

1. Kuidas defineeritakse sündmuse tõenäosus? Kuidas leida sündmuste korrutise ja summa tõenäosust?

ülesanne: Urnis on **8 rohelist**, **7 punast** ja **5 sinist** kuuli. Urnist võetakse välja **3** kuuli. Kui suur on tõenäosus, et:

- 1) kõik need kuulid on **rohelised**,
- 2) kõik need kuulid on erinevat värvi,
- 3) Vähemalt üks kuul on **punane**.
- 4) Vähemalt üks kuul on **punane** ja vähemalt üks kuul on **sinine**.
- 5) Vähemalt üks kuul on **punane** või vähemalt üks kuul on **sinine**.

2. Kuidas on defineeritud pideva juhusliku suuruse **tihedusfunktsioon** $p(x)$ ja **jaotusfunktsioon** $F(x)$? Kuidas on nad omavahel seotud?

ülesanne: Juhusliku suuruse jaotusfunktsioon $F(x)$ on defineeritud allpooltoodud kujul.

$$F(x) := \begin{cases} x^{-2} & \text{if } x \leq -1 \\ 1 & \text{otherwise} \end{cases}$$

Koostada tihedusfunktsiooni $p(x)$ avaldis ning teha mõlema funktsiooni graafikud.

Leida tõenäosused selleks (**nii tihedusfunktsiooni kui ka jaotusfunktsiooni kaudu**), et selle juhusliku suuruse väärtused:

- 1) ei ületa arvu $x = -5$;
- 2) satuvad vahemikku $(-10, -1)$;
- 3) ei satu vahemikku $(-3, -2)$.

3. Mis on punkthinnangud ja vahemikhinnangud?

ülesanne: Leia antud mõõteseeria järgi uuritud suuruse täpse väärtuse mõõteriista mõõtmistäpsuse punkthinnangud ja 90% usaldusnivooga vahemikhinnangud.

(2.34 2.38 2.32 2.36 2.35 2.31 2.35 2.29 2.30 2.35)^T

4. Mismoodi püstitatakse statistilised hüpoteesid ja mismoodi neid kontrollitakse?

ülesanne: Kontrolli, kas antud mõõteseeriade aritmeetiliste keskmiste erinevus on statistiliselt oluline 5% olulisusnivooga. Mõlemad mõõteseeriad on tehtud sama mõõteriistaga, mille mõõtmistäpsus 0.05.

$(2.34 \ 2.38 \ 2.32 \ 2.36 \ 2.35 \ 2.31 \ 2.35 \ 2.29 \ 2.41 \ 2.35)^T$

$(2.24 \ 2.27 \ 2.22 \ 2.33 \ 2.29 \ 2.30 \ 2.24)^T$

5. Mis on regressioon ja korrelatsioon?

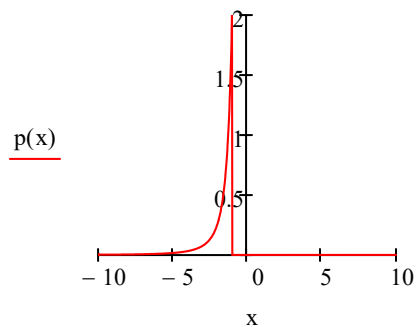
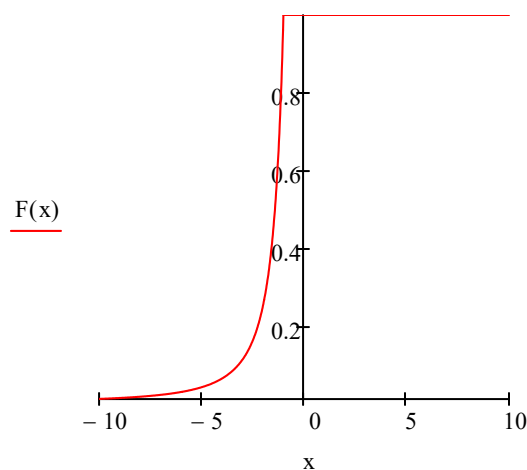
ülesanne: Kuidas on omavahel seotud suurused **x** ja **y**? Koosta regressioonisirge. Kas peaks kasutama kõrgemat järku regressioonipolünoomi?

