Министерство образования и науки Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
—
Институт компьютерных наук и технологий
Кафедра «Информационная безопасность компьютерных систем»

ЛАБОРАТОРНАЯ РАБОТА № 2

по дисциплине «Теория вероятностей и математическая статистика»

Выполнила
студентк гр. 23508/4 Проценко М.А.
<*подпись*>
Проверила ассистент Лаврова Д.С.
<*подпись*>

Санкт-Петербург
2016

Формулировка задания 1

1. Используя таблицу значений функции$f(x)$, записать 100 цифр, выбирая из каждого значения функции второй знак справа(указать точность округления). С помощью критерия хи-квадрат проверить для такой выборки гипотезу о случайности цифр 0, 1, …, 9. Уровень значимости положить равным: а) 0,05; б) 0,01. Функцию $f(x)$брать в соответствии с вариантом.

Решение:

1. Была взята функция lnx в соответствии с вариантом.
2. Гипотеза H0: вторая цифра справа случайна.
3. Закон редких событий – распределение Пуассона: $P_{(S=k)} = \frac{a^k}{k!} e^{-a}$,
4. Рассчитали количество встречаемости каждой цифры (0, 1,...,9)
5. Вычислили параметр a, который должен получиться равен среднему значению.
6. Вычислим Pi по формуле: $P_{(S=k)} = \frac{a^k}{k!} e^{-a}$;
7. Затем вычислили n*Pi
8. Далее посчитали (Ni - N*pi)^2
9. Получившиеся значения разделим на N*pi и сложим. Это и будет критерий хи-квадрат. У нас он получился равным 92.5 (наблюдаемое значение). Это больше чем 15.5 (хи-квадрат критическое при a=0.01) и больше чем 20.1 (хи-квадрат наблюдаемое при a=0.05)
10. Таким образом наша гипотеза не подтвердилась, отсюда делаем вывод, что вторая цифра в числе не случайна.

Формулировка задания 2

Взять три текста из разных областей знаний, объемом 2000 знаков (включая пробелы) каждый. На основе анализа частот встречаемости букв проверить гипотезу об однородности этих текстов.

Решение:

Были взяты три текста из разны областей знаний (текст из курса философии, статья про теннис из википедии, технический текст).

Для оценки однородности текстов был использован критерий Колмогорова-Смирнова.

Однако этим критерием можно сравнить только 2 выборки. Поэтому, так как у нас 3 текста, а следовательно и три выборки, нам придется применить этот критерий 3 раза (т.к. три пары выборок).

Вычисления:

1. Вычисляем относительные частоты f, равные частному от деления частот на объём выборки, для двух имеющихся выборок.
2. Далее определяем модуль разности соответствующих относительных частот для контрольной и экспериментальной выборок.

3. Среди полученных модулей разностей относительных частот выбираем наибольший модуль, который обозначается dmax.
4. Эмпирическое значение критерия λэмп определяется с помощью формулы:

$$\lambda_{эмп} = d_{\max} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

5. Чтобы сделать вывод о схожести по рассматриваемому критерию между двумя текстами, сравним экспериментальное значение критерия с его критическим значением, определяемым по специальной таблице, исходя из уровня значимости . В качестве нулевой гипотезы примем утверждение о том, что сравниваемые тексты незначительно отличаются друг от друга. При этом нулевую гипотезу следует принять в том случае, если наблюдаемое значение критерия не превосходит его критического значения.
6. По таблице определяем критическое значение критерия: λкр(0,05)=1,36.
7. Таким образом, λэмп<1,36= λкр только в первом и третьем случае. Следовательно, нулевая гипотеза принимается для текстов (1,2) и (2,3), и группы по рассмотренному признаку отличаются не существенно. Текста 1 и 3 отвергают гипотезу о том что тексты однородны и принимают конкурирующую гипотезу о том, что текста неоднородны.

S	T	U
Аэмп1	Аэмп2	Аэмп3
1,309896	1,865389	1,074085

Формулировка задания № 3

Построить критерий отношения правдоподобия для проверки двух простых гипотез о неизвестном параметре распределения из расчетных заданий № 1.3 и 1.4.

3.1) Для показательного распределения (задание 1.3) сформулировать гипотезы:

• для значения a = 0,134;

• для значения a = 3,13.

3.2) Для нормального распределения (задание 1.4) сформулировать гипотезы:

• для математического ожидания (M = 9);

• для дисперсии (σ 2 = 3,13).

Выполнение задания № 3.1

Для выполнения задания 3.1 из результатов пункта 3 первой лабораторной работы были извлечены данные о вариационных рядах выборок:

	A	B
1	0,013311	0,030815
2	0,045817	0,060681
3	0,077253	0,089626
4	0,107653	0,11768
5	0,137051	0,144868
6	0,165481	0,17122
7	0,192974	0,196759
8	0,219562	0,221511
9	0,245274	0,2455
10	0,270138	0,26875
11	0,294184	0,291284
12	0,317437	0,313123
13	0,339924	0,334289
14	0,36167	0,354803
15	0,3827	0,374685

И т.д. Всего по 100 значений в выборках.

Были сформулированы две простые гипотезы:

• Гипотеза H0: элементы выборки имеют показательное распределение с параметром a0 = 3, 13 (т.е. xi ∈ exp(a0), a0 = 3, 13)

• Гипотеза H1: элементы выборки имеют показательное распределение с параметром a1 = 0, 134 (т.е. xi ∈ exp(a1), a1 = 0, 134)

Для проверки сформулированных гипотез, необходимо построить критерий отношения правдоподобия. Сделать это можно, вычислив функцию отношения правдоподобия Λ:

$$\Lambda = \frac{p_1(x_1) \cdot p_1(x_2) \cdot \ldots \cdot p_1(x_n)}{p_0(x_1) \cdot p_0(x_2) \cdot \ldots \cdot p_0(x_n)}$$

Под x_i понимается i-ый элемент выборки, под $p_j(x_i)$ - вероятность случайной величине, распределенной согласно гипотезе H_j, принять значение x_i, n - кол-во элементов в выборке и в нашем случае $n = 100$.

Что же отображает Λ? Зададим некоторое число $c > 0$. Если значение Λ оказывается не больше, чем c, то принимается гипотеза H_0, иначе - H_1. Подставим в вычисление Λ вероятности, которые соответствуют сформулированным нами гипотезам:

$$\Lambda = \frac{a_1^n e^{-\alpha_1 \sum x_i}}{a_0^n e^{-\alpha_0 \sum x_i}}$$

Нам нужно выразить $\sum x_i$ - для этого, очевидно, нам нужно прологарифмировать наше равенство. Подставляя результат в дальнейшем в наше неравенство с параметром c, будем считать, что $c_1 = lnc$. Итак, логарифмируем:

$$\ln\Lambda = n\ln\frac{a_1}{a_0} - (a_1 - a_0)\sum x_i$$

Теперь, как уже и говорилось, подставим $\ln\Lambda$ в неравенство $\ln\Lambda \leq \ln c$ и выразим $\sum x_i$:

$$\sum x_i \leq \frac{c_1 - n\ln\frac{a_1}{a_0}}{a_0 - a_1}$$

Далее нам потребуются некоторые свойства гамма-распределения. Гамма-распределение - двухпараметрическое семейство абсолютно непрерывных распределений. Оно обладает следующими интересующими нас свойствами:

- Экспоненциальное распределение - частный случай гамма-распределения: $\Gamma(1/\theta, 1) = Exp(\theta)$
- χ^2-распределение - частный случай гамма-распределения: $\Gamma(2, \frac{n}{2}) = \chi^2(n)$

В соответствии с этим, получаем следующее:

$$\sum x_i \in \Gamma(a; n), b\sum x_i \in \Gamma(\tfrac{a}{b}; n)$$

$$2a\sum x_i \in \chi^2(2n)$$

Теперь выразим вероятность ошибки первого рода:

$$\alpha = P\left\{\sum x_i > \frac{c_1 - n\ln\frac{a_1}{a_0}}{a_0 - a_1}|H_0\right\}$$

$$\alpha = P\left\{2a_0\sum x_i > 2a_0\frac{c_1 - n\ln\frac{a_1}{a_0}}{a_0 - a_1}|H_0\right\} = 1 - F(2a_0\frac{c_1 - n\ln\frac{a_1}{a_0}}{a_0 - a_1}; 2n)$$

Здесь функция $F(x; 2n)$ - функция $\chi^2(2n)$-распределения. Выходит следующее:

$$\frac{c_1 - n\ln\frac{a_1}{a_0}}{a_0 - a_1} = \frac{X(1-\alpha; 2n)}{2a_0}$$

Здесь, очевидно, $X(1-\alpha; 2n)$ - α-квантиль $\chi^2(2n)$-распределения. Вычислим это отношение при $\alpha = 0,01$:

$$\frac{c_1 - n\ln\frac{a_1}{a_0}}{a_0 - a_1} = \frac{X(1-\alpha; 2n)}{2a_0} = \frac{249{,}445}{2\cdot 2{,}06} \cong 60,545$$

Посчитаем суммы первой и второй выборок. Для выборки 1 сумма равна 71.101, для выборки 2 сумма равна 69.923.

Выводы по заданию № 3.1

Т.к. сумма элементов выборок меньше, чем вычисленное нами отношение (60,545), принимается гипотеза H0. Результаты выполнения данного задания согласуются с результатами, полученными в ходе первой лабораторной работы.

Выполнение задания № 3.2

Для выполнения задания 3.2 из результатов пункта 4 первой лабораторной работы были извлечены данные о вариационном ряде следующей выборки:

C	D
0,006	0,005
0,006	0,008
0,003	0,004
0,008	0,007
0,009	0,004
0,003	0,003
0,004	0,009
0,003	0,005
0,004	0,004
0,005	0,005
0,004	0,003
0,004	0,008
0,005	0,006
0,008	0,004
0,006	0,003

И т.д.

Были сформулированы две простые гипотезы о мат.ожидании:

- Гипотеза H0: элементы выборки имеют нормальное распределение мат. ожиданием a0 = 6 и дисперсией σ 2 = 4, 136
- Гипотеза H1: элементы выборки имеют нормальное распределение мат. ожиданием a0 = 25 и дисперсией σ 2 = 4.136.

Для проверки сформулированных гипотез, необходимо построить критерий отношения правдоподобия. Сделать это можно, вычислив функцию отношения правдоподобия Λ:

$$\Lambda = \frac{e^{-\frac{1}{2\sigma^2}\sum(x_i-a_1)^2}}{e^{-\frac{1}{2\sigma^2}\sum(x_i-a_0)^2}}$$

Нам нужно выразить $\sum x_i$ - для этого, очевидно, нам нужно прологарифмировать наше равенство. Подставляя результат в дальнейшем в наше неравенство с параметром c, будем считать, что $c_1 = lnc$. Итак, логарифмируем:

$$\ln\Lambda = \frac{a_1-a_0}{\sigma^2}\left(\sum x_i - n\frac{a_1+a_0}{2}\right)$$

Теперь, как уже и говорилось, подставим $\ln\Lambda$ в неравенство $\ln\Lambda \leq \ln c$ и выразим $\sum x_i$:

$$\sum x_i \leq \frac{c_1\sigma^2}{a_1-a_0} + n\frac{a_1+a_0}{2}$$

Теперь выразим вероятность ошибки первого рода:

$$\alpha = P\left\{\sum x_i > \frac{c_1\sigma^2}{a_1-a_0} + n\frac{a_1+a_0}{2} | H_0\right\}$$

$$\alpha = P\left\{\sum x_i > \frac{c_1\sigma^2}{a_1-a_0} + n\frac{a_1+a_0}{2} | H_0\right\} = 1 - \Phi\left(\frac{\frac{c_1\sigma^2}{a_1-a_0} + n\frac{a_1+a_0}{2} - na_0}{\sqrt{n\sigma^2}}\right)$$

Здесь функция $\Phi(x)$ - функция стандартного нормального распределения. Выходит следующее:

$$\frac{c_1\sigma^2}{a_1-a_0} + n\frac{a_1+a_0}{2} = Q(1-\alpha)\sqrt{n\sigma^2} + na_0$$

Здесь, очевидно, $Q(\alpha)$ - α-квантиль стандартного нормального распределения. Вычислим это отношение при $\alpha = 0,01$:

$$\frac{c_1\sigma^2}{a_1-a_0} + n\frac{a_1+a_0}{2} = Q(1-\alpha)\sqrt{n\sigma^2} + na_0 =$$

= 647.304

Сумма представленной выборки оказалась равно 0.561.

Используя ту же выборку были сформулированы две простые гипотезы о дисперсии:

• Гипотеза H0: элементы выборки имеют нормальное распределение мат. ожиданием a0 = 6 и дисперсией σ 2 = 4. 136.

• Гипотеза H1: элементы выборки имеют нормальное распределение мат. ожиданием a0 = 6 и дисперсией σ 2 = 25

Для проверки сформулированных гипотез, необходимо построить критерий отношения правдоподобия. Сделать это можно, вычислив функцию отношения правдоподобия Λ:

$$\Lambda = \frac{\left(\frac{1}{2\pi\sigma_1^2}\right)e^{-\frac{1}{2\sigma_1^2}\sum(x_i-a)^2}}{\left(\frac{1}{2\pi\sigma_0^2}\right)e^{-\frac{1}{2\sigma_0^2}\sum(x_i-a)^2}}$$

Нам нужно выразить $\sum x_i$ - для этого, очевидно, нам нужно прологарифмировать наше равенство. Подставляя результат в дальнейшем в наше неравенство с параметром c, будем считать, что $c_1 = lnc$. Итак, логарифмируем:

$$\ln\Lambda = n\ln\frac{\sigma_0^2}{\sigma_1^2} - \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\sum(x_i-a)^2$$

Теперь, как уже и говорилось, подставим $\ln\Lambda$ в неравенство $\ln\Lambda \le \ln c$ и выразим $\sum(x_i-a)^2$:

$$\sum(x_i-a)^2 \le 2\cdot\frac{c_1 - n\ln\frac{\sigma_0^2}{\sigma_1^2}}{\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}}$$

Заметим, что $\frac{x_1-a}{\sigma} \in N(0;1) \Rightarrow \frac{\sum(x_i-a)^2}{\sigma^2} \in \chi^2(n)$. В соответствии с этим выразим теперь вероятность ошибки первого рода:

$$\alpha = P\left\{\sum(x_i-a)^2 > 2\cdot\frac{c_1 - n\ln\frac{\sigma_0^2}{\sigma_1^2}}{\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}}|H_0\right\}$$

$$\alpha = P\left\{\frac{1}{\sigma_0^2}\sum(x_i-a)^2 > \frac{2}{\sigma_0^2}\cdot\frac{c_1 - n\ln\frac{\sigma_0^2}{\sigma_1^2}}{\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}}|H_0\right\} = 1 - F\left(\frac{2}{\sigma_0^2}\cdot\frac{c_1 - n\ln\frac{\sigma_0^2}{\sigma_1^2}}{\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}};n\right)$$

Здесь функция $F(x)$ - функция $\chi^2(n)$-распределения. Выходит следующее:

$$2\cdot\frac{c_1 - n\ln\frac{\sigma_0^2}{\sigma_1^2}}{\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}} = \sigma_0^2 X(1-\alpha;n)$$

Здесь, очевидно, $X(1-\alpha;n)$ - α-квантиль $\chi^2(n)$-распределения. Вычислим это отношение при $\alpha = 0,01$:

$$2\cdot\frac{c_1 - n\ln\frac{\sigma_0^2}{\sigma_1^2}}{\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}} = \sigma_0^2 X(1-\alpha;n) = 3,06\cdot 135,807 \cong 415,569$$

Сумма квадратов (xi − a) равна 0.03.

Выводы по заданию № 3.2

Для выборки 1, имеем следующее: в случае с проверкой гипотез для мат. ожидания, сумма элементов выборки: Pxi = 0.561, а вычисленное ранее отношение, включающее квантиль стандартного нормального распределения: Q(1 − α) √ nσ2 + na0 = 2, 326 · √ 100 · 3, 06 + 100 · 9 ∼= 940, 688. Следовательно, принимаем гипотезу о мат. ожидании H0 при α = 0, 01. В случае проверки гипотез для дисперсии, P(xi − a) = 0.03, в то время как отношение, содержащее α-квантиль χ 2 (n)-распределения: σ 2 0X(1 − α; n) = 3, 06 · 135, 807 ∼= 415, 569. Следовательно, принимаем гипотезу о дисперсии H0 с заданным уровнем ошибки первого рода.