Высшая инженерная школа РУТ (МИИТ)

СТАТГИПОТЕЗЫ

ОДНА ВЫБОРКА

1. Гипотеза о значении матожидания в ГС в случае известного стандартного отклонения

$$z_{\text{набл}} = \frac{\bar{X}-m_0}{\sigma/\sqrt{n}}$$

2. Гипотеза о значении матожидания в ГС в случае неизвестного стандартного отклонения

$$t_{\text{набл}} = \frac{\bar{X}-m_0}{s/\sqrt{n}} \sim T(n-1)$$

3. Гипотеза о значении матожидания в ГС в случае неизвестного стандартного отклонения большая выборка

$$z_{\text{набл}} = \frac{\bar{X}-m_0}{s/\sqrt{n}}$$

4. Гипотеза о значении доли в ГС (одна большая выборка) (гипотеза о вероятности успеха в единичном испытании)

$$z_{\text{набл}} = \frac{w_{\text{в}}-w_0}{\sqrt{\frac{w_0(1-w_0)}{n}}} \sim N(0;1)$$

5. Гипотезы о значении дисперсии (стандартного отклонения) в ГС – одна малая выборка

$$\chi^2_{\text{набл}} = \frac{(n-1)\cdot s^2}{\sigma_0^2} \sim \chi^2(n-1)$$

6. Гипотезы о значении дисперсии (стандартного отклонения) в ГС – одна большая выборка

$$z_{\text{набл}} = \frac{\frac{(n-1)\cdot s^2}{\sigma_0^2}-(n-1)}{\sqrt{2(n-1)}}$$

ДВЕ ВЫБОРКИ

7. Гипотезы о равенстве долей (две большие выборки)

$$z_{\text{набл}} = \frac{\frac{k1}{n1}-\frac{k2}{n2}}{\sqrt{\tilde{w}(1-\tilde{w})(\frac{1}{n1}+\frac{1}{n2})}}, \tilde{w} = \frac{k_1+k_2}{n_1+n_2}$$

8. Гипотезы о равенстве генеральных средних (матожиданий) при известных стандартных отклонениях

$$z_{\text{набл}} = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_1^2}{n1}+\frac{\sigma_2^2}{n2}}} \sim N(0;1)$$

9. Гипотезы о равенстве генеральных средних (матожиданий) при неизвестных равных стандартных отклонениях (неизвестное σ_1 равно неизвестному σ_2)

$$t_{\text{набл}} = \frac{\bar{X}-\bar{Y}}{\sqrt{\tilde{s}^2\cdot(\frac{1}{n_1}+\frac{1}{n_2})}} \sim T(n_1+n_2-2), \text{ где } \tilde{s}^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}.$$

10. Гипотезы о равенстве генеральных средних (матожиданий) в парных выборках.
Для выборки $d_1 = X_1 - Y_1, d_2 = X_2 - Y_2 \dots d_n = X_n - Y_n$ проверяем гипотезу о равенстве матожидания нулю.

11. Гипотеза о значимости кк Пирсона

$$t_{\text{набл}} = \frac{r\cdot\sqrt{n-2}}{\sqrt{1-r^2}} \sim T(n-2),$$

НЕПАРАМЕТРИЧЕСКИЕ ГИПОТЕЗЫ

12. Критерий хи-квадрат независимости номинальных признаков

$$\chi^2_{\text{набл}} = \sum \frac{(\text{наблюдаемая}-\text{ожидаемая частота})^2}{\text{ожидаемая}} \sim \chi^2((k-1)\cdot(m-1))$$

13. Критерий согласия хи-квадрат

$$\chi^2_{\text{набл}} = \sum \frac{(\text{наблюдаемая}-\text{ожидаемая частота})^2}{\text{ожидаемая}} \sim \chi^2(k-1)$$

В каждой задаче должны быть записаны гипотезы, к каждой задаче должен даваться текстовый вывод, привязанный к условию задачи:

Т.к. наблюдаемое значение статистики попало в область естественных значений, то на уровне значимости *** выборочные данные не противоречат основной гипотезе, т.е. мы не отвергаем гипотезу о том, что *******

Т.к. наблюдаемое значение статистики не попало в область естественных значений, то на уровне значимости *** выборочные данные не соответствуют нулевой гипотезе, то есть мы отвергаем нулевую гипотезу и принимаем альтернативную о том, что *******

Для нормального закона (аналогично для распределения Стьюдента)
для правосторонней альтернативы $p.v. = P(Z > Z_{\text{набл}})$,
для левосторонней $p.v. = P(Z < Z_{\text{набл}}) = P(Z > |Z_{\text{набл}}|)$,
для двусторонней альтернативы $p.v. = P(Z > |Z_{\text{набл}}|)\cdot 2$ – отсекаемый хвост нужно умножить на 2.