Milline on nendevaheline korrelatsioonikordaja? Mida see

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -18 & 10 & 22 & 55 & 2 & -30 & -15 & 5 & 10 & 18 \\ 6 & 12 & 7 & 10 & 1 & 16 & 8 & 2 & 4 & -1 \end{pmatrix}$$

1.ülesanne

$$F(x) := \begin{cases} \frac{1}{x^2} & \text{if } x \leq -1 \\ 1 & \text{if } x > -1 \end{cases} \quad p(x) := \begin{cases} \frac{-2}{x^3} & \text{if } x \leq -1 \\ 0 & \text{if } x > -1 \end{cases}$$



Juhusliku suuruse väärtused:

-- ei ületa arvu -5:

$$\int_{-\infty}^{-5} p(x) dx = 0.04 \quad F(-5) = 0.04$$

-- satuvad vahemikku (-10,-1): $\int_{-10}^{-1} p(x) dx = 0.99 \quad F(-1) - F(-10) = 0.99$

Ei satu vahemikku (-3,-2): $1 - \int_{-3}^{-2} p(x) dx = 0.861 \quad 1 - F(-2) + F(-3) = 0.861$

5.ülesanne

$r_0 := -18$ $r_1 := 10$ $r_2 := 22$ $r_3 := 55$ $r_4 := 2$ $r_5 := -30$
 $r_6 := -15$ $r_7 := 5$ $r_8 := 10$ $r_9 := 18$
 $s_0 := 6$ $s_1 := 12$ $s_2 := 7$ $s_3 := 10$ $s_4 := 1$ $s_5 := 16$
 $s_6 := 8$ $s_7 := 2$ $s_8 := 4$ $s_9 := -1$

KATA := $\begin{pmatrix} r_0 & s_0 \\ r_1 & s_1 \\ r_2 & s_2 \\ r_3 & s_3 \\ r_4 & s_4 \\ r_5 & s_5 \\ r_6 & s_6 \\ r_7 & s_7 \\ r_8 & s_8 \\ r_9 & s_9 \end{pmatrix}$ $i := 0..9$ $rr := -30..55$
vs -- interpolateerimine ruutpolünoomiga
vk -- interpolateerimine kuuppolünoomiga

$r_i := \text{KATA}_{i,0}$, $s_i := \text{KATA}_{i,1}$

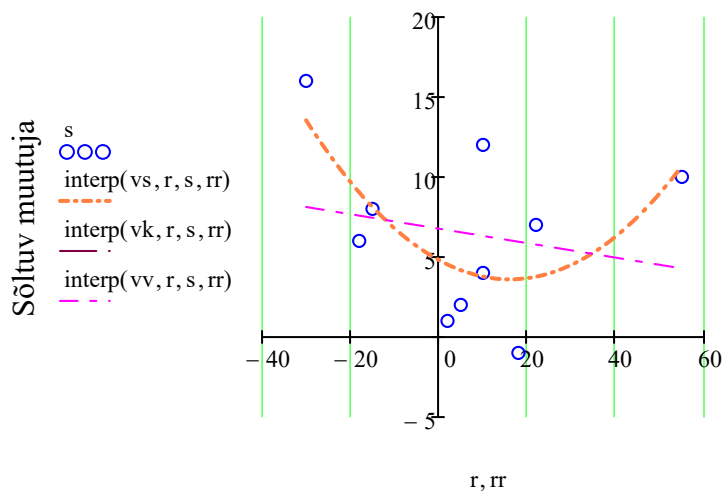
$\text{corr}(r,s) = -0.206$ langev trend

