Домашнее задание 1

Дородный Дмитрий СКБ172

12 декабря 2019 г.

Содержание

1 Проверка гипотез о виде распределения для гамма распределения

1.1 Критерий Колмогорова-Смирнова

1.1.1 Теоретическое введение

Критерий, или, скорее его статистика определяется формулой:
$D_n = D_n(X) = -\infty \overset{sup}{<} x < \infty |\hat{F_n}(x) - F(x)|$

Известно, что относительная частота произвольного события в n независимых испытаниях является оптимальной несмещенной оценкой для вероятности этого события. Отсюда следует, что значение эмпирической функции распределения в каждрй точке является оптимальной несмещенной оценкой для значения в этой точке теоретической функции распределения. Это подтверждает, что статистика критерия подходит для проверки нулевой гипотезы. Так же, из неравенства Чебышева
$P\{|T_n - g| > \epsilon\} \leq \frac{DT_n}{\epsilon^2} \to 0, \forall \epsilon > 0$
следует, что эмпирическая функция распределения является состоятельной оцнкой теоретической функции распределения, а это значит, что с увеличением выборки значение D_n должно быть близко к нулю.
Делать выводы будем на основании теоремы Колмогорова:
$\lim_{n\to\infty} P\sqrt{D_n} \leq Tt = K(t)$
Где $K(t)$ - распределение Колмогорова:
$K(t) = \sum_{j=-\infty}^{\infty}(-1)^j e^{-2j^2t^2}$
Из которой и будем получать критическую границу:
$t_\alpha = \frac{\lambda_\alpha}{\sqrt{n}}, K(\lambda_\alpha) = 1 - \alpha$
Следовательно
$PD_n \in_{1\alpha} |H_0 = P\sqrt{n}D_n \geq \lambda_\alpha | H_0 \approx 1 - K(\lambda_\alpha) = \alpha$ (При достаточно больших выборках)
Таким образом, критерий согласия Колмогорова для $n \geq 20$, при выбранном уровне значимости α, определяющего число $\lambda_\alpha : K(\lambda_\alpha) = 1 - \alpha$:
H_0 отвергается $\iff \sqrt{n}D_n \geq \lambda_\alpha$
Так как распределение статистики D_n свободно от неизвестного распределения выборки и в качестве меры используется максимальное отклонение, какова бы ни была функция распределения, истиная функция распределения будет с вероятностью $1 - \alpha$ лежать в некоторой доверительной области вокруг выборочной функции распределения.

1.1.2 Практические результаты:

$n = 1000$
1) 0.465
2) 1.0554
3) 0.318
4) 0.591
5) 0.176

$n = 100000$
1) 1.063
2) 0.401
3) 0.934
4) 0.836
5) 0.736

Табличные значения для уровней значимости 0.05 и 0.1: 1.36 и 1.22 соответственно. можно заметить, что для данных выборок критерий выполняется, то есть не отвергается нулевая гипотеза для этих уровней значимости.

1.2 Критерий согласия хи-квадрат К.Пирсона

1.2.1 гамма-распределение

Данный метод считается одним из наиболее универсальных, так как любые исходные данные можно свести к дискретным при помощи группировки наблюдений, а именно, перейти от выборки к частотам попаданий элементов выборки в определенные интервалы, на которые эта самая выборка разбивается (в отличие от использованного в 1м пункте критерия Колмогорова-Смирнова, который работает лишь в случае выборок из одномерного непрерывного распределения). В чем же заключается данный метод: в эксперименте наблюдается некоторая дискретная величина ζ, принимающая некоторые значения с некоторыми вероятностями p_i. Тогда
$v_j = \sum_{i=1}^{n} I(\zeta_i = j), j = 1, 2...M$
частоты исходов. В случае непрерывного распределения знак равенства в сумме заменяется на принадлежность j-тому интервалу (отсюда очевидный недостаток - потеря данных при группировке, также, выбор интервалов вносит некоторую погрешность). Тогда вектор этих частот $\bar{v} = (v_1...v_N)$ имеет полиномиальное распределение. Метод хи-квадрат представляет из себя параметрическую модель, работающую с вектором частот и вектором вероятностей. Вообще говоря, для проверки простой гипотезы H_0 нужно измерить отклонение эмпирических данных от теоритических, и в данном методе в качестве меры этих отклонений используется мера хи-квадрат.
$\overset{o}{X_n}{}^2 = \sum_{j=1}^{N} \frac{v_j - np_j}{np_j}$
Метод основан на том, что если гипотеза верна, то относительная частота события будет состоятельной оценкой его вероятности, тогда с ростом размера выборки отклонения будут достаточно малы, и значение выбранной статистики будет также мало. Так как при больших n выполняется следующее:
$L(X_N^2|H_0) \to \Xi^2(N-1)$
Можно определить критическую облать как
$\{X_n^2 > \Xi_{1-\alpha,N-12}\}$
Тогда сам критерий согласия можно сформулировать как:
H_0 отвергается $\Longleftrightarrow \{X_n^2 > \Xi_{1-\alpha,N-12}\}$
К достоинствам метода, кроме его универсальности и простоты, можно отнести его "неприхотливость нет необходимости учитывать точные значения наблюдений, или же если наблюдения имеют не числовой характер (например, селекция семян). Практические результаты
Результаты для гамма-распределения. Изначально, разбиение проводилось на равные интервалы по различным критериям (н-р формула Стерджесса), однако, вероятности для интервалов различались слишком сильно, и в ис-

