bar{y})$ и $x - m_\xi = \rho \cdot \frac{\sigma_\xi}{\sigma_\eta}(y - m_\eta)$.

Решая систему уравнений $\begin{cases} y - m_\eta = \rho \cdot \frac{\sigma_\eta}{\sigma_\xi}(x - m_\xi) \\ x - m_\xi = \rho \cdot \frac{\sigma_\xi}{\sigma_\eta}(y - m_\eta) \end{cases}$, получаем, что точкой пересечения прямых линий регрессии будет точка с координатами $(m_\xi ; m_\eta)$. Что вполне ожидаемо, так как точка с этими координатами интерпретируется нами как центр тяжести единичной массы распределённой на плоскости. Распределение этой массы определяется законом распределения вероятностей двумерной случайной величины $\zeta = (\xi ; \eta)$.

Очевидно, что, применяя метод наименьших квадратов, мы можем всегда найти статистические оценки коэффициентов функции регрессии любого вида: параболического, экспоненциального, гиперболического и т.д.

Но справедлива теорема о том, что если многомерная случайная величина подчиняется нормальному распределению вероятностей, то функция регрессии любой её компоненты на другую, или на другие компоненты, всегда будет линейной.

§ 4.4 Остаточная дисперсия

Комментарий

Если между компоненты двумерной случайной величины $\zeta = (\xi;\eta)$ существует статистическая зависимость, то разброс возможных значений любой из компонент около её математического ожидания вызывается как «внутренними» свойствами самой компонентами, так и «влиянием» на неё другой компоненты. Возникает вопрос оценки мер разброса значений компоненты, вызванные этими причинами.

Пусть статистическая зависимость между компонентами ξ и η имеет линейный характер. Введём случайную величину $\hat{\eta} = \rho \cdot \frac{\sigma_\eta}{\sigma_\xi}(\xi - m_\xi) + m_\eta$. Ясно, что случайная величина $\hat{\eta}$ является линейной функцией случайной величины ξ и распределение её возможных значений полностью определяется этой случайной величиной. Случайную величину η запишем так: $\eta = (\eta - \hat{\eta}) + \hat{\eta}$ и вычислим дисперсии этих трёх случайных величин.

$D\eta = \sigma_\eta^2$, $D\hat{\eta} = D\left[\rho \cdot \frac{\sigma_\eta}{\sigma_\xi}(\xi - m_\xi) + m_\eta\right] = \rho^2 \cdot \frac{\sigma_\eta^2}{\sigma_\xi^2} \cdot D\xi = \rho^2 \cdot \sigma_\eta^2$, а третью дисперсию определим, используя формулу: $D[\eta - \hat{\eta}] = M[\eta - \hat{\eta}]^2 - M^2[\eta - \hat{\eta}]$. Ясно, что $M[\eta - \hat{\eta}] = 0$. Вычислим $M[\eta - \hat{\eta}]^2$.

Так как $[\eta - \hat{\eta}]^2 = (\eta - m_\eta)^2 - 2\rho\frac{\sigma_\eta}{\sigma_\xi}(\xi - m_\xi)\cdot(\eta - m_\eta) + \rho^2\frac{\sigma_\eta^2}{\sigma_\xi^2}(\xi - m_\xi)^2$

и $M[(\xi - m_\xi)\cdot(\eta - m_\eta)] = \mu_{11} = \sigma_\xi \cdot \sigma_\eta \cdot \rho$, то

$D[\eta - \hat{\eta}] = M[\eta - \hat{\eta}]^2 = D\eta - \rho^2 \cdot D\eta$, то есть $D[\eta - \hat{\eta}] = (1 - \rho^2)\cdot\sigma_\eta^2$.

Получили равенство: $\sigma_\eta^2 = (1 - \rho^2)\cdot\sigma_\eta^2 + \rho^2 \cdot \sigma_\eta^2$.

Наблюдаемые значения случайной величины η являются результатом влияния двух факторов. Первый фактор – это влияние случайной величины ξ, которая определяет значения случайной величины $\hat{\eta}$. Другой фактор – это внутренние особенности случайной величины η, которые определяют значения случайной величины $\eta - \hat{\eta}$. Следовательно, разброс значений случайной величины η складывается из разброса значений, вызванных влиянием на неё случайной величины ξ, и разброса значений, вызванных внутренними свойствами случайной величины η. Величину разброса, вызванного влиянием ξ оценивает величина $\rho^2 \cdot \sigma_\eta^2$. Величину разброса, вызванного внутренними свойствами η оценивает величина $(1 - \rho^2)\cdot\sigma_\eta^2$, которая называется *остаточной дисперсией*. То есть величина $(1 - \rho^2)\cdot\sigma_\eta^2$

оценивает внутренние особенности случайной величины η, когда влияние случайной величины ξ исключено.