$vs := \text{regress}(r,s,2)$

$vv := \text{regress}(r,s,1)$

$vk := \text{regress}(r,s,3)$

Polünoomiaalne regressioon



Sõltumatu muutuja

$r_{\text{ruutkordaja}} := \text{corr}(r, s)^2 = 0.042$

- Andmed
- Ruutpolünoom
- Kuuppolünoom
- - - - - lineaarne lähendus

4. ülesanne -- mõõtmistäpsuste erinevuse hindamine. Olgu usaldusnivoo $\alpha = 0.05$.
 H_0 : Mõõteseeriate aritmeetiliste keskmiste erinevus ei ole statistiliselt oluline olulisusnivool 0.05 (eeldame seda, sest teada on, et mõlemad mõõteseriad on tehtud sama mõõteriistaga).
 H_1 : H_0 väär.

$x := (2.34 \ 2.38 \ 2.32 \ 2.36 \ 2.35 \ 2.31 \ 2.35 \ 2.29 \ 2.41 \ 2.35)$ $n_1 := 10$

$y := (2.24 \ 2.27 \ 2.22 \ 2.33 \ 2.29 \ 2.30 \ 2.24)$ $n_2 := 7$

$s := 0.05$ see on antud mõõteriista mõõtetäpsus -- üksikmõõtmise standardhälve H_0 järgi

$m_x := \text{mean}(x) = 2.346$ $m_y := \text{mean}(y) = 2.27$

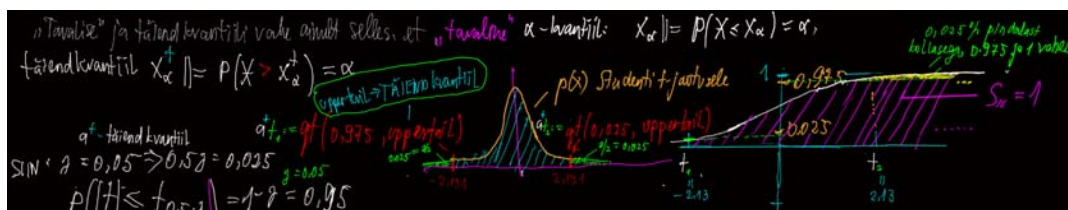
$s_1 := \text{Stdev}(x) = 0.034$ $s_2 := \text{Stdev}(y) = 0.039$

$$\frac{m_x - m_y}{s} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} < t_{0.5 \cdot \gamma} \quad \text{vaja uurida seda}$$

$$t := \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \cdot \frac{m_x - m_y}{s} = 3.084 \quad qt(0.975, n_1 + n_2 - 2) = 2.131$$

Kriitilise teststatistiku väärtuse saame t-jaotuse (täiend)kvantiilfunktsiooni qt kasutades. Hindame kumulatiivse jaotusfunktsiooni pöördfunktsiooni (kvantiilfunktsiooni) valimi abil. Praegune tulemus tähendab, et 97.5% tõenäosusega on statistiku väärtus antud vabadusastmete korral suurem kui 2.131, sest leidsime 0.975-täiendkvantiili. Kuna t-jaotus on sümmeetriline, siis analoogiliselt on tõenäosus 97.5% mõõta statistiku väärtuseks vähem kui -2.131. Kuna valimi põhjal leitud statistik on oluliselt suurem kui 2.131 (kuulub kriitilisse piirkonda), siis peame hüpoteesi H_0 "mõõteseeriate aritmeetiliste keskmiste erinevus ei ole statistiliselt oluline olulisusnivool 0.05" ümber lükkama.

Märkus: praeguses ülesandes on kumulatiivse jaotusfunktsiooni asemel statistikuna kasutusel empiiriline jaotusfunktsioon, mis on kumulatiivse jaotusfunktsiooni nihketa, konsistentne ja ilmselt ka mõjus punkthinnang.

