Материалы региональной конференции (с участием иностранных ученых) "Вероятностные идеи в науке и философии". – Новосибирск: Ин-т философии и права СО РАН / НГУ. 2003. – С. 102-105.

Исследование распределений статистик, используемых при проверке гипотез о математическом ожидании и дисперсии, в случае принадлежности наблюдаемых величин экспоненциальному семейству распределений[1]

Лемешко Б.Ю., Помадин С.С.
г. Новосибирск, ser@fpm.ami.nstu.ru

В классической статистике при выводе предельных распределений статистик, используемых при проверке гипотез о математическом ожидании и дисперсии, в качестве основного предположения рассматривается принадлежность наблюдаемой выборки нормальному закону. На практике нормальный закон далеко не всегда является наилучшей моделью для описания реально наблюдаемых случайных величин. Что произойдет с распределениями данных статистик, если наблюдаемый закон в той или иной степени отличается от нормального? Будут ли корректны статистические выводы, базирующиеся на классических результатах, если нарушено предположение о нормальности?

В большинстве случаев отсутствие необходимых теоретических результатов объясняется сложностью и трудоемкостью получения решений аналитическими методами. Поэтому возникает необходимость развития компьютерных методов исследования статистических закономерностей, компьютерных методов исследования статистик различных критериев проверки статистических гипотез, построения вероятностных моделей для исследуемых закономерностей. Это позволяет с меньшими интеллектуальными затратами получать фундаментальные знания, и, следовательно, осуществлять корректные статистические выводы при анализе данных в различных прикладных областях. Поэтому проводимые исследования были основаны на методике компьютерного анализа статистических закономерностей.

В [1] было показано, что в многомерном случае распределения статистик, используемых при проверке гипотез о векторе математических ожиданий, оказались устойчивыми к отклонениям наблюдаемого закона от нормального. В данной работе продолжены и расширены исследования, начатые в [2] относительно статистик одномерного анализа.

В качестве наблюдаемого закона, отличного от нормального, использовалось экспоненциальное семейство распределений, общий вид функции плотности которого определяется выражением:

$$f(x;\theta_0,\theta_1,\lambda)=\frac{\lambda}{2\sqrt{2}\theta_1\Gamma\left(1/\lambda\right)}\exp\left(-\left(\frac{|x-\theta_0|}{\sqrt{2}\theta_1}\right)^{\lambda}\right),$$

где: θ_0 – параметр сдвига, θ_1 – параметр масштаба, λ – параметр формы.

[1] Работа выполнена при поддержке Минобразования РФ (проект № ТО2-3.3-3356)

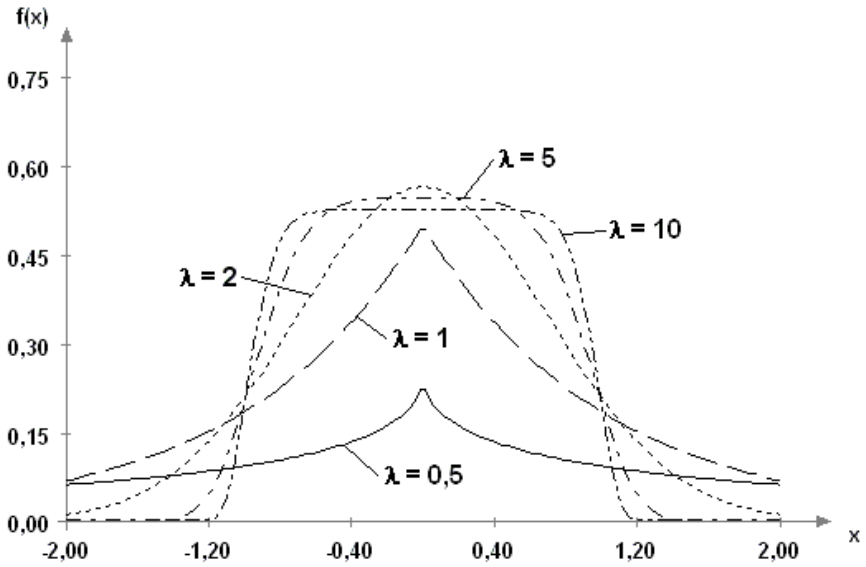

Рис 1. Функции плотности двухстороннего экспоненциального распределения при различных параметрах формы

Выбор экспоненциального семейства обусловлен тем, что оно представляет собой целый класс симметричных распределений (частными случаями являются нормальный закон при параметре формы, равном 2, и распределение Лапласа при параметре формы 1). Примечательно и то, что предельными случаями являются распределение Коши (параметр формы стремится к нулю) и равномерное распределение (параметр формы стремится к бесконечности). На рис. 1 проиллюстрировано, как изменяется функция плотности у класса экспоненциальных распределений при изменении параметра формы от 0,5 до 10.