§ 5.4 Корреляционное отношение

Комментарий

Коэффициент линейной корреляции оценивает силу статистической связи между случайными величинами ξ и η в том случае, когда статистическая зависимость имеет линейный характер, то есть когда функция регрессии – линейная функция. Если функция регрессии не будет линейной, то равенство нулю коэффициента линейной корреляции может привести к ошибочному решению об отсутствии статистической зависимости между случайными величинами. В таких случаях используется *корреляционное отношение.*

Теория

Представим дисперсию случайной величины η в виде суммы двух слагаемых. Значения одного из них определяются влиянием ξ и η, а значения другого определяются внутренними свойствами самой случайной величины η.

Пусть функция регрессии $M\left[\eta/_{\xi=x}\right]=f(x)$ - произвольная функция любого типа, не обязательно линейного, параболического, гиперболического экспоненциального. Определим случайную величину $\hat{\eta}=f(\xi)$ - функцию случайной величины ξ. Возможные значения этой случайной величины определяются случайной величиной ξ.

Будет справедливо следующее представление дисперсии случайной величины η: $D\eta=M\left[\eta-m_\eta\right]^2=M\left[(\eta-f(\xi))+(f(\xi)-m_\eta)\right]^2$.

Преобразуем правую часть этого равенства:

$$M\left[\eta-m_\eta\right]^2=M\left[\eta-f(\xi)\right]^2+2M\left[(\eta-f(\xi))\cdot(f(\xi)-m_\eta)\right]+M\left[f(\xi)-m_\eta\right]^2$$

.

Вычислим второе слагаемое:

$$M\left[(\eta-f(\xi))\cdot(f(\xi)-m_\eta)\right]=\iint_{R^2}(y-f(x)\cdot(f(x)-m_\eta)dF(x,y)).$$

Из определения условного распределения вероятностей следует, что $F(x,y)=F_1(x)\cdot F\left(y/_{\xi=x}\right)$, где $F\left(y/_{\xi=x}\right)$ - функция распределения условной случайной величины $\eta/_{\xi=x}$. Переходя от кратного интеграла к повторному интегралу, получим:

$$M\left[(\eta - f(\xi))\cdot(f(\xi)-m_\eta)\right]=\int_{R_1}(f(x)-m_\eta)\left[\int_{R_2}(y-f(x))dF\left(y/\xi=x\right)\right]dF_1(x).$$

Рассмотрим внутренний интеграл:

$$\int_{R_2}(y-f(x))dF\left(y/\xi=x\right)=\int_{R_2}ydF\left(y/\xi=x\right)-f(x)\cdot\int_{R_2}dF\left(y/\xi=x\right).$$

Уменьшаемое этой разности, согласно определению функции регрессии, будет равно $M\left[\eta/\xi=x\right]=f(x)$, а так как изменение любой функции распределения на числовой оси равно единице, то вычитаемое этой разности $f(x)\cdot\int_{R_2}dF\left(y/\xi=x\right)=f(x)\cdot 1$ так же будет равно $f(x)$. Итак, второе слагаемое $M\left[(\eta - f(\xi))\cdot(f(\xi)-m_\eta)\right]$ равно нулю и дисперсия случайной величины η представлена как сумма двух слагаемых $D\eta = M\left[\eta - f(\xi)\right]^2 + M\left[f(\xi)-m_\eta\right]^2$.

Очевидно, что первое слагаемое оценивает разброс возможных значений η около функции регрессии $y=f(x)$, то есть разброс возможных значений η, объясняемый «внутренними» свойствами этой случайной величины. Второе слагаемое оценивает разброс значений функции регрессии около математического ожидания случайной величины η. Величина этого разброса зависит от «силы» влияния случайной величины ξ на случайную величину η. Поэтому в роли оценки силы этого влияния естественно взять величину $\delta^2=\dfrac{M\left[f(\xi)-m_\eta\right]^2}{D\eta}$, которую называют *корреляционным отношением.* Ясно, что $0\le\delta^2\le 1$ и чем сильнее влияние случайной величины ξ на случайную величину η, ближе будет значение корреляционного отношения δ^2 к единице. Если функция регрессии $M\left[\eta/\xi=x\right]=f(x)$ линейная, то будет выполняться равенство $\rho^2=\delta^2$. В общем случае для любой функции регрессии будет справедливо неравенство $\rho^2\le\delta^2$.

Комментарий

Применяя метод моментов, определим значение точечной оценки $(\delta^2)^*$ корреляционного отношения по двумерной числовой выборке $\{(x_i, y_i)\}$, где $i=1,2,...,n$.