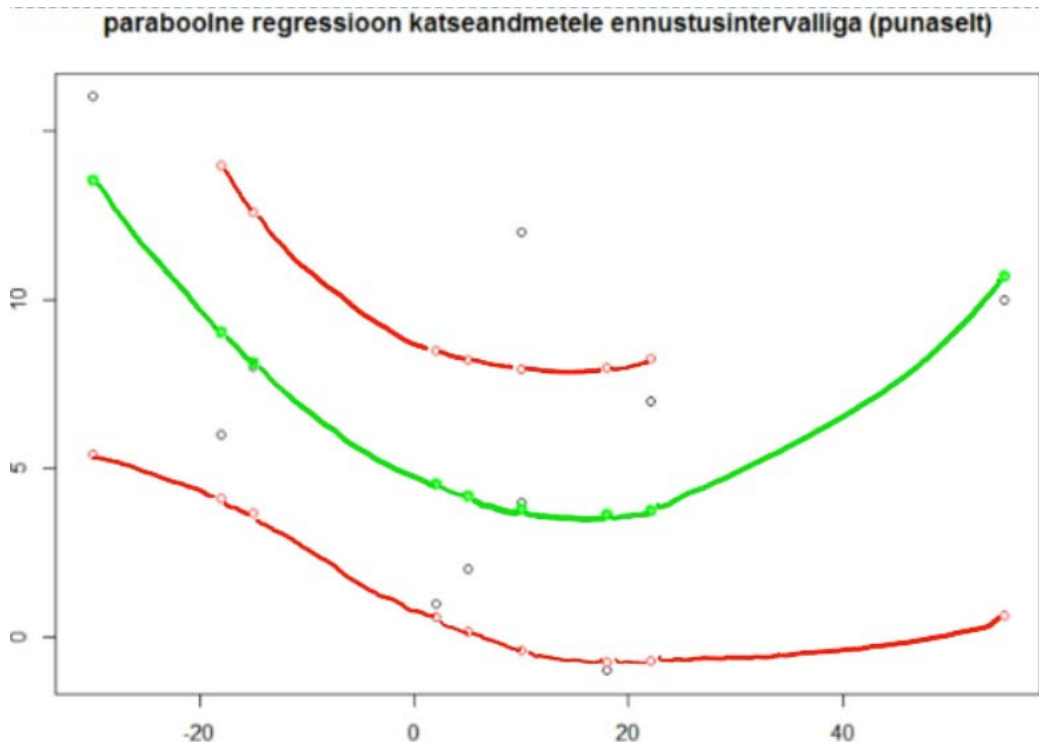


Kindlasti peaks kasutama kõrgemat järku regresioonipolünoomi, sest:

1. antud andmete graafik ei tundunud graafilisel uurimisel kaugeltki mitte lineaarne -- jäädid oleksid liiga suured võrreldes selle osaga, mida mudel seletab.
2. Korrelatsioonikordaja väärtus on ainult -0.045, mis on põhimõtteliselt nii väike seepärast, et x-i standardhälve on suur ja x-y korrelatsioon väike -- seega korrelatsioonikordaja $b = r \cdot \text{sd}(y) / \text{sd}(x)$ väärtus on samuti väga väike.
3. Lineaarse mudeli tõusu p-väärtus on kõrge -- see tähendab seda, et 1. liiki vea tegemise võimalus on lubatust (näiteks olulisusnivool 5%) oluliselt kõrgem -- seetõttu ei ole lineaarne sõltuvus statistiliselt oluline.

Võimalused edasiseks analüüsiks: mõistlik oleks kasutada 2-järku polünomiaalset regressiooni, st lähendada katseandmed parabooliga.

Tehes analüüsi 2. järku polünomiaalsele regressioonmudelile, osutuvad vabaliige ja ruutliikme kordajad statistiliselt olulisteks ja r^2 (r -ruutkordaja ehk korrelatsioonikordaja ruudus) väärtus on 0.467, st x-i varieeruvus sellise mudeli korral seletab ligikaudu 47 % y-i varieeruvusest -- tihti eeldatakse seda, et kui r^2 on 0.4 ja 0.8 vahel, siis on mudel huvitav. Antud ruutparabooli korral nii ongi.