Для хи-квадрат для правосторонней альтернативы $p.v. = P(\chi^2 > {\chi^2}_{\text{набл}})$,
для левосторонней $p.v. = P\left(\chi^2 < {\chi^2}_{\text{набл}}\right)$,
для двусторонней альтернативы отсекаемый хвост нужно умножить на 2:

$$p.v. = 2\cdot\min\left(P\left(\chi^2 < {\chi^2}_{\text{набл}}\right); P\left(\chi^2 > {\chi^2}_{\text{набл}}\right)\right)$$

Основная гипотеза будет приниматься на любом уровне значимости α, меньшем, чем p.v., а на больших уровнях значимости она будет отвергаться.

Все задачи ниже объединены одной историей: один из магазинов торговой сети анализирует свою работу, и нам надо изучить поведение покупателей этого магазина, сравнить поведение разных покупателей друг с другом и сравнить этот магазин с остальными магазинами.
Предполагая, что все интересующие нас данные распределены нормально в генеральной совокупности, выборки репрезентативны, генеральные совокупности где надо независимы, определить:

14.1	Мы хотим выяснить – можно ли считать, что в данном магазине в среднем за неделю совершается больше покупок, чем в остальных магазинах сети (по сети среднее 1450 покупок в неделю). а) На уровне значимости 3% проверить соответствующую гипотезу, если по результатам наблюдений за 20 недель среднее количество покупателей за неделю в данном магазине оказалось равно 1500 покупок, и при этом известно, что генеральное стандартное отклонение количества покупок за неделю равно 100. б) Найти $p.v.$ На каких уровнях значимости нулевая гипотеза будет отвергаться, а на каких не будет? *Указание: по тексту задачи речь идет о гипотезе «в данном магазине среднее больше 1450», это альтернативная, а основная должна быть в виде равенства, то есть «в данном магазине среднее равно 1450».*
14.2	Известно, что по всей сети среднее время между входом покупателя в магазин и оплатой товара равно 19 минутам. По данным, полученным по 25 случайным покупателям данного магазина, среднее время оказалось равным 24 минутам, стандартное отклонение – 7 минутам. а) На уровне значимости 5% проверить гипотезу о том, что в данном магазине покупатель тратит больше времени на совершение покупки, чем в среднем по сети. б) Оценить р-значение в соответствии с имеющейся таблицей критических точек.
14.3	Известно, что по всей сети среднее время между входом покупателя в магазин и оплатой товара равно 19 минутам, стандартное отклонение равно 5.5 минутам. По данным, полученным по 25 случайным покупателям данного магазина, среднее время оказалось равным 24 минутам, стандартное отклонение – 7 минутам. а) На уровне значимости 1% проверить гипотезу о том, что стандартное отклонение времени, потраченного на покупки в данном магазине, больше чем во всей сети. б) Оценить р-значение в соответствии с имеющейся таблицей критических точек, ответ записать в процентах.
14.4	По выборке из 9 мужчин найдены средний чек и стандартное отклонение, оказавшиеся равными 1000 и 250 рублей. По выборке из 7 женщин соответствующие параметры оказались равны 1500 и 300.

а) Считая, что стандартные отклонения чеков на самом деле одинаковы, проверить гипотезу о равенстве средних чеков для мужчин и женщин. В качестве альтернативной взять двустороннюю, уровень значимости 1%.
б) Оценить p-значение в соответствии с имеющейся таблицей критических точек.