Пусть первые координаты x_i двумерной выборки принимают μ различных значений. Переберём эти возможные значения. Значение $\xi=x_k$

встречается в выборке n_k раз, $k = 1,...,\mu$, значит, в двумерной выборке имеется n_k пар (x_k, y_j), в которых первая координата равна одному и тому же значению x_k, а вторая координата принимает n_k значений y_j, среди которых могут быть и одинаковые значения. Ясно, что $n_1 + ... + n_k + ... + n_\mu = n$. Для каждого встречающегося в выборке значения $\xi = x_k$ вычислим значение условного среднего арифметического $\bar{y}_{/\xi = x_k} = \frac{1}{n_k}\sum_{j=1}^{n_k} y_j$. Тогда дисперсионная сумма $\frac{1}{\mu - 1}\sum_{k=1}^{\mu}\left(\bar{y}_{/\xi = x_k} - \bar{y}\right)^2 = \vartheta^*$ будет состоятельной и несмещённой оценкой дисперсии функции регрессии $M\left[f(\xi) - m_\eta\right]^2 = \vartheta$. Следовательно, точечной оценкой $(\delta^2)^*$ корреляционного отношения будет статистика $(\delta^2)^* = \frac{\frac{1}{\mu - 1}\sum_{k=1}\left(\bar{y}_{/\xi = x_k} - \bar{y}\right)^2}{s_\eta^2}$.

«Статистика – совокупность методов,
которые дают нам возможность принимать
решения в условиях неопределённости»
Абрахам Вальд

Модуль пятый

Элементы дисперсионного анализа

§1.5 Планирование эксперимента. Однофакторный комплекс

Теория

Дисперсионный анализ один из методов статистической проверки гипотез о равенстве значений математических ожиданий и планирования эксперимента.

Предположим, что мы изучаем некоторый количественный признак (*урожайность ячменя, риса; больных некоторой болезнью, которую выявляем путём проведения анализов)*. Изучаемый количественный признак мы называем случайной величиной ξ.

Мы хотим изучить влияние некоторого фактора F на случайную величину ξ. Это влияние проявляется в том, что значения случайной величины ξ изменяются при действии фактора F. Этот фактор имеет k уровней F_i $(i = 1,2,...,k)$. Например: k различных уровней концентрации применяемых удобрений, k уровней засолённости почвы, k схем приёма лекарственных препаратов, k различных лекарственных средств, k методик обучения учащихся или персонала предприятия и т.п. Экспериментатор желает выяснить влияют ли различные уровни фактора на средние значения случайной величины, то есть, выяснить наилучший уровень концентрации применяемых удобрений, выбрать наилучшую схему лечения заболевания, выбрать наилучшую методику обучения. Предполагается провести эксперимент, опыт и наблюдать получающиеся значения изучаемой случайной величины ξ.

Составим план проведения эксперимента.

Общая конфигурация планируемого эксперимента может быть представлена в виде таблицы 1.5.

Каждому анализируемому уровню фактора F_i соответствует i-тая строка. Предполагается провести эксперимент, опыт и наблюдать получающиеся значения изучаемого количественного признака.

Таблица 1.5.

уровни	Результаты эксперимента						
фактора	**1**	**2**	**3**	**…**	**j**	**…**	**n**
F_1	x_{11}	x_{12}	x_{13}	…	x_{1j}	…	x_{1n}
F_2	x_{21}	x_{22}	x_{23}	…	x_{2j}	…	x_{2n}
…	…	…	…	…	…	…	…
F_i	x_{i1}	x_{i2}	x_{i3}	…	x_{ij}	…	x_{in}
…	…	…	…	…	…	…	…
F_k	x_{k1}	x_{k2}	x_{k3}	…	x_{kj}	…	x_{kn}

Элементы i-той строки таблицы рассматриваются как выборка возможных значений $\{x_{ij}\}$, $j = 1,2,...,n$ случайной величины ξ_i. Таких выборок у нас будет k штук и каждая выборка независима от остальных выборок. Объём каждой выборки равен n. Эксперимент, в котором все выборки имеют одинаковый объём, называется *сбалансированным*.

По элементам каждой из k штук выборок мы можем вычислить среднее арифметическое $\bar{x}_i$ и статистическую оценку дисперсии s_i^2, которые будем называть *групповыми*: $\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$; $s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$.

В результате будем иметь k штук групповых средних арифметических $\bar{x}_i$ и k штук оценок групповых дисперсий s_i^2. Кроме этих оценок вычисляем общее среднее арифметическое и общую оценку дисперсии по элементам объединённой выборки $\{x_{ij}\}, j=1,...,n; i=1,...,k$. Объём этой объединённой выборки равен $k \cdot n$.

$$\bar{x} = \frac{1}{k \cdot n} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}; \quad s^2 = \frac{1}{k \cdot n - 1} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} - \bar{x} \right)^2 .$$

Основная гипотеза H_0 формулируется так:

H_0: *Фактор F не влияет на изучаемый количественный признак.*

То есть, различные уровни фактора F_i не вызывают заметных различий в распределениях вероятностей случайных величин ξ_i.

