УДК 519.22

Проблемы интерпретации результатов статистической обработки данных при выполнении лабораторных работ

А. М. Ревякин[1], *И. В. Бардушкина*[2], *А.М. Терещенко*[3]

На примере трех лабораторных работ по курсу теории вероятностей и математической статистики с экономическим содержанием, которые выполняются студентами 2 курса в Национальном исследовательском университете «Московский институт электронной техники» с применением пакетов прикладных программ, проиллюстрированы ошибки, которые могут возникнуть при интерпретации результатов статистической обработки данных.

Ключевые слова: лабораторные работы по математической статистике, статистическая обработка данных, пакеты прикладных программ.

Problems of interpretation of the results of statistical data processing during laboratory work

A.M.Revyakin, I.V.Bardushkina, A.M.Tereshchenko

Using the example of three laboratory works on the course of probability theory and mathematical statistics with economic content, which are performed by 2nd-year students at the National Research University "Moscow Institute of Electronic Technology" using application software packages, the errors that may arise when interpreting the results of statistical data processing are illustrated.

Keywords: laboratory work on mathematical statistics, statistical data processing, application software packages.

Статистические методы широко проникают в нашу повседневную жизнь. Разработано огромное количество специализированных и не специ-

[1]**Ревякин Александр Михайлович,** канд. физ.-мат. наук, доц., доцент института физики и прикладной математики, ФГАОУ ВО «Национальный исследовательский университет «Московский электронной техники», e-mail: arevyakin@mail.ru

[2] **Бардушкина Ирина Вячеславовна**, канд. физ.-мат. наук, доц., доцент института физики и прикладной математики, ФГАОУ ВО «Национальный исследовательский университет «Московский институт электронной техники», e-mail: i_v_bars@mail.ru

[3] **Терещенко Анатолий Михайлович**, доктор тех. наук, доц., профессор института физики и прикладной математики, ФГАОУ ВО «Национальный исследовательский университет «Московский институт электронной техники», e-mail: 4443655@mail.ru

ализированных пакетов прикладных программ [1–3], позволяющих проводить статистическую обработку данных даже тем, кто не обладает знаниями по теории вероятностей и математической статистике.

Основная задача математической статистики – по выборке, максимально используя содержащуюся в ней информацию, сделать то или иное научно обоснованное заключение о генеральной совокупности X. Для надёжности этого заключения выборка должна быть репрезентативной. Выборка репрезентативна, если её объём достаточно велик, а её значения независимы, т.е. получены при независимых измерениях X в одних и тех же условиях.

Однако на практике в силу разных причин получить репрезентативную выборку не удается. Тогда на генеральную совокупность накладывают дополнительное условие, что она имеет нормальное распределение. Проверить это практически невозможно. Чаще легче опровергнуть это условие. Например, используя маломощные критерии: проверки гипотез либо о равенстве среднего, медианы и моды, либо о равенстве нулю коэффициентов асимметрии и эксцесса, либо выполнения правила трех сигм, либо используя вероятностную бумагу или критерий согласия хи-квадрат. Мощных критериев в курсе теории вероятностей и математической статистике в вузах не изучают. Более того, многие показатели, такие как доходы фирм, зарплата трудящихся, надои молока, урожайность и другие, заведомо не подчиняются нормальному закону распределения [4–6]. Часто мы не можем измерить саму генеральную совокупность, а наблюдаем какую-нибудь функцию от нее (любая функция кроме линейной портит нормальное распределение).

Пример. В отделении больницы средняя температура больных 36.6, но живых не осталось. Это говорит о том, что выборка не однородна (получена в результате наблюдения нескольких генеральных совокупностей).

Другой пример неоднородности, приводящий к неправильному выводу, можно найти в [4]. Содержание гемоглобина в крови и размеры кровяных шариков не показывают корреляции ни у новорожденных, ни у мужчин, ни у женщин. Значения коэффициента корреляции равны соответственно 0.06, –0.03 и 0.07. Если статистический материал объединить, то коэффициент корреляции получается равным 0.75.

В [5] на большом количестве примеров продемонстрировано неправильное оценивание средних скорости надоев молока, урожайности, размера яиц и т.п. Для их оценки предлагается использовать показатели средней гармонической, средней квадратичной, средней кубической или средней геометрической. Заметим, что эти оценки легко получаются, если исходные выборки преобразовать с помощью монотонных функций.

