Тестване на хипотези за два параметъра

Сравняване на две средни стойности

Зависими извадки

Зависими извадки - когато могат да комбинират по двойки (в някакъв смисъл)

Пример:

При измерване ефективност на нова диета, се претеглят едни и същи хора, подложени на диетата, преди и след прилагането й.

Проверка на хипотези за **две зависими** извадки

Комбинираме двете извадки в една= разликата от двете

- Използваме следната статистика

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

където $\bar{d}$ е средната стойност на разликата , s_d е стандартното отклонение на разликата, и n е броя на двойките (разликите)

Пример

- За да се изучат дневните тарифи за коли под наем на леки автомобили на компаниите Hertz и Avis в САЩ е направена случайна извадка от 8 големи града и е записана информацията за тарифата в следната таблица. При ниво на значимост 0,05 може ли да се твърди, че има разлика в тарифата на двете компании?

Пример

Град	Hertz (цена в $)	Avis (цена в $)
Атланта	42	40
Чикаго	56	52
Кливланд	45	43
Денвър	48	48
Хонолулу	37	32
Канзас	45	48
Маями	41	39
Сеатъл	46	50

Решение:

Образуваме нова извадка от разликите

Град	Hertz	Avis	*d*	*(d-средно)²*
Атланта	42	40	2	1
Чикаго	56	52	4	9
Кливланд	45	43	2	1
Денвър	48	48	0	1
Хонолулу	37	32	5	16
Канзас	45	48	-3	16
Маями	41	39	2	1
Сиатъл	46	50	-4	25
			Сума= 8	сума=70

Средно =d=1 дисп.=s^2_d=10

1. Нулева и алтернативна хипотеза

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

или $H_0 : \mu_d = 0$ $H_1 : \mu_d \neq 0$

2. Ниво на значимост

$\alpha = 0{,}05$

3. Статистика, извадково разпределение

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

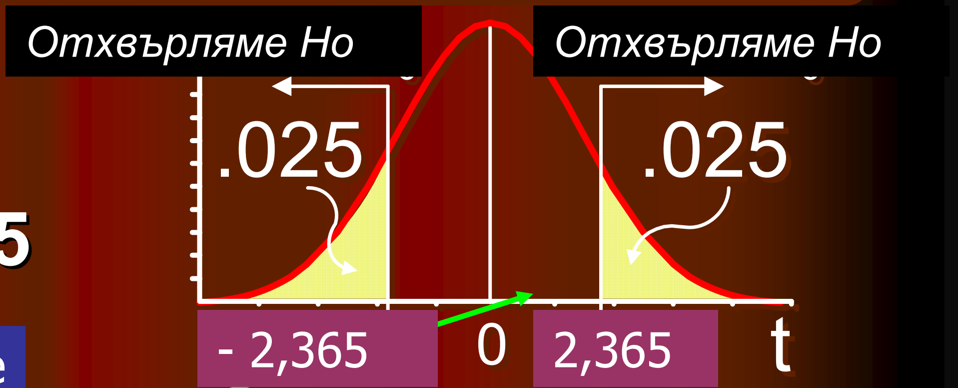
4. Критична област

Критична област

(- ∞, -2,365) и (2,365 ; ∞)

$\alpha = 0{,}05$ $\alpha/2 = 0{,}025$

t (7) разпределение



5. Извод

t=0,894 не е в критичната област, затова не отхвърляме H_0

6. Интерпретация на извода

Няма достатъчно основание на считаме, че има разлика в цените на Hertz и Avis.

Независими извадки (голям обем)

Предположения

1. Двете извадки са независими

2. Обемите на двете извадки са големи

$n_1 > 30 \qquad n_2 > 30$

Алтернативи

отхвърляме H_0 ако:

$H_1 : \mu_1 - \mu_2 > D_0$	$z > z_\alpha$
$H_1 : \mu_1 - \mu_2 < D_0$	$z < -z_\alpha$
$H_1 : \mu_1 - \mu_2 \neq D_0$	$\lvert z \rvert > z_{\alpha/2}$, т.е. $z > z_{\alpha/2}$ или $z < -z_{\alpha/2}$

Статистика

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Използваме s_1 и s_2 ако σ_1 и σ_2 не са известни

Пример:

Твърди се, че средната възраст на студентите от хуманитарните специалности е различна от средната възраст на студентите от техническите специалности. За да се провери твърдението, са направени две случайни извадки от по 50 студенти от хуманитарни специалности и 50 студенти от технически специалности и е записана възрастта им. От данните е получено, че средната възраст на първата група е 21 година, а средната възраст на втората е 20 години със стандартни отклонения съответно 4 и 2,5 години. При ниво на значимост 0,04, тествайте твърдението за различие в средната възраст на двете групи студенти .