Мы считаем, что количественный признак ξ, который имеет нормальное распределение вероятностей,

Сделать вывод о влиянии, или не влиянии уровней фактора F_i на количественный признак ξ мы сможем, анализируя и сравнивая значения математических ожиданий $M\xi_i = m_i$ этих случайных величин ξ_i. Если фактор F не влияет на количественный признак ξ, то все групповые математические ожидания должны быть равны между собой и равны математическому ожиданию $M\xi = m$. Значит, основная гипотеза записывается так:

$H_0: m_1 = m_2 = ... = m_k = m$.

Построение критерия проверки справедливости этой гипотезы предполагает, что все групповые дисперсии σ_i^2 случайных величин ξ_i равны между собой. Значит, прежде чем рассматривать критерий проверки основной гипотезы H_0 о равенстве групповых математических ожиданий, мы сначала проверяем справедливость гипотезы $H_0: \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$.

Можно применить критерий, используемый при сравнении дисперсий двух генеральных совокупностей $H_0: D\xi = D\eta$, с которым мы познакомились при рассмотрении задач второго типа статистической проверки гипотез. Выбираем максимальную $max\, s_i^2$ и минимальную $min\, s_i^2$ оценки дисперсий. В качестве критерия проверки гипотезы о равенстве дисперсий берём статистику $f_{k-1,k-1} = \frac{max\, s_i^2}{min\, s_i^2}$. Если разница между максимальной и минимальной оценками дисперсий будет признана незначительной, то мы можем считать, что гипотеза о равенстве всех k групповых дисперсий принята, то есть все групповые дисперсии σ_i^2 равны между собой и равны числу σ^2.

Обратимся к оценкам, вычисляемым по элементам объединённой выборки объёма $k \cdot n$. Вычисляя общее среднее арифметическое $\bar{x}$, мы последовательно будем суммировать сначала элементы первой выборки $\{x_{1j}\}$ (*здесь* $i=1, \quad j=1,2,...,n$), затем будем суммировать элементы второй выборки $\{x_{2j}\}$(*здесь* $i=2, \quad j=1,2,...,n$), и так далее.

То есть мы можем записать: $\bar{x}=\frac{1}{k \cdot n}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n}x_{ij}\right)=\frac{1}{k}\sum_{i=1}^{k}\bar{x}_i$,

или - среднее арифметическое элементов объединённой выборки $\bar{x}$ равно среднему арифметическому групповых средних арифметических $\bar{x}_i$ $(i=1,2,...,k)$.

Из формулы статистической оценки дисперсии s^2:

$$s^2=\frac{1}{k \cdot n-1}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}\right)^2 ,$$

вычисляемую по элементам объединённой выборки, возьмём только двойную сумму и обозначим её Q, то есть $\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}\right)^2=Q$.

§2.5 Дисперсионные суммы. Критерий проверки основной гипотезы

Комментарий

Идея метода дисперсионного анализа, предложенная Фишером, состоит в представлении дисперсионной суммы Q в виде суммы двух дисперсионных сумм $Q=Q_1+Q_2$. Одна сумма зависит от внутригрупповых особенностей случайных величин ξ_i, а они у них практически одинаковые. Другая сумма зависит от межгрупповых отличий случайных величин ξ_i. Эти отличия являются следствием действия уровней фактора F и проявляются в различии значений групповых средних $M\xi_i=m_i$.

Теория

Очевидно, что каждое слагаемое $\left(x_{ij}-\bar{x}\right)^2$, составляющее Q, не изменится, если мы в скобках отнимем и прибавим одно и то же число $\bar{x}_i$:

$$Q=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\left(x_{ij}-\bar{x}_i\right)+\left(\bar{x}_i-\bar{x}\right)\right)^2 .$$

Вспомнив, что $(a+b)^2=a^2+2ab+b^2$ и, проведя упрощения, получим:

$$Q=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2+n\sum_{i=1}^{k}\left(\bar{x}_i-\bar{x}\right)^2 .$$

Обозначим первую дисперсионную сумму: $\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2=Q_1$, а вторую дисперсионную сумму: $n\sum_{i=1}^{k}\left(\bar{x}_i-\bar{x}\right)^2=Q_2$.

Запишем: $Q=Q_1+Q_2$ и проанализируем отдельно каждое слагаемое.

