

ЧАСТ V

ПРОВЕРКА НА СТАТИСТИЧЕСКИ ХИПОТЕЗИ

Учебни цели:

- Проверка на хипотеза за средна аритметична на генерална съвкупност
- Проверка на хипотеза за относителен дял на генерална съвкупност
- Проверка на хипотеза за разлика между средни аритметични от две независими извадки
- Проверка на хипотеза за разлика между относителни дялове от две независими извадки
- Еднофакторен дисперсионен анализ

1. Проверка на хипотеза за средна на генерална съвкупност

Пример 1.

Проведено е репрезентативно извадково изследване за оценка на средния доход на лице в трудоспособна възраст в едно населено място. Направена е извадка от 25 лица чрез прост случаен възвратен подбор. Установен е среден месечен доход на лицата в извадката 540 лв и оценка на стандартното отклонение по дохода в размер на 160 лв. Да се провери при риск за грешка 5% дали средният доход на лице в населеното място се различава съществено от 600 лв.

Решение:

1) Формулиране на хипотезите.

$H_0: \bar{x}_0 = 600$ /Средният доход на лице в трудоспособна възраст в населеното място не се различава съществено от 600 лв./

$H_1: \bar{x}_0 \neq 600$ /Средният доход на лице в трудоспособна възраст в населеното място се различава статистически значимо от 600 лв./

2) Определяне на критичното равнище на значимост.

Проверката ще се осъществи при риск за грешка $\alpha=0,05$.

3) Определяне на метода (теста) на проверката.

Тъй като обемът на извадката е под 30 единици, то емпиричната характеристика на хипотезата има разпределение на Стюдънт (t-критерий):

$$t_{em} = \frac{|\bar{x} - x_0|}{\mu_{\bar{x}}} = \frac{|\bar{x} - x_0|}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

4) Осигуряване на данни от представителна извадка.

$\bar{x} = 540$ $n=25$ $\hat{\sigma} = 160$ $x_0 = 600$

5) Изчисляване на емпиричната характеристика на хипотезата.

$$t_{EM} = \frac{\frac{|\bar{x} - x_0|}{\hat{\sigma}}}{\frac{1}{\sqrt{n}}} = \frac{\frac{|540 - 600|}{160}}{\frac{1}{\sqrt{25}}} = \frac{40}{160:5} = \frac{40}{32} = 1,25$$

6) Определяне на теоретичната характеристика на хипотезата.

Критичната област на хипотезата (област на отхвърляне на H_0) е двустранна поради двупосочното неравенство, дефинирано в H_1 . Тогава табличната стойност ще се определи при избрания риск за грешка (5%), двустранна критична област и 24 степени на свобода за t-разпределение:

$$t_T \left[\begin{matrix} \alpha=0,05 \\ \text{КО-двустр.} \\ \phi=n-1=24 \end{matrix} \right] = 2,06$$

6) Вземане на решение.

Тъй като $t_{EM} < t_T$ ($1,25 < 2,06$), то по данни от тази извадка не може да се твърди, че средният доход на лице в трудоспособна възраст в населеното място се различава съществено от 600лв.

2. Проверка на хипотеза за разлика между средни аритметични величини от две независими извадки.

Пример 2.

Проведено е репрезентативно извадково изследване за оценка на потреблението на една марка кашкавал в две населени места – “А” и “Б”. В едното от тях (А) се провежда локална рекламна кампания, а в другото (Б) не се провежда. Направени са две случайни извадки от домакинства в двете населени места, съответно 60 и 50 домакинства. Установени са месечните количества, потребявани средно от домакинство в двете извадки, съответно 2,5 кг и 2,2 кг, при стандартни отклонения 0,5 кг и 0,6 кг. При риск за грешка 5% да се провери дали потреблението на кашкавал от тази марка средно на домакинство в населено място “А” е по-високо от това в населено място “Б”.

Решение:

1) Формулиране на хипотезите.

$H_0: \bar{x}_{1(0)} = \bar{x}_{2(0)}$ / Средното потребление в “А” (насел.място 1) не се различава съществено от средното потребление в “Б”/

$H_1: \bar{x}_{1(0)} > \bar{x}_{2(0)}$ / Средното потребление в “А” (населено място 2) е по-високо от средното потребление в “Б”/

2) Определяне на критичното равнище на значимост.

Проверката ще се осъществи при риск за грешка $\alpha=0,05$.

2) Определяне на критичното равнище на значимост.

Проверката ще се осъществи при риск за грешка $\alpha=0,05$.

3) Определяне на метода на проверката.

Тъй като общият обемът на извадката е под 30 единици следва да се използва t-критерий при Стюдънтово разпределение на емпиричната характеристика на хипотезата:

$$t_{em} = \frac{|p - p_0|}{\mu_p} = \frac{|p - p_0|}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{|p - p_0|}{\sqrt{\frac{p(1-p)}{n} \cdot \frac{n}{n-1}}} = \frac{|p - p_0|}{\sqrt{\frac{p(1-p)}{n-1}}}$$

4) Осигуряване на данни от представителна извадка.

$$n=25 \quad m=5 \quad p = \frac{m}{n} = \frac{5}{25} = 0,2 \quad (20\% \text{ безработни в извадката})$$

5) Изчисляване на емпиричната характеристика на хипотезата.

$$t_{em} = \frac{|p - p_0|}{\sqrt{\frac{p(1-p)}{n-1}}} = \frac{|0,2 - 0,15|}{\sqrt{\frac{0,2 \cdot 0,8}{24}}} = \frac{0,05}{0,0817} = 0,612$$

6) Определяне на теоретичната характеристика на хипотезата.

Критичната област на хипотезата е едностранна поради еднопосочното неравенство, дефинирано в H_1 . Тогава табличната стойност ще се определи при избрания риск за грешка (5%), едностранна критична област и 24 степени на свобода за t-разпределение:

$$t_T \left[\begin{array}{l} \alpha=0,05 \\ \text{КО-едностр.} \\ \phi=n-1=24 \end{array} \right] = 1,71$$

7) Вземане на решение.

Тъй като $t_{em} < t_T$ ($0,612 < 1,71$), то при данните от тази извадка няма основание да се твърди, че относителният дял на безработните лица в населеното място е по-голям от 15%.

4. Проверка на хипотеза за разлика между относителни дялове при две независими извадки.

Пример 4.

При проведеното репрезентативно изследване за оценка на потреблението (пример 2) в двете населени места е поставена задачата да се провери дали в населено място “А” (където се провежда локалната рекламна кампания) е налице по-висок дял на домакинствата, които потребяват кашкавал от изследваната марка. От 60-те наблюдавани домакинства в “А” марката се потребява от 9 домакинства, а от 50-те домакинства в “Б” – от 5 домакинства. Проверката да се направи при 1% риск за грешка.

Решение:

1) Формулиране на хипотезите.

$H_0: p_{1(0)} = p_{2(0)}$ / Относителният дял на домакинствата-потребители на марката в “А” не се различава от този дял в “Б” /

$H_1: p_{1(0)} > p_{2(0)}$ / Относителният дял на домакинствата-потребители в “А” е съществено по-висок от този дял в “Б” /

2) Определяне на критичното равнище на значимост.

Проверката ще се осъществи при риск за грешка $\alpha=0,01$.

3) Определяне на метода на проверката.

Тъй като общият обем на двете извадки е над 30 единици, може да се използва Z-критерий при нормално разпределение на емпиричната характеристика на хипотезата:

$$Z_{em} = \frac{|p_1 - p_2|}{\sigma_{p_1 - p_2}} = \frac{|p_1 - p_2| \sqrt{n_1 + n_2 - 2}}{\sqrt{(p_1 q_1 n_1 + p_2 q_2 n_2) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{където } q = 1 - p$$

5) Определяне на наличната информация от извадката.

$$n_1=60 \quad m_1=9 \quad p_1 = \frac{m_1}{n_1} = \frac{9}{60} = 0,15 \quad q_1 = 1 - 0,15 = 0,85$$

$$n_2=50 \quad m_2=5 \quad p_2 = \frac{m_2}{n_2} = \frac{5}{50} = 0,10 \quad q_2 = 1 - 0,1 = 0,9$$

6) Изчисляване на емпиричната характеристика на хипотезата.

$$Z_{em} = \frac{|0,15 - 0,10| \sqrt{60 + 50 - 2}}{\sqrt{(0,15 \cdot 0,85 \cdot 60 + 0,1 \cdot 0,9 \cdot 50) \left(\frac{1}{60} + \frac{1}{50} \right)}} = \frac{0,05 \sqrt{108}}{\sqrt{12,15 \cdot 0,0367}} = \frac{0,05 \cdot 10,4}{\sqrt{0,4455}} = \frac{0,52}{0,667} = 0,78$$

7) Определяне на теоретичната характеристика на хипотезата.

Критичната област на хипотезата е едностранна поради еднопосочното неравенство, дефинирано в H_1 . Тогава табличната стойност ще се определи при избрания риск за грешка (1%) и едностранна критична област за нормалното разпределение:

$$Z_T \left[\begin{matrix} \alpha=0,01 \\ \text{КО-едностр.} \end{matrix} \right] = 2,34$$

8) Вземане на решение.

Тъй като $Z_{em} < Z_T$ ($0,78 < 2,34$), то по данни от тези извадки няма основание да се твърди, че относителният дял на домакинствата-потребители на марката кашкавал в населено място “А” е различен от този в населено място “Б”.

5. Еднофакторен дисперсионен анализ

Пример 5.

Поставена е задачата да се провери при риск за грешка 5% дали съществува зависимост между отрасъла и броя на подразделенията на търговските дружества. Известно е, че извадката от 25 дружества е направена по метода на простия случаен подбор.

Решение:

Тъй като факторният признак (отрасъл) е неметриран, а резултативният (брой на подразделенията) е метриран, то наличието на зависимост може да се установи чрез еднофакторен дисперсионен анализ въз основа на данни от случайна извадка. Етапите на работа при еднофакторния ДА са следните:

Първо. Търговските дружества (единиците от съвкупността) се разпределят в групи, формирани по значенията на факторния признак “отрасъл”. Във всяка група се обособяват значенията на резултативния признак “брой на подразделенията” за попадналите в нея единици. За удобство на работата се съставя следната таблица:

Отрасъл	Брой подразделения
Промисленост	2, 1, 1, 0, 0, 1, 3, 1, 2, 1, 1, 2
Транспорт	0, 2, 3, 0, 2, 3, 0, 2
Строителство	1, 1, 0, 1, 0

Второ. Изчисляват се средните равнища по резултативния признак за всяка група, формирана по значенията на факторния. Използва се следната формула:

$$\bar{y}_i = (1/n_i) \sum_{j=1}^{n_i} y_{ij}$$

Използваните означения са: n – обем на извадката (общ брой на единиците); n_i – брой на единиците в група “ i ”; m – брой на категориите (значенията) на факторния признак; y_{ij} – значение на резултативния признак при единица “ j ” от група “ i ”; $\bar{y}_i$ – средна аритметична по резултативния признак в група “ i ”.

i	Отрасъл	y_{ij}	n_i	$\sum y_{ij}$	$\bar{y}_i$
1	Промисленост	2, 1, 1, 0, 0, 1, 3, 1, 2, 1, 1, 2	12	15	1,25
2	Транспорт	0, 2, 3, 0, 2, 3, 0, 2	8	12	1,50
3	Строителство	1, 1, 0, 1, 0	5	3	0,60
$m=3$		Σ	25	30	-

Общото средно равнище за цялата извадка е:

$$\bar{y}_0 = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij}}{n} = \frac{30}{25} = 1,2 \quad \text{където } n = \sum_{i=1}^m n_i$$

Трето. За всяка група се изчисляват стойностите на вътрешно-груповата девиация, като сума от квадратите на отклоненията на значенията на резултативния признак при отделните единици от тяхната обща (групова) средна. За изчисляването на вътрешногруповите девиации се използва подходяща работна таблица. Последната колона на тази таблица съдържа тези девиации за всяка от формираните групи:

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

i													Σ
1	2	1	1	0	0	1	3	1	2	1	1	2	
	0,75	-0,25	-0,25	-1,25	-1,25	-0,25	1,75	-0,25	0,75	-0,25	-0,25	0,75	
	0,56	0,06	0,06	1,56	1,56	0,06	3,06	0,06	0,56	0,06	0,06	0,56	8,22
2	0	2	3	0	2	3	0	2					
	-1,5	0,5	1,5	-1,5	0,5	1,5	-1,5	0,5					
	2,25	0,25	2,25	2,25	0,25	2,25	2,25	0,25					12,0
3	1	1	0	1	0								
	0,4	0,4	-0,6	0,4	-0,6								
	0,16	0,16	0,36	0,16	0,36								1,20

Четвърто. Чрез получените девиации (суми от квадратите на отклоненията) се изчислява оценката на вътрешногруповата дисперсия. Тя се получава при степени на свобода, равни на обема на извадката, намален с броя на категориите на факторния признак:

$$\sigma_B^2 = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n - m} = \frac{8,22 + 12,0 + 1,2}{25 - 3} = \frac{21,44}{22} = 0,975$$

Пето. Въз основа на груповите средни се изчислява междугруповата девиация. За целта се използва подходяща работна таблица, която е продължение на таблицата за изчисляване на груповите средни:

i	...	n _i	$\bar{y}_i$	$\bar{y}_i - \bar{y}_0$	$(\bar{y}_i - \bar{y}_0)^2$	$(\bar{y}_i - \bar{y}_0)^2 n_i$
1	...	12	1,25	0,05	0,0025	0,03
2	...	8	1,50	0,30	0,0900	0,72
3	...	5	0,60	-0,60	0,3600	1,80
Σ		25	-	-	-	2,55

Чрез получената девиация се изчислява оценката на междугруповата дисперсия. Тя се получава при степени на свобода, равни на броя на категориите на факторния признак, намален с 1:

$$\sigma_M^2 = \frac{\sum_{i=1}^m (\bar{y}_i - \bar{y}_0)^2 n_i}{m-1} = \frac{2,55}{3-1} = \frac{2,55}{2} = 1,275$$

Шесто. Формулира се статистическа хипотеза, чрез която се проверява статистическата значимост на разликата между двете извадкови оценки на дисперсията - междугруповата и вътрешногруповата.

H₀: Липсва статистически значима разлика между двете дисперсии (това е свързано с липсата на статистически значима разлика между груповите средни: $\bar{y}_1 = \bar{y}_2 = \bar{y}_3$).

Нулевата хипотеза предполага липса на зависимост /влияние/ между признака “отрасъл” и признака “брой на подразделенията”.

H₁: Налице е статистически значима разлика между двете дисперсии (поради наличието на статистически значима разлика между груповите средни: $\bar{y}_1 \neq \bar{y}_2 \neq \bar{y}_3$).

Алтернативната хипотеза предполага наличие на закономерна зависимост между факторния и резултативния признак (т.е. поне за един отрасъл се наблюдава съществено различие на съответния му среден брой подразделения от средните нива в другите отрасли).

Проверката на хипотезата изисква изчисляване на емпиричната стойност на критерия на Фишер (F_{ем} величина):

$$F_{em} = \frac{\sigma_M^2}{\sigma_B^2} = \frac{1,275}{0,975} = 1,31$$

Седмо. Определя се теоретичната (таблична) стойност на F-критерия при следните условия:

- равнище на значимост $\alpha=0,05$;
- степени на свобода, равни на знаменателите на двете дисперсии:

$$\phi_1 = m-1 = 3-1 = 2 \quad \text{и} \quad \phi_2 = n-m = 25-3 = 22.$$

$$F_T \left[\begin{array}{l} \alpha = 0,05 \\ \phi_1 = 2 \\ \phi_2 = 22 \end{array} \right] = 3,44$$

Осмо. Решение относно хипотезите се взима след сравняване на емпиричната и теоретичната стойност на F-критерия:

Тъй като $F_{em} = 1,31 < 3,44 = F_T$, то по данни от тази извадка няма достатъчно основание да се отхвърли нулевата хипотеза. Приема се, че влиянието на признака “отрасъл” върху признака “брой на подразделенията” е несъществено от статистическа гледна точка, т.е. липсва статистически значима зависимост между двата признака.

 Решете самостоятелно следната задача

А) Да се провери при риск за грешка 5% дали съществува зависимост между отрасъла и броя на персонала в търговските дружества. Налице са данните за случайна извадка от 10 дружества:

Отрасъл	Брой наети
Промисленост	38, 17, 30, 63
Транспорт	10, 27, 50
Строителство	20, 35, 20

Б) Да се провери при риск за грешка 5% дали съществува зависимост между вида на дружеството и размера на неговите дълготрайни активи. Да се използват данните за първите 10 компании от част I, т.1 и 2.

В) Да се провери при риск за грешка 1% дали съществува зависимост между териториалното местоположение на дружеството (областта) и броя на неговите подразделения. Да се използват данните за всичките 25 компании от част I, т.1 и 2.

 ДОПЪЛНИТЕЛНИ ЗАДАЧИ

Задача 5.1.

(А) По време на репрезентативното изследване на доходите в населеното място (пример 1) е установен и средният брой дни, през които са работили лицата от извадката през обхванатия едномесечен период. За тези лица той възлиза на 15,5 дни, при стандартно отклонение 4,5 дни. При риск за грешка 5% да се провери дали средният брой дни, отработени от лице в трудоспособна възраст през обхванатия период в населеното място, е различен от 17,5 дни.

(Б) По време на същото изследване (на доходите и заетостта в населеното място) е установено, че четири от десетте жени и двама от петнадесетте мъже в извадката са безработни. При риск за грешка 5% да се провери дали може да се счита, че безработицата при мъжете е по-ниска от тази при жените.

Задача 5.2.

(А) Направена е случайна извадка от 100 апартамента, предлагани за продажба на пазара на недвижимите имоти в гр.София. За тези жилища е установена средна цена на квадратен метър 835 €/м² при стандартно отклонение 120 €/м². При риск за грешка 5% да се провери дали средната цена (на кв.м) на предлаганите на пазара недвижими имоти се различава съществено от 800 €/м².

(Б) Проведено е национално представително изследване за оценка на относителния дял на малките търговски обекти, в които се предлага една стока в страната. За две от областите са получени следните резултати:

в област “М” извадката наброява 80 обекта, от които 8 предлагат стоката, а в област “Н” извадката възлиза на 60 обекта, от които 12 обекта предлагат стоката. При риск за грешка 1% да се провери дали относителният дял на търговските обекти в област “М”, предлагащи стоката, е по-нисък от този в област “Н”.

Задача 5.3.

(А) Производител гарантира, че делът на нестандартните изделия (брак) в произвежданите от него изделия е не повече от 5%. Ако някоя от партидите съдържа по-висок такъв дял производителят се задължава да я замени с друга. Купувач на едро провежда извадков контрол на качеството на една партида чрез случайна извадка от 200 изделия, от които 16 се оказват нестандартни. При риск за грешка 1% може ли да се направи извода, че бракът в партидата надхвърля договорираното допустимо ниво?

(Б) По време на репрезентативното проучване на потреблението (пример 2) са получени и оценки за средните равнища на потреблението на краве сирене. За извадката в населено място “А” е получено потребление от 4,5 кг средно на домакинство при стандартно отклонение на потреблението 1,5 кг, а за извадката в населено място “Б” съответно 4,7 кг при стандартно отклонение 1,8 кг. При риск за грешка 1% да се провери дали потреблението на краве сирене средно на домакинство в населено място “А” е по-ниско от това в населено място “Б”.

Задача 5.4.

(А) Социологическа агенция трябва да провери (при риск за грешка 1%) дали на предстоящите парламентарни избори партия “АВС” ще прескочи 4 процентната бариера. За целта агенцията провежда представително допитване чрез извадка от 1000 пълнолетни лица, от които 45 декларираат подкрепата си за партията.

(Б) По време на извадковото изследване (пример 1) на доходите и заетостта на лицата, от които 10 жени и 15 мъже, са получени оценки за средномесечните възнаграждения за труд – за мъжете 250 лв при стандартно отклонение 50 лв, а за жените 225 лв при стандартно отклонение 72 лв. При риск за грешка 5% да се провери дали има основания да се счита, че заплащането на мъжете е /средно взето/ по-високо от това на жените.