Matemātiskās statistikas pamatjēdzieni

Uzskatīsim, ka ξ - gadījuma lielums, kas apraksta pētāmā objekta uzvedību (rādītāji par vienu, vai vairākām objekta ***pazīmēm***).

Gadījuma lielums ξ pieņem vērtības no kādas kopas **X**. Kopu **X** sauksim par ***ģenerālo kopu*** (***pazīmes vērtību kopu***).
Pieņemsim, ka doti gadījuma lieluma ξ n mērījumu (novērojumu, mēģinājumu) rezultāti $x_1, x_2, \ldots x_n$, kas veido virkni, kuru sauc par ***empīrisko rindu***.
Reģistrētās pazīmes vērtības, kas veido empīrisko rindu, sauksim par ***novērojumiem (variantiem)*** **.**

Modelis
Vienkāršākais un ticamākais modelis ir gadījumā, kad novērojumi tika iegūti veicot atkārtotus neatkarīgus eksperimentus nemainīgos apstākļos. Tas nozīmē, ka skaitļus $x_1, x_2, \ldots x_n$ var uzskatīt kā vienu gadījuma vektora $X = (X_1, X_2, \ldots X_n)$ realizāciju $x = (x_1, x_2, \ldots x_n)$, kur gadījuma lielumi $\{X_i, i = 1, 2, \ldots, n\}$ - savstarpēji neatkarīgi un vienādi sadalīti,

t.i., $$(F_{X_i}(x) \overset{def}{=} P(X_i < x),\ F_{X_i}(x) = F_{\xi}(x),\ i = 1, 2, \ldots, n,$$
$$i \neq j \Rightarrow P(X_i < x, X_j < y) = P(X_i < x) \cdot P(X_j < y))\ .$$

Vektora X (jeb tā realizāciju x) sauc par ***izlasi*** . Izlases kopas elementu skaitu ***n*** sauc par ***izlases apjomu***.

Par ***statistiku*** sauc attēlojumu $T : \prod_{i=1}^{n} \mathrm{X}_i \to \mathrm{R}^l, l \in \mathrm{N}$.

Formalizēt ***statistisko nenoteiktību*** pētāmā objektā nozīmē konstruēt objekta stohastisko modeli:

$< \mathrm{X}, F_\xi(x,\theta) >,\ \theta \in \Theta, F_\xi(x,\theta) \overset{def}{=} P_\theta(\xi < x), x \in \mathrm{R}$,
kur mūs interesējošā gadījuma lieluma ξ sadalījuma funkcija $F_\xi(x,\theta)$ ir atkarīga no parametra θ, kas var pieņemt vērtības no kopas Θ.

Par ***variāciju rindu*** sauc novērojumu rezultātu virkni, ja šie rezultāti sakārtoti augošā kārtībā. Apzīmējot k-to pēc vērtības novēroto lielumu ar $x_k^{'}, k = 1,2,...,n$, varam uzdot apskatāmā gadījuma lieluma ξ novēroto vērtību virkni variāciju rindas veidā $x_1^{'} \le x_2^{'} \le ... \le x_n^{'}$.

Empīriskā sadalījuma funkcija

Par empīrisko sadalījuma funkciju sauc funkciju

$$F_n(x) = \begin{cases} 0, & x \le x_1' \\ \dfrac{k}{n}, & x_k' < x \le x_{k+1}' \\ 1, & x > x_n' \end{cases}.$$

Empīriskā sadalījuma funkcija ir monotona, nepārtraukta no kreisās puses un tai ir pārtraukuma punkti tikai pie argumenta vērtībām, kas vienādas ar variāciju rindas locekļiem. Lēcienu lielumi pārtraukuma punktos ir $\frac{1}{n}$ daudzkārtņi. Pie katra x $F_n(x)$ ordināta ir gadījuma lielums ar iespējamām vērtībām $0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1$. Varbūtību notikumam $\left\{ \varpi : F_n(x) = \frac{k}{n} \right\}$ pie jebkuras x vērtības aprēķina pēc formulas

$$P\left(F_n(x) = \frac{k}{n} \right) = C_n^k \left[P(\xi < x) \right]^k \left[1 - P(\xi < x) \right]^{n-k} =$$

$$= C_n^k \left[F_\xi(x) \right]^k \left[1 - F_\xi(x) \right]^{n-k}, \quad k = 1, 2, ..., n.$$

$F_\xi(x)$ nosaka notikuma $\{\omega : \xi(\varpi) < x\}$ varbūtību dotam *x*, bet $F_n(x)$ dod notikuma $\{\omega : \xi(\varpi) < x\}$ biežumu.