Внутренняя сумма $\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2$ первого слагаемого Q_1 вычисляется при определении статистической оценки дисперсии s_i^2 по элементам i-той выборки $\{x_{ij}\}$, значения которой зависят от действия i-того уровня фактора F_i *(это данные опыта, которые записываются в* i*-той строке)*:

$$s_i^2=\frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2.$$

То есть, все k штук слагаемых, составляющих величину Q_1 , учитывают только внутригрупповые разбросы значений n элементов i-той группы, и каждое из k возможных уровней фактора F_i на эти значения никак не влияет, так как мы приняли гипотезу о том, что все внутригрупповые дисперсии равны.

Рассмотрим оценку дисперсии:

$$S_1^2=\frac{Q_1}{k\cdot(n-1)}=\frac{1}{k\cdot(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2.$$

Ясно что $S_1^2=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2\right)=\frac{1}{k}\sum_{i=1}^{k}s_i^2$, то есть, S_1^2 является средним арифметическим групповых оценок дисперсий s_i^2 . Так как мы считаем, что все групповые дисперсии σ_i^2 равны одному и тому же числу, которое мы обозначили σ^2 , то математическое ожидание этой оценки дисперсии будет равно: $M\left[S_1^2\right]=\frac{1}{k}\sum_{i=1}^{k}M\left[s_i^2\right]=\frac{1}{k}\cdot k\sigma^2=\sigma^2$. Из записи оценки дисперсии S_1^2 видно, что она зависит только от оценок внутри групповых дисперсий s_i^2, и каждый из k возможных уровней F_i фактора F на эти значения оценок никак не влияет, так как мы приняли гипотезу о том, что все внутригрупповые дисперсии равны σ^2. Оценку дисперсии S_1^2 будем называть «внутригрупповой дисперсией».

Во втором слагаемом Q_2 величина суммы $\sum_{i=1}^{k}\left(\bar{x}_i-\bar{x}\right)^2$ будет зависеть от величины отклонений групповых средних арифметических $\bar{x}_i$ от общего среднего арифметического $\bar{x}$. Ясно, что величины этих отклонений (*величины квадратов разностей* $\left(\bar{x}_i-\bar{x}\right)$) зависят от силы влияния каждого

уровня F_i фактора F на каждое групповое среднее арифметическое $\bar{x}_i$. Оценка $\frac{1}{k-1}\sum_{i=1}(\bar{x}_i-\bar{x})^2$ является статистической оценкой дисперсии групповых средних арифметических. Эту статистическую оценку $S_2^2=\frac{Q_2}{k-1}$ будем называть «межгрупповой дисперсией». Можно проверить, что:

$$M\left[S_2^2\right]=\sigma^2+\frac{n}{k-1}\sum_{i=1}^{k}(m_i-m)^2 .$$

Отсюда видно, что величина «межгрупповой дисперсии» S_2^2 значительно зависит от влияния уровней F_i фактора F на значения количественного признака ξ. Это влияние проявляется в отклонении значений групповых средних $M\xi_i=m_i$, $i=1,2,...,k$, от $M\xi=m$ среднего значения количественного признака ξ.

Если фактор F не влияет на случайную величину ξ, то групповые средние m_i будут равны общему среднему m, то есть будет справедлива основная гипотеза H_0. Значит, в качестве критерия проверки справедливости гипотезы H_0 можно взять статистику: $f_{k-1,k(n-1)}=\frac{S_2^2}{S_1^2}$, которая подчиняется "$F$-распределению" Фишера-Снедекора.

Если значения этой статистики мало отличаются от единицы, то фактор F не оказывает значительного влияния на исследуемую случайную величину ξ. Если значения этой статистики значимо отличаются от единицы, то мы делаем вывод о том, фактор F заметно влияет на случайную величину ξ. Выбрав уровень значимости α, для значения параметров $k-1$ и $k\cdot(n-1)$ по таблицам "F-распределения" определяем критическое значение $f_{кр}(\alpha)$. Областью допустимых значений $Q_{доп}$ критерия $f_{k-1,k(n-1)}$ будет интервал $(1;f_{кр}(\alpha))$, а критической областью $S_{кр}$ - интервал $(f_{кр}(\alpha);+\infty)$.

Далее процедура обычная: построив области $Q_{доп}$ и $S_{кр}$, вычисляем наблюдаемое значение критерия $f_{набл}=\frac{S_2^2}{S_1^2}$. Промежуточные результаты вычислений значения $f_{набл}$ удобно записывать таблице.

Таблица 2.5.

Изменчивость	Число степеней свободы	Дисперсионные суммы	Оценки дисперсий
Внутригрупповая	$k(n-1)$	$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2$	$S_1^2 = \frac{Q_1}{k(n-1)}$
Межгрупповая	$k-1$	$Q_2 = n \cdot \sum_{i=1}^{k}\left(\bar{x}_i - \bar{x}\right)^2$	$S_2^2 = \frac{Q_2}{k-1}$
Общая	$k \cdot n - 1$	$Q = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}\right)^2$	$S^2 = \frac{Q}{k \cdot n - 1}$

Равенство $Q = Q_1 + Q_2$ можно использовать как промежуточный контроль точности вычислений.