следовании разбиение проводилось по методу равных вероятностей.

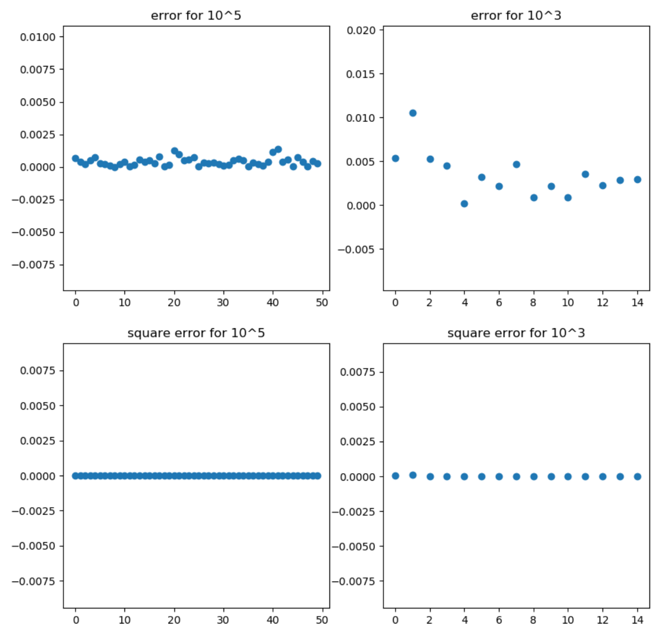
$n = 1000$
7.249
4.999
5.388
5.313
5.820

Табличные критические значения:
0.05: 22.362
0.1: 19.812

$n = 100000$
13.444
17.472
15.564
17.151
18.801

Табличные критические значения:
0.05: 66.339
0.1: 62.008

Видно, что полученные значения подтверждают гипотезу на заданных уровнях значимости.
Хорошей иллюстрацией служит визуализация числителей статистики, т.е. квадратичной разницы между теоретическими и эмпирическими данными, видно что даже не квадратичная ошибка крайне мала:



1.2.2 Распределение Бореля-Таннера

Хотя распределение Бореля-Таннера и является дискретным, была также применена группировка данных, так как из-за больших объемов выборок было большое количество значений, которым соответствовало меньше 5 наблюдений (иногда даже 0), что не позволяет использовать критерий хи-квадрат с необходимой точностью. Группировка проводилась с таким рассчетом, чтобы ни у одного интервала не было двух одинаковых границ, а так как большинство значений сконценстированно (что можно было заметить и ранее, при получении квантилей в одной из предыдущих работ) в начале числовой оси, интервалов получилось относительно мало, что не должно, в теории, никак повлиять на результаты, ведь считается интеграл теоретической функции на границах, и если гипотеза правильная, то критерий будет работать и при малом количестве интервалов (при правильной группировке)
Значения статистики:

$n = 100000$
1.5010274192219615
1.8877170873304872

0.6228811484030287
2.545503311294103
2.566438722510217
Разбиение на 8 интервалов, следовательно, $N-1=7$
Табличные значения для квантилей с соответствубщими параметрами: $\alpha = 0.05$: 14.067, $\alpha = 0.1$: 12.017

$n = 100000$
4.02454063728003
1.6898977183728239
13.49921567055036
3.4364109151041227
10.143030902904052
Разбиение на 10 интервалов, следовательно, $N-1=9$
Табличные значения для квантилей с соответствубщими параметрами: $\alpha = 0.05$: 16.919, $\alpha = 0.1$: 14.684

2 Проверка сложной гипотезы

2.1 Критерий Колмогорова-Смирнова

2.1.1 Теоретическое введение

Сложная гипотеза H_0 вида
$F(x) \in \{F(x,\theta), \theta \in \Theta\}$
, то есть, предельное распределение критерия согласия зависит от вида наблюдаемого закона $F(x,\theta)$, соответствующего проверяемой гипотезе, но и от типа оцениваемого параметра, их числа и метода оценивания а также конкретного значения параметра формы закона. Критерий Колмогорова-Смирнова - непараметрический, поэтому, чтобы сохранить независимость от распределения, оценку параметра нельзя проводить по той же выборке. В данном исследовании будет использоваться поправка Большева:
$S_k = \frac{6nD_n+1}{b\sqrt{n}}$

2.1.2 Вычисления

Для начала, получим оценку методом максимального правдоподобия для гамма-распределения:
$\grave{}(x,\theta) = \frac{1}{\lambda^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\lambda}}$, $\theta = (\lambda, \alpha)$
$L(\lambda, \alpha | X^{(n)}) = -n\alpha \ln\lambda - n\ln\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n} \ln X_i - \frac{1}{\lambda}\sum_{i=1}^{n} X_i$
Составим уравнения правдоподобия:
$\frac{\partial L}{\partial \lambda} = -\frac{n\alpha}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{n} X_i = 0$
$\frac{\partial L}{\partial \alpha} = -n\ln\lambda - n\psi(\alpha) + \sum_{i=1}^{n} \ln X_i = 0$
Где $\psi(x) = \frac{d\ln\Gamma(x)}{dx}$. Решая систему получаем:

$\ln\alpha - \psi(\alpha) = \ln\overline{X} - \frac{1}{n}\sum_{i=1}^{n}\ln X_i$
$\ln\alpha - \psi(\alpha)$ очевидно монотонна, значит существует единственное решение. Используя табулированные значения пси-функции Эйлера и асимптотическую формулу можно получить приближенные формулы для α:
$\alpha^* = \frac{8.899+9.06z+0.978z^2}{z(17.797+11.168z+z^2)}$
Где $0.5572 \leq z \leq 17$ - среднее геометрическое по выборке.Зная оценку для α можно получить оценку для λ:
$\lambda^* = \frac{\alpha^*}{\overline{X}}$
Что, как и ожидалось, совпадает с оценкой для экспоненциальной модели, полученной ранее, так как эта оценка была эффективная, а по свойству оценки ММП, эффективная оценка параметра будет совпадать с его оценкой ММП.
Теперь найдем значения оценок параметров для выборок. Для начала, проверим средние геометрические для всех рассматриваемых выборок, чтобы можно было использовать приближенную формулу оценки:

$n = 1000$
1) 1.673
2) 1.659
3) 1.671
4) 1.667
5) 1.704

$n = 100000$
1) 1.666
2) 1.669
3) 1.666
4) 1.670
5) 1.667

Все значения попадают в необходимые пределы. Тогда вычислим значения оценок:
Теперь вычислим предельное значение статистики и применим поправку Большева. И на основании выбранного метода оценивания, теоретического распределения можно выбрать одну из табулированных моделей распределения. Для моего случая такой моделью оказалось Бета распределение третьего рода с параметрами: $B_{III}(6.1957, 6.1114, 2.8894, 1.13140, 0.2801)$. Бета распределение третьего рода имеет следующую плотность:
$B_{III}(\theta_0,\theta_1,\theta_2,\theta_3,\theta_4) = \frac{\theta_2^{\theta_1}}{\theta_3 B(\theta_0,\theta_1)} \frac{(\frac{(x-\theta_4)}{\theta_3})^{\theta_0-1}(1-\frac{(x-\theta_4)}{\theta_3})^{\theta_1-1}}{(1+(\theta_2-1)\frac{(x-\theta_4)}{\theta_3})^{\theta_0+\theta_1}}$.
Теперь чтобы получить значение для выбранного уровня значимости надо вычислить значение
$P(\{S > S^*\}) = \int_{S^*}^{+\infty} g(s|H_0)ds = 1 - G(S^*|H_0)$
Если результат будет больше требуемого уровня значимости, на нем проверяемая гипотеза отвергается.

Значения статистики Колмогорова:
$n = 1000$
1.5640985645225103 1.489169546099757 1.4781893343274591 1.3488705706409894
1.6405572252214618

$n = 100000$
1.595700095315776 1.3676064320557732 1.268733903919745 1.45878910170314
1.386265873459554
На графике видно что при таких значениях значение функции действительно довольно близко к нулю. Проведем соответствующие вычисления, чтобы подтвердить:

$n = 1000$
1) 0.0019294472822242974
2) 0.000927363975125187
3) 0.0011279563146415215
4) 0.004411114265498286
5) 0.00036414086687442828

$n = 100000$
1) 0.0007676835629112642
2) 0.013374401303275376
3) 0.020010422214396958
4) 0.0038153397848997554
5) 0.007915214519788046

Очевидно, все полученные величины меньше $\alpha = 0.05$, откуда можно сделать вывод о хорошем согласии выборок с гипотезой.

2.2 Критерий хи-квадрат Пирсона для сложной гипотезы

Сложные гипотезы для полиномиального распределения в общем случае имеют вид:
$H_0 : p = p(\theta), \theta = (\theta_1 ... \theta_r) \in \Theta, r < N - 1$
Получается, что при такой гипотезе вероятности наблюдений являются функциями от параметра. Таким образом, для построения критерия можно воспользоваться аналогичной простой гипотезе статистикой:
$X_n^2(\theta) = \sum_{j=1}^{n} \frac{(v_j - np_j(\theta))^2}{np_j(\theta)}$
Т.к. статистика зависит от неизвестного параметра, нужно заменить его некоторой его оценкой, получая статистику
$\overline{X_n^2} = X_n^2(\overline{\theta_n})$
Которая будет зависеть только от наблюдений, следовательно, ее можно од-

нозначно вычислить для каждой реализации. Для построения самого критерия можно воспользоваться теоремой Фишера, который показал, что при определенных методах оценивания параметра (в частности ММП) предельное распределние будет иметь вид $^2(N-1-r)$. Сформулируем саму теорему: Пусть функции вероятностей от параметра удовлетворяют следующим свойствам:
$\sum_{j=1}^{N} p_J(\theta) = 1, \forall\theta \in \Theta$
$p_j(\theta) \geq c > 0 \forall j$ и существуют непрерывные производные
$\frac{\partial p_j(\theta)}{\partial \theta_k}, \frac{\partial^2 p_j(\theta)}{\partial \theta_k \partial \theta_l}, k, l = 1...r$
Nxr - матрица первых производных - имеет ранг $r \forall \theta \in \theta$
Тогда оценка статистики будет иметь предельное распределение хи-квадрат:

$L(\hat{X}_n^2 \to^2 (N-1-r))$
Тогда сам критерий будет имет вид
H_0 отвергается $\iff \hat{X}_n^2 >^2_{1-\alpha,N-1-r}$
Аналогично с простой гипотезой, этим критерием можно проверять и непрерывные распределения, применив группировку.

2.2.1 сложная гипотеза для гамма-распределния по критерию хи-квадрат

Воспользуемся оценками, полученными ранее, чтобы получить значения статистики:
$\hat{X}_n^2$:

$n = 1000$
10.456968617014969
13.194173566849162
2.46073205246025
10.583963108193537
8.061811400877453

Разбиение на 12 интервалов, следовательно, $N-1-r=9$
Табличные значения для квантилей с соответствубщими параметрами: $\alpha = 0.05$: 16.919, $\alpha = 0.1$: 14.684

$n = 100000$
17.06768238030598
16.984231241943064
4.70125987147686
18.350966493460465
17.74126667450131

Разбиение на 20 интервалов, следовательно, $N-1-r=17$
Табличные значения для квантилей с соответствубщими параметрами: $\alpha =$

0.05: 27.587, $\alpha = 0.1$: 24.769
Можно сделать вывод, что данные гипотезы подтвердились
Разбиение методом равных вероятностей для определения оптимальной длины интервала и количества.

2.2.2 Сложная гипотеза для распределения Бореля-Таннера по критерию хи-квадрат

При неизвестном параметре r метод максимального правдоподобия не имеет смысла, поэтому будем использовать его только для оценки второго параметра и рассматривать распределение как однопараметрическое.
Воспользуемся оценками, полученными ранее, чтобы получить значения статистики:
$\hat{X}_n^2$:

$n = 1000$
1.5612209471763667
2.845553626885467
1.1433066782277925
1.6849507771345686
3.597088989492619
Разбиение на 12 интервалов, следовательно, $N - 1 - r = 10$
Табличные значения для квантилей с соответствубщими параметрами: $\alpha = 0.05$: 18.307, $\alpha = 0.1$: 15.987

$n = 100000$
4.038207286414831
1.6893728432008763
13.65337420972201
4.013425216352789
10.137580492235074
Разбиение на 30 интервалов, следовательно, $N - 1 - r = 28$
Табличные значения для квантилей с соответствубщими параметрами: $\alpha = 0.05$: 41.337, $\alpha = 0.1$: 37.916
Можно сделать вывод, что данные гипотезы подтвердились

3 ВАЖНО

Как отмечалось ранее, было бы серьезной ошибкой брать оценки параметров из выборки, к которой применется критерий на согласие с распределением. Поэтому, были сгенерированы по 5 выборок каждого размера для расчета оценок (с теми же параметрами, конечно), а по изначальным выборкам проводилось исследование соответствие распределению.