Apzīmējot ar n_x -notikuma $\{\omega : \xi(\varpi) < x\}$ novērojumu skaitu $\forall x$, kur *n*- kopējais novērojumu skaits, iegūstam $F_n(x) = \frac{n_x}{n}$.

No pastiprinātā lielo skaitļu likuma (PLS) seko

$$F_n(x) \underset{n\to\infty}{\to} F_\xi(x), \ \forall x \in R \text{ ar varbūtību } 1.$$

Glivenko teorēma:**[1],[2]**

Ja $F_\xi(x)$ - gadījuma lieluma ξ sadalījuma funkcija, $F_n(x)$ - gadījuma lieluma ξ empīriskā sadalījuma funkcija, kas iegūta, izmantojot lieluma ξ *n* neatkarīgus novērojumus, tad

$$P(\sup_{-\infty<x<\infty} \left| F_n(x) - F_\xi(x) \right| \underset{n\to\infty}{\to} 0) = 1.$$

Kolmogorova teorēma: **[1],[2]**

Ja $F_\xi(x)$ –nepārtraukta gadījuma lieluma ξ sadalījuma funkcija, $F_n(x)$ - gadījuma lieluma ξ empīriskā sadalījuma funkcija,

$D_n := \sup_{-\infty<x<\infty} \left| F_n(x) - F_\xi(x) \right|$, tad

$$P(\sqrt{n} D_n < z) \underset{n\to\infty}{\to} K(z), \ K(z) = \begin{cases} 0, & z \le 0 \\ \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, & z > 0 \end{cases}.$$

Piezīme. Funkcija $K(z), \ z > 0$ ir tabulēta.

Statistiskā materiāla noformēšana

- ξ - diskrēts gadījuma lielums

ξ ***statistiskais sadalījums*** –tabula, kurā ierakstītās augošā kārtībā sakārtotas novērotās vērtības $x_i, i = 1,...,k$, un tām atbilstošie biežumi $n_i, i = 1,2,...,k,\ \sum_{i=1}^{k} n_i = n$, vai relatīvie biežumi $\frac{n_i}{n}, i = 1,2,...,k$.

- ξ - nepārtraukts gadījuma lielums

Variācijas rindas variantus apvieno k ***intervālos*** $(x_{i+1}, x_i), i = 1,2,...,k,$ kur $x_1 :\leq \min\{x_i, i = 1,...,n\}, x_{k+1} :\geq \max\{x_i, i = 1,...,n\}$, un saskaita cik novēroto vērtību atrodas katrā intervālā. Novēroto vērtību skaitu katrā *i*-tajā intervālā apzīmē ar $n_i, i = 1,2,...,k$, $\sum_{i=1}^{k} n_i = n$.

ξ ***statistiskā (variācijas) rinda***- tabula, kurā doti intervāli augošā kārtībā un tiem atbilstošie biežumi n_i , vai relatīvie biežumi $\frac{n_i}{n}, i = 1,2,\ldots,k$.

Tālākā statistiskās (variācijas) rindas apstrādē izvēlas katra intervāla pārstāvi $x_i^*, i = 1,2,...,k$, parasti ***intervāla centru*** (vai ***intervāla robežu***), un katrā i-tajā intervālā nonākušos variantus nosacīti uzskata par vienādiem x_i^* .