Смотрим, в какую область попало это наблюдаемое значение критерия $f_{набл}$, и принимаем решение об отклонении, или принятии гипотезы H_0.

На этом завершается первый этап дисперсионного анализа. Можно провести второй этап анализа, на котором определяются уровень фактора F_μ (или уровни), который наиболее существенно влияет на количественный признак, что проявляется в наиболее существенном отклонении группового математического ожидания m_μ от общего значения математического ожидания $m = \frac{1}{k}\sum_{i=1}^{k} m_i$, или от среднего значения остальных $k-1$ групповых средних. Если m' - среднее значение групповых математических ожиданий, из числа которых исключено одно слагаемое m_μ, то определение наиболее существенно влияющего уровня фактора F_μ осуществляется так же, как и при решении задачи проверки гипотезы о равенстве математических ожиданий двух случайных величин ξ и η.

Можно провести определение уровней фактора F_μ и F_ν, дающих наиболее значимую разницу математических ожиданий m_μ и m_ν.

Практика

Пример 1.5. Предлагаются четыре методики изучения некоторого материала:

1. сокращённая программа (краткий конспект);
2. конспектирование учебника;
3. использование программированного учебника;
4. использование обучающей программы.

Планируется проверить качество усвоения материала в зависимости от используемой методики изучения. Выбранные сорок студентов случайным образом делятся на четыре группы по десять человек. Все студенты изучают один и тот же материал, каждая группа использует одну из предлагаемых методик. Качество усвоения материала измеряется тестом с многовариантным выбором.

В описываемом эксперименте рассматривается фактор F – «активность учащихся в изучении нового материала», а четыре предлагаемые методики изучения – уровни фактора.

Формулируем основную гипотезу H_0: *Качество усвоения материала студентами не зависит от методики изучения.*

Результаты эксперимента – результаты тестирования сорока студентов представлены в таблице.

Таблица 3.5.

Уровни фактора F		Наблюдения									
		1	2	3	4	5	6	7	8	9	10
F_1	Сокращённая программа	26	34	46	48	42	49	74	61	51	53
F_2	Конспектирование учебника	51	50	33	28	47	50	48	60	71	42
F_3	Программированный учебник	52	64	39	54	58	53	77	56	63	59
F_4	Обучающая программа	41	49	56	72	65	63	87	77	62	64

Мы имеем выборку $\{x_{ij}\}$, где $i = 1,2,3,4$; $j = 1,2,...,10$, объём которой равен $k \cdot n = 4 \cdot 10 = 40$. Элементы i-той строки таблицы являются выборкой значений случайной величины ξ_i - результат тестирования при изучении нового материала по методике F_i. Вычислим групповые средние арифметические и оценки дисперсий по экспериментальным данным каждой строки результатов тестирования:

$\bar{x}_1 = 48,4; s_1^2 = 180,50$ $\bar{x}_2 = 48,0; s_2^2 = 150,22$ $\bar{x}_3 = 57,5; s_3^2 = 95,83$ и $\bar{x}_4 = 63,6; s_4^2 = 176,04$. Общее среднее арифметическое и общая оценка дисперсии сорока элементов выборки будут равны $\bar{x}_{общ.} = 54,38$ и $s_{общ}^2 = 182,29$. Групповые средние арифметические $\bar{x}_i$ являются точечными оценками математических ожиданий $M\xi_i = m_i$. Основную гипотезу можно переписать H_0: $m_1 = m_2 = m_3 = m_4$.

Перед вычислением наблюдаемого значения критерия проверки основной гипотезы мы должны убедиться, что групповые дисперсии $D\xi_i = \sigma_i^2$, $i = 1,2,3,4$, равны. То есть мы должны проверить справедливость гипотезы H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$.

Применим критерий, используемый при сравнении дисперсий двух генеральных совокупностей H_0: $D\xi = D\eta$, с которым мы познакомились

при рассмотрении задач второго типа статистической проверки гипотез. Выбираем максимальную и минимальную оценки дисперсий $\max s_i^2$ и $\min s_i^2$. В качестве критерия проверки гипотезы о равенстве дисперсий берём статистику $f_{k-1,k-1} = \frac{\max s_i^2}{\min s_i^2}$. Тогда наблюдаемое значение критерия будет равно $f_{набл.} = \frac{180,5}{95,83} = 1,884$. По таблицам "$F$-распределения" при уровне значимости $\alpha = 0,05$ определяем для параметров $(k-1, k-1)$, где k=10, критическое значение $f_{кр.} = 3,18$. Так как $f_{набл.} < f_{кр.}$, то гипотеза о равенстве групповых дисперсий принимается.

