ИДЗ № 4

По дисциплине

«Мат.статистика»

Выполнил: Федоров Н.С.

Группа №5362

Преподаватель: Медведев А.Н.

***Задание №1:***

1. Необходимо провести тестирование и анализ работы алгоритма. А именно, сгенерировать выборки длины 10, 100, 1000 из нормального распределения (параметры выбираете сами, желательно “пооригинальней”) .
   1. В качестве основной гипотезы берете H\_0=(среднее равно тому, которое брали при генерации), альтернатива H\_1=(среднее другое). Необходимо для уровней α=0.25, 0.1, 0.05, 0.01, 0.001 построить и применить следующие критерии проверки простой гипотезы согласия H\_0:
      1. с известной дисперсией (берете ту, которую задали при генерации)
      2. с неизвестной

Под построением критерия подразумевается то, что для данного уровня α вы вычисляете соответствующее пороговое значение С. В отчете необходимо будет выписать общую формулу для критерия и формулу для вычисления С. Применить критерий к вашей выборке означает - вычислить значения статистики отклонения для вашей выборки, сравнить с пороговым значением, и сделать вывод: отвергнуть гипотезу на данном уровне значимости, или же данные не противоречат гипотезе.

1. Провести анализ зависимости результата от уровня α.
2. Для вашей выборки вычислить реально достигнутый уровень значимости. Интерпретировать его.
3. Провести анализ зависимости результатов от длины выборки.
   1. Пусть при генерации вы взяли среднее a. Проделать пункт 1.1. для гипотез: H\_0=(среднее равно a+ε), где ε=10, 0.1, 0.01. Интерпретировать зависимость от ε.

В качестве примера взято нормальное распределение с длинами выборок **n** = 10,100 и 1000 элементов и параметрами **a** = 7, **s2** = **σ2** = 5.76 = (2.4)2.

Возьмём в качестве гипотезы H0 предположение о том, что выбор.среднее имеющейся выборки равно а и проверим правдоподобие данной гипотезы. Известно, что если гипотеза Н0 верна, то величина будет распределена по стандартному нормальному закону N(0,1). Тогда должно удовлетворять равенству , где С1-*а*/2 – квантиль стандартного нормального распределения для известного уровня значимости . В качестве примера сгенерируем три выборки, которые можно найти в файлах "Distr\_for\_10.csv", "Distr\_for\_100.csv" и "Distr\_for\_1000.csv".

Сначала рассмотрим выборку длиной в 10 элементов. Рассчитаем значения квантилей для каждого уровня значимости(α=0.25, 0.1, 0.05, 0.01, 0.001), а также значение для данной выборки и занесём их в таблицу:

|  | **C1** | **C2** | **C3** | **C4** | **C5** | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **1** | 1.150349 | 1.644854 | 1.959964 | 2.575829 | 3.290527 |

Стоит напомнить, что критерий согласия выглядит как:

, а С вычисляется вручную, исходя из уравнения = 1 – Ф(С), где Ф(С) – функция распр-я N(0,1) в точке С(либо в пакете R с помощью базовой функции qnorm(1- /2), где – имеющееся значение уровня значимости).

В нашем случае значения величины равны: 1) 0.8004051(**n**=10), 2) 0.1569805(**n**=100), 3) 0.5678165 (**n**=1000). Как видно, все значения удовлетворяют неравенствам критерия согласия, для каждого уровня значимости, согласно которому мы можем принять гипотезу H0 как верную(на самом деле, при больших значениях , таких как =0.25, например, довольно часто случается, что гипотеза не удовлетворяет неравенству, выписанному в критерии, но анализ этого будет представлен ниже).

Теперь рассчитаем реально достигнутые уровни значимости для каждой выборки. Пусть **p** – реально достигнутый уровень значимости. Чтобы его найти, необходимо в формулу = 1 – Ф(С) подставить значение вместо С. Тогда полученное = **p** и есть реально достигнутый уровень значимости для данной выборки. Получим следующие значения **р** имеющихся выборок: 1) p0.42(**n**=10), 2) p0.88 (**n**=100), 3) p0.56 (**n**=1000).

Если считать дисперсию неизвестной, а гипотезу – верной, то имеет распределение Стьюдента Tn-1(по следствию из леммы Фишера). Чтобы найти значение константы C при распределении Стьюдента, необходимо решить уравнение = 1 – FTn-1(С), где – выбранный уровень значимости. Дисперсия же рассчитывается по известной формуле , а функция отклонения = . Рассчитаем получившиеся значения , С, , а также реально достигнутые уровни значимости. Дисперсия равна: 1) 2.608411 = (1.62)2 (**n**=10), 2) 5.23302 = (2.29)2 (**n**=100), 3) 5.834038 = (2.41)2 (**n**=1000) – как видно, из-за особенностей первой выборки дисперсия значительно отличается от изначально заданного значения (2.4)2, хотя остальные значения отличаются ненамного. Значения С практически не отличаются от полученных при нормальном распределении(при = 0.25 и 0.10 различие порядка 0.001, при = 0.05, 0.01 и 0.001 – порядка 0.01), а значения равны 1)1.189414(**n**=10), 2)0.1646952(**n**=100), 3)0.564202(**n**=1000) – т.к. дисперсия в первой выборке оказалась более чем в два раза меньше изначально заданной, значения для **n** =10 возросло, значения же остальных практически не отличаются от аналогичных цифр, полученных в случае с известной дисперсией. Это можно объяснить тем, что при достаточно больших размерах выборки минимизируется(хотя и присутствует) риск того, что значение дисперсии будет «испорчено» из-за случайно сгенерированных значений выборки. Реально достигнутые уровни значимости **p** равны:

1) p0.23(**n**=10), 2) p0.87(**n**=100), 3) p0.56(**n**=1000).