4. Mismoodi püstitatakse statistilised hüpoteesid ja mis moodi neid kontrollitakse?

Kuna otsuseid saab reaalses maailmas teha vaid teatud kindlusega, siis peame püstitama statistilised hüpoteesid – nullhüpotees H_0 ja alternatiivne hüpotees H_1 . Statistilisi teste kontrollitakse enamasti riskiprotsenti või usaldusnivood α kasutades (enamasti $\alpha=0.05$), mis määrab ära maksimaalse tõenäosuse teha 1. liiki viga – st H_0 hüpotees ümber lükata, kui see on tegelikult õige.

Oluline ongi mõista, et klassikaline statistika ehk frequentist statistika ei võimalda andmetest otse ilma populatsiooni kohta täiendavaid eeldusi tegemata populatsiooni kohta ka järeldusi teha ja seetõttu põhinevad arvutused nullhüpoteesi kehtivusel, näiteks et kasside toortoidu ja purgitoidu jaotuste keskmiste erinevus populatsioonis (üldkogumis) on 0. See vastaks näiteks normaaljaotusel mediaanile (keskmisele) jaotuse keskel ja ühtlasi määrab siis ära populatsiooni kirjeldava jaotuse asukoha arvteljel. Nullhüpotees valitakse tavaliselt analoogselt "süütu kuni pole süüdi mõistetud" kaalutlustele – see tähendab, et otsitav efekt ei loeta tõestatuks. Mõtleme: Kui aga nullhüpoteesiks valida "efekt tõestatud" juhus, siis tekiks kohe küsimus, et kui suur see efekt on – ja tõenäosusele on raske anda mingit täpset hinnangut, sest eelnevalt ei tea ju, kui suur efekt tegelikult ikkagi on – seetõttu valitaksegi nullhüpotees "algselt eeldame, et üldkogumis efekti pole" alusel – see tähendab, et näiteks kassi purgitoidu ja toortoidu vaheline keskmiste erinevus on 0 – seda pole võimalik valesti arvata, sest kui efekti pole, siis ka keskmiste erinevus peaks olema 0. Loodetakse leida valimi põhjal arvutatud valimi-statistiku väärtus, mis H_0 kehtivusel on võimalikult ekstremaalne – loodetavasti üle kriitilise teststatistiku väärtuse, mille korral on leitud vastuolu H_0 -ga sel olulisusnivool ja seega H_0 lükatakse ümber.

Kui nullhüpoteesi alusel on populatsiooni kirjeldava jaotuse asukoht paigas, siis sellega on määratud teststatistiku-jaotus – selle keskmise x -asukoht on x -telje 0-is. Valimi põhjal arvutatakse välja valimi-statistik ja eeldades H_0 hüpoteesi kehtivust, arvutatakse välja või vaadatakse tabelist (vastava vabadusastmete jaoks) kriitiline teststatistiku väärtus (ja sellega seonduv p -väärtus, mis annab tõenäosuse (valimi statistikuks) mõõta kriitilise teststatistikuga võrdne või ekstremaalsem väärtus (st valimi-statistik kuulub kriitilisse piirkonda)). Kui valimi-statistiku väärtus on väiksem kui kriitiline teststatistiku väärtus, siis on selge, et peab jääma H_0 kehtivuse juurde ja vastupidi – kui see on suurem, siis lükatakse H_0 ümber. Tihti kasutatakse kriitilise statistiku väärtuse asemel p -väärtusi, aga selle matemaatiline sisu on tegelikult hüpoteesi tõesuse hindamise koha pealt samaväärne, sest kui kriitilise teststatistiku väärtusega seotud p -väärtus (ehk kriitilise piirkonnaga seotud jaotusealune pindala) jääb valitud usaldusnivoo-väärtusest (näiteks $\alpha=0.05$) väiksemaks, siis järeldatakse, et maksimaalne risk teha 1. liiki viga on lubatud piirides ja otsustatakse H_0 hüpotees ümber lükata. Kui aga saadav p -väärtus on suurem kui α , siis hinnatakse riski liiga suureks ja hüpoteesi H_0 ümber ei lükata.