Piemēram, lai konstruētu gadījuma lieluma ξ empīrisko sadalījuma funkciju, vai aprēķinātu statistikas vērtības, gadījumā, kad dati uzdoti statistiskās rindas formā, jāizmanto ξ statistiskais

sadalījums **,** t.i**.,** jākonstruē tabula, kurā ierakstītas augošā kārtībā sakārtotas novērotās vērtības $x_i^*, i = 1, 2, ..., k$ un tām atbilstošie biežumi n_i. Apstrādājot intervālu statistiskās rindas, jālieto šādi lielumi:

$x_{\min}$ - pazīmes vismazākā reģistrētā vērtība, bet, ja tā nav zināma, pirmā intervāla apakšējā robeža,

x_{maks} - pazīmes lielākā reģistrētā vērtība vai pēdējā intervāla augšējā robeža,

$x_{maks} - x_{\min}$ - variācijas ***amplitūda*** jeb apjoms.

Parasti lieto vienāda garuma intervālus, bet ja vienāda garuma intervālu rindas malējos intervālos biežumi ir mazi vai parādās tukši intervāli, tad lieto arī nevienāda garuma intervālus. Ja nav nekādu citu apsvērumu par vēlamo intervālu garumu, var izmantot ***Sterdžesa formulu:***

$$\Delta = \frac{x_{maks} - x_{\min}}{1 + 3{,}2 \cdot \lg n}.$$

Variācijas rindu grafiskie attēli

Variācijas rindu attēlo ar ***poligonu*** vai ***histogrammu***.

Poligonu izmanto galvenokārt diskrētu variācijas rindu attēlošanai. To izveido šādi: koordinātu sistēmā atliek punktus, kuru koordinātas attiecīgi ir pazīmes vērtības (varianti) un to biežumi, vai relatīvie biežumi. Blakus esošos punktus savienojot ar taisnes nogriežņiem, iegūst poligonu.

Histogrammu lieto galvenokārt intervālu variācijas rindu attēlošanai. Uz abscisu ass atliek nogriežņus, kas atbilst variācijas rindas intervāliem. Pieņemot tos par pamatiem, virs katra intervāla konstruē taisnstūri, kura laukums vienāds ar dotā intervāla relatīvo biežumu $\frac{n_i}{n}, i = 1, 2, ..., k$, vai biežumu n_i.

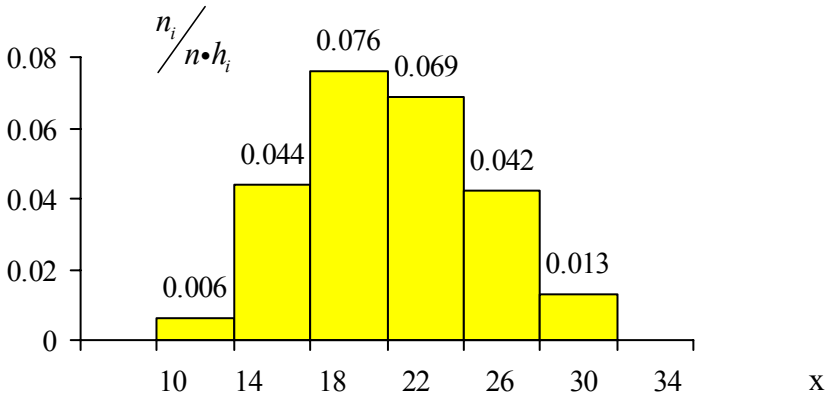
Ja katrā intervālā relatīvo biežumu (biežumu) dala ar intervāla garumu $h_i = x_{i+1} - x_i, i = 1,...,k$, tad iegūtais skaitlis ir taisnstūra augstums $\frac{n_i}{nh_i}$, $i = 1,...,k$ $\left(\frac{n_i}{h_i}, i = 1,2,...,k\right)$. Pilnais relatīvo biežumu histogrammas laukums ir vienāds ar 1. Biežumu histogrammas laukums ir vienāds ar *n*.