Интерпретация на данните:

$$n_1 = 50 \qquad n_2 = 50$$

$$\bar{x}_1 = 21 \qquad \bar{x}_2 = 20$$

$$s_1 = 4 \qquad s_2 = 2,5$$

$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

статистика

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

$$\frac{21 - 20}{\sqrt{\frac{4^2}{50} + \frac{2,5^2}{50}}} = 1,499$$

Критична област : z<-2,05 или z>2,05

Z=1.499 не е в критичната област

Не отхвърляме H₀

Няма достатъчно основание да се твърди, че има съществена разлика във възрастта.

Независими извадки с малък обем

Случай1: Равни дисперсии

Предположения :

1. Двете популации са нормално разпределени
2. Обемът поне на едната от двете извадки е малък (n< 30 или m<30)
3. Двете популации имат еднакви (макар и неизвестни) дисперсии σ
4. Двете извадки, взети от тези популации са независими

Разликата на двете извадкови средни е $N(\mu_X - \mu_Y, \sigma\left(\frac{1}{n} + \frac{1}{m}\right))$

$$H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 > D_0$$

$$H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 < D_0$$

$$H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 \neq D_0$$

Статистика: $$t = \frac{(\overline{x} - \overline{y}) - D_0}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$ където $$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{(n+m-2)}$$

- Използва се **t-разпределението с n+m-2 степени на свобода.**

Пример:

Ботаник се интересува от влиянието на определен вид тор върху растежа на стъблото на грах. Използвайки 16-дневни растения, той измерва дължината на стеблата на 11 растения, на които средното изменение на дължината е 1,03 с дисперсия 0,24. Той третира 13 растения със съответния тор в продължение на 16 дни, измерва стеблата им и намира, че средното изменение в дължините е 1,66 с дисперсия 0,35. Може ли да се твърди, че този вид тор подобрява растежа? Предполагаме една и съща популационна дисперсия.

Интерпретация на данните:

$$n = 11,\ \bar{x} = 1{,}03,\ s_X^2 = 0{,}24, \quad m = 13,\ \bar{y} = 1{,}66,\ s_Y^2 = 0{,}35$$

$H_0: \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 - \mu_2 < 0$

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{(n+m-2)} \rightarrow s_p^2 = \frac{(10)(0{,}24) + (12)(0{,}35)}{(11+13-2)} = 6{,}6$$

$$\mathrm{t} = \frac{(\overline{\mathrm{x}} - \overline{\mathrm{y}}) - \mathrm{D}_0}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$\mathrm{t} = \frac{(1{,}03 - 1{,}66) - 0}{\sqrt{0{,}3\left(\frac{1}{11} + \frac{1}{13}\right)}} = -0{,}47$$

Критична област:

Нека α=0,05

Използваме t-рапределението при степени на свобода= 22

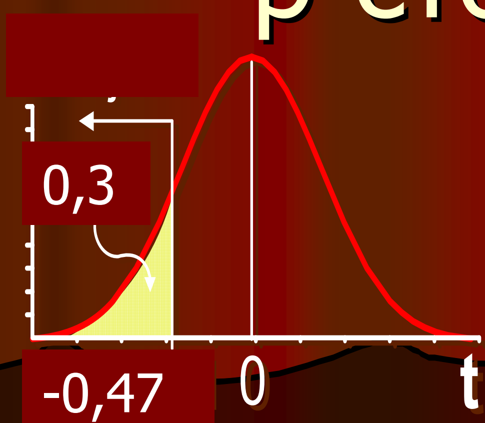
Критична област: t<-1,7171

извод:

t= - 0,47 не е в критичната област, затова не отхвърляме хипотезата, т.е. Няма статистическо основание да отхвърлим твърдението, че този вид тор подобрява растежа.

p-стойност:

t= - 0,47



Лявостранен тест

p-стойността на теста =0,3

извод: 0,3>0,1, затова не отхвърляме хипотезата.

Независими извадки с малък обем

Случай 2: Различни дисперсии

Предположения :

1. Двете популации са <u>нормално разпределени</u>
2. Обемът поне на едната от двете извадки е малък (n< 30 или m<30)
3. Двете популации имат <u>различни неизвестни дисперсии</u>
4. Двете извадки, взети от тези популации са <u>независими</u>

$$H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 > D_0$$
$$H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 < D_0$$
$$H_0 : \mu_1 - \mu_2 = D_0 \qquad H_1 : \mu_1 - \mu_2 \neq D_0$$

Статистика:

$$t = \frac{(\overline{x} - \overline{y}) - D_0}{\sqrt{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)}}$$

- Използва се **t-разпределението с к степени на свобода,** където к=min(n-1,m-1).

Пример:

Два града, А и Б, се разделят от една река. Местната преса публикува, че леките автомобили в град А са на повече километри от тези в град Б. За да се провери твърдението се избират по случаен начин 40 автомобила от А и се установява, че средно те са на 38 000 км със стандартно отклонение 6 000 км. Избрани са по случаен начин и 35 автомобила от Б и се намира, че средно те са на 35 000 км със стандартно отклонение 7 000 км. При ниво на значимост 0,01 може ли да се търди, че колите в А са на повече километри, ако километрите са нормално разпределени?

Стъпка 1: Нулева и алтернативна хипотеза:

$$H_0:\ \mu_А = \mu_Б\ ; \qquad H_1:\ \mu_А > \mu_Б$$

Стъпка 2: Ниво на значимост: α=0,01

стъпка 3: Статистика:

Обемите и на двете извадки са >30, използваме *z*.

$$z = \frac{38000 - 35000}{\sqrt{\frac{(6000)^2}{40} + \frac{(7000)^2}{35}}} = 1,98$$

стъпка 4: Критична област: $z > 2,33$

Критична стойност $z_{0,01} = 2,33$

1,98 не попада в критичната област, не отхвърляме нулевата хипотеза. Няма основание да твърдим, че колите в град А са на повече километри.

Пример:

Разглеждаме предишния пример, но при избрани по 15 автомобила от всеки град.

- Стъпка 1: Нулева и алтернативна хипотеза.
- $H_0: \ \mu_А = \mu_Б; \qquad H_1: \ \mu_А > \mu_Б$

Стъпка 2: Ниво на значимост: α=0,01

Стъпка 3: Статистика:

Тъй като обемите и на двете извадки са <30, то е необходимо да сравним статистически популационните дисперсии, дали са равни или не, т.е да тестваме хипотезата $H_0: \ \sigma_А = \sigma_Б; \qquad H_1: \sigma_А \neq \sigma_Б$

Как да тестваме хипотеза за две дисперсии

$H_0: \sigma_А = \sigma_Б$; $H_1: \sigma_А \neq \sigma_Б$

Статистика:

$$F = \frac{s_1^{\ 2}}{s_2^{\ 2}}$$

F-разпределение, има две степени на свобода- в числителя и в в знаменателя

Изчисляване на p-стойността може да използвате :

davidmlane.com/hyperstat/**F_table**.html

Обратно към задачата: извадковите стандартни откл. са 7000 и 6000

$$F = \frac{7000^2}{6000^2} = \frac{49}{36} = 1{,}361$$

Степени на свобода:

в числител 35-1=34

В знаменател 40-1=39

P-стойност = 0,49697 няма основание да отхвърлим хипотезата

Тогава използваме теста за равни дисперсии, т.е.

Използваме *t-разпределение и статистиката*

$$t = \frac{(\overline{x} - \overline{y}) - D_0}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{(n+m-2)}$$

$$s_p^2 = \frac{(14)(6000)^2 + (14)(7000)^2}{(15+15-2)} = 42\ 500\ 000$$

S_p=6519

$$t = \frac{(38000 - 35000) - 0}{6519\sqrt{\left(\frac{1}{15} + \frac{1}{15}\right)}} = 1,2605$$

Стъпка 4: Намираме *t(28)* от таблицата и построяваме критичната област.

$$t_{0,01}(28) = 2,467$$

$$t > 2,467$$

t=1,2605 не в критичната област; не отхвърляме нулевата хипотеза. Няма основание да се счита, че автомобилите в А са на повече километри.

Тестване на хипотези за средните на две популации

начало

Двете извадки са зависими?

да

t тест за двойки (извадките са от нормални популации):

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

където ст. на свобода = n – 1.

не

И двете n_1 и n_2 са големи (> 30)

да

z тест (нормално разпр.):

$$z = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(ако σ_1^2 и σ_2^2 са неизвестни, използвай оценките им s_1^2 и s_2^2)

не

Двете популации са нормално разпределени

не

Използвай непараметрични методи (не са включени в този курс).

да

следващ слайд

продължение

Двете популации са нормално разпределени

не → Използвай непараметрични методи (не са включени в този курс).

да ↓

σ_1 σ_2 са известни

да → z тест (нормално разпр.):

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

(s_1^2 и s_2^2 не могат да бъдат използвани за оценки на σ_1^2 и σ_2^2.)

не ↓

Проверяваме хипотеза за $\sigma_1^2 = \sigma_2^2$

отхвърляме $\sigma_1^2 = \sigma_2^2$ → t тест (извадките са от нормална популация):

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

където ст. на св. = по-малкото от $n_1 - 1$ и $n_2 - 1$.

Не отхвърляме $\sigma_1^2 = \sigma_2^2$ ↓

t тест (извадките са от нормална популация):

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

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

и ст. на свобода= $n_1 + n_2 - 2$

Сравняване на две пропорции

Независими извадки

Разглеждаме опити на Бернули и две биномно разпределени популации (алтернативни популации)

Първа извадка: обем n и брой успехи в нея x =>намираме $\hat{p}_1 = \dfrac{x}{n}$

Втора извадка: **обем m и брой успехи в нея y =>намираме** $\hat{p}_2 = \dfrac{y}{m}$

Предположения :

$$n\hat{p}_1 \geq 10, \quad n(1-\hat{p}_1) \geq 10$$

$$m\hat{p}_2 \geq 10, \quad m(1-\hat{p}_2) \geq 10$$

H_0: $p_1 - p_2 = D_0$ H_1: $p_1 - p_2 \neq D_0$

статистика

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n-1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{m-1}}}$$

- "Несемейните служители отсъстват по-често от работа отколкото семейните".

За целта се избират 250 семейни служители, от които се оказва че 22 са отсъствали повече от 5 дни последната година, докато при случайно избрани 300 несемейни служители се оказало, че 35 са отсъствали повече от 5 дни. При ниво на значимост 0,05 какво може да кажете за твърдението?

Интерпретация на данните:

Първа извадка: обем n=250 и брой успехи в нея x =22>намираме

$$\hat{p}_1 = \frac{x}{n} = \frac{22}{250} = 0{,}088$$

Втора извадка: **обем m=300 и брой успехи в нея y =35>намираме**

$$\hat{p}_2 = \frac{y}{m} = \frac{35}{300} = 0{,}1167$$

$H_0: p_1 = p_2$ $H_1: p_1 < p_2$

Предположения :

$$n\hat{p}_1 = 22 \geq 10, \quad n(1-\hat{p}_1) = 228 \geq 10$$

$$m\hat{p}_2 = 35 \geq 10, \quad m(1-\hat{p}_2) = 265 \geq 10$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1}}} = \frac{0{,}088 - 0{,}1167}{\sqrt{\frac{0{,}088(0{,}912)}{249} - \frac{0{,}1167(0{,}8833)}{299}}} = -1{,}1112$$

- α=0,05 и критичната област е

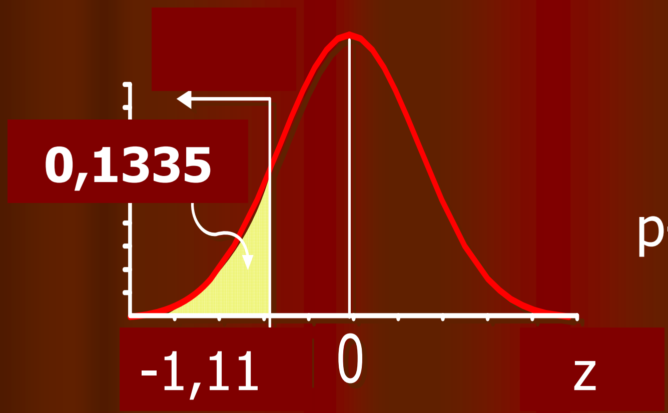
  $(-\infty, \ -1{,}65)$.

Не отхвърляме хипотезата, т.е. няма основание да смятаме, че несемейните отсъстват по-често от работа.

p-стойност на теста

$$z=\frac{(\hat{p}_1-\hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}}}=-1,1112$$

Лявостранен тест



p-стойност на теста=0,1335>0,1

Извод: не отхвърляме хипотезата

Ако отхвърлим хипотезата, то ще допуснем грешка от първи род= 0,1335