**Лекция 19**

## §19. Регресионен и сравнителен анализ.

**1. Двумерно нормално разпределение. Коефициент на корелация.** Главната задача на статистическите анализи е установяване на количествени връзки от статистически характер между наблюдаваните величини. Да разгледаме основния случай на две метрични величини $X$ и $Y$. Извадката трябва да включва ***едновременни*** наблюдения $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$. В този случай ***първично*** теоретично понятие се явява ***съвместното разпределение*** на $X$ и $Y$, което съвместно разпределение включва цялата индивидуална и взаимна информация за тези величини. Когато $X$ и $Y$ са метрични, то ние винаги ще предполагаме, че тяхното съвместно разпределение е ***нормално***, което означава, че тяхната съвместна плътност се дава по формулата

$$(19.1)\quad f(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(x-\mu_X)}{\sigma_X}\frac{(y-\mu_Y)}{\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]},$$

където $\mu_X$, $\mu_Y$ са двете популационни средни, $\sigma_X$, $\sigma_Y$ са двете популационни стандартни отклонения, а $\rho$ представлява популационният коефициент на корелация, $\sigma_X>0$, $\sigma_Y>0$, $-1<\rho<1$. Информацията за взаимното отношения между двете величини $X$ и $Y$ изцяло е заложена в корелационния коефициент $\rho$, за когото по-нататък ще получим статистическа оценка. Разпределението (19.1) се описва напълно от вектора на средните $\boldsymbol{\mu}$ и ковариационната матрица $\boldsymbol{\Sigma}$,

$$\boldsymbol{\mu}=\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix},\ \Sigma=\begin{pmatrix}\sigma_{XX} & \sigma_{XY}\\ \sigma_{XY} & \sigma_{YY}\end{pmatrix},\ \sigma_{XX}=\sigma_X^2,\ \sigma_{YY}=\sigma_Y^2,\ \sigma_{XY}=\operatorname{cov}(X,Y)=\frac{\rho}{\sigma_X\sigma_Y}.$$

По стойностите на наблюденията може да се построи ***тримерна хистограма*** (Рис. 19.1), следвайки същия принцип както при обикновената хистограма. При нормално разпределение тримерната хистограма има ясно изразен център на струпване, а останалите наблюдения са струпани по елипси около този център.

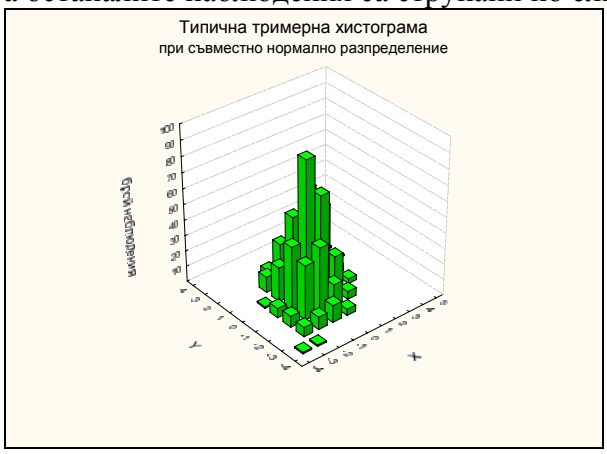


Рис. 19.1.

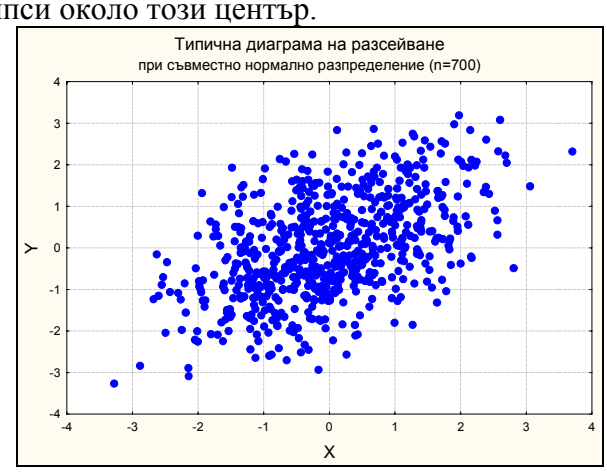


Рис. 19.2.

Формата на взаимно разпределение личи от диаграмата на разсейване (Рис. 19.2), която се получава като отбележим наблюденията в една правоъгълна координатна система. За горните две рисунки са използвани илюстративни данни, генерирани от компютър, които следват стриктно някакво двумерно нормално разпределение.

На следващата диаграма е показан случай на много съществено отклонение от нормално разпределение.Данните са от валутните курсове USD и GBP за работните дни от 1993г. до месец февруари 2006г. В този случай е безпредметно да правим каквито и

да било валидни изводи с помощта на техники, характерни за случая на двумерно нормално разпределение.


Рис. 19.3.

Да разположим данните от извадката за величините $X$ и $Y$ в таблица.

Таблица 19.1.

| $X$ | $Y$ |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| ... | ... |
| $x_n$ | $y_n$ |

В този случай говорим за случаен вектор $(X,Y)$ със стойности от наблюденията $(x_k,y_k)$, $k=1,2,\ldots,n$, и да разсъждаваме аналогично както при оценка на параметрите на едно нормално разпределение. Тук функцията на правдоподобие има вида

$$L=\prod_{k=1}^{n} f(x_k,y_k)=\frac{1}{\left(\sqrt{2\pi\left(1-\rho^2\right)}\right)^n\sigma_X^n\sigma_Y^n}e^{-\frac{1}{2\left(1-\rho^2\right)}\sum_{k=1}^{n}\left[\frac{(x_k-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(x_k-\mu_X)(y_k-\mu_Y)}{\sigma_X\ \sigma_Y}+\frac{(y_k-\mu_Y)^2}{\sigma_Y^2}\right]},$$

а логаритмичната функция на правдоподобие $l=\ln L$ има вида

$$l=-\frac{1}{2\left(1-\rho^2\right)}\sum_{k=1}^{n}\left[\frac{(x_k-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(x_k-\mu_X)}{\sigma_X}\frac{(y_k-\mu_Y)}{\sigma_Y}+\frac{(y_k-\mu_Y)^2}{\sigma_Y^2}\right]-$$

$$-n\ln\sigma_X-n\ln\sigma_Y-\frac{n}{2}\ln\left(1-\rho^2\right)-\frac{n}{2}\ln 2\pi$$

След решаване уравненията на максимално правдоподобие се получават следните оценки

$$\hat{\mu}_X=\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n},\ \hat{\mu}_Y=\bar{y}=\frac{y_1+y_2+\cdots+y_n}{n},$$

$$\hat{\sigma}_{XX}=\hat{\sigma}_X^2=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n},$$

$$\hat{\sigma}_{YY}=\hat{\sigma}_Y^2=\frac{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2}{n},$$

$$\hat{\sigma}_{XY}=\frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{n}.$$

Оценките за елементите на ковариационната матрица обаче се оказват изместени и по тази причина вместо тях се предпочитат следните неизместени оценки

$$s_{XX} = s_X^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1},$$
$$s_{YY} = s_Y^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1},$$
$$s_{XY} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n-1},$$

които формират ***извадковата ковариационна матрица***

$$S = \begin{pmatrix} s_{XX} & s_{XY} \\ s_{XY} & s_{YY} \end{pmatrix}.$$

Тази матрица се явява неизместена оценка за популационната ковариационна матрица $\Sigma$, при което за достатъчно голям обем на извадката $n$ матрицата $S$ се различава несъществено от максимално правдоподобната си оценка.

Последното дава основание да определим ***извадковия коефициент на линейна корелация*** по формулата

$$r = r_{XY} = r[X, Y] = \frac{s_{XY}}{\sqrt{s_{XX}}\sqrt{s_{YY}}},$$
$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{(n-1)s_y s_y},$$

който се явява точкова оценка за популационния коефициент на корелация $\rho$.

Това, че цялата взаимна връзка между $X$ и $Y$ в този случай се изразява с едно единствено число се явява един от многото благоприятни факти, които правят статистическите модели сравнително лесни за прилагане в сложни ситуации. Ако обаче съвместното разпределение не е нормално (съществено се отклонява от нормалното), то линейният корелационен коефициент не съдържа вече цялата взаимна информация за двете величини даже в определени ситуации не носи никаква информация.

При нормално съвместно разпределение независимостта означава, че популационният коефициент на линейна корелация $\rho$ е равен на нула. Това разбира се не означава, че извадковият коефициент $r$ ще бъде винаги равен на нула, понеже във всяка извадка винаги има елементи на случайност. Например ако сме получили $r = 0.125$, това все още не е убедителен аргумент, че популационният коефициент $\rho$ е наистина различен от нула. По-нататък ще приведем статистически аргументи в полза на приемане или на отхвърляне на подобно заключение. Ако обаче $\rho = 0$, то с нарастване обема на извадката стойностите на $r$ сигурно ще се стремят към нула.

Тук отново се срещаме с характерния начин на мислене, според който зависимостта е понятие, отнесено към популацията, а данните от извадката служат за нейното оценяване.

Коефициентите на линейна корелация $\rho$ и $r$ представляват числа между $-1$ и $1$, като самите крайни стойности $-1$ и $1$ на практика не се достигат. Линейният коефициент на корелация представлява индикатор за линейната статистическа зависимост между величините, която зависимост се характеризира с ***посока*** и ***сила***. По тази причина корелационният коефициент се интерпретира по ***знак*** и ***абсолютна стойност***. Знакът "+" означава наличие на ***право пропорционална*** връзка – нарастването при едната величина е свързано с нарастване при другата, а знакът "**–**" означава наличие на ***обратно пропорционална*** връзка – нарастването при едната величина е свързано с намаляване при другата. Колкото е по-голяма абсолютната стойност на корелационния коефициент (максимална абсолютна стойност 1), толкова е

по-силно изразена съответната връзка. Някои автори приемат различни условни класификация в този смисъл, но най-добре е в качеството си на първично понятие да оставим корелационният коефициент да говори сам за себе си чрез своята стойност.

Линейните преобразования и на двете променливи (с положителни множители) не променят стойността на корелационния коефициент, което всъщност представлява неговото най-важно математическо свойство.

В приложната статистика се употребяват множество различни коефициенти на корелация според типа на употребените величини, между които се пресмятат. Всички те обаче имат една обща характеристика – тяхната интерпретация по знак и абсолютна стойност е напълно аналогична на интерпретацията на линейния коефициент на корелация за случая на съвместно нормално разпределени метрични величини.

***Пример 19.1.*** Да разгледаме величината $MAT$-оценка от конкурсния изпит по математика и величината $GU$-среден годишен успех от семестриалните изпити на група обучаеми-студенти и величината за извадка от $n=135$ курсанта. Получени са следните данни, групирани по двойки $(mat_k, gu_k)$, $k=1,2,\ldots,n$,

(5.25, 5.06), (5.50, 5.54), (5.50, 5.78), (4.25, 4.77), (4.50, 4.02), (4.75, 5.31), (4.00, 4.38), (4.75, 4.03), (5.00, 4.30), (3.75, 4.32), (5.00, 4.61), (5.00, 4.05), (4.25, 4.71), (5.00, 4.48), (4.75, 5.24), (5.00, 4.60), (6.00, 5.59), (4.25, 4.11), (3.00, 4.55), (4.75, 4.70), (4.00, 4.01), (4.50, 5.44), (5.50, 4.14), (5.50, 5.51), (5.75, 4.91), (3.50, 4.32), (4.75, 4.84), (3.75, 4.45), (4.00, 4.34), (5.25, 3.69), (5.00, 4.28), (5.00, 3.88), (4.50, 4.40), (4.75, 4.00), (3.75, 3.74), (4.75, 3.94), (5.25, 4.26), (5.50, 5.34), (3.70, 4.25), (5.00, 4.69), (3.75, 4.99), (4.75, 5.36), (3.50, 4.86), (5.50, 4.40), (3.50, 5.00), (4.50, 4.63), (4.75, 4.39), (5.25, 4.76), (5.00, 4.76), (5.00, 5.07), (5.00, 4.56), (5.00, 5.42), (5.25, 4.34), (4.00, 4.14), (5.50, 5.91), (5.25, 6.00), (4.50, 5.02), (5.50, 4.55), (5.25, 5.02), (5.50, 4.51), (4.75, 4.84), (5.50, 4.53), (4.25, 5.30), (3.00, 4.42), (5.25, 4.70), (4.75, 4.18), (3.00, 4.76), (5.25, 5.34), (5.25, 5.34), (5.50, 5.13), (5.50, 4.96), (4.75, 5.52), (5.00, 5.65), (4.50, 3.78), (5.75, 5.59), (5.00, 4.77), (4.50, 4.68), (4.50, 4.66), (4.50, 3.73), (3.25, 4.55), (4.50, 4.30), (4.50, 4.41), (3.75, 4.39), (6.00, 4.93), (5.25, 4.45), (3.50, 4.23), (5.00, 4.86), (4.50, 4.55), (4.50, 5.45), (4.25, 4.31), (5.25, 4.57), (4.25, 4.89), (4.75, 3.85), (5.25, 5.63), (5.00, 4.47), (4.75, 4.35), (4.75, 4.93), (6.00, 5.51), (5.50, 5.61), (4.00, 4.45), (4.00, 4.99), (3.50, 4.48), (4.50, 4.14), (5.00, 4.14), (4.25, 4.87), (4.75, 4.84), (5.25, 4.94), (6.00, 5.16), (5.75, 5.07), (3.00, 4.94), (5.50, 5.11), (4.50, 4.89), (6.00, 5.31), (5.75, 5.16), (5.50, 5.45), (5.00, 5.10), (5.50, 5.28), (5.25, 5.60), (5.25, 4.65), (3.25, 4.05), (5.50, 5.72), (5.00, 4.72), (5.50, 4.59), (5.00, 5.89), (5.75, 5.53), (4.75, 5.29), (5.00, 4.85), (4.75, 4.40), (5.50, 5.11), (4.75, 4.90), (5.75, 5.72), (5.50, 5.15), (5.25, 4.72), (5.25, 4.89), (5.50, 6.00)

Диаграмата на разсейване от Рис. 19.4 показва несъществено отклонение от съвместно нормално разпределение, което дава основание да анализираме двете величини въз основа на предположението за нормално разпределение.

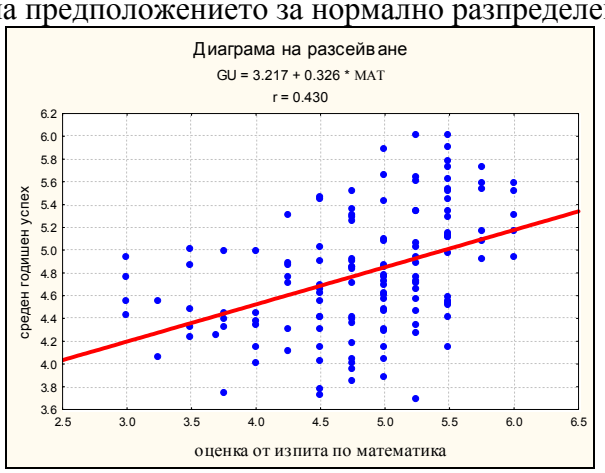


Рис. 19.4.

Изчисленията дават следните резултати за средните $\overline{mat}=4.811$ и $\overline{gu}=4.789$. За елементите на извадковата ковариационна матрица $S$ имаме $\sigma^2_{mat}=0.505$, $\sigma^2_{gu}=0.292$, $\sigma_{mat,gu}=0.165$. За корелационния коефициент намираме стойността

$$r = \frac{0.165}{\sqrt{0.505}\sqrt{0.292}} = 0.430 .$$

Матрицата $S$ има вида

$$S = \begin{pmatrix} 0.505 & 0.165 \\ 0.165 & 0.292 \end{pmatrix}.$$

Статистическият характер на връзката между дадени величини $X$ и $Y$ е сравнително трудна за детайлна интерпретация в общия случай. Да разгледаме задачата за установяване на най-добра в определен смисъл функционална връзка $y = \varphi(x)$ между двете величини. Очевидно такава една връзка ще представлява абстракция от статистическия модел, оценена въз основа на данните. Оказва се, че тази задача има естествено и лесно за разбиране решение, което се нарича ***уравнение на регресия*** на $Y$ върху $X$. За стойност на функцията $y = \varphi(x)$, при дадено $x$, се приема условното математическо очакване на стойностите на $Y$, при това фиксирано $x$.

При нормално съвместно разпределение, популационното уравнение на регресия на $Y$ върху $X$ има вида

$$\frac{y - \mu_y}{\sigma_y} = \rho \frac{x - \mu_x}{\sigma_x},$$

което има извадков аналог

$$\frac{y - \bar{y}}{s_y} = r \frac{x - \bar{x}}{s_x} .$$

Последното в разгърнат вид изглежда така

$$(19.2) \quad y = r \frac{s_y}{s_x} x + \left( \bar{y} - r\bar{x} \frac{s_y}{s_x} \right),$$

което представлява линейно уравнение. Естествената функционална връзка между две величини със нормално съвместно разпределение е линейна. Уравнението (19.2) се нарича още уравнение на ***линейна регресия*** на $Y$ върху $X$ при всяко взаимно разпределение на $X$ и $Y$, но има ясен смисъл само когато съвместното разпределение е нормално.

В общия случай до уравнението на линейна регресия (19.2) може да се стигне по различни начини, най известният от които носи името метод на най-малките квадрати (***least squares***) и може да бъде обоснован извън вероятностните съображения. При този метод търсим линейна връзка от вида

$$(19.3) \quad y = \alpha x + \beta ,$$

където коефициентите $\alpha$ и $\beta$ подлежат на определяне съгласно изискването за минимум на сумата

$$(19.4) \quad LS = \sum_{k=1}^{n} (\alpha x_k + \beta - y_k)^2 .$$

Всяко от събираемите $(\alpha x_k + \beta - y_k)^2$, $k = 1,2,\ldots,n$, показва квадрата на разликата между предсказаната от модела (19.3) стойност $\alpha x_k + \beta$ и стойността от съответното наблюдение $y_k$. Ако разгледаме израза $LS$ от (19.4) като функция на двата параметъра на модела, $LS = LS(\alpha, \beta)$, и намерим нейния минимум, ще получим точно уравнението (19.2). Друг начин за постигане на същата цел се състои в разглеждането на вероятностен модел за $X$ и $Y$ по формулата

$$Y = \alpha X + \beta + E ,$$

където $E$ е случайна величина, разпределена нормално с нулево средно и някаква дисперсия $\sigma^2$, $E \in N(0,\sigma^2)$. Тогава $E = Y - \alpha X - \beta$ и можем да разгледаме експеримента спрямо $E$ като случайна извадка над разпределение $N(0,\sigma^2)$ с наблюдения

(19.5) $e_k = y_k - \alpha x_k - \beta$, $k = 1,2,\ldots,n$.

Случайната величина $E$ има плътност

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}},$$

следователно функцията на правдоподобие има вида

$$L = \frac{1}{\sigma^n\left(\sqrt{2\pi}\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k-\alpha x_k-\beta)^2},$$

а логаритмичната функция на правдоподобие $l = \ln L$ има вида

$$l = -\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k - \alpha x_k - \beta)^2 - n\ln\sigma - \frac{n}{2}\ln(2\pi).$$

Намирането на максимума на $l$ относно $\alpha$ и $\beta$ очевидно е еквивалентно на намирането на минимума на сумата в израза, което отново води до метода на най-малките квадрати. При този подход обаче има възможност за уточняване вида на разпределението на $E$ въз основа данните от експеримента и след това да проверим доколко данните от наблюденията за $E$ се съгласуват с този извод.

Уравнението (19.2) може да послужи за предсказване стойността на $Y$ при известна стойност на $X$.

В случая когато $r = 0$, уравнението (19.2) приема вида $y = \bar{y}$, което показва, че при всяка стойност на $x$, най-добрата предсказана стойност на $y$ е просто средното $\bar{y}$. С други думи промяната в стойностите на $X$ не влияе върху прогнозата за $Y$, което разбира се е напълно очаквано, понеже $r = 0$ означава независимост на двете величини.

***Пример 19.2.*** Да намерим уравнението на линейна регресия на величината *MAT* върху *GU* от пример 19.1. Изчисленията дават следното уравнение

(19.6) $GU = 0.326 * MAT + 3.217$.

Върху диаграма от Рис. 19.4 е отбелязана и линията на регресия, определена от (19.6). Последното уравнение дава възможност да прогнозираме средният годишен успех при известна оценка по математика. Например ако оценката по математика от конкурсния изпит е 4.25, то прогнозирания среден годишен успех е

$4.604 = 0.326 * 4.25 + 3.217$.

Следващата диаграма показва ***разпределението на остатъците*** за модела, т.е. разпределението на стойностите (19.5).

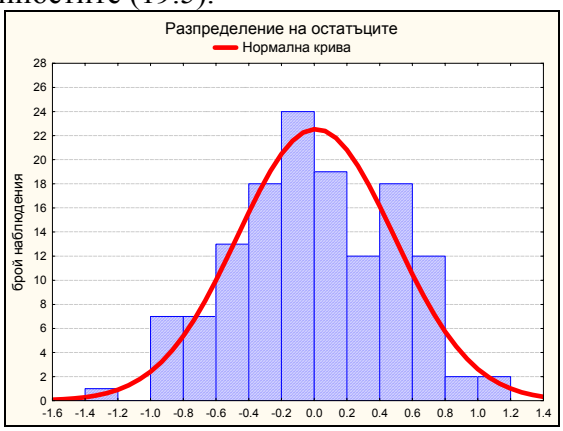


Рис. 19.5.

Полученото разпределение се съгласува с нормалното което представлява белег за адекватност на модела.

**2. Линеен регресионен анализ.** Описаният в предишния раздел подход може да се обобщи за повече величини. Тук основната цел е даване количествен израз на ефектите на дадена група метрични величини $X_1$, $X_2$, ..., $X_p$, които условно се наричат ***независими*** (***independent***) върху друга величина $Y$, която условно се нарича ***зависима*** (***dependent***). Независимите величини се наричат понякога и ***фактори***. Предметният контекст на регресионния анализ предполага каузални връзки между факторите и зависимата величини. Основната идея се състои в следното. Въз основа на взаимното популационно разпределение на всичките величини се търси естествена функционална връзка от вида

**(19.**7) $y = f\left(x_1, x_2, \ldots, x_p\right),$

която по статистически обоснован начин дава прост израз на ефектите на отделните независими величини върху зависимата. Уравнението (9.1) се нарича ***уравнение на регресия*** на $Y$ върху $X_1$, $X_2$, ..., $X_p$ и се получава като условното математическо очакване на стойностите на $Y$ при фиксирани стойности на $X_1$, $X_2$, ..., $X_p$. Когато взаимното популационно разпределение на всичките величини е ***нормално***, уравнението на регресия (19.7) се оказва ***линейно***

(19.8) $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p,$

където $b_0$, $b_1$, $b_2$, ..., $b_p$ са коефициентите на уравнението. Тези коефициенти лесно се интерпретират по знак. Ако например $b_1 > 0$, то нарастването на $x_1$ води до нарастване на $y$ и ако $b_1 < 0$, то нарастването на $x_1$ води до намаляване на $y$. По-големите по абсолютна стойност коефициенти са свързани с по-голяма промяна при зависимата величина. Съпоставката по абсолютна стойност на тези коефициенти като критерий доколко е голям ефектът на отделните независими величини трябва да отчита и други обстоятелства, свързани с дисперсиите на величините. По тази причина е по-удобно всичките величини да бъдат приведени към $z$-стойности (нормално стандартно разпределение) вместо (19.8), при което се получава уравнението

(19.9) $$\frac{y - \bar{y}}{s_y} = \beta_1 \frac{x_1 - \bar{x}_1}{s_1} + \beta_2 \frac{x_2 - \bar{x}_2}{s_2} + \cdots + \beta_p \frac{x_p - \bar{x}_p}{s_p},$$

в което свободният коефициент е равен на нула, а коефициентите $\beta_1$, $\beta_2$, ..., $\beta_p$ се наричат ***стандартизирани коефициенти на регресия***. Стандартизираните коефициенти се интерпретират по знак както преди но вече са съпоставими и по абсолютна стойност. Параметрите на (19.8-9) се оценяват въз основа ***метода на най-малките квадрати***.

За всеки от коефициентите на регресия се пресмята ***значимост***, като фактически се проверява нулева хипотеза, че съответният популационен коефициент е равен на нула. При тези хипотези проверяващата статистика е има разпределение на Стюдънт $t\left(n - p - 1\right)$, където $n$ е обемът на извадката. Ако някой от коефициентите се получи с пренебрежима значимост, то съответната независима величина може да бъде изключена от анализа без съществена загуба на информация. Освен това за целия модел се пресмята величина $R^2$ (квадрата на множествения коефициент на корелация), която показва каква пропорция от изменчивостта на зависимата величина се обяснява посредством изменението на независимите величини. В тази връзка се проверява и нулева хипотеза, че линейният регресионен модел не обяснява по същество

изменчивостта при зависимата величина, като проверяващата статистика е разпределена $F(p, n-p-1)$.

***Пример 19.3.*** Изследва се група от $n=135$ края на първи курс по различни учебни показатели. Наблюдаваните величини са средния успех от семестриалните изпити и балообразуващите показатели – оценки от дипломата по математика, физика и български език и средния успех от самата диплома, както и разбира се оценката от конкурсния изпит по математика. Зависимата величина представлява средния годишен успех, а в качеството на независими се вземат изброените пет величини $(p=5)$. След провеждане на регресионния анализ се получават следните резултати.

Таблица 19.2.

| **фактори** | **$R^2$=0.372; F(5,129)=17.424; p=0.000** | | |
|---|---|---|---|
| | **Beta** | **t(129)** | **p** |
| **оценка по математика от дипломата** | -0.021 | -0.252 | 0.801 |
| **оценка по физика от дипломата** | 0.055 | 0.644 | 0.521 |
| **оценка по български език от дипломата** | 0.137 | 1.109 | 0.270 |
| **оценка от конкурсния изпит по математика** | 0.429 | 5.957 | 0.000 |
| **среден успех от дипломата** | 0.294 | 2.205 | 0.029 |

Тук статистически значими в рамките на традиционния критерий $p<0.05$ се оказаха само коефициентът оценката от конкурсния изпит $[beta=0.429;\ p=0.000]$ и коефициентът на средния успех от дипломата $[beta=0.294;\ p=0.029]$, при което и двата фактора имат положителен ефект понеже знакът на съответните коефициенти е плюс. За останалите коефициенти не разполагаме с достатъчно основания да приемем, че техните популационни стойности фактически са равни на нула.

Моделът като цяло обяснява $R^2.100\% \approx 37\%$ от изменчивостта на зависимата величина "адаптация", което не е много добър резултат и означава, че за по изчерпателно обяснение на годишния успех трябва да се привлекат и други независими величини. Изненадата в този случай се явява липсата на прогностична сила на оценките по математика и физика от дипломата относно академичната успеваемост.

**3. Сравнителен анализ на независими извадки.** ***Тестове на Стюдънт и Фишер.*** Да предположим например, че трябва да извършим статистическо сравнение на резултатите от различни личностни черти или постижения между групите на жени и мъже. В този случай говорим за сравняване между две ***независими*** извадки, което сравнение от гледна точна на статистическия механизъм всъщност се отнася към средните на популациите, от които произхождат извадките. Изборът на средното като средство за сравнение е напълно естествен, понеже средното се явява основната мярка за централна тенденция на метрични величини. Именно средното е онова единствено число, което по най-добър и икономичен начин представя поведението на цялата група. Всяка от двете извадки носи определена информация за своята популация.

Схемата за провеждане на това сравнение ще покажем върху следния пример.

***Пример 19.4.*** Върху $n=175$ на брой курсанти е проведено психологическо изследване с цел установяване нивото на "адаптация към средата". Тук цялата извадка

се състои от $n=175$ лица, от които $n_f=61$ жени и $n_m=114$ мъже. Индексирането чрез $f$ и $m$, което служи единствено за различаване на групите, е предпочетено въз основа на английската лексика за наименование на половете female и male. Статистическият анализ на данните показва следните резултати

$$\bar{x}_f=177.491,\ \bar{x}_m=179.061,\ s_f=27.255,\ s_m=27.769.$$

Различията между емпиричните стандартните отклонения са видимо пренебрежими. Между емпиричните средните съществува известно различие, за което трябва да преценим дали отразява някаква закономерна тенденция или можем да го отнесем към игра на случайността. И двата извода имат определена стойност за изследователя. В този случай статистиката предлага специфичен подход за решаване на поставената задача, който се нарича ***t-тест на Стюдънт*** за сравнение на средните от две независими извадки. Статистическата рамка на задачата е представена на следната диаграма.

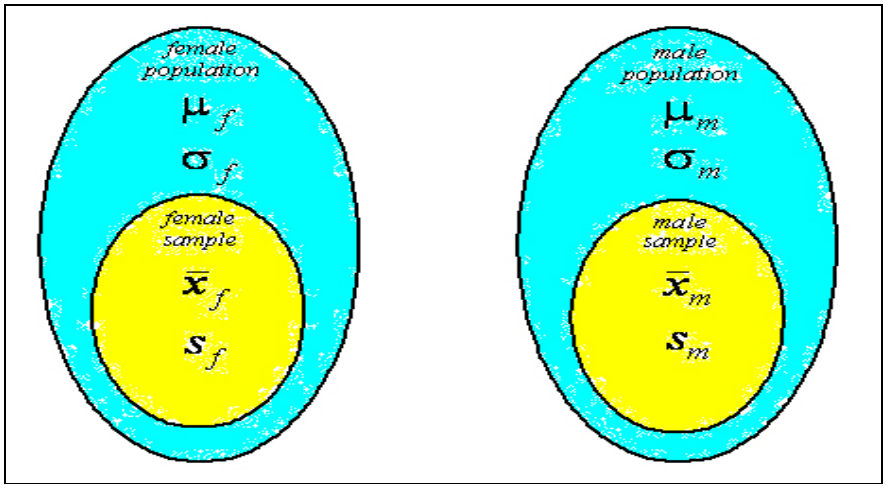


Рис. 19.6.

В дадения случай наблюдаваният ефект, който предизвиква интерес, представлява различието между двете емпирични (извадкови) средни, а хипотезата за нулев ефект се състои в предположението, че това различие има случаен характер и не се явява белег за някаква закономерна разлика между половете, което на езика на статистиката се формулира, че средните за двете популации са равни. По тази причина при класическия тест на Стюдънт се проверява нулевата хипотеза

$$H_0:\mu_f=\mu_m,$$

срещу двустранната алтернатива

(19.10) $H_{alt}:\mu_f\neq\mu_m,$

където се предполага, че дисперсиите на двете популации са равни, $\sigma_f=\sigma_m$. При валидна нулева хипотеза $H_0$, ***проверяващата статистика***

$$(19.11)\ t_{emp}\left(n_f+n_m-2\right)=\frac{\bar{x}_f-\bar{x}_m}{\sqrt{\left(\frac{1}{n_f}+\frac{1}{n_m}\right)\left(\frac{\left(n_f-1\right)s_f^2+\left(n_m-1\right)s_m^2}{n_f+n_m-2}\right)}},$$

се подчинява на $t$-разпределение на Стюдънт с $n_f+n_m-2$ степени на свобода. Отхвърлянето на нулевата хипотеза би довело до извода за наличие на съществена закономерна разлика между средните нива на адаптация за двата пола.

Краят на тази процедура е напълно аналогичен на случая за хипотеза относно средното на една популация. При зададено ниво на значимост $\alpha$, нулевата хипотеза се отхвърля когато

$$\left|t_{emp}\left(n_f + n_m - 2\right)\right| \geq t_{1-\frac{\alpha}{2};n_f+n_m-2},$$

и се приема в противен случай. И тук обаче на практика единственото нещо което трябва да се направи е да се пресметне оцененото ниво на значимост $p$ в условие на двустранна алтернатива.

За изследвания пример имаме $t_{emp}(173) = -0.359$ и $p = 0.720$. Както вече знаем нулевата хипотеза $H_0$ се отхвърля, когато стойността на $p$ се получи достатъчно малка за целта, понеже $p$ представлява вероятността за грешка при такова решение. Тук получената стойност $p = 0.720$ разбира се не дава никакви основания за отхвърляне на $H_0$. Нещо повече, такава голяма стойност на $p$ показва, че решението за приемане на нулевата хипотеза $H_0$ е свързано с малък риск за грешка от втори род. Описаният пример представлява типична ситуация, при която приемането на нулевата хипотеза не буди никакви възражения. Резултатите от анализа можем да запишем в изречението

"Различието между средното ниво на адаптация при жените $\bar{x}_f = 177.491$ и мъжете $\bar{x}_m = 179.061$ е статистически незначимо, [$t(173) = -0.359$ ; $p = 0.720$]."

Когато са налице достатъчно основания, вместо двустранната алтернатива (19.10) може да се използва едностранна алтернатива $H_{alt}: \mu_f > \mu_m$ или $H_{alt}: \mu_f < \mu_m$. Да припомним, че при едностранна алтернатива оцененото ниво на значимост се получава два пъти по-малко отколкото при двустранна такава, което открива съществено повече шансове за отхвърляне на нулевата хипотеза $H_0$ и предполага наличие на убедителни аргументи в полза на такова предпочитание.

Коректното прилагане на теста на Стюдънт предполага нормални разпределения с равни дисперсии в двете популации, както и достатъчно голям обем на двете извадки.

Въпросът дали дисперсиите в двете популации могат да се разглеждат като равни въз основа на данните от извадките се решава посредством ***F -теста на Фишер*** за сравняване дисперсиите на две независими извадки. $F$-тестът на Фишер представлява статистическа процедура на проверка на нулевата хипотеза

$$H_0: \sigma_f = \sigma_m,$$

срещу двустранната алтернатива

$$H_{alt}: \sigma_f \neq \sigma_m.$$

При вярна нулева хипотеза $H_0$, проверяващата статистика

$$F_{emp}\left(n_f - 1, n_m - 1\right) = \frac{s_f^2}{s_m^2}$$

се подчинява на $F$-разпределение на Фишер със степени на свобода $n_f - 1$ и $n_m - 1$. При избрано ниво на значимост $\alpha$, нулевата хипотеза $H_0$ се отхвърля когато

$$F_{emp}\left(n_f - 1, n_m - 1\right) \leq F_{\frac{\alpha}{2};n_f-1,n_m-1} \text{ или } F_{emp}\left(n_f - 1, n_m - 1\right) \geq F_{1-\frac{\alpha}{2};n_f-1,n_m-1},$$

и се приема в противен случай. Тук също единственото нещо което трябва да се направи е да се пресметне оцененото ниво на значимост $p$ в условие на двустранна алтернатива.

За този конкретен пример пресмятанията дават $F_{emp}(60,113)=0.963$ и $p=0.887$, което показва (и без това очевидното), че трябва да приемем нулевата хипотеза $H_0$ без особен риск за грешка от втори род.

"Двете групи на мъжете и жените показват съществено равни дисперсии в резултатите от теста за адаптация, [ $F(60,113)=0.963$ ; $p=0.887$ ]."

Резултатите от сравнението между средните и дисперсиите може да се илюстрира посредством следната диаграма.

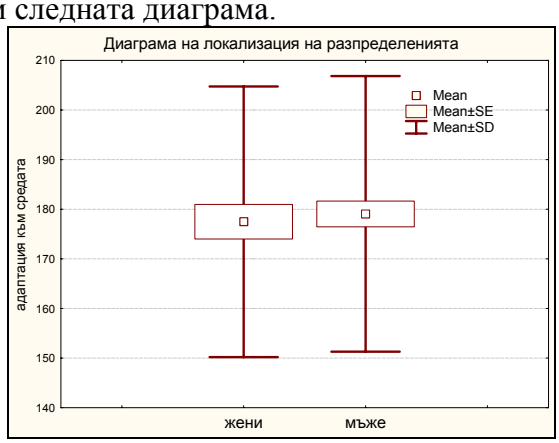


Рис. 19.7.

Точката по средата показва разположението на средното $\bar{x}$ за съответната група, горната и долната страна на кутията сочат $\bar{x}\pm\frac{s}{\sqrt{n}}$, а "мустачките" сочат $\bar{x}\pm s$, което обхваща около 68% от разпределението.

***Пример 19.5.*** Сега ще извадка от същите лица, но в този случай анализираната величина ще бъде техният среден успех от семестриалните изпити през дадена учебна година. В този случай имаме

$\bar{x}_f=5.579$, $\bar{x}_m=4.920$, $s_f=0.353$, $s_m=0.596$.

За да проверим хипотезата за нулев ефект $H_0:\mu_f=\mu_m$ срещу двустранна алтернатива пресмятаме $t_{emp}(173)=7.293$ и $p=0.000$ $(p<0.000001)$, което показва изключително достоверно отхвърляне на нулевата хипотеза. Проверката на хипотезата за нулев ефект $H_0:\sigma_f=\sigma_m$ срещу двустранна алтернатива дава $F_{emp}(60,113)=0.123$ и $p=0.000$ $(p<0.000001)$, което също води до практически достоверно отхвърляне на тази нулева хипотеза. Тези изводи могат да се оформят например в следното изречение.

"Групата на жените показва значително по-висок среден годишен успех, [ $t(173)=7.923$ ; $p=0.000$ ], като същевременно показва и значително по-ниска дисперсия, [ $F(60,113)=0.123$ ; $p=0.000$ ]."

Двата резултата по сравнение на средните могат да бъдат приведени в една обща таблица, която може да изглежда например по следния начин.

Таблица 19.3.

| | средно жени | средно мъже | $t(173)$ | $p$ |
|---|---|---|---|---|
| адаптация към средата | 177.491 | 179.061 | -0.359 | 0.720 |
| среден годишен успех | 5.579 | 4.921 | 7.923 | 0.000 |

При статистически значими различия в дисперсиите се препоръчва използване на друга проверяваща статистика вместо (19.11), която няма да обсъждаме, понеже както се сочи в литературата, тестът на Стюдънт е устойчив към отклоненията от

изискването за равенство на дисперсиите. Средите за статистическа обработка обикновено предлагат резултатите и от двата варианта на изпълнение.

***Еднофакторен дисперсионен анализ.*** Дисперсионният анализ (***ANOVA*** – ***analysis of variance***) представлява специфична форма на сравнителен анализ (при който проверяващите статистики за нулевите хипотези се основават на пресмятане на дисперсии). Сега ще разгледаме еднофакторния дисперсионен анализ (***One-Way ANOVA***), който се явява непосредствено обобщение на изложените в предишния раздел тестове и се отнася за случая на ***едновременно сравняване*** на две или повече независими извадки. Когато извадките са точно две, еднофакторният дисперсионен анализ се покрива напълно по смисъл с теста на Стюдънт, затова неговото използване предполага наличие на поне три независими извадки. В типичния случай на прилагане на този анализ присъстват две ясно обособени величини. Първата величина, която се нарича още ***фактор*** (откъдето идва и наименованието на процедурата) трябва да бъде номинална. Другата се нарича ***зависима величина*** и трябва да бъде метрична. Факторът и зависимата величина се наблюдават едновременно върху извадката. По този начин цялата извадка може да се раздели на отделни подгрупи според категориите на фактора, които групи формират споменатите по-горе независими извадки. Тези подгрупи могат да се интерпретират като независими извадки от нормални популации с едни и същи дисперсии, но евентуално с различни средни. Отделните групи се характеризират със собствени обеми $n_j$, емпирични средни $\bar{x}_j$ и емпирични дисперсии $s_j^2$, $j = 1,2,\ldots,J$, където $J$ означава броя на групите, което е и броят на категориите (***нивата***) на фактора. Популациите за отделните групи се характеризират със собствени средни $\mu_j$, $j = 1,2,\ldots,J$, и равни дисперсии

(19.12) $\sigma_1 = \sigma_2 = \cdots = \sigma_J = \sigma$.

Хипотезата за нулев ефект има вида

(19.13) $H_0 : \mu_1 = \mu_2 = \cdots = \mu_J$,

което гласи, че всичките популационни средни са равни. Алтернативната хипотеза $H_{alt}$ гласи, че някои две от тези средни са различни, без да уточнява кои точно двойки средни са различни.

Отхвърлянето на нулевата хипотеза $H_0$ в типичния случай се интерпретира като наличие на статистически значим ефект от страна на фактора върху зависимата променлива, откъдето и произтича нейното често срещано название "зависима". Приемането на нулевата хипотеза се разглежда като липса на статистическа значимост на въпросния ефект. Например от анализа на успеха можем да направим извод, че факторът пол има статистически значимо влияние върху средният успех от семестриалните изпити и не показва значимо влияние върху адаптацията към средата. В такъв контекст величината фактор и зависимата величина могат да се подразбират в каузална връзка, но анализът в общия случай представлява интерес и се провежда извън рамките на акцентирана каузална връзка между фактора и метричната величина.

***Проверяващата статистика*** на нулевата хипотеза $H_0$ има вида

$$F_{emp}\left(J-1, N-J\right) = \frac{\dfrac{\sum_{j=1}^{J} n_j \left(\bar{x}_j - \bar{x}\right)^2}{J-1}}{\dfrac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}\right)^2}{N-J}} = \frac{\dfrac{SS_b}{J-1}}{\dfrac{SS_w}{N-J}},$$

където $SS_b$ се нарича сума на квадратите ***между групите***, $SS_w$ се нарича сума на квадратите ***вътре в групите***. Тук $\bar{x}$ и $N = n_1 + n_2 + \cdots + n_J$ са емпирично средното и обемът на обединената извадка, а $x_{1j}$, $x_{2j}$, ..., $x_{n_j j}$ са стойностите на наблюденията в извадка с номер $j$. Ако нулевата хипотеза $H_0$ е валидна, то величината $F_{emp}(J-1, N-J)$ се подчинява на $F$ -разпределение на Фишер със степени на свобода $J-1$ и $N-J$. При зададено ниво на значимост $\alpha$, нулевата хипотеза $H_0$ се отхвърля когато $F_{emp}(J-1, N-J) \geq F_{1-\alpha; J-1, N-J}$ и се приема в противен случай.

***Пример 19.6.*** В този пример факторът ще бъде поредният курс на обучение, а зависимата променлива ще бъде средният годишен успех от семестриалните изпити при жените. Факторът има четири нива, $J = 4$, които отговарят на четирите курса на обучение. Данните от описателния анализ са приведени в таблица 19.4.

Таблица 19.4.

| **курс** | **брой** | **средно** | **стандартно отклонение** |
|---|---|---|---|
| първи | $n_1 = 12$ | $\bar{x}_1 = 5.670$ | $s_1 = 0.273$ |
| втори | $n_2 = 16$ | $\bar{x}_2 = 5.557$ | $s_2 = 0.358$ |
| трети | $n_3 = 11$ | $\bar{x}_3 = 5.571$ | $s_3 = 0.379$ |
| четвърти | $n_4 = 22$ | $\bar{x}_4 = 5.551$ | $s_4 = 0.389$ |
| общо | $N = 61$ | $\bar{x} = 5.579$ | $s = 0.353$ |

Тези данните показват видимо незначителни разлики в средния успех между курсовете. За да потвърдим това наблюдение ще си послужим с техниката на еднофакторния дисперсионен анализ. Резултатите от обработката показват $F_{emp}(3,57) = 0.326$ с оценено ниво на значимост $p = 0.807$, което предлага сигурно основание за приемане на нулевата хипотеза.

"Различията в средния годишен успех за жените между четирите курса са статистически незначими, [ $F(3,57) = 0.326$ ; $p = 0.807$ ]."

Този резултат може да бъде придружен от диаграма, на която са показани средните стойности и 95% доверителни интервали за всяка от тях.

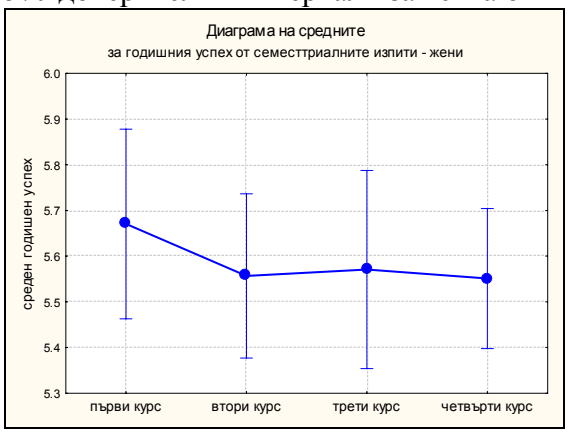


Рис. 19.8.

Коректното прилагане на горната процедура изисква някаква проверка на условието за равенство на дисперсиите между отделните групи. Това може да се стане посредством ***теста на Бартлет*** или ***теста Левин***, които проверяват нулева хипотеза за такова равенство $H_0 : \sigma_1 = \sigma_2 = \cdots = \sigma_J$ и в този смисъл се явяват

***обобщение*** на $F$ -теста на Фишер. Проверяващата статистика при теста на Бартлет има $\chi^2(J-1)$ разпределение, а при теста на Левин проверяващата статистика има $F(J-1, N-J)$ разпределение. При теста на Бартлет нулевата хипотеза за равенство на дисперсиите се отхвърля когато $\chi^2_{emp}(J-1) \geq \chi^2_{1-\alpha;J-1}$, а при теста на Левин, когато $F_{emp}(J-1, N-J) \geq F_{1-\alpha;J-1,N-J}$. За последния пример тестът на Бартлет дава $\chi^2(3) = 1.655$ с оценено ниво на значимост $p = 0.647$, а теста на Левин дава $F(3,57) = 0.529$ с $p = 0.664$. И двата теста не дават основание за отхвърляне на нулевата хипотеза за равенство на дисперсиите.

Освен равенство на дисперсиите между отделните групи, коректното прилагане на еднофакторния дисперсионен анализ изисква достатъчен брой наблюдения за всяка група и нормално разпределение във всяка от групите. В литературата се сочи, че този вид анализ се влияе малко от нарушаването на изискванията за неговото прилагане.

***Пример 19.7.*** Тук примера е аналогичен на пример 19.6 с тази разлика, че извадката се състои само от мъже. Описателният анализ показва следните резултати.

Таблица 19.5.

| курс | брой | средно | стандартно отклонение |
|---|---|---|---|
| първи | $n_1 = 50$ | $\bar{x}_1 = 4.631$ | $s_1 = 0.520$ |
| втори | $n_2 = 20$ | $\bar{x}_2 = 4.831$ | $s_2 = 0.523$ |
| трети | $n_3 = 18$ | $\bar{x}_3 = 5.061$ | $s_3 = 0.545$ |
| четвърти | $n_4 = 26$ | $\bar{x}_4 = 5.450$ | $s_4 = 0.427$ |
| общо | $N = 114$ | $\bar{x} = 4.921$ | $s = 0.596$ |

Данните показват промяна в средния успех между курсовете. Еднофакторният дисперсионен анализ дава резултатите $F_{emp}(3,110) = 15.669$ с оценено ниво на значимост $p = 0.000$, което предлага много сигурно основание за отхвърляне на нулевата хипотеза.

"Различията в средния годишен успех за мъжете между четирите курса са статистически значими, [$F(3,110) = 15.669$ ; $p = 0.000$]."

Тестът на Бартлет дава $\chi^2(3) = 1.570$ с оценено ниво на значимост $p = 0.667$, а теста на Левин дава $F(3,110) = 0.854$ с $p = 0.468$. И двата теста не дават основание за отхвърляне на нулевата хипотеза за равенство на дисперсиите.

Резултатът от анализа е илюстриран на следната диаграма.

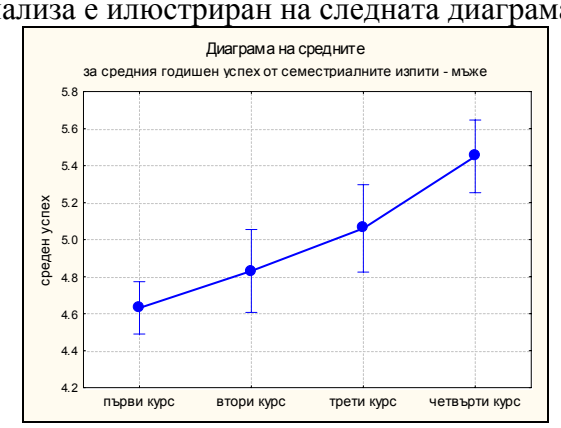


Рис. 19.9.

Отхвърлянето на нулевата хипотеза за равенство на средните открива нова задача за уточняване къде точно е съсредоточено различието. За тази цел има разработени различни тестове за така наречения ***Post-Hoc анализ***. За психологическите изследвания се препоръчва използването на ***HSD-теста на Тюки*** или ***теста на Нюман-Коулс***. За последния пример HSD-тестът дава следната таблица за статистически значими разлики при ниво на значимост $\alpha = 0.05$.

Таблица 19.6.

| **курс** | **{2}** | **{3}** | **{4}** |
|---|---|---|---|
| **{1}** | 0.445 | 0.013 | 0.000 |
| **{2}** | | 0.502 | 0.001 |
| **{3}** | | | 0.063 |

Еднофакторният дисперсионен анализ не е еквивалентен на провеждането на съответния брой тестове на Стюдънт с разглеждане на всяка срещу всяка група, които за последните два примера са шест на брой отделни теста на Стюдънт. Еднофакторният дисперсионен анализ дава "поглед отгоре" върху проблема за съществуването на различия и се разглежда като методически по-добър отколкото провеждането на отделни анализи всяка срещу всяка група, което се явява "поглед върху детайлите". Тълкуването на неговите резултати обаче, когато се отхвърля главната хипотеза за нулев ефект е свързано с повече елементи на неопределеност.

***Тест на Ман-Уитни. Непараметрични тестове.*** $U$-тестът на Ман-Уитни (който понякога се отбелязва и като $W$-тест на Ман-Уитни) има същата цел и познавателно значение, както теста на Стюдънт за независими извадки, но постигането на тази цел се извършва на базата на други статистически съображения. Първата основна характеристика на $U$-теста е, че неговото провеждане не предявява изисквания към формата на разпределение на изследваната величина. Такива тестове се наричат ***непараметрични*** (***non-parametric, distribution free***). Да припомним, че едно от изискванията за приложимост на теста на Стюдънт беше свързано с нормално разпределение на величината. Влагайки известни несъществени за практиката елементи на неточност, може да се каже, че $U$-тестът на Ман-Уитни използва ***медианата*** вместо средното като цел и средство за извършване на сравнението. Медианата $Me$ представлява другата следваща по важност мярка за централна тенденция на метрични величини след средното $\mu$. Отклоненията във формата на разпределение на дадена метрична величина се получават най-вече като следствие от малкия обем на извадката, например по-малък от $30$. Затова използването на непараметрични тестове се препоръчва именно в случаи на извадки с малки обеми.

Втората важна характеристика на $U$-теста се състои в това, че обработката на резултатите от извадките не се извършва директно върху натуралните стойности на величината, а върху техните поредни номера – рангове. Подреждането става по следния начин. Най-малката стойност получава ранг $1$, следващата по големина получава ранг $2$ и т.н. докато се изчерпят наблюденията.

Нека за величината $X$ разполагаме с извадка с обем $m$, а за величината $Y$ разполагаме с извадка с обем $n$. Като обединим двете извадки получаваме една обединена извадка с обем $m+n$, за която пресмятаме поредния номер на всяко наблюдение. Нека $W_X$ означава сумата от ранговете на стойностите от извадката за величината $X$. Хипотезата за нулев ефект гласи, че двете извадки произхождат от еднакво разпределени непрекъснати величини. При валидна нулева хипотеза величината

$$Z_{emp} = \frac{W_X - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}$$

следва приблизително нормално стандартно разпределение, $Z_{emp} \in N(0,1)$, което открива следната възможност за проверка. При зададено ниво на значимост $\alpha$, нулевата хипотеза се отхвърля когато $|Z_{emp}| \geq z_{1-\frac{\alpha}{2}}$ и се приема в противен случай. Нулевата хипотеза може да се запише в опростен вид като предположение за равенство на медианите

$H_0 : Me_X = Me_Y$

при което двустранната алтернативна хипотеза се свежда до

$H_{alt} : Me_X \neq Me_Y$ .

При насочена алтернатива $H_{alt} : Me_X > Me_Y$ , критичната област за отхвърляне на $H_0$ има вида $Z_{emp} \geq z_{1-\alpha}$, а при насочена алтернатива $H_{alt} : Me_X < Me_Y$, критичната област за отхвърляне на $H_0$ има вида $Z_{emp} \leq z_{\alpha}$ .

За примера 19.4 със сравнение между мъже и жени по скалата за адаптация към средата, пресмятанията показват $Z_{emp} = -0.297$ с оценено ниво на значимост $p = 0.766$, което също както при теста на Стюдънт води до приемане на нулевата хипотеза.

"$U$ -тестът на Ман-Уитни показва несъществено различие между нивото на адаптация при жените и мъжете, [ $Z = -0.297$ ; $p = 0.766$ ]."

За примера 19.5 със сравнение между мъже и жени по средния годишен успех имаме $Z_{emp} = 6.953$ с оценено ниво на значимост $p = 0.000$, което отново както при теста на Стюдънт води до отхвърляне на нулевата хипотеза.

"$U$ -тестът на Ман-Уитни показва съществено различие за средния годишен успех в полза на жените, [ $Z = 6.953$ ; $p = 0.000$ ]."

В типичния случай $t$ -тестът на Стюдънт и $U$ -тестът на Ман-Уитни водят до едно и също решение за отхвърляне или приемане на нулевата хипотеза. Когато обемите на извадките са малки или по други причини се наблюдава отклонение от нормалното разпределение, за по-голяма убедителност на извода могат да се прилагат и двата теста, надявайки се, че както е в типичния случай, те ще доведат до един едно и също заключение.

Различните програмни среди за статистическа обработка не се придържат към единен формализъм при описване на резултатите от $U$ -теста, поради което потребителят трябва внимателно да прочете помощната информация, за да може да приведе резултатите в правилен вид.

***Тест на Кръскал-Уолис.*** Тестът на Кръскал-Уолис се явява непараметричен аналог на еднофакторния дисперсионен анализ и се явява своеобразно обобщение на теста на Ман-Уитни. Нулевата хипотеза тук се състои в предположение за равенството на популационните медиани за отделните групи

$H_0 : Me_1 = Me_2 = \cdots = Me_J$ ,

която се проверява срещу алтернативата, че някои от тези медиани са различни. Отхвърлянето или приемането на $H_0$ се интерпретира по същия начин както при еднофакторния дисперсионен анализ. Както при $U$ -теста на Ман-Уитнии отначало

обединяваме извадките и пресмятаме поредния номер на всяко наблюдение в обединената извадка. Проверяващата статистика за $H_0$ има вида

$$H_{emp} = \frac{12}{N(N+1)} \sum_{j=1}^{J} \frac{W_j^2}{n_j} - 3(n+1),$$

където $W_j$ е сумата от ранговете на наблюденията за група с номер $j$. Ако хипотезата $H_0$ е валидна, то величината $H_{emp}$ се подчинява приблизително на $\chi^2(J-1)$ разпределение. При зададено ниво на значимост $\alpha$, нулевата хипотеза се отхвърля когато $H_{emp} \geq \chi^2_{1-\alpha;J-1}$ и се приема в противен случай.

За двата примера с успеха 19.6-7 имаме следните резултати

Таблица 19.7.

| | $\chi^2(3)$ | $p$ |
|---|---|---|
| жени | 0.682 | $p = 0.887$ |
| мъже | 34.649 | $p = 0.000$ |

Последните резултати водят до аналогични на направените вече заключения за отхвърляне или приемане на нулевата хипотеза.

"Тестът на Кръскал-Уолис показва несъществено различие при групата на жените между средния успех от четирите курса, $[\chi^2(3) = 0.682\,;\, p = 0.887]$, и статистически значимо различие при групата на мъжете, $[\chi^2(3) = 34.649\,;\, p = 0.000]$."

Изобщо в типичния случай тестът на Кръскал-Уолис води до заключения аналогични на тези от еднофакторния дисперсионен анализ, затова когато желаем повече аргументация в спорни ситуации можем да приведем резултатите и от двата теста.

Получените резултати могат да се илюстрират посредством диаграми за локализация на разпределенията в отделните групи. При тези диаграми локализацията е фиксирана чрез разположението на трите квартила, вторият от които е медианата.

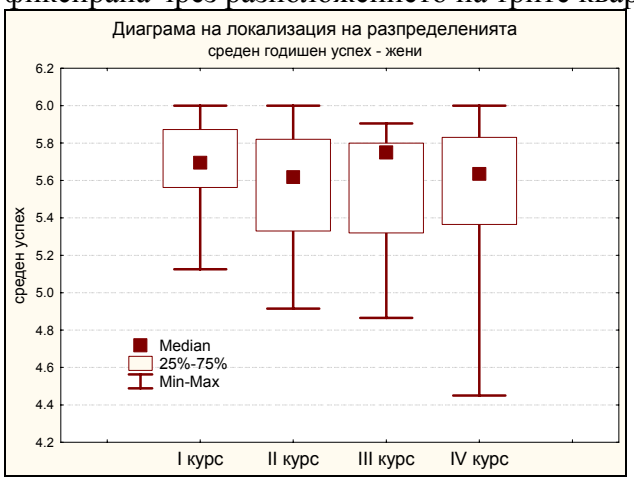


Рис. 19.10.

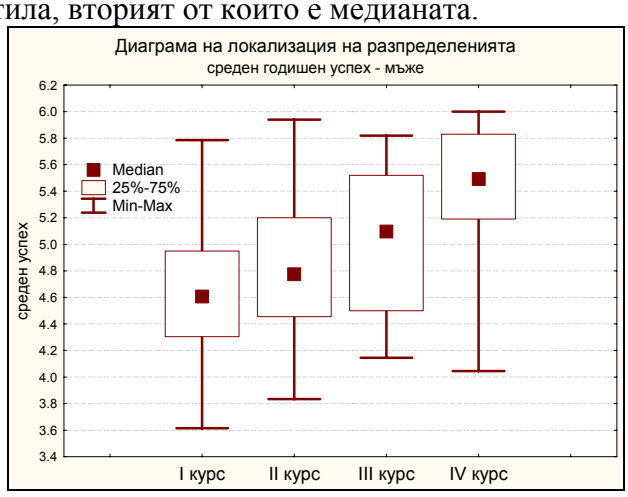


Рис. 19.11.

Тестът на Кръскал-Уолис предлага и аналог на Post-Hoc анализа в случай на отхвърляне на нулевата хипотеза.

Когато подреждаме наблюденията в поредни номера възниква технически проблем какво да се прави за наблюдения с равни стойности. В такъв случай на всяко от тях се присвоява среден ранг. Всъщност за това както и за почти всичко останало се грижи компютърът. От потребителят се иска само добро познаване основната схема на метода както и възможните интерпретации на резултатите в предметната област.