NB! See ei tähenda nullhüpoteesi tõestamist, sest see on frequentist-statistikas sisuliselt nagu aksioom – pole võimalik tõestada.

Näide: Olgu H_0 ja sealttulenevalt ka teststatistiku kuju ja asukoht paigas, määrame ka alternatiivse hüpoteesi tüübi (ühepoolne vs kahepoolne). Saadakse standardiseeritud keskmiste erinevus (valimi-statistik) suurusega 1.12. Teame, et see vastab siis ühele konkreetsele punktile "jaotuse all" – mis selle teadmisega peale hakata? See on kasulik, sest kuna teststatistiku kuju ja asukoht on teada, siis on ka selle jaotuse (kvantiilid) ja kriitilised teststatistiku väärtused vastavate vabadusastmete korral teada ja saame kohe võrrelda, kas saadud valimi-statistik 1.12 on sellest kriitilisest väärtusest suurem või väiksem. Kui on väiksem, siis tehakse järeldus jääda H_0 paikapidavuse juurde (kuigi seda ei tõestata), aga kui on suurem, siis lükatakse H_0 ümber, sest siis kuulub valimi-statistik kriitilisse piirkonda.

Miks on vaja H_0 hüpoteesi ja miks ei piisa järelduste tegemiseks lihtsalt sellest, kui teame protsessi kirjeldava jaotuse kuju? Vastus: Sest me ei tea, kus kohas arvteljel uuritavat protsessi kirjeldav jaotus asub. Jaotuse asukoht on üheselt määratud näiteks keskvväärtusega ja sümmeetrilise jaotuse korral nagu näiteks normaaljaotus, langeb see kokku mediaaniga ja asub täpselt jaotuse keskel.

TEOORIA

2. Pideva juhusliku suuruse X tihedusfunktsioon $f(x)$ ja jaotusfunktsioon $F(x)$

$$f(x) = \lim_{\Delta x \rightarrow 0^+} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}$$

$F(x) := P(X \leq x)$ jaotusfunktsioon. Nende vahel valitseb seos: $\frac{d}{dx}(F(x)) = f(x)$

See tähendab seda, et tihedusfunktsioon on jaotusfunktsiooni tuletis vastava muutuja järgi.

3. Punkthinnangud ja vahemikhinnangud.

Olgu antud valim, millele elemendid on kõik sõltumatute ja samade jaotustega juhuslike suuruste realisatsioonid.

Praktikas ei huvita meid mitte juhuslikku suurust esindav valim, vaid valimi abil üldpopulatsiooni kohta üldistuste tegemine. Märgime, et üldkogum on klassikalise statistika eelduste põhjal on lõpmatu hulk. Kuna statistikas ja tõenäosusteoorias eeldame, et juhuslik suurus on täielikult määratud enda tõenäosusjaotusega, huvitavadki meid selle jaotuse parameetrid.

Olgu eesmärgiks hinnata jaotuse F , mille loeme meie poolt uuritavate juhuslike suuruste generaatoriks, parameetreid. Need parameetrid võtame lühidalt kokku parameetrite vektorina $\Theta = (\theta_1, \theta_2, \dots, \theta_n)$. Tähistame jaotuse tegeliku parameetri θ_j ja sellele vastava statistiku O_j .

Statistik O_j on valimist sõltuv juhuslik suurus hindamaks parameetri θ_j tegelikku väärtust. Tavaliselt on jaotuse parameetri hinnanguid kahte tüüpi: punkthinnang ja vahemikhinnang.

