Описательная часть расчетной работы №1 по математической статистике за 6 триместр

Вихляев Егор, ММТ-2

June 27, 2023

1 Задание №1

Проверить гипотезу случайности на 5%-ном уровне значимости с помощью критерия серий или инверсий.

H_0: Результаты наблюдений представляют реализацию независимой повторной выборки.

1. Задаем уровень значимости $\alpha = 0{,}05$ единиц (5%).
2. Воспользуемся критерием инверсий (см. все вычисления в расчетной части).
3. Поскольку, по результатам вычислений с помощью критерия инверсий, имеем неравенство $|T_{\text{набл}}| < T_{\text{кр}}$, делаем вывод, что нулевая гипотеза H_0 принимается. Результаты наблюдений представляют реализацию независимой повторной выборки.

2 Задание №2

Построить гистограмму относительных частот. Определить выборочные характеристики: среднее, дисперсию, моду, медиану, асимметрию и эксцесс.

На основе визуального анализа гистограммы, а также выборочных характеристик, выдвинуть гипотезу о виде закона распределения исследуемого набора данных. Сделать выводы о свойствах гипотетического распределения (наличие симметрии, близость к нормальному распределению, близость среднего к медиане и т.д.).

1. Построить гистограмму относительных частот.
   Относительная частота определяется следующей формулой:
   $$w_i = \frac{n_i}{n},$$
   где n_i – частоты, n – количество вариант выборки.

   (a) Для начала построим вариационный ряд.

   (b) Находим минимальное значение $x_{min} = 0.5$.

   (c) Находим максимальное значение $x_{max} = 4.84$.

   (d) Находим размах вариации $R = x_{max} - x_{min} = 4.84 - 0.5 = 4.34$.

   (e) Находим оптимальное количество интервалов $k = 1 + [log(2, n)]$ (где n – объём выборки) $= 1 + [log(2, 200)] = 8$.

   (f) Находим длину интервала $h = \frac{R}{k} = 0.5425$.

   (g) Далее строим таблицу, солержащую наши интервалы, среднее значение интервалов x_i, частоты n_i, относительные частоты w_i, плотности $\frac{w_i}{h}$.

   (h) На основании данных таблицы, а именно первого столбца с интервалами и столбца с плотностями, строим гистограмму относительных частот.

2. Определить выборочные характеристики: среднее, дисперсию, моду, медиану, асимметрию и эксцесс.

   (a) Среднее (выборочное среднее) – это среднее арифметическое всех значений выборки. В нашем случае, $\overline{x_в} = 1.51$.

   (b) Дисперсия (выборочная дисперсия) – это среднее арифметическое квадратов отклонений всех вариант выборки от её средней. В нашем случае, $S_в^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x_в})^2}{n} = 0.93$.

   (c) Мода M_0 дискретного вариационного ряда – это варианта с максимальной частотой. В нашем случае, $M_0 = 0.76$.

(d) Медиана m_e вариационного ряда – это значение, которое делит его на две равные части (по количеству вариант). В нашем случае, $m_e = 1.21$.

(e) Ассиметрия – характеризует меру скошенности полигона или гистограммы влево / вправо относительно самого высокого участка. В нашем случае, $A_3 = \frac{m_3}{\sigma_{\text{в}}^3} = 1.44 > 0$, то есть, распределение обладает существенной правосторонней асимметрией, что, хорошо видно по гистограмме.

(f) Эксцесс – показатель остроты пика графика распределения. В нашем случае, $E_k = \frac{m_4}{\sigma_{\text{в}}^4} = 1.84 > 0$, то есть, распределение заметно выше, чем нормальное распределение с параметрами $\overline{x_{\text{в}}} = 1.51, \sigma_{\text{в}} = \sqrt{S_{\text{в}}^2} = 0.9643....$

3. На основе визуального анализа гистограммы, а также выборочных характеристик, выдвинуть гипотезу о виде закона распределения исследуемого набора данных.

   На основе визуального аналига гистограммы, можно сделать вывод, что перед нами нечто похожее на показательный закон распределения, а именно на закон F4. Потому выдвинем следущую нулевую гипотезу.

