ПРЕДПОЛОЖЕНИЕ О НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ В СТАТИСТИЧЕСКОМ АНАЛИЗЕ ДАННЫХ

С.Я. Шатских

Самарский государственный аэрокосмический университет имени академика С.П. Королева (национальный исследовательский университет) (СГАУ), Самара, Россия.

Приводится краткий обзор истории методов анализа данных, основанных на предположении о нормальности распределений. Рассматриваются вопросы, связанные с практической проверкой нормальности, а также последствия нарушения нормальности распределений в стандартных методах анализа данных.

Ключевые слова: нормальное распределение, критерии согласия, последствия нарушения гипотезы нормальности, большие выборки.

Стандартное обоснование использования нормального распределения вероятностей в задачах обработки наблюдений основано на «механизме» центральной предельной теоремы. Согласно этой теореме нормальное распределение служит хорошим приближением в том случае, когда рассматриваемая случайная величина представляет собой сумму большого числа независимых случайных переменных, каждая из которых вносит лишь небольшой вклад во всю сумму (теоремы Ляпунова и Линдеберга - Феллера).

Этот аргумент к концу XIX-го века был подкреплен большим числом опытов, результаты которых показали хорошее приближенное согласие распределения наблюдаемых частот с нормальным распределением. Следует отметить, что общее признание нормального закона в то время основывалось на результатах наблюдений (или измерений) в основном в физических науках и, в частности, в астрономии и геодезии. В этих исследованиях, основным источником появления случайности были *ошибки измерений*.

В астрономии, где движения небесных тел определяется уравнениями классической механики, измерения можно было делать с высокой точностью. Таким образом, многое было известно об измерениях, ошибках и уравнениях.

Нормальные распределения потеряли свою исключительную позицию к началу XX-го века, в результате применений статистических методов для обработки результатов (в основном) биологических исследований. Встречающие там распределения часто обладали значительной асимметрией и другими отклонениями от нормальности

Карлом Пирсоном была предложена система непрерывных распределений, состоящая из 12 типов распределений (и нормального распределения), с помощью которых можно проводить сглаживание эмпирических данных. В настоящее время для системы К. Пирсона существуют аналоги дискретных распределений.

В первых десятилетиях XX-го века нормальное распределение восстановило свою значимость в результате глубоких работ Р. Фишера, который показал, что на основе предположений о нормальности распределений, можно делать выводы самой широкой практической полезности.

Однако после выхода замечательной книги Р. Фишера "Статистические методы для исследователей" (1925 г.) Эгоном Пирсоном (сыном Карла Пирсона) были высказаны критические замечания о правомерности предположения нормальности в статистическом анализе данных. По мнению Э. Пирсона большое число тестов в книге Р. Фишера осно-

ваны на предположении о нормальности распределений популяций, из которых извлекаются выборки. Не рассматривается вопрос точности тестов, когда распределения популяций отклоняются от нормального, нет четкого указания о необходимой осторожности в применении тестов в такой ситуации.

Отвечая на критику Э. Пирсона, Р. Фишер отстаивал свою точку зрения, исходя из статистических данных, полученных в результате опытов в области селекции сельскохозяйственных растений. Биологи проверяют адекватность своих методов с помощью контрольных экспериментов. Таким образом, по мнению Фишера, предположения нормальности популяций проверяется с помощью опытов, а не теории.

Следует заметить, что ко времени этой дискуссии уже были известны некоторые последствия нарушения предположения о нормальности распределений. Такого рода нарушения оказывают незначительное влияние на выводы о средних значениях, но опасны для выводов о дисперсиях.

К середине XX века дальнейшее применение статистических методов в биологических, медицинских, социологических и экономических исследованиях привело к большому разнообразию распределений, встречающихся в этих науках. Помимо нормальных, использовались распределения имеющие «тяжелые» хвосты, а иногда и явно выраженную асимметричность.

Это было связано с теми обстоятельствами, что во многих задачах этих наук, наличие «механизма» центральной предельной теоремы было весьма проблематичным и, в отличие от физических наук, слабее выражена воспроизводимость результатов экспериментов выполненных, казалось бы, в одинаковых условиях. Поэтому основным источником появления случайности (помимо ошибок измерений) стало *влияние неучтенных факторов,* которые интерпретировались как случайные.

Сложившееся положение вещей привело к необходимости разработки робастных методов анализа данных, а также методов, не использующих предположения о нормальности распределений, например, непараметрических методов статистики [2].

Следует сказать, что в последние десятилетия указанные выше ненормальные устойчивые распределения широко используются в моделях экономики, финансовой математики и биологии [6, 7]. Кроме того, необходимо отметить удачное применение устойчивого ненормального распределения Леви в теории лазерного охлаждения (К. Коэн – Таннуджи, Нобелевская премия по физике 1997 г.). В этой теории использовалась предельная теорема Леви - Гнеденко о сходимости к устойчивым распределениям [8].

Дж. Тьюки: Гауссовское распределение используется, прежде всего, для следующих целей: а) как стандарт для сравнения, относительно которого оценивается истинное поведение реальных данных с целью выявления и анализа отклонений; б) часто (но с осторожностью) как грубую аппроксимацию действительного распределения самих данных и величин, вычисляемых на их основе. Используя гауссовскую форму распределения в качестве аппроксимации, нужно иметь в виду, что распределение реальных данных, как правило, во многих отношениях от нее отличается, и поэтому анализ нормального приближения — это только начало анализа [1].

1. Проверка нормальности выборочного распределения с помощью критериев согласия

Иногда встречается неправильное использование критериев Колмогорова, омега-квадрат и хи-квадрат Фишера для проверки нормальности выборочного распределения. Напомним, что с помощью критерия Колмогорова проверяют гипотезу о том, что выборка извлечена из генеральной совокупности с *известной и полностью определенной непрерывной* функцией распределения. Проверяя нормальность распределения, мы можем не знать точные значения математического ожидания и дисперсии. Известно, что при замене неизвестных значений этих параметров их выборочными оценками*, гипотеза нормальности принимается чаще, чем следовало бы*.

Кроме того, в этом случае для надежной проверки нормальности требуются выборки большого объёма (несколько сотен наблюдений) [2, 3]. Подобные проблемы возникают при использовании критериев омега-квадрат и хи-квадрат Фишера. Для выборок такого большого объёма трудно гарантировать однородность наблюдений.

Рекомендации: См. [4], а также уточнение Лиллиефорса (Lilliefors H.) критерия Колмогорова. Критерии нормальности Колмогорова - Лиллиефорса и критерий согласия Шапиро – Уилка (Shapiro – Wilk`s test) реализованы в системах **STATISTICA** и **R**.

2. Последствия нарушения предположения о нормальности

Распределения t-статистики Стьюдента и F-статистики Фишера относятся к случаю, когда наблюдаемые значения имеют нормальное распределение, а корреляция между наблюдениями тождественно равна нулю. Если распределение наблюдений не является нормальным, то распределение статистик t и F отличается от описанных выше, в особенности для F.

Сравнение средних значений двух выборок

Наиболее распространенный тест для сравнения средних двух выборок, имеющих *равные* дисперсии, основан на статистике Стьюдента *t*. В этом случае все наблюдения должны быть независимыми и при нулевой гипотезе должны иметь одинаковые нормальные распределения.

В том случае, когда распределения *не являются нормальным*, уровень значимости теста *t* является *почти точным* для выборок объёма больше 12.

Если дисперсии двух выборок *различны,* то тест Стьюдента *t* не будет давать точные значения уровней значимости даже для нормальных распределений (проблема Беренса – Фишера, для которой на сегодняшний день нет точного решения, но есть приближенные).

Сравнение дисперсий двух выборок

Традиционная проверка равенства дисперсий двух независимых нормальных выборок основано на *F*-статистике Фишера. Критерий Фишера, основанный на статистике *F* очень чувствителен к отклонениям от нормальности.

Более подробное изложение материала параграфа 2 можно найти в справочнике [5], монографии [11] и энциклопедиях [10, 12].

3. Распределение выборочного коэффициента корреляции r

Обозначим через **ρ** коэффициент корреляции пары случайных величин X и Y, а через ***r*** коэффициент корреляции Пирсона, построенный по двумерной выборке наблюдений над этими величинами. Для случайных величин X и Y имеющих двумерное нормальное распределение при $\rho \neq 0$ функция распределения и плотность коэффициента корреляции Пирсона ***r*** не выражается через элементарные функции, но может быть представлена с помощью гипергеометрической функции. При $\rho = 0$ известны представления плотности коэффициента корреляции Пирсона ***r*** с помощью элементарных функций.

В случае $\rho \neq 0$ *для* больших выборок, (когда объём n стремиться к бесконечности) имеет место асимптотическая нормальность распределения коэффициента корреляции Пирсона. Однако сходимость распределения коэффициента ***r*** к нормальному распределению является слишком медленной. Не рекомендуется пользоваться нормальным приближением при $n < 500$.

В этом случае обычно используют преобразование Фишера коэффициента ***r,*** приводящее к величине ***z***, распределение которой значительно ближе к нормальному. С помощью этого распределения можно найти приближенные доверительные интервалы для коэффициента ***r.***

Изучение проблемы чувствительности к отклонениям от нормальности распределения коэффициента ***r*** к настоящему времени ещё нельзя считать завершенным.

Одна из причин: распределения ***r*** для ненормальных выборок детально разработаны для относительно небольшого числа частных случаев. Существуют примеры как повышенной, так и незначительной чувствительности к отклонениям от нормальности распределения коэффициента ***r*** [11].

Литература

1. Тьюки Дж. Анализ результатов наблюдений. / Дж. Тьюки, пер. с англ. – М.: МИР, 1981, 694 с. (J.W. Tukey Exploratory data analysis, Pearson. – 1977).
2. Лагутин М.Б. Наглядная математическая статистика. – М.: БИНОМ, 2007, 472 с.
3. Орлов А.И. Прикладная статистика. – М.: ЭКЗАМЕН, 2004, 1069 с.
4. Тюрин Ю.Н. / Статистический анализ данных на компьютере. / Тюрин Ю.Н., Макаров А.А., – М.: ИНФРА – М, 1998, 528 с.
5. Кобзарь А.И. Прикладная математическая статистика. – М.: ФМ, 2006, 814 с.
6. Ширяев А.Н. Основы стохастической финансовой математики, т. 1. – М.: ФАЗИС, 1998, 489 с.
7. Прохоров Ю.В. (гл. ред.) Вероятность и математическая статистика. Энциклопедия. – М.: БРЭ, 1999, 910 с.
8. Bardou F. Levy statistics and laser cooling. / Bardou F., Bouchaud J., Aspect A., Cohen-Tannoudji C. /. – Cambridge university press, 2002. 198 p.
9. Lemann E. On the history and use of some standart statistical models. Probability and Statistics: Essays in Honor of D. Freedman. V. 2 (2008). 421 p.
10. Good P. / Common errors in statistics, and how to avoid them. / Good P., Hardin J., Wiley, 2003.). 222 p.
11. Johnson N. Continuous univariate distribution. Vol. 2. / Johnson N., Kotz S., Balakrishnan N. – Wiley, 1995. 717 p.
12. Kotz S. Encyclopedia of statistical science. / Kotz S., (et al.) // 16 volumes. 2ed, Wiley, 2005.