Punkthinnang jaotuse tegelikule parameetrile leitakse näiteks suurima tõepära funktsiooni kasutades – see paneb paika parameetri tõenäoseima väärtuse, jaotuse kohta antud andmeid arvestades. Punkthinnangu nimi tuleneb ilmselt sellest, et ühe parameetri hinnang kujutub arvteljel ühe punktina. Et punkthinnang oleks üldkogumi jaoks piisavalt hea, peavad olema rahuldatud kolm järgmist nõuet:

1. Punkthinnang peab olema nihketa (hinnangu keskväärtsus võrdne vastava parameetri väärtusega üldkogumis)
2. Punkthinnang peab olema efektiivne (hinnangu dispersioon peab olema minimaalne kõikvõimalikest nihketa hinnangutest)
3. Punkthinnang peab olema konsistentne (veenev) – st uute elementide lisamisel läheneb hinnang otsitava parameetri väärtusele.

Need kolm omadust vastavad kindlatele matemaatilistele määrangutele.

Vahemikhinnang jaotuse parameetritele θ_j defineeritakse enamasti α -usaldusintervalli (või usaldusvahemiku) abil.

Definitsioon. Parameetri θ_j α -usaldusintervalliks nim sellist lõiku I_α , mille korral kehtib järgmine võrdus:

$$P(\theta_j \in I_\alpha) = \alpha.$$

Kui α -usaldusintervall on sümmeetriline ja u on alumine usalduspiir ja U ülemine usalduspiir, siis $P(\theta_j < u) = P(\theta_j > U) = (1 - \alpha)/2$.

Vahemikhinnangu nimi tuleb ilmselt sellest, et see kujutub arvteljel vahemikuna.

5. Mis on regressioon ja korrelatsioon?

Liigitus -- lineaarne, mittelineaarne // parameetriline, mitteparameetriline // ühemõõtmeline, mitmemõõtmeline // kontrollmuutujatega, ilma.

Regressioon on pidevate juhuslike suuruste (erandjuhul võivad olla ka binaarsed kontrollmuutujad) (sõltumatu juhuslik suurus ja sellest sõltuv juhuslik suurus) vahelise seose kirjeldamiseks kasutatav statistiline meetod. Tihti kasutatakse regressiooni hindamiseks sõltuva juhusliku suuruse tinglikku keskvärtust sõltuvust sõltumatute muutujate kindlate väärtuste korral (ing. k. regression to the mean). Tüüpiliselt kasutatakse regressioonifunktsioone, mis on defineeritud polünoomide, splineide, regressioonipuude või kernelite abil.

Näide: multiplikatiivne mitteparameetriline regressioon, mida on ideaalne kasutada, kui sõltuva muutuja varieeruvust kirjeldava pinna võrrand on teada ja sõltumatud muutujad on omavahelises vastastikulises vastasmõjus. Kasutatakse näiteks selleks, et modelleerida organismi reaktsiooni keskkonna(tingimuste)le.

Korrelatsioon on kahe juhusliku standardiseeritud suuruse korrutise keskvärtus, mis on alati 0 ja 1 vahel ning mõõdab nendevahelise lineaarse seose suurust ja ka 'suunda' -- näiteks kui korrelatsioon kahe muutuja vahel on negatiivne, siis sõltumatu muutuja suurenedes sõltuva muutuja tinglik keskvärtus väheneb -- seega korrelatsiooni arvuline väärtus näitab muutujatevahelise lineaarse seose tugevust.

Märgime, et lineaarne seos (korrelatsioon) kahe juhusliku muutuja vahel ei tähenda, et need muutujad oleksid vastastikusel põhjuslikulises seoses (causal relation).

Õppematerjalis on toodud ka teine üldisem ajalooline mõiste korrelatsioonile: korrelatsiooni võib mõista ka stohhastilise sõltuvusena, kus nii sõltumatu kui sõltuv muutuja pole deterministlikud, vaid antud enda tõenäosusjaotustega.