   H_0: Исследуемый набор данных имеет вид закона распределения

$$F4 : f(x) = 2\lambda x e^{-\lambda x^2}, \text{при } x \geq 0, \lambda > 0$$

   .

4. Сделать выводы о свойствах гипотетического распределения (наличие симметрии, близость к нормальному распределению, близость среднего к медиане и т.д.).

   (a) Наличие симметрии.

   Основываясь как на гистограмме гипотетического распределения, так и на его выборочных характеристиках, делаем вывод, что симметрия у распределения отсутствует. Гистограмма явно убывает по направлению оси OX, о чем говорит и ассиметрия A_3, согласно которой гистограмма значительно скошена вправо относительно самого высокого участка. Коэффициент эксцесса E_k это подтверждает.

(b) Близость к нормальному распределению.

На основании гистограммы относительных частот и вычисленных выборочных характеристик, можно сделать вывод, что различие между исходным распределением и нормальным распределением статистически значимо и вряд ли объяснимо случайными факторами.

(c) Близость среднего к медиане.

Среднее $\overline{x_в} = 1.51$, медиана $m_e = 1.21$. Очевидно, разность между средним и медианой не столь велика, а потому они достаточно близки.

3 Задание №3

С помощью метода максимального правдоподобия и метода моментов оценить неизвестные параметры гипотетического распределения. Построить график плотности гипотетического распределения на том же рисунке, что и гистограмма, используя вместо неизвестного параметра его статистическую оценку (ОМП и ОММ).

1. Оценим неизвестные параметры гипотетического распределения методом максимального правдоподобия (ОМП).

   Установим исходные данные. Имеем закон распределения

   $$F4 : f(x) = 2\lambda x e^{-\lambda x^2}, \text{при } x \geq 0, \lambda > 0,$$

   и, следовательно, один неизвестный параметр λ.

   (a) Строим $L(\lambda; x_1, ..., x_n) = \prod_{i=1}^{n} 2\lambda x_i e^{-\lambda x_i^2} I(x_i \geq 0) = (2\lambda)^n e^{-\lambda \sum_{i=1}^n x_i^2} \prod_{i=1}^n x_i I(x_{(1)} \geq 0)$.

   (b) Логарифмируем $L(\lambda; x_1, ..., x_n) = ln(L(\lambda; x_1, ..., x_n)) = ln((2\lambda)^n e^{-\lambda \sum_{i=1}^n x_i^2} \prod_{i=1}^n x_i)$ $ln((2\lambda)^n) + ln(e^{-\lambda \sum_{i=1}^n x_i^2}) + ln(\prod_{i=1}^n x_i) = n * ln(2\lambda) - \lambda \sum_{i=1}^n x_i^2 + ln(\prod_{i=1}^n x_i)$.

(c) Строим уравнение $\frac{\partial lnL(\lambda;x_1,...,x_n)}{\partial\lambda}$, которое решаем относительно параметра λ.

$$\frac{\partial lnL(\lambda;x_1,...,x_n)}{\partial\lambda} = \frac{\partial}{\partial\lambda}(n*ln(2\lambda)-\lambda\sum_{i=1}^{n}x_i^2+ln(\prod_{i=1}^{n}x_i)) = \frac{n}{\lambda}-\sum_{i=1}^{n}x_i^2 = 0$$

$$\widetilde{\lambda_{\text{ОМП}}} = \frac{n}{\sum_{i=1}^{n}x_i^2} = \frac{200}{640.7856} \approx 0.312$$

(d) Проверяем полученное значения на максимум:

$$\frac{\partial^2}{\partial\lambda^2} = -\frac{n}{\lambda^2} < 0, \forall\lambda \Rightarrow$$

$\Rightarrow$ найденное решение максимизирует $L(\lambda;x_1,...,x_n)$ относительно параметров, найден максимум на границе области их изменения.

Таким образом, $\widetilde{\lambda_{\text{ОМП}}} \approx 0.312$.

2. Оценим неизвестные параметры гипотетического распределения методом моментов (ОММ).