Вычисляем значения дисперсионных сумм и записываем их в таблицу.

Таблица 2.5.

Изменчивость	Число степеней свободы	Дисперсионные суммы	Оценки дисперсий
Внутригрупповая	$k(n-1)$ 36	$Q_1 = 5397,30$	$S_1^2 = 149,93$
Межгрупповая	$k-1$ 3	$Q_2 = 1712$	$S_2^2 = 570,67$
Общая	$k \cdot n - 1$ 39	$Q = 7109,30$	$S^2 = 182,29$

Наблюдаемое значение критерия будет равно $f_{набл} = \frac{S_2^2}{S_1^2} = \frac{570,67}{149,93} = 3,806$. По таблицам "$F$-распределения" при уровне значимости $\alpha = 0,05$ для параметров $(k-1, k \cdot (n-1))$, где k=4, n = 10, определяем критическое значение $f_{кр.} = f_{3,36}(0,05) = 2,86$. Так как $f_{набл.} > f_{кр.}$, то гипотеза о равенстве групповых математических ожиданий отклоняется. То есть мы не можем утверждать, что «*качество усвоения материала студентами не зависит от методики изучения*».

Таблица значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

	.,.0	.,.1	.,.2	.,.3	.,.4	.,.5	.,.6	.,.7	.,.8	.,.9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3652	3637	3621	3605	3589	3572	3555	3538
0,5	0,3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	0,1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0,0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0,0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003

3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Таблица значений функции $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}dz$

	...,...0	...,...1	...,...2	...,...3	...,...4	...,...5	...,...6	...,...7	...,...8	...,...9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0754
0,2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	2258	2291	2324	2356	2389	2422	2454	2486	2518	2549
0,7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2996	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3906	3925	3944	3962	3980	3997	4015
1,3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4430	0,4441
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4648	4656	4664	4671	4678	4686	4693	4700	4606
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4762	4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2,2	4861	4864	4868	4871	4874	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4924	4927	4929	4930	4932	4934	4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	4953	4955	4956	4957	4958	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2,9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3,0	0,4986	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	4990	4991	4991	4991	4992	4992	4992	4992	4993	4993
3,2	4993	4993	4994	4994	4994	4994	4994	4995	4995	4995
3,3	4995	4995	4996	4996	4996	4996	4996	4996	4996	4996
3,4	4997	4997	4997	4997	4997	4997	4997	4997	4998	4998
3,5	0,49977	0,49978	0,49978	0,49979	0,49980	0,49981	0,49981	0,49982	0,49983	0,49983

3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997
4,0	0,49997			4,5	0,499997				5,0	0,499997

Распределение Стьюдента

$P\left(|t| \geq t_n(\alpha_1)\right) = \alpha_1$ (1-я строка)

$P\left(t \geq t_n(\alpha_2)\right) = \alpha_2$ (2-я строка)

n \ α_1 / α_2	0,20	0,10	0,05	0,02	0,01	0,002	0,001
	0,10	0,05	0,025	0,010	0,005	0,001	0,0005
1	3,078	6,314	12,706	31,821	63,657	318,31	636,62
2	1,886	2,920	4,303	6,965	9,925	22,326	31,598
3	1,638	2,353	3,182	4,541	5,841	10,213	12,924
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	1,323	1,721	2,080	2,519	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	1,303	1,684	2,021	2,423	2,704	3,307	3,551

60	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	1,294	1,667	1,993	2,379	2,641	3,207	3,430
100	1,289	1,658	1,980	2,358	2,617	3,160	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Распределение Пирсона («Хи-квадрат»)

$$P\left(\chi_n^2 \geq \chi^2(\alpha)\right) = \alpha$$

n α	0,995	0,990	0,975	0,950	0,050	0,025	0,010	0,005
1	0,00004	0,00016	0,00098	0,00393	3,841	5,024	6,635	7,879
2	0,0100	0,0201	0,0506	0,1026	5,991	7,378	9,21	10,60
3	0,0717	0,1148	0,2158	0,3518	7,815	9,348	11,34	12,84
4	0,2070	0,2971	0,4814	0,7107	9,488	11,14	13,28	14,86
5	0,4117	0,5543	0,8312	1,145	11,07	12,83	15,09	16,75
6	0,6757	0,8721	1,2373	1,635	12,59	14,45	16,81	18,55
7	0,9893	1,239	1,690	2,167	14,07	16,01	18,48	20,28
8	1,344	1,646	2,180	2,733	15,51	17,53	20,09	21,96
9	1,735	2,088	2,700	3,325	16,92	19,02	21,67	23,59
10	2,156	2,558	3,247	3,940	18,31	20,48	23,21	25,19
11	2,603	3,053	3,816	4,575	19,68	21,92	24,72	26,76
12	3,074	3,571	4,404	5,226	21,03	23,34	26,22	28,30
13	3,565	4,107	5,009	5,892	22,36	24,74	27,69	29,82
14	4,075	4,660	5,629	6,571	23,68	26,12	29,14	31,32
15	4,601	5,229	6,262	7,261	25,00	27,49	30,58	32,80
16	5,142	5,812	6,908	7,962	26,30	28,85	32,00	34,27
17	5,697	6,408	7,564	8,672	27,59	30,19	33,41	35,72
18	6,265	7,015	8,231	9,390	28,87	31,53	34,81	37,16
19	6,844	7,633	8,907	10,12	30,14	32,85	36,19	38,58
20	7,434	8,260	9,591	10,85	31,41	34,17	37,57	40,00
21	8,034	8,897	10,28	11,59	32,67	35,48	38,93	41,40
22	8,643	9,542	10,98	12,34	33,92	36,78	40,29	42,80
23	9,260	10,20	11,69	13,09	35,17	38,08	41,64	44,18
24	9,886	10,86	12,40	13,85	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	42,56	45,72	49,59	52,34

30	13,79	14,95	16,79	18,49	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	90,53	95,02	100,40	104,20
80	51,17	53,54	57,15	60,39	101,90	106,60	112,30	116,30
90	59,20	61,75	65,65	69,13	113,10	118,10	124,10	128,30
100	67,33	70,06	74,22	77,93	124,30	129,60	135,80	140,20

Распределение Фишера-Снедекора (F-распределение)

$$P\left(f_{n,m} \geq f(\alpha)\right) = \alpha$$

$\alpha = 0{,}05$ - обычный шрифт $\alpha = \mathbf{0{,}01}$ - **жирный шрифт**

n – число степеней свободы для большей дисперсии
m – число степеней свободы для меньшей дисперсии

m \ n	20	30	40	60	100	120	200	∞
20	2,12 **2,94**	2,04 **2,77**	1,99 **2,69**	1,95 **2,21**	1,90 **2,53**	1,89 **2,52**	1,87 **2,47**	1,84 **2,42**
30	1,93 **2,55**	1,84 **2,38**	1,79 **2,29**	1,75 **2,21**	1,69 **2,13**	1,68 **2,12**	1,66 **2,07**	1,62 **2,01**
40	1,84 **2,37**	1,74 **2,20**	1,69 **2,11**	1,65 **2,03**	1,59 **1,94**	1,58 **1,93**	1,55 **1,88**	1,51 **1,81**
60	1,75 **2,20**	1,65 **2,03**	1,59 **1,93**	1,57 **1,85**	1,48 **1,74**	1,47 **1,73**	1,44 **1,68**	1,39 **1,60**
100	1,68 **2,06**	1,57 **1,89**	1,51 **1,79**	1,47 **1,76**	1,39 **1,59**	1,38 **1,57**	1,34 **1,51**	1,28 **1,43**
125	1,65 **2,03**	1,55 **1,85**	1,49 **1,75**	1,43 **1,65**	1,36 **1,54**	1,35 **1,52**	1,31 **1,46**	1,25 **1,37**
200	1,62 **1,97**	1,52 **1,79**	1,45 **1,69**	1,40 **1,60**	1,32 **1,48**	1,30 **1,46**	1,26 **1,39**	1,19 **1,28**
∞	1,57 **1,87**	1,46 **1,69**	1,40 **1,59**	1,35 **1,49**	1,24 **1,36**	1,23 **1,34**	1,17 **1,25**	1,00 **1,09**

Литература

1. Боровков А.А. Теория вероятностей. – М.: Наука, 1976.
2. Боровков А.А. Математическая статистика. – М.: Наука,1984.
3. Боровков А.А. Математическая статистика. Дополнительные главы – М.: Наука,1984.
4. Гланц С. Медико-биологическая статистика. – М.: Практика, 1999.
5. Глотов Н.В. и др. Биометрия. – Л.: ЛГУ, 1982.
6. Климов Г.П. Теория вероятностей и математическая статистика. – М.: МГУ, 1983.
7. Крамер Г. Математические методы статистики. – М.: Мир, 1975.
8. Леман Э. Проверка статистических гипотез. – М.: Наука, 1964.
9. Терентьев П.В., Ростова Н.С. Практикум по биометрии.- Л.: ЛГУ, 1977.
10. Уилкс С. Математическая статистика. – М.: Наука, 1967.
11. Урбах В.Ю. Биометрические методы. – М.: Наука, 1964.
12. Феллер В. Введение в теорию вероятностей и её приложения, т.1,2. – М.: Мир, 1984.
13. Шметтерер Л. Введение в математическую статистику. – М.: Наука, 1976
14. Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980.
15. Ширяев А.Н. Вероятность. – М.: Наука, 1980.

Оглавление