Приведем примеры лабораторных работ [7–9], при выполнении которой студенты учатся правильно пользоваться готовыми программами для статистической обработки данных.

Лабораторная работа 1. Моделирования работы банка.

1. Сгенерировать под именем *V* массив объема 66 из нормальной генеральной совокупности с математическим ожиданием 3000 и среднеквадратическим отклонением 1000. Пусть *A* это массив 1/*V*, *B* – 10000*A*, а *T* – *B*/10000. Найти средние и медианы массивов *V*, *T*, *B*.

2. В результате анализа деятельности 66 пунктов продажи валюты Вашего коммерческого банка установлено время реализации 1 у.е. в долях рабочего дня. Определить время необходимое банку для реализации 6 600 000 у.е., если все пункты получат валюту поровну. Оценить степень достоверности результата с помощью доверительных интервалов. Проверить гипотезу о том, что для этого хватит 33 рабочих дней.

Указание: для моделирования работы банка взять подготовленный массив *T*.

3. Проверить гипотезу о нормальности выборок *T* и *V* с помощью критерия согласия хи-квадрат. Для выборки не согласующейся с нормальным распределением подобрать функциональное монотонное преобразование, нормализующую выборку (близость к нормальному закону проверять визуально с помощью гистограмм и оценок коэффициенты асимметрии и эксцесса).

Лабораторная работа 2. Моделирование деятельности страховой кампании.

При наступлении страхового случая страховая компания выплачивает страховую премию в размере 500 у.е. Клиент страховой кампании выкупает страховой полис на год страхования.

Вид страхования	Вероятность страхового случая
1	0.0178409
2	0.0089262
3	0.0063148
4	0.0025136
5	0.0025136
6	0.0009505
7	0.0004283

По каждому виду страхования произвести расчет стоимости страхового полиса в предположении, что договоры страхования заключат по тысяче человек и сбор компании должен превышать ожидаемую сумму выплат на 20%.

Как измениться цена полиса, если для расчета используется наиболее вероятная сумма выплат и ее превышение на 20%?

Определить размер сумм выплат страховых премий, гарантируемых с вероятностями 0.8, 0.9, 0.95. Результаты представить в виде заполненной таблицы в виде:

Вид	Цена полиса	Ожид. выплата	Гарантируемые выплаты с вероятностью		
			P<0.8	P<0.9	P<0.95
1					
2					
3					
4					
5					
6					
7					
Итого доход:					

Оценить вероятность разорения страховой кампании для первого вида страхования. При возникновении проблем с определением вероятностей биномиального распределения для тысячи участников страхования вспомните о теореме Пуассона, сравните биномиальное распределение (при n=250) и соответствующее пуассоновское распределение, сделайте выводы.

Лабораторная работа 3. Исследование логарифмически-нормального распределения, его применение в экономической статистике.

1. Построить графики плотности логарифмически-нормального распределения с параметрами $\text{mean} = 2000$ и $\text{STD} = 100\,000$; 200 000 и 300 000. Оценить для каждого случая: моды, медианы, а также вероятности попадания в области: ниже моды, ниже медианы, ниже mean.

Вопрос: верно ли, что с увеличением параметра STD арифметическое среднее смещается вправо от медианы, а мода – влево?

2. Сгенерировать выборки *X, Y* и Z объёма 200, 400 и 600 из генеральных совокупностей с указанными параметрами mean и STD. Провести сравнение выборок с теоретическим законом распределения. Найти по выборке точечные оценки среднего (average, median, mode, geometric mean), меры рассеяния (std, range, interquartile range) и меры формы (skewness, kurtosis).

Согласуются ли полученные Вами оценки с ответом на вопрос п.1? Объясните причину того, что полученные оценки mode и median противоречат замеченной Вами тенденции в п.1.

3. Прологарифмировать выборки *X, Y* и Z. Проверить нормальность распределения полученных массивов (по коэффициентам асимметрии и эксцесса, гистограммам, на вероятностной бумаге и критерию согласия хи-квадрат).

Верно ли, что оценку параметра среднего логарифмически-нормального распределения X можно получить по оценке среднего выборки X_1 по формуле exp(mean X_1), где X_1=log X?

4. Построить 95% доверительный интервал для среднего геометрического генеральной совокупности *X*, используя доверительный интервал для математического ожидания нормально распределенной выборки X_1. С помощью операции обратной к логарифмированию найти искомые границы.

Исследовательская часть.

5. Известно, что месячная заработная плата служащих фирмы подчиняется логарифмически-нормальному распределению. Фирма состоит из 6 отделов численностью соответственно 200, 400, 150, 180, 220 и 310 сотрудников. Средняя зарплата (средняя геометрическая) сотрудников первого отдела 183 569 у.е. (в сумме с одним среднеквадратическим отклонением это составляет 317 760 у.е.); при 95% доверительном интервале (170 020; 198 166). Средняя месячная заработная плата сотрудников второго отдела 198 278 у.е. при 95% доверительном интервале для математического ожидания, построенного в предположении о нормальном законе распределения, равен (175 757; 220 800). Параметры распределения (mean и STD) месячной зарплаты сотрудников других отделов соответственно равны: 180 000, 100 000 (третий отдел); 210 000, 300 000 (четвёртый отдел); 200 000, 350 000 (пятый отдел) и 150 000, 200 000 (шестой отдел). Массивы зарплат сотрудников этих отделов либо заданы, либо предлагается сгенерировать самостоятельно.

6. Проверить гипотезы о равенстве средних зарплат различных отделов, проводя попарное сравнение средних. Ранжировать отделы по уровню заработной платы их сотрудников (от большего к меньшему).

7. Полагая уровень бедности равным 125 000 у.е., определить для каждого отдела число сотрудников, получающих зарплату ниже этого уровня. Ранжировать отделы по проценту низкооплачиваемых сотрудников (от меньшего к большему). Сравнить полученные результаты. Объяснить замеченные противоречия.

8. Выполнить п.6 для средних геометрических. Сравнить с результатами п.7. Дать объяснение.

9. Полагая признаком принадлежности к среднему классу получение зарплаты в диапазоне от 125 000 до 250 000 у.е., определить этот показатель для каждого отдела.

10. Оценить среднюю и среднюю геометрическую зарплату в фирме. Верно ли, что более 50% служащих живут ниже черты бедности (уровень значимости 0.05)?

11. Выполнить исследовательскую часть методами дисперсионного анализа. Сравнить полученные результаты.

Список литературы

1. Теория и практика статистических исследований / *Под ред. А.М. Ревякина* и *В.В. Костылёва*. – М.: МГАДА, 2007. – 354 с.

2. Теоретико-вероятностные и статистические методы и модели анализа внешнеэкономической деятельности предприятий / *И. Н. Абанина, В. В.*

Бардушкин, Э. А. Вуколов [и др.]; под общ. ред. *И. Н. Абаниной, А. М. Ревякина*. М.: МГАДА, 2014. 214 с.
3. Задания для выполнения лабораторных и индивидуальных работ по курсу «Теория вероятностей и математическая статистика» с использованием пакета MATLAB / *В. В. Бардушкин, И. В. Бардушкина, В. В. Лесин, А. М. Ревякин* // Проектирование инженерных и научных приложений в среде MATLAB: мат-лы V Междунар. науч. конф. (г. Харьков, 11—13 мая 2011 г.) / Сост. В. В. Замаруев. Харьков: БЭТ, 2011. С. 471—533.
4. *Закс Л.* Статистическое оценивание. – М.: Статистика, 1976. – 598 с.
5. *Лакин Г.Ф.* Биометрия: Учеб. пособие для биол. спец. вузов. – М.: Высш. шк., 1990. – 352 с.
6. *Ревякин А. М., Бардушкина И.В.* Об особенностях выполнения курсовой работы по статистике с применением электронного компонента // Экономические и социально-гуманитарные исследования. – М.: 2017. № 1 (13). – С. 112 – 122.
7. *Ревякин А.М., Бардушкина И.В., Бардушкин В.В.* Сборник задач для самостоятельной работы студентов по курсу «Статистика»: учеб. пособие. – М.: МИЭТ, 2016. – 160 с.
8. *Бардушкин В.В., Ревякин А.М., Бардушкина И.В.* Теория вероятностей и математическая статистика. Часть 1: Теория вероятностей: учеб. пособие. – М.: МИЭТ, 2017. – 180 с.
9. *Ревякин А.М., Бардушкин В.В., Бардушкина И.В.* Теория вероятностей и математическая статистика. Часть 2: Математическая статистика: учеб. пособие. – М.: МИЭТ, 2017. – 224 с.