1.*piemērs.*

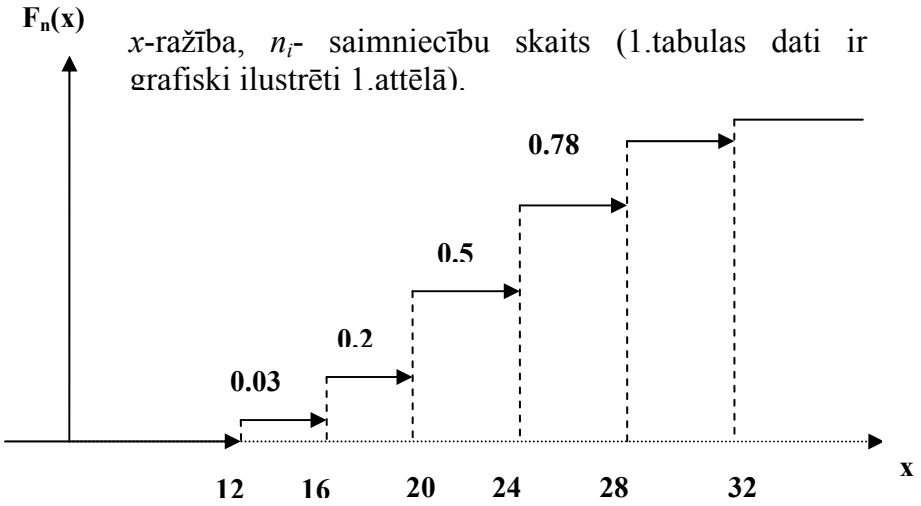
Lauksaimniecības uzņēmumu sadalījums pēc graudaugu ražības.

Grupas Nr.	Graudaugu ražība, cnt/ha		Saimniecību skaits		
	$x_i - x_{i+1}$	x_i^*	n_i	relatīvais biežums (%)	$\frac{n_i}{n}$
1	10,0…14,0	12	4	3	0,03
2	14,0…18,0	16	27	17	0,17
3	18,0…22,0	20	47	30	0,30
4	22,0…26,0	24	43	28	0,28
5	26,0…30,0	28	26	17	0,17
6	30,0…34,0	32	8	5	0,05
Kopā	×	×	n =155	100	1

1.tabula.



1.zīm. Graudaugu ražības relatīvo biežumu histogramma.

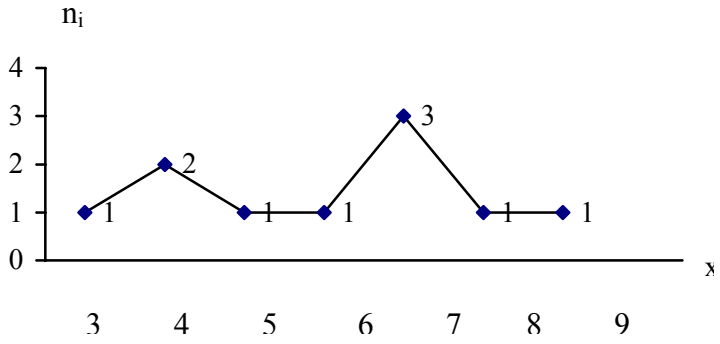


2.zīm. Graudaugu ražības empīriskā sadalījuma funkcija.

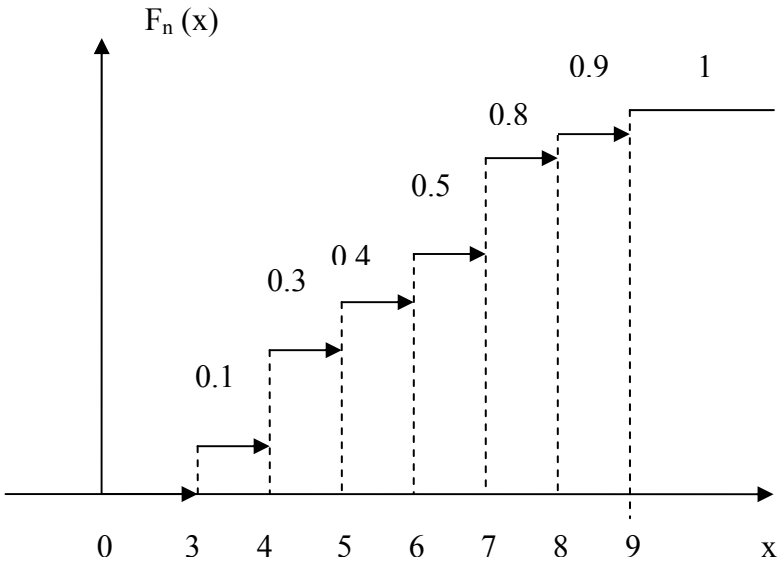
2.*piemