Лекция 7. Сравнение двух выборок (дополнительная лекция)

Курбацкий А. Н.

МШЭ МГУ

30 марта 2020

Содержание

Содержание

Идея

В предыдущей теме мы занимались одной генеральной совокупностью и делали выводы о её параметрах. Но часто исследователю приходится сравнивать две выборки.

Пример

Как сравнить, у кого средний доход на душу населения больше, у жителей Одессы или Ростова-на-Дону? У мамы или у папы?

В первой части лекции мы будем иметь дело с независимыми выборками.

Содержание

Равенство средних при известных дисперсиях

В случае, когда дисперсии известны, для проверки гипотезы о равенстве разности средних некоторому значению применяется статистика:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

где
$\bar{x}_1$ и $\bar{x}_2$ - выборочные средние,
μ_1 и μ_2 - гипотетические генеральные средние,
n_1 и n_2 - объемы выборок,
σ_1^2 и σ_2^2 - известные генеральные дисперсии.

Задача

Докажите, что статистика z имеет стандартное нормальное распределение.

Пример

Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	дисперсия
X	9	25	2
Y	6	21	1

Проверьте гипотезу о равенстве средних значений этих выборок на 95% уровне доверия против односторонних альтернатив.

Решение

- Сформулируем основную и альтернативную гипотезы:
$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 > \mu_2.$$
- Критическая область является правосторонней. По таблице z-распределения находим $1 - \alpha = 0.95$ и определяем критическая точка $z_{cr} = 1.64$. Критическая область имеет вид $(1.64; +\infty)$.
- Значение статистики критерия равно
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{25 - 21}{\sqrt{\frac{2}{9} + \frac{1}{6}}} \approx 6.45.$$
- Вывод. Так как $z \in (1.64; +\infty)$, то основная гипотеза H_0 отвергается.
- Значение статистики уже посчитано и равно 6.45. Поэтому минимальный уровень значимости составляет $(1 - z^{-1}(6.45)) \approx 0.000$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

Содержание

Дисперсии неизвестны, но равны

Теорема

В случае, когда дисперсии неизвестны, но равны, для проверки гипотезы применяется статистика:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \boldsymbol{t}(n_1 + n_2 - 2)$$

где
$\bar{x}_1$ *и* $\bar{x}_2$ *– выборочные средние,*
μ_1 *и* μ_2 *– гипотетические генеральные средние,*
n_1 *и* n_2 *– объемы выборок,*
s_p^2 *— объединённая оценка дисперсии.*
Вычисляется объединённая оценка дисперсии по формуле:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

где s_1^2 *и* s_2^2 *- выборочные дисперсии.*

Пример

Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	выборочная дисперсия
X	10	15	2
Y	7	12	1

С помощью критерия Стьюдента проверить гипотезу о равенстве средних значений этих выборок (считая их дисперсии равными) при 95% уровне доверия против двусторонних альтернатив.

Решение

- Сформулируем основную и альтернативную гипотезы:
$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 \neq \mu_2.$$
- Критическая область является двусторонней. По таблице t-распределения находим $\alpha/2 = 0.025$, $1 - \alpha/2 = 0.975$, число степеней свободы $n_1 + n_2 - 2 = 10 + 7 - 2 = 15$ и определяем критические точки $t_{cr} = \pm 2.13$. Критическая область имеет вид $(-\infty; -2.13) \cup (2.13; +\infty)$.
- Значение статистики критерия равно
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{15 - 12}{\sqrt{\frac{1.6}{10} + \frac{1.6}{7}}} \approx 4.81,$$
так как $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{2\cdot 9+1\cdot 6}{15} = 1.6$.
- Вывод. Так как $t \in (-\infty; -2.13) \cup (2.13; +\infty)$, то основная гипотеза H_0 отвергается.

Ответ: при данном уровне значимости и такой альтернативе гипотеза отвергается.

Дисперсии неизвестны и не предполагаются равными

- В самом общем случае, когда дисперсии неизвестны и не равны, точный критерий для проверки гипотезы о равенстве средних указать трудно. В этом случае пользуются приблизительными формулами.
- Как и следовало ожидать для проверки гипотезы применяется t-статистика, в которой вместо теоретических значений дисперсий стоят выборочные оценки, то есть статистика критерия имеет вид

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Статистика критерия

Важно!

В случае, когда дисперсии неизвестны и не предполагаются равными, для проверки гипотезы о равенстве разности средних некоторому значению применяется статистика

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

Статистика близка к t-распределению с числом степеней свободы $\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$.

Для упрощения число степеней свободы вычисляют по формуле $\min(n_1 - 1, n_2 - 1)$.

Содержание

Доверительный интервал для разности средних(дисперсии известны)

Для полноты картины построим доверительный интервал для разности средних двух генеральных совокупностей.

Теорема

Доверительный интервал для разности средних, когда дисперсии генеральных совокупностей известны имеет вид

$$((\bar{x}_1 - \bar{x}_2) - \Delta; (\bar{x}_1 - \bar{x}_2) + \Delta),$$

где $\Delta = z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, z_α*-это квантиль нормального распределения уровня* $1 - \frac{\alpha}{2}$.

Пример

Даны две нормальные выборки со следующими характеристиками

	объем выборки	выборочное среднее	дисперсия
X	9	25	2
Y	6	21	1

Построим 90%-доверительный интервал для разности средних.

Решение.

- Выборочные средние $\bar{x}_1 = 25$ и $\bar{x}_2 = 21$, а дисперсии равны $\sigma_1^2 = 2$ и $\sigma_2^2 = 1$.
- По таблице нормального распределения находим $1 - \frac{\alpha}{2} = 0.95$ и определяем квантиль $z_\alpha = 1.64$. Теперь можем найти точность $\Delta = z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.64\sqrt{\frac{2}{9} + \frac{1}{6}} \approx 1.02$.
- Искомый 90%-доверительный интервал имеет вид $((\bar{x}_1 - \bar{x}_2) - \Delta; (\bar{x}_1 - \bar{x}_2) + \Delta) = (4 - 1.02; 4 + 1.02) = (2.98; 5.02)$.

Ответ: $(2.98; 5.02)$.

Дисперсии неизвестны, но предполагаются равными

Теорема

В случае, когда дисперсии неизвестны, но предполагаются равными, точность доверительного интервала находится по формуле

$$\Delta = t_\alpha \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}},$$

где $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$, t_α*-это квантиль распределения Стьюдента уровня* $1 - \frac{\alpha}{2}$ *с* $n_1 + n_2 - 2$ *степенью свободы.*

Содержание

Равенство средних в случае зависимых (парных) выборок

- В предыдущих параграфах выборки были независимыми, а здесь выборка часто одна и та же, просто в разные моменты времени. При выводе формул факт независимости существенно использовался, например, дисперсия разности средних в сумму дисперсий распадалась.
- Как правило, парные данные возникают, когда работают с одним и тем же набором объектов и наблюдения над ними производят дважды (до и после некоторого воздействия/эксперимента). Требуется выяснить, есть ли эффект от этого воздействия.
- Формализуем задачу следующим образом. Пусть имеется совокупность n пар наблюдений $(x_1, y_1), \ldots, (x_n, y_n)$. Составим разности $d_i = y_i - x_i$ и проверим гипотезу о равенстве нулю среднего разностей μ_d:

$$H_0 : \mu_d = 0; \quad H_1 : \mu_d \neq 0.$$

Статистика критерия

Теорема

Для проверки гипотезы о равенстве средних в парных выборках применяется следующая статистика

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \sim t(n-1),$$

где
d - разность между двумя значениями в одной паре,
$\bar{d}$ - выборочное среднее для парных разностей,
μ_d - среднее для парных разностей генеральной совокупности,
s_d - стандартное отклонение разностей для выборки,
n - количество пар.

Стандартное отклонение разностей для выборки можно вычислить не только по обычной формуле, для ручного счёта может пригодиться:

$$s_d = \sqrt{\frac{1}{n-1}\left(\sum d^2 - n\bar{d}^2\right)} = \sqrt{\frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right)}.$$

Допущения

Важно помнить условия (ограничения) на использование вышеприведённого теста.

- Все d_i взаимно независимы;
- Предположим, что $d_i = \theta + \varepsilon_i$, где θ - неизвеснтый оцениваемый параметр(эффект воздействия), а $\varepsilon_i \sim i.i.d. N(0; \sigma^2)$ (белый шум).

При этих допущениях задача сводится к проверке гипотезы о среднем для одной выборки.

Попробуйте самостоятельно

Пример

10 абитуриентов пришли на подготовительные курсы по ЕГЭ и написали тестирование в начале обучения и после. Результаты теста приведены в таблице

	1	2	3	4	5	6	7	8	9	10
До	7	6	5	4	6	2	10	3	8	5
После	9	6	4	5	7	4	10	6	9	6
Разность d	2	0	-1	1	1	2	0	3	1	1

Проверим гипотезу об отсутствии влияния подготовительных курсов на подготовку абитуриентов на уровне значимости 0.01.

Решение

- Формулируем гипотезы: $H_0 : \mu_d = 0; \quad H_1 : \mu_d \neq 0.$
- Критическая область является двусторонней. По таблице t-распределения находим $\alpha/2 = 0.005$, $1 - \alpha/2 = 0.995$, число степеней свободы $n - 1 = 9$ и определяем критические точки $t_{cr} = \pm 3.25$. Критическая область имеет вид $(-\infty; -3.25) \cup (3.25; +\infty)$.
- Вычислим значение статистики критерия. $\bar{d} = \frac{\Sigma d}{n} = \frac{10}{10} = 1$, $s_d = \sqrt{\frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right)} = \sqrt{\frac{22 - \frac{1}{10}(10)^2}{10-1}} = 1.15$. Значение статистики критерия равно $t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{1-0}{\frac{1.15}{\sqrt{10}}} = 2.75$.
- Вывод. Так как $t \notin (-\infty; -3.25) \cup (3.25; +\infty)$, то основная гипотеза H_0 не отвергается.
  Так как альтернативы была двухсторонней, то минимальный уровень значимости равен $2 \cdot (1 - t^{-1}(2.75)) \approx 0.022$.

Ответ: при данном уровне значимости и такой альтернативе гипотеза не отвергается.

Содержание

Доверительный интервал для разности средних

Теорема

Доверительный интервал для среднего разностей имеет вид:

$$\bar{d} - \Delta < \mu_d < \bar{d} + \Delta.$$

Точность оценки находится по формуле $\Delta = t_\alpha \frac{s_d}{\sqrt{n}}$, *где* d *- разность между двумя значениями в одной паре,* $\bar{d}$*- среднее для парных разностей для выборки,* μ_d *- среднее для парных разностей генеральной совокупности,* s_d *- стандартное отклонение разностей для выборки,* n *- количество пар.*

Пример

10 абитуриентов пришли на подготовительные курсы по ЕГЭ и написали тестирование в начале обучения и после. Результаты теста занесены в таблице

	1	2	3	4	5	6	7	8	9	10
До	7	6	5	4	6	2	10	3	8	5
После	9	6	4	5	7	4	10	6	9	6
Разность d	2	0	-1	1	1	2	0	3	1	1

Построим 99%-доверительный интервал для разности средних.

Решение

- Среднее значение разностей равно $\bar{d} = \frac{\sum d}{n} = \frac{10}{10} = 1$, а выборочное стандартное отклонение
$$s_d = \sqrt{\frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right)} = \sqrt{\frac{22-\frac{1}{10}(10)^2}{10-1}} \approx 1.15.$$
- По таблице распределения Стьюдента находим $1 - \frac{\alpha}{2} = 0.995$, у нас 9 степеней свободы, поэтому квантиль $t_\alpha = 3.25$. Теперь можем найти точность $\Delta = t_\alpha \frac{s_d}{\sqrt{n}} = 3.25\frac{1.15}{\sqrt{10}} \approx 1.18$.
- Искомый 99%-доверительный интервал имеет вид $(1 - 1.18; 1 + 1.18) = (-0.18; 2.18)$.

Ответ: $(-0.18; 2.18)$.

Замечание

Так как доверительный интревал накрывает ноль, то гипотеза о равенстве разности средних нулю не отвергается.

Содержание

Проверка гипотезы о равенстве дисперсий

- Сейчас мы узнаем, каким образом проверить гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей.
- Это интересно не только само по себе, но и необходимо нам, когда мы проверяем гипотезу о равенстве средних и предполагаем, что дисперсии равны. Такие предположения надо проверять.
- Если генеральные совокупности имеют нормальное распределение, то гипотезу о равенстве их дисперсий можно проверить с помощью F-критерия, называемый также критерием Фишера.

Статистика критерия

Теорема

Для проверки гипотезы о равенстве дисперсий $H_0 : \sigma_1^2 = \sigma_2^2$ *используется статистика* $F = \frac{s_1^2}{s_2^2}$, *которая имеет распределение Фишера с числом степеней свободы числителя* $n_1 - 1$ *и знаменателя* $n_2 - 1$.

Здесь σ_1^2 *и* σ_2^2*- дисперсии генеральных совокупностей,* s_1^2 *и* s_2^2 *- выборочные дисперсии,* n_1 *и* n_2 *- объемы выборок. Без ограничения общности считаем, что* $s_1^2 > s_2^2$.

Пример

Для нормальных выборок объемами 9 и 17 известны выборочные дисперсии 5 и 4 соответственно. Проверим гипотезу о равенстве дисперсий, на уровне значимости $\alpha = 0.05$.

Решение

- *Формулируем гипотезы:* $H_0 : \sigma_1^2 = \sigma_2^2; \quad H_1 : \sigma_1^2 > \sigma_2^2.$
- *Критическая область является правосторонней. По таблице F-распределения находим* $\alpha = 0.05$*, число степеней свободы* $n_1 - 1 = 8$ *и знаменателя* $n_2 - 1 = 16$ *и определяем критические точки* $F_{cr} = 2.59$*. Критическая область имеет вид* $(2.59; +\infty)$.
- *Значение статистики критерия равно* $F = \frac{s_1^2}{s_2^2} = 1.25$.
- *Так как* $F \notin (2.59; +\infty)$*, то основная гипотеза* H_0 *не отвергается.*
- *Минимальный уровень значимости равен* ≈ 0.33.

Ответ: *при данном уровне значимости и такой альтернативе гипотеза не отвергается.*

Содержание

Проверка гипотезы о равенстве долей

- Перейдем к сравнению долей признака в двух генеральных совокупностях. То есть мы хотим сравнить долю p_1 некоторого признака в первой генеральной совокупности с долей этого признака p_2 во второй генеральной совокупности. И для этого мы научимся проверять гипотезу $H_0 : p_1 = p_2$.
- Здесь мы предполагаем, что выборки независимы и для них выполняются условия $n\hat{p} \geq 5$ и $n\hat{q} \geq 5$. Иначе выводы будут ненадёжными.
- Проверка гипотезы о равенстве долей осуществляется с помощью z-статистики.

Статистика критерия

Теорема

Для проверки гипотезы о равенстве двух долей некоторому значению используется статистика

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}},$$

распределение которой стремится к нормальному закону при $n \to \infty$.

Здесь p_1 и p_2 - это генеральные доли признака,
n_1 и n_2 - объемы выборок,
m_1 и m_2 - число «успехов» в каждой выборке,
$\hat{p}_1 = \frac{m_1}{n_1}$ и $\hat{p}_2 = \frac{m_2}{n_2}$ - доля «успехов» в каждой выборке,
$\hat{p} = \frac{m_1+m_2}{n_1+n_2}$ - общая доля «успехов» в двух выборках.

Пример

Пусть из 200 случайно отобранных студентов экономического факультета 86 ездят в университет на велосипеде в тёплое время года, а из 300 студентов химического факультета таких оказалось 135. Проверим гипотезу о равенстве соответствующих долей ($\alpha = 0.05$).