В данном случае естественно, что при **n** = 10 реально достигнутый уровень значимости **p** отличается от уровня, достигнутого при известной дисперсии, из-за их отличия.

Проведём анализ, насколько сильно итоговый результат зависит от того или иного параметра. Насколько понятно из обширного количества проведённых симуляций, длина выборки может оказать существенное влияние на итоговое решение: принимать или не принимать гипотезу. Так, при больших **n** значение весьма «чувствительно» к тому, какую гипотезу мы выдвигаем(в данном случае - значение a0). Чем сильнее реальное среднее отличается от предполагаемого, тем больше будет итоговое значение , тогда при большом значении константы С(и соответственно, малом уровне значимости ) вероятность принять гипотезу H0 будет выше, и полученный критерий согласия будет весьма слабым инструментом оценивания. Однако и при низких значениях С также существует вероятность совершить ошибку первого рода, если длина выборки небольшая(например, n = 10), так как теперь значение будет менее восприимчиво к длине выборки(как известно, = ), и поэтому даже при достаточно сильном критерии согласия вполне возможно выдвинуть неверную гипотезу и при этом получить достаточно высокое значение реального уровня значимости **p**. Принято считать, что чем меньше значение , тем сильнее аргументы против нулевой гипотезы, т.е. в итоге мы получим недостаточно аргументов для того, чтобы отвергнуть гипотезу H0, если при небольшой длине выборки сделаем предположение, ненамного отличающееся от истинного, но тем не менее неверное. Подводя итог, можно сказать, что наиболее высокой вероятность совершить ошибку первого рода будет лишь при одновременном выполнении нескольких факторов:

1) Средняя и большая длина выборки, но низкий уровень значимости .

2) Любой, даже довольно высокий уровень значимости при соблюдении малой длины выборки.

***1.2.***

Теперь выдвинем гипотезу H0, что среднее a0 = **a**+, где соответственно равен 10, 0.1 и 0.01. Рассчитаем, как изменится ф-я в этом случае. Для известной и неизвестной дисперсий получим следующие значения:

1.1) Дисперсия известна:

|  |  |  |  |
| --- | --- | --- | --- |
|  | = 10 | = 0.1 | = 0.01 |
| **n** = 10 | 12.37575 | 0.6686435 | 0.7872289 |
| **n** = 100 | 13.01918 | 0.02521896 | 0.14380144 |
| **n** = 1000 | 13.23294 | 0.1885432 | 0.06951248 |

1.2) Дисперсия неизвестна:

|  |  |  |  |
| --- | --- | --- | --- |
|  | = 10 | = 0.1 | = 0.01 |
| **n** = 10 | 18.39055 | 0.9936143 | 1.169834 |
| **n** = 100 | 13.65899 | 0.02645832 | 0.1508715 |
| **n** = 1000 | 13.1487 | 0.187343 | 0.06995781 |

Если сравнить данные значения со значениями, полученными при гипотезе **a0**= **a**, то выяснится следующее:

1) Вполне естественно, что для всех выборок при = 10 реально достигнутый уровень значимости фактически сведён к нулю, т.к. выдвинутая гипотеза очевидным образом не верна(изначальное **a0** равно 7 против гипотезы **a0** = 17, которое мы выдвинули в этот раз), а значение увеличилось в несколько раз.

2) При = 0.1 можно наблюдать увеличение реально достигнутого уровня значимости для всех без исключения выборок, т.е. можно считать, что если значение e достаточно мало(110%), то функция уменьшается по абсолютному значению и из-за этого гипотеза H0=**a0**+ будет удовлетворять критерию чаще, чем при гипотезе H0=**a0**.

3) При = 0.01 значения для выборок длины **n** = 10 и 100 практически не изменились, зато реально достигнутый уровень значимости для **n**=1000 намного улучшился. Тот факт, что при большой длине выборки значение функции сильно зависит от корректности и адекватности выбранной гипотезы, уже было отмечено ранее.

Стоит отметить, что во всех случаях, кроме = 10, где всё критически ухудшилось, реально достигнутые уровни значимости улучшились для каждой из рассматривавшихся выборок и составили:

1.1) **p** при известной дисперсии:

|  |  |  |  |
| --- | --- | --- | --- |
|  | = 10 | = 0.1 | = 0.01 |
| **n** = 10 |  |  |  |
| **n** = 100 |  |  |  |
| **n** = 1000 |  |  |  |

1.2) **p** при неизвестной дисперсии:

|  |  |  |  |
| --- | --- | --- | --- |
|  | = 10 | = 0.1 | = 0.01 |
| **n** = 10 |  |  |  |
| **n** = 100 |  |  |  |
| **n** = 1000 |  |  |  |

***Задание № 2:***

Рассмотрим файл “zz\_norm\_x.csv”. В нем содержится выборка из предполагаемого нормального распределения со средним x, дисперсией 1. Необходимо проделать пункты a,b,d из задания 1.1. для данной выборки(уровни в пунктах a и b те же) Интерпретировать полученные результаты.

Для начала сделаем предположение, что в данном распределении мат.ожидание **a0**= -75.3, т.е. гипотеза H0 такова, что a= **a0**. Теперь, когда у нас есть гипотеза, можно её проверить.

Проделав те же действия, что и в задании №1, получим следующие результаты:

1) Для известной дисперсии(2 = 1):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C ( | ) | C ( | C ( | C ( |
| 0.25124 | 1.15035 | 1.6449 | 1.96 | 2.57583 | 3.29053 |

2) Для неизвестной дисперсии(2 = s2):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C ( | ) | C ( | C ( | C ( |
| 0.421374 | 1.15102 | 1.6464 | 1.96234 | 2.58076 | 3.3003 |

Значения были рассчитаны по той же формуле, что и в задании №1. Разница в значениях функции связана с тем, что реальная дисперсия распределения оказалась равна 2 = 0.3555. Реальные уровни значимости **p1** и **p2** будут соответственно равны: 1) **p1** =0.80, 2) **p2** = 0.67. Понижение уровня значимости случилось из-за того, что в случайно сгенерированной выборке реальное значение квадрата среднеквадратического отклонения заметно отличалось от заданного. Можно сделать вывод, что особенности каждой выборки как положительным, так и отрицательным образом могут повлиять на итоговый результат и величину реально достигнутого уровня значимости **p**.

***Задание №3, 4:***

Проделать пункт 1.1. для задачи о проверки простой гипотезы согласия о параметрах распределения Бернулли (генерируете выборку со своим параметром p; берете те же уровни)

Проделать пункт 1.2 для задачи о проверки простой гипотезы согласия о параметрах распределения Бернулли (для ε=0.2, 0.1, 0.01; следите, чтобы сумма p+ ε была меньше 1)

Сгенерируем случайную выборку из распределения Бернулли со следующими параметрами: **n**=100 - количество событий, где «событие» – результат подбрасывания монетки, **p**=0.7 – вероятность успеха, где «успех» для события – монета упала решкой вверх, а «неудача» – монета упала орлом вверх. Сгенерированная выборка сохранена в файле «Bernulli\_100.csv». Рассчитаем выборочное среднее , которое оказалось равно 0.71. Теперь сделаем предположение, что **p0**= 0.5, т.е. монета без дефектов и вероятность выпадения орла и решки одинакова. Тогда гипотеза H0 такова, что p=**p0**. Для построения критерия используем центральную предельную теорему(ЦПТ), тогда , значит, по Th Муавра-Лапласа распределение случайной величины с ростом n приближается к нормальному, поэтому эту функцию можно взять в качестве функции . Построим критерий :

,

где С выбирается как квантиль уровня 1 - стандартного нормального распределения, т.е. ( критерии были указаны выше). Рассчитаем функцию отклонения и получим, что = 4.2, т.е. очевидно, что выдвинутое предположение неверно. Тогда сделаем предположение, что монета имеет дефект, причём такой, что в 3 случаях из 4 выпадает решка, т.е. и гипотеза имеет вид p=. Проверим правдоподобность этой гипотезы и рассчитаем значение :

= 0.9237604,

а значит, что гипотезу вполне можно считать верной, и реально достигнутый уровень значимости при этом будет равен 0.355. Теперь изменим гипотезу на , которая такова что p =**= +** , где = 0.2, 0.1, 0.01. Получим следующие значения :

|  |  |  |  |
| --- | --- | --- | --- |
|  | = 0.2 | = 0.1 | = 0.01 |
|  |  |  |  |

Как видно, в данной выборке сделанные предположения уменьшили значение . Как упоминалось ранее < C при уровне значимости , говорят, что недостаточно аргументов для того чтобы отвергнуть гипотезу. В нашем случае изменение основной гипотезы в первых двух случаях привело к тому, что от гипотезы пришлось отказаться ввиду её малой правдоподобности. В третьем же случае добавление = 0.01 ухудшило реально достигнутый уровень значимости, но гипотеза при этом осталась правдоподобна. В реальности **p** оказалось равным 0.7, т.е. мы были довольно близки к истине, когда предположили, что . Анализируя значимость для построения критерия стоит сказать, что незначительные изменения основной гипотезы(увеличение/уменьшение **p0** на 1-2%) могут повлиять лишь в том случае, когда берётся довольно высокий уровень значимости , однако при стандартном гипотеза всё равно остаётся правдоподобной для большинства случаев.