14.5 В выборке из 800 покупателей, совершавших покупки в будни, постоянными покупателями оказались 180 человек, в выборке из 200 покупателей, совершавших покупки в выходные, постоянными оказались 40 человек.
а) На уровне значимости 2% проверить гипотезу о том, что в будни доля постоянных покупателей меньше 25%.
б) Найти p-значение.
в) На уровне значимости 1% проверить гипотезу о том, что в будни доля постоянных покупателей больше, чем в выходные.
г) Найти p-значение.

14.6 По выборке из 200 покупателей, совершавших покупки в выходные, найдены средний чек и стандартное отклонение, оказавшиеся равные 1700 и 400. Есть ли у нас основания утверждать, что средний чек данной категории покупателей выше среднего по сети, равного 1650?
а) Проверить соответствующую гипотезу на уровне значимости 10%
б) Найти p-значение.

14.7 По выборке из 800 покупателей, совершавших покупки в будни, найдены средний чек и стандартное отклонение, оказавшиеся равные 1600 и 500. Есть ли у нас основания утверждать, что стандартное отклонение чека для данной категории покупателей больше стандартного отклонения по сети, равного 480?
а) Проверить соответствующую гипотезу на уровне значимости 10%
б) Найти p-значение.

11.8 В таблице приведены данные по 5 покупателям: первая строка – количество позиций в чеке, вторая – величина чека.
а) Найти коэффициент корреляции и на уровне 1% проверить гипотезу о его значимости. В качестве альтернативной взять двустороннюю.
б) При каких значениях выборочного коэффициента корреляции будет принята гипотеза о его значимости? (такое не решали, подумайте)

количество позиции	7	8	3	5	4
общая стоимость	400	500	400	200	300

14.9 В таблице приведены данные по 5 покупателям: первая строка – количество покупок в течение месяца до получения карточки постоянного покупателя, вторая – в течение месяца после. Можно ли считать, что получение карточки постоянного покупателя приводит к изменению среднего количества покупок? Проверить соответствующую гипотезу на уровне значимости 5%. (Указание: надо сравнить матожидания до и после получения)

количество покупок до	7	8	3	5	4
количество покупок после	8	10	4	3	6

14.10 В таблице приведены данные по покупателям – возрастная категория и время, в которое была совершена покупка. На уровне значимости 1% проверить гипотезу о независимости данных признаков.

возраст / время	<25	25-35	>35
9:00 – 13:00	70	90	120
13:00 – 18:00	150	130	100
18:00 – 21:00	50	140	50

14.11 Известно, что число покупателей в магазинах сети за неделю распределяется таким образом – по 15% в каждый из рабочих дней, 20% в субботу, остальные в воскресенье.
По результатам наблюдений в течение недели количество покупателей в магазине оказалось таким:

понедельник	вторник	среда	четверг	пятница	суббота	воскресенье
200	240	270	220	215	275	80

На уровне значимости 2.5% проверить гипотезу о том, что распределение покупателей по дням недели в данном магазине такое же, как и в остальных магазинах.

дополнительная задача

14.12 Рассмотрим гипотетическую ситуацию – мы хотим выяснить, как связано количество времени, которое студенты тратят на выполнение домашнего задания за каждые выходные (в минутах), и их уровень счастья. Мы отдельно опросили студентов первых трех курсов, и получили такие данные (коэффициенты корреляции найдены по стандартной формуле):

Первый курс

количество времени	175	185	170
уровень счастья	30	40	20

выборочный коэффициент корреляции равен $r = 0.982$

Второй курс

количество времени	155	150	160
уровень счастья	40	30	55

выборочный коэффициент корреляции равен $r = 0.993$

Третий курс

количество времени	130	120	115
уровень счастья	60	50	40

выборочный коэффициент корреляции равен $r = 0.982$

Мы видим, что у нас по трем этим выборкам имеется четкая положительная корреляция – чем больше студент тратит времени на выполнение дз, тем он более счастлив.
а) без дополнительных расчетов, просто на уровне интуиции – как вы думаете: если мы возьмем и найдем коэффициент корреляции сразу по всем девяти имеющимся студентам вместе – чему он будет равен? Усреднится, будет равен самому большому из этих трех, самому маленькому и тд?
б) Вычислите этот коэффициент корреляции и объясните произошедшее.

так как это доп задача, то можно все считать и не вручную – например в excel есть встроенная функция, которая считает корреляцию между двумя наборами данных: