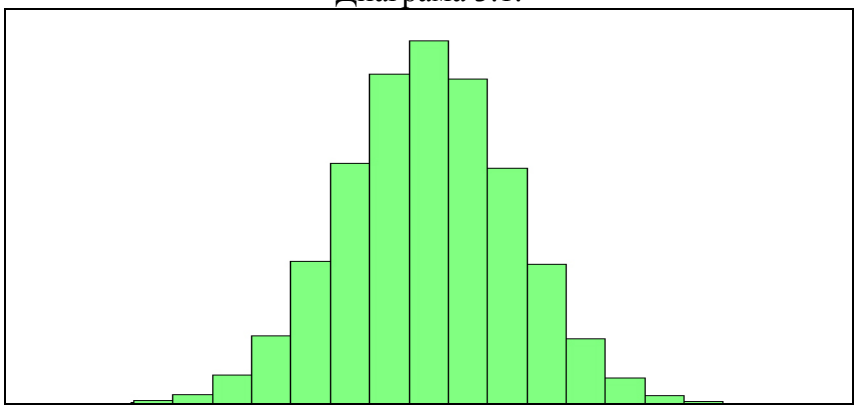
§3. Нормални разпределения. Линейна корелация

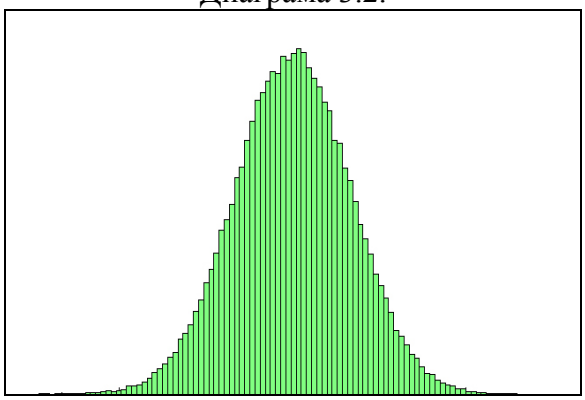
1. Определение за нормално разпределение. Практиката показва, че типичното разпределение на дадена метрична величина X има следната характерна форма, при която се наблюдава ясно изразен ***център*** на разпределението (***с единствена мода)***, както и две ***симетрично*** разположени ***опашки***.

Диаграма 3.1.

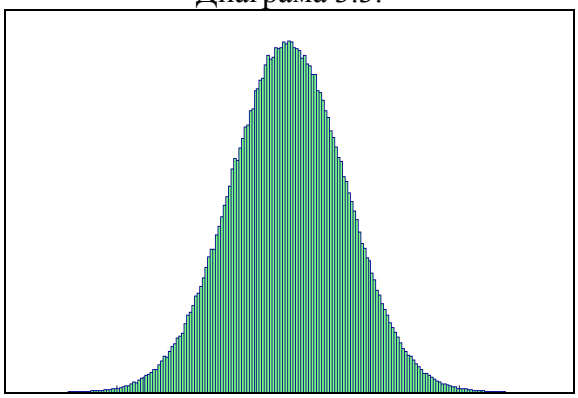


В такъв случай се казва, че наблюдавана величина X е разпределена нормално или още, че X следва ***нормално разпределение*** (***normal distribution***). Ако въпросната величина наистина следва някакво нормално разпределение, то с увеличаване броя на наблюденията хистограмата се стреми да приеме своеобразно гранично положение, при което горните страни на нейните стълбове се превръщат постепенно в непрекъсната линия, както е показано на следващите две диаграми.

Диаграма 3.2.



Диаграма 3.3.



Въпросната непрекъсната линия представлява графиката на ***плътността*** (***probability density function – pdf***) за конкретното нормално разпределение. В общия случай плътността се задава от формулата

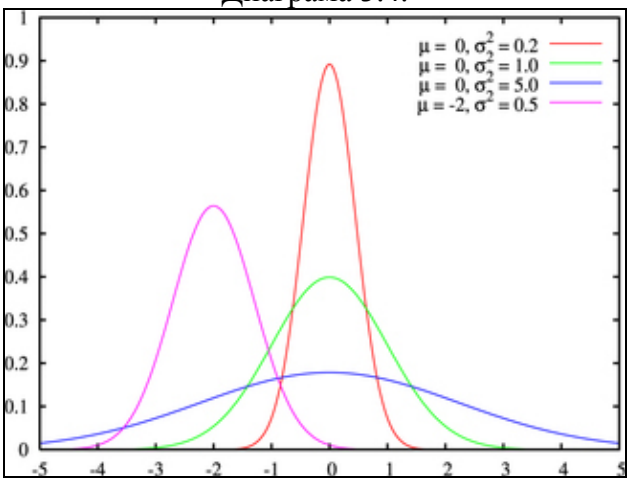
$$(3.1) \qquad f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

където числата μ и $\sigma>0$ са параметри на теоретичното нормално разпределение $N(\mu,\sigma^2)$. Посредством записа $X\in N(\mu,\sigma^2)$ указваме, че наблюдаваната величина следва теоретично нормално разпределение с параметри μ и σ^2. Наличието на два параметъра показва, че всъщност разполагаме с цяла съвкупност от нормални разпределение.

Параметърът μ се нарича ***теоретично средно*** на разпределението и на практика представлява стойността, около която се стабилизира емпиричното средно $\bar{x}$ при неограничено нарастване обема на извадката. Параметърът μ може да се интерпретира

като средната стойност на X, отнесено към цялата популация и по тази причина μ се нарича още ***средно за популацията***. По аналогичен начин се разглежда и интерпретира параметърът σ – като ***теоретично*** (***популационно***) ***стандартно отклонение***, съответно σ^2 – като ***теоретична*** (***популационна***) ***дисперсия***. Популационната дисперсия σ^2 представлява стойността, около която се стабилизира емпиричната дисперсия s^2 при неограничено нарастване обема на извадката. На следващата графика са показани различни форми на нормални разпределения.

Диаграма 3.4.



При статистическите анализи на една метрична променлива обикновено се предполага, че тя се подчинява на някакво нормално разпределение. Нормалното разпределение е оптимално по отношение на математическата структура, понеже се определя само от два параметъра – един за популационното средно и един за популационната дисперсия. Всеки гъвкав модел от общ характер трябва да съдържа поне един параметър за определяне на централна тенденция и поне един параметър за изменчивост около въпросната централна тенденция. В този смисъл нормалното разпределение се явява максимално икономично, понеже съдържа точно два параметъра, които представят непосредствено централната тенденция и изменчивостта относно нея. В типичния случай са налице достатъчно основания да предполагаме, че метричните величини в психологията се подчиняват на нормално разпределение или поне не се отклоняват съществено от него. Сравнително точни критерии за въпросната нормалност ще бъдат приведени по-нататък. Благоприятен за този начин на мислене се явява фактът, че отклоненията от нормалното разпределение губят постепенно своето значение при нарастване обема на извадката. Прецизният анализ обаче изисква отчитането на това съображение.

Строгите аргументи в полза на нормалността на дадена величина в типичния случай се базират на известната ***централна гранична теорема*** (***central limit theorem***) от теорията на вероятностите, която в свободен текст гласи, че ако дадена величина X със случаен характер може да се разглежда като сбор от достатъчно много (например повече от 30) други такива, то X се подчинява (приблизително) на нормално разпределение.

Казаното дотук дава основание да говорим за ***основен статистически модел за една метрична величина*** X, който се състои в предположението, че X следва нормално разпределение с някакво популационно средно μ и някаква популационна дисперсия σ^2. За да бъде детерминиран този модел е необходимо да познаваме с известна точност стойностите на μ и σ^2. Информацията за тези параметри е скрита донякъде в стойностите от наблюденията, т.е. в резултатите от извадката. Може да се

обоснове математически, че емпиричното средно от наблюденията $\bar{x}$ се явява напълно удовлетворителна оценка за популационното средно μ, а емпиричната дисперсия s^2 се явява такава оценка за популационната дисперсия σ^2.

Таблица 3.1.

	статистика	***теоретичен (популационен) параметър***
средно	$\bar{x}$	μ
дисперсия	s^2	σ^2
стандартно отклонение	s	σ

В този случай се казва, че статистиката $\bar{x}$ се явява ***точкова оценка*** за μ (означава се още $\hat{\mu} = \bar{x}$). Аналогично статистиката s^2 се явява точкова оценка за σ^2 $\left(\hat{\sigma}^2 = s^2\right)$ и s се явява точкова оценка за σ $\left(\hat{\sigma} = s\right)$.

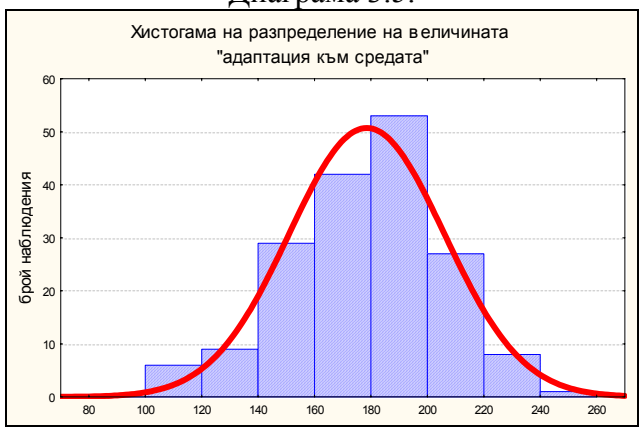
С нарастване обема на извадката n, точковите оценки, получени от емпиричните стойности стават все по близки до теоретични стойностите и по тази причина най-важната характеристика на дадена извадка е нейният обем.

За примера с величината "адаптация към средата", разгледана в **§2** имаме $\bar{x} = 178.514$ и $s = 27.523$, което задава конкретна крива на плътност от вида (3.1)

$$f(x) = \frac{1}{27.523\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-178.514}{27.523}\right)^2}.$$

Като изчертаем графиката на тази плътност върху хистограмата (отчитайки мащаба) се получава следната диаграма.

Диаграма 3.5.



В този случай се наблюдава добро визуално съвпадение, което е белег за валидност на предположението за нормално разпределение.

Основните отклонения от нормалното разпределение се изразяват чрез термините на ***изкривяване*** (***skewness***) и ***ексцес (kurtosis)***, за които въз основа на данните от извадката се пресмятат коефициенти

$$skewness = \frac{n}{(n-1)(n-2)} \sum_{k=1}^{n} \left(\frac{x_k - \bar{x}}{s}\right)^3,$$

$$kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{k=1}^{n} \left(\frac{x_k - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

Стойност на коефициента *skewness* близка до нула показва, че разпределението е симетрично, а стойност на коефициента *kurtosis* близка до нулата показва, че центърът на разпределението е "полегат" по характерния за нормално разпределение начин. В този смисъл стойности близки до нула на тези два коефициента представляват белег за нормално разпределение. За примера имаме $skewness = -0.460$ и $kurtosis = -0.104$.

На практика винаги се налага да се примиряваме с известни отклонения от нормалното разпределение, за което вече споменахме че в определен контекст не представляват заплаха за качеството на статистическите анализи. Трудно може да се надяваме, че в една реална ситуация визуалното съвпадение между хистограмата и съответната нормална крива може да изглежда по-добре от това на диаграма 3.5.

Всички параметри на емпиричното разпределение имат своя теоретичен еквивалент и обратно. В определен смисъл основната задача на статистиката е да се оценят теоретичните параметри въз основа на емпиричните, т.е. да детерминираме теоретичния статистическия модел, който се отнася за популацията, според данните, които разполагаме от извадката. Тук на преден план излиза фактът, че статистическият модел има абстрактен математически характер, който модел се отнася към определяне на специфичните количествени закономерности в популацията. Тази математическа насоченост не може да бъде отхвърлена напълно, понеже математиката е онази наука, която се занимава систематично с изграждане на количествени модели. Освен това още веднъж се вижда ясно колко е важно извадката да бъде представителна (непреднамерена), понеже изводите направени върху преднамерена извадка не могат да бъдат отнесени строго към популацията.

Поради изложените по-горе причини, средното и дисперсията се наричат ***основни статистически параметри*** и имат важно значение даже за величини, които не се подчиняват на нормално разпределение.

2. Преобразуване на нормални разпределения. Линейните преобразования не променят нормалния характер на разпределението. Ако $X \in N\left(\mu_x, s_x^2\right)$, а $\alpha \neq 0$ и β са някакви числа, то величината $Y = \alpha X + \beta$ също следва нормално разпределение, със средно $\mu_y = \alpha\mu_x + \beta$ и дисперсия $\sigma_y^2 = \alpha^2\sigma_x^2$, т.е. $Y \in N\left(\alpha\mu_x + \beta, \alpha^2 s_x^2\right)$. Освен това при произволно разпределение на величината X с наблюдавани в някаква извадка стойности x_1, x_2 , ..., x_n, преобразуваните по формулата $y_k = \alpha x_k + \beta$, стойности имат средно $\bar{y} = \alpha\bar{x} + \beta$ и дисперсия $s_y^2 = \alpha^2 s_x^2$. Тези свойства на практика дава възможност да преминаваме от едно нормално разпределение към друго.

Най-често се ползва преобразуване към ***нормално стандартно разпределение***, което има средно нула и дисперсия единица. Отнесено към данните, това означава да пресметнем z-стойностите по формулата

$$z_k = \frac{x_k - \bar{x}}{s},\ k = 1,2,\ldots,n.$$

Тези z-стойности имат средно $\bar{z} = 0$ и дисперсия $s_z^2 = 1$. В някои приложения се извършва преминаване към T-стойности по формулата

$$t_k = 10z_k + 50 = 10\frac{x_k - \bar{x}}{s} + 50,\ k = 1,2,\ldots,n,$$

които имат средно $\bar{t} = 50$ и стандартно отклонение $s_t = 10$ (дисперсия $s_t^2 = 100$).

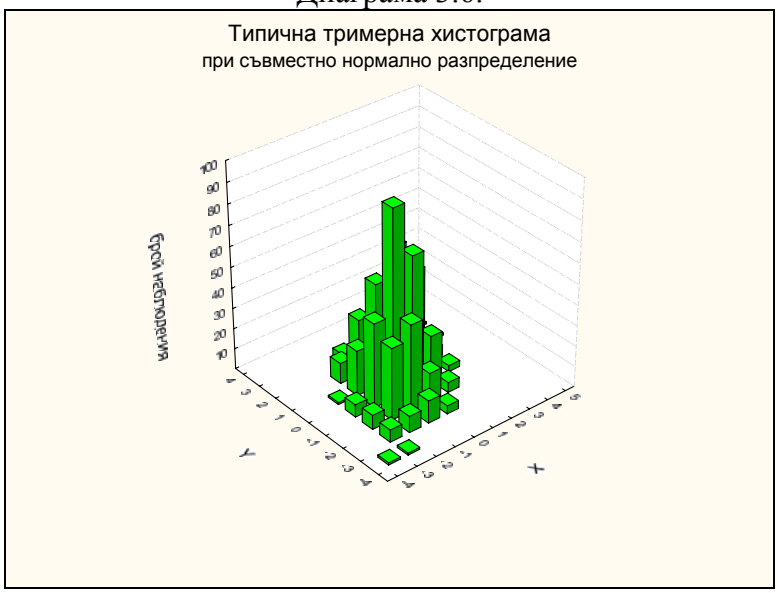
Преминаването от натурални стойности към z-стойности се нарича ***стандартизиране***. Стандартизирането от своя страна включва две последователни действия – ***центриране***, което се състои в изваждане на средното и ***нормиране***, което се състои в разделяне на стандартното отклонение.

Линейните преобразувания на дадена величина фактически означават ***промяна на центъра и мащаба*** на интервалната скала, с която е свързана въпросната величина, както е в типичния случай на статистически анализ на психологични величини. Тези преобразования са продиктувани от съображения за техническо удобство при следващите анализи и сами по себе се не представляват съдържателен анализ.

3. Двумерно нормално разпределение. Статистическа зависимост. Главната задача на статистическите анализи е установяване на количествени връзки от статистически характер между наблюдаваните величини. Да разгледаме основния случай на две 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.

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

Диаграма 3.6.

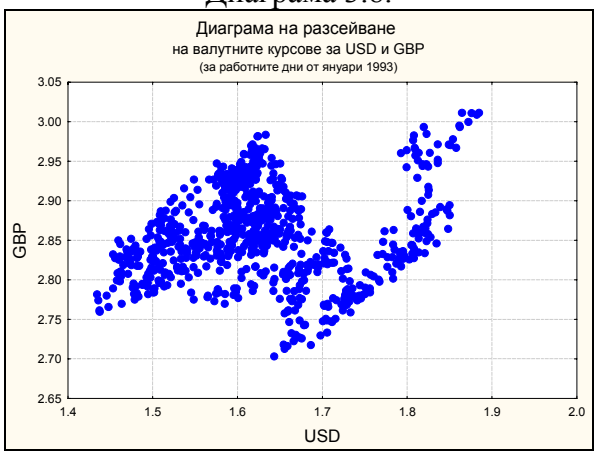


Формата на взаимно разпределение личи от диаграмата на разсейване, която се получава като отбележим наблюденията в една правоъгълна координатна система.

Диаграма 3.7.

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.

Диаграма 3.8.



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

Таблица 3.1.

X	Y
x_1	y_1
x_2	y_2
...	...
x_n	y_n

По тази таблица можем да пресметнем индивидуалните средни $\bar{x}$ и $\bar{y}$, а също така и индивидуалните дисперсии s_x^2 и s_y^2. Остава да решим главния проблем за изразяване взаимната зависимост между X и Y. Ако съвместното разпределение на метричните величини X и Y е нормално, то се оказва, че ***цялата взаимна статистическа зависимост*** между тях, заложена в данните от извадката, се изразява с едно единствено число

$$(3.2) \quad r = r[X,Y] = r[Y,X] = \frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{(n-1)s_y s_y},$$

което се нарича ***емпиричен (извадков) коефициент на линейна корелация*** или още извадков корелационен коефициент на Пирсън. Величината

$$\operatorname{cov}[X,Y] = \operatorname{cov}[Y,X] = \frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})}{(n-1)}$$

се нарича ***емпирична*** (***извадкова***) ***ковариация*** на X и Y. По този начин

$$r = \frac{\operatorname{cov}[X,Y]}{s_x s_y}.$$

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.

Пресметнатият по формула (3.2) коефициент представлява всъщност ***точкова оценка*** за ***теоретичния*** (***популационния***) коефициент на корелация ρ. Двете величини X и Y са ***статистически независими***, когато са разпределени независимо, т.е. когато разпределението на едната не съдържа информация за разпределението на другата. Статистическата независимост означава, че по наблюденията над едната

величина не можем да правим заключения за поведението на другата величина. Статистическата независимост има ***взаимен*** характер. При нормално съвместно разпределение независимостта означава, че популационният коефициент на линейна корелация ρ 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 ρ 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 ρ 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. Тук е добре да се има предвид, че ако корелационният коефициент между суровите балове на два психологически теста надвишава по абсолютна стойност 0.8, то има достатъчно основания да се приеме, че двата теста фактически измерват един и същ психологически конструкт.

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.

4. Уравнение на регресия. Статистическият характер на връзката между дадени величини X и Y е сравнително трудна за детайлна интерпретация. 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 средното на стойностите на Y, при това фиксирано x. Последното твърдение изисква известно уточнение от технически характер, което няма да привеждаме.

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 статистически еквивалент

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

Последното в разгърнат вид изглежда така

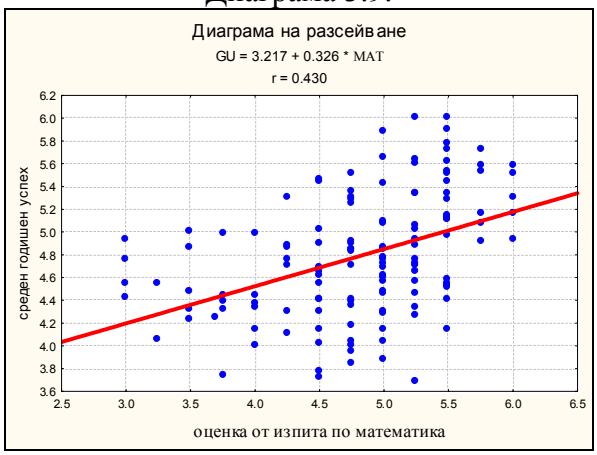
(3.3) $$y=r\frac{s_y}{s_x}x+\left(\bar{y}-r\bar{x}\frac{s_y}{s_x}\right),$$

което представлява линейно уравнение. Естествената функционална връзка между две величини с нормално съвместно разпределение се оказа ***линейна***. Уравнението (3.3) 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. Уравнението (3.3) може да послужи за предсказване стойността на Y при известна стойност на X.

В случая когато $r=0$, уравнението (3.3) 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.

В качеството на пример да разгледаме величината GU -среден годишен успех от семестриалните изпити на група обучаеми-студенти и величината MAT -оценка от конкурсния изпит по математика. Данните са от файла NEVENA22, а обемът на извадката е $n=135$. Диаграмата на разсейване 3.9 показва несъществено отклонение от съвместно нормално разпределение.

Диаграма 3.9.



В този случай имаме $r=0.430$, а статистическото уравнение на линейна регресия има вида

$$GU=0.326*MAT+3.217.$$

Върху диаграма 3.9 е отбелязана и линията на регресия. 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 $4.604=0.326*4.25+3.217$.

5. Причинно-следствени връзки. Последният пример по естествен начин асоциира върху извода, че успехът от конкурсният изпит MAT влияе съществено върху средния годишен успех от семестриалните изпити GU. С други думи налице е причинно-следствена връзка между психичните конструкти, които обуславят проявата на тези две величини.

Установяването на ***причинно следствени*** (***каузални***) отношения между две или повече величини в много случаи се явява основна цел на дадено изследване. Тук обаче трябва да се имат предвид няколко неща. Първо статистиката сама по себе се не притежава (и няма как да притежава) никакво средство за ***доказване*** на причинно-следствени отношения. Всичките статистически зависимости между величините имат симетричен – взаимен характер. Статистиката обаче разполага със средства за придаване ***количествен израз*** на причинно-следствени връзки, когато те действително съществуват. Ако величините X и Y се намират в причинно-следствени отношения, то математическият израз на това отношение представлява уравнението на регресия на Y върху X . По-нататък ще разгледаме по-сложни модели на каузални схеми.

Причинно-следствените отношения могат да бъдат обосновавани единствено в рамките на предметната научна област – в нашия случай в областта на науката психология. Единственият твърд индикатор, който различава безспорно причината от следствието е фактът, че причината предхожда следствието във времето. Този индикатор обаче не се явява достатъчен. Могат да се приведат прости примери, в които е налице времево предхождане и силна корелация, но между величините очевидно да няма каузална връзка. В такива случаи обикновено и двете величини могат да се разгледат като следствие от трета ***явна величина*** (***manifest variable***) или ***скрита величина*** (***latent variable***).

В психологическите изследвания в качеството на величини-причини обикновено се разглеждат личностните черти, етническата и религиозната идентификация, полът, възрастта или други физиологични черти. Тези величини имат изразен устойчив характер във времето и напълно естествено е да предположим, че имат определено влияние върху неустойчивите прояви на личността като емоционални реакции, нива на адаптация към средата, динамични поведенчески стереотипи и т.н. Към настоящия момент не съществуват общоприети единни теории за личността, както и прости таксономии на психичните конструкти, което осигурява известна свобода на изследователя в съставянето на каузални схеми.