В классической математической статистике получены предельные распределения статистик, используемых при проверке гипотез о математическом ожидании и дисперсии, при условии наблюдения случайных величин $\xi_i \in N(\mu_{ист}, \sigma^2_{ист}),\ i = \overline{1,n}$.

Если проверяется гипотеза вида $H_0 : \mu = \mu_0$ и известна дисперсия, то вычисляется статистика $T_1 = \frac{1}{n}\sum_{i=1}^{n}\xi_i$, предельным распределением которой при истинности гипотезы H_0 является нормальное $G(T_1 \mid H_0) = N\left(\mu_0, \frac{\sigma^2_{ист}}{n}\right)$.

Если проверятся та же гипотеза, но при неизвестной дисперсии, тогда вычисляется статистика $T_2 = \frac{\bar{\xi} - \mu_0}{\bar{\sigma}}\sqrt{n}$, где $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ и $\bar{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^2$, которая подчиняется $G(T_2 \mid H_0) = t_{n-1}$– распределению Стьюдента.

Когда проверяемая гипотеза имеет вид $H_0 : \sigma = \sigma_0$ и математическое ожидание известно, то вычисляется статистика $T_3 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(\xi_i - \mu_{ист}\right)^2$, предельным распределением которой будет $G(T_3 \mid H_0) = \chi^2_n$-распределение. В случае неизвестного математического ожидания и проверке этой же гипотезы вычисляется статистика $T_4 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(\xi_i - \bar{\xi}\right)^2$, которая в пределе подчиняется $G(T_4 \mid H_0) = \chi^2_{n-1}$-распределению.

Для подтверждения работоспособности применяемых далее методов было проведено исследование эмпирических распределений статистик, используемых для проверки гипотез о математическом ожидании и дисперсии, при наблюдении нормального закона. Полученные результаты показали близость получаемых эмпирических распределений статистик, в данном случае, известным предельным законам. Соответствие в такой ситуации эмпирических распределений, получаемых в процессе моделирования,

предельным классическим распределениям статистик служат доводом, подчеркивающим достоверность результатов в случае наблюдения экспоненциального семейства распределений.

Далее будем рассматривать выборку случайных величин $\xi_i \in f(x;\theta_0,\theta_1,\lambda),\ i=\overline{1,n}$. В общем случае, когда параметр формы не равен 2, предельные распределения статистик T_1, T_2, T_3, T_4, используемых при проверке гипотез о математическом ожидании и дисперсии, неизвестны.

Результаты компьютерного моделирования выборок статистик T_1 и T_2 в случае принадлежности наблюдаемых величин экспоненциальному семейству распределений (параметр формы менялся в диапазоне от 1 до 10) показали, что значимого изменения предельных распределений статистик T_1 и T_2, используемых в критериях проверки гипотез о значениях математического ожидания (при известной и неизвестной дисперсии), не происходит.

На рис. 2 в качестве примера представлены графики теоретических (соответствующих классическому случаю) и полученных эмпирических функций распределений, используемых при проверке гипотезы о равенстве математического ожидания нулевому значению при известной дисперсии. Использованы следующие обозначения: S_λ – статистика T_1, смоделированная по экспоненциальному семейству с параметром формы λ; σ_λ – известная дисперсия экспоненциального семейства распределений с параметр формы λ. Визуальная близость распределений статистики, построенной по экспоненциальному семейству, к предельному (классическому) распределению, полученному для нормального закона, позволяет отметить, что значимого изменения распределений статистик не произошло. Это же подтверждает и применение критериев согласия для проверки значимости отклонений смоделированных эмпирических распределений статистики T_1 от классических предельных (нормальных). Для статистики T_2, используемой при проверке гипотезы $H_0:\mu=\mu_0$ в случае неизвестной дисперсии, наблюдается подобная же картина. Данные результаты позволяют утверждать, что в случае отклонений наблюдаемого закона от нормального (при сохранении симметричности), использование классических предельных распределений статистик для статистик T_1 и T_2 не нарушает корректности выводов статистического анализа при проверке гипотез вида $H_0:\mu=\mu_0$.

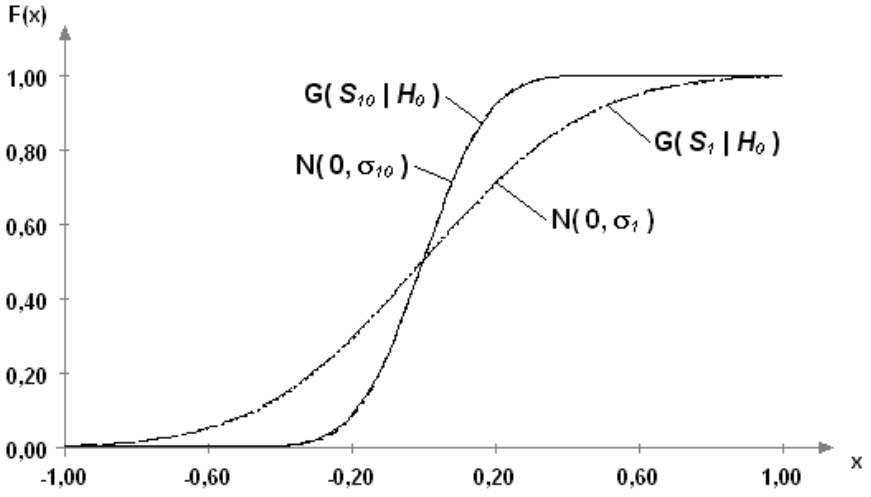

Рис 2. Эмпирические и теоретические функции распределения при проверке гипотезы $H_0:\mu=0$ дисперсия известна