Решение

- *Формулируем гипотезы:* $H_0 : p_1 = p_2; \quad H_1 : p_1 \neq p_2$.
- $\hat{p}_1 = \frac{m_1}{n_1} = \frac{86}{200} = 0.43$, $\hat{p}_2 = \frac{m_2}{n_2} = \frac{135}{300} = 0.45$,
  $\hat{p} = \frac{m_1+m_2}{n_1+n_2} = \frac{86+135}{200+300} = 0.442$.
- *Значение статистики критерия равно*
  $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} = \frac{0.42-0.45}{\sqrt{\frac{0.442\cdot 0.558}{200} + \frac{0.442\cdot 0.558}{300}}} \approx -0.441$.
- *По таблице z-распределения находим $\alpha/2 = 0.025$, $1 - \alpha/2 = 0.975$ и определяем критические точки $z_{cr} = \pm 1.96$. Критическая область имеет вид $(-\infty; -1.96) \cup (1.96; +\infty)$.*
- *Вывод. Так как $z \notin (-\infty; -1.96) \cup (1.96; +\infty)$, то основная гипотеза H_0 не отвергается.*

Содержание

Доверительный интервал для разности двух долей

Теорема

Асимптотический доверительный интервал для разности между долями некоторого признака в двух независимых нормальных генеральных совокупностях имеет вид

$$((\hat{p}_1 - \hat{p}_2) - \Delta) < p_1 - p_2 < ((\hat{p}_1 - \hat{p}_2) + \Delta).$$

Точность оценки Δ *вычисляется по формуле:* $\Delta = z_\alpha \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$*, где* $\hat{q}_1 = 1 - \hat{p}_1$*,* $\hat{q}_2 = 1 - \hat{p}_2$*, а* z_α*-это квантиль нормального распределения уровня* $1 - \frac{\alpha}{2}$*.*

Пример

Пример

Пусть объемы выборок равны $n_1 = 100$ *и* $n_2 = 200$, *выборочные доли* $\hat{p}_1 = 0.2$, $\hat{p}_2 = 0.25$. *Построим* 95%*-доверительный интервал для разности долей генеральных совокупностей.*

Решение

- *Выборочные доли* $\hat{p}_1 = 0.2$ *и* $\hat{p}_2 = 0.25$, *поэтому* $\hat{q}_1 = 1 - \hat{p}_1 = 0.8$ *и* $\hat{q}_2 = 1 - \hat{p}_2 = 0.75$. *Проверьте, что условия надёжности использования этих формул* $n\hat{p}_i \geq 5$ *и* $n\hat{q}_i \geq 5$ *выполнены!*
- *По таблице нормального распределения находим* $1 - \frac{\alpha}{2} = 0.975$ *и определяем квантиль* $z_\alpha = 1.96$. *Теперь можем найти точность* $\Delta = z_\alpha \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = 1.96 \sqrt{\frac{0.2 \cdot 0.8}{100} + \frac{0.25 \cdot 0.75}{200}} \approx 0.099$.
- *Искомый* 95%*-доверительный интервал имеет вид* $((\hat{p}_1 - \hat{p}_2) - \Delta; (\hat{p}_1 - \hat{p}_2) + \Delta) = (0.05 - 0.099; 0.05 + 0.099) = (-0.049; 0.0149)$.

Содержание

Где и что почитать?

Тема. Проверка гипотез для двух выборок (зависимые и независимые выборки). ([Ф-Л], главы 16; [Т-М], §5.4.2-5.4.3).

- Фадеева Л. Н., Лебедев А. В., Теория вероятностей и математическая статистика: учебное пособие. - 2-е изд., перераб. и доп. - М.: Эксмо, 2010. - 496 с. – (Новое экономическое образование).
- Тюрин Ю. Н., Макаров А.А., Анализ данных на компьютере: учебное пособие. - 4-е изд., перераб. - М.: ИД Форум, 2008. - 368 с., ил. - (Высшее образование).