Итак, $M[X] = \overline{X}$. Выборочное среднее $\overline{X} = \overline{x_{\text{в}}} = 1.51$.

$$M[X] = \int_0^{+\infty} 2\lambda x^2 e^{-\lambda x^2}dx = 2\lambda\int_0^{+\infty} x^2 e^{-\lambda x^2}dx = \frac{\sqrt{\pi}}{2\sqrt{\lambda}}$$

$$\frac{\sqrt{\pi}}{2\sqrt{\lambda}} = 1.51$$

$$\sqrt{\lambda} = \frac{\sqrt{\pi}}{3.02}$$

$$\widetilde{\lambda_{\text{ОММ}}} = \frac{\pi}{9.1204} \approx 0.344.$$

4 Задание №4

С помощью критерия хи-квадрат проверить гипотезу о виде распределения с уровнем значимости α, приведя все промежуточные расчеты. Вычислить p-значение критерия (реальный уровень значимости критерия).

$H_0 : X \sim 2\lambda x e^{-\lambda x^2}$ – выборка подчиняется показательному закону

Пусть уровень значимости $\alpha = 0.05$, число интервалов $k = 8$, число неизвестных параметров $r = 2$. $\widetilde{m_{\text{ОМП}}} = \overline{x} = 1.51, \widetilde{\sigma^2_{\text{ОМП}}} = S^2_{\text{в}} = 0.93$.

$$\chi^2_{\text{крит}} = x_{1-\alpha}[\chi^2_{k-r-1}] = x_{1-0.05}[\chi^2_{8-2-1}] = x_{0.95}[\chi^2_5] = 11.07.$$

Вычислим $\chi^2_{\text{выб}} = \sum_{i=1}^{k} \frac{(n_i - n*p_i)^2}{n*p_i}$. Прежде всего, ищем вероятности p_i по формуле $p_i = P(X_i \leqslant X \leqslant X_{i+1}) = \Phi(\frac{X_{I+1}-m}{\sigma_x}) - \Phi(\frac{X_I - m}{\sigma_x})$. Для упрощения расчетов, воспользуемся функцией ЭКСП.РАСП в Excel (все расчеты на 4 листе), подставляя в качестве параметра $\lambda = \widetilde{\lambda_{\text{ОМП}}} = 0.312$. Далее ищем значения $n * p_i$ для облегчения поиска итоговой суммы, а затем и элементы $\frac{(n_i - n*p_i)^2}{n*p_i}$ самой суммы.

Таким образом, $\chi^2_{\text{выб}} = 105.12$. Отсюда ясно, что

$$(\chi^2_{\text{выб}} = 105.12) > (\chi^2_{\text{крит}} = 11.07) \Rightarrow$$

$\Rightarrow$ нулевая гипотеза не принимается $\Rightarrow$ наша исходная выборка не подчиняется показательному закону.

5 Задание №5

Разделить набор данных на 2 части и, проверить гипотезу однородности этих частей.

Итак, поделим исходную выборку пополам, получив, таким образом, две выборки X и Y равных размеров. Воспользуемся критерием Манна-Уитни, установим уровень значимости $\alpha = 0.05$ и выдвинем следующую нулевую гипотезу:

$$H_0 : F_x(x) = F_y(y).$$

Получаем, что $n = 100$ и $m = 100$ – объёмы выборок X и Y соответственно. Ищем статистику $T_{\text{кр}}$:

$$T_{\text{кр}} = x_{1-\frac{\alpha}{2}}[N(0;1)] = x_{0.975}[N(0;1)] = 1.96.$$

Чтобы найти $T_{\text{набл}} = \frac{|U - \frac{nm}{2}|}{\sqrt{\frac{nm(n+m+1)}{12}}}$, необходимо прежде найти вспомогательную функцию $U = \sum_{i=1}^{n} \sum_{j=1}^{m} I(x_i < y_j) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} I(x_i = y_j)$. По итогам расчетов (см. расчетную часть), $U = 5457$, тогда:

$$T_{\text{набл}} = \frac{|5457 - \frac{100*100}{2}|}{\sqrt{\frac{100*100(100+100+1)}{12}}} = 1.117.$$

Таким образом, переходим к сравнению найденных статистик:

$$(T_{\text{набл}} = 1.117) < (T_{\text{кр}} = 1.96) \Rightarrow$$

$\Rightarrow H_0$ – принимается $\Rightarrow$ обе части X и Y из исходной выборки однородны.