Напротив, распределения статистик T_3 и T_4, используемых в критериях проверки гипотез о дисперсии, как в случае известного математического ожидания, так и в случае неизвестного очень чувствительны к наблюдаемому закону распределения.

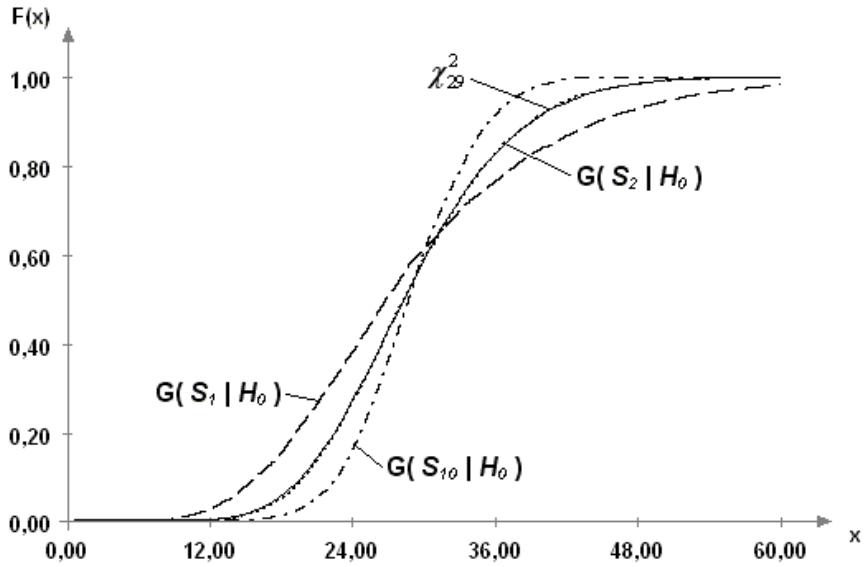

Рис 3. Теоретическая и эмпирические функции распределения при проверке гипотезы $H_0 : \sigma = \sigma_0$ математическое ожидание неизвестно, $n = 30$

Об этом свидетельствует рис. 3, на котором изображены графики эмпирических функций распределений статистик, смоделированных при выборках наблюдаемых случайных величин объемом $n = 30$, принадлежащих экспоненциальному семейству с параметрами формы равным 1, 2 (нормальный закон) и 10. На рисунке представлен также предельный закон распределения статистики T_4 в случае нормального закона (χ^2_{29}-распределение). В данном случае S_λ – статистика T_4, смоделированная при наблюдении случайных величин, принадлежащих экспоненциальному семейству с параметром формы λ.

Из рисунка видно, что распределения статистики T_4, смоделированные при выборках случайных величин, принадлежащих экспоненциальному семейству с параметром формы не равным 2, существенно отличаются от предельного распределения, полученного для нормального закона. Аналогичная картина наблюдается и для статистики T_3 в случае проверки гипотез о дисперсии при известном математическом ожидании. Полученные результаты говорят о том, что распределения статистик, используемых при проверке гипотез о дисперсии (математическое ожидание известно или неизвестно), значимо отличаются от классических предельных при отклонениях наблюдаемого закона от нормального. Поэтому перед проверкой гипотез о дисперсии настоятельно рекомендуется убедиться в том, что наблюдаемый закон является нормальным.

Заключение. Таким образом, проведенные исследования показали, с одной стороны, высокую устойчивость критериев проверки гипотез о математических ожиданиях наблюдаемых величин к отклонениям от нормальности. А, с другой стороны, - неустойчивость критериев, используемых при проверке гипотез о дисперсиях. В то же время исследования подтвердили надежность развиваемой методики исследований и возможность построения моделей предельных распределений для статистик T_3 и T_4 при произвольных наблюдаемых законах случайных величин, что актуально для различных приложений задач статистического анализа данных.

1. Лемешко Б.Ю., Помадин С.С. Корреляционный анализ наблюдений многомерных случайных величин при нарушении предположений о нормальности // Сибирский журнал индустриальной математики. 2002. - Т.5. - № 3. - С.115-130.
2. Лемешко Б.Ю., Ванюкевич О.Н. Проверка гипотез о дисперсии при нарушении предположений о нормальности // Сб. научных трудов НГТУ. – 2002. – № 3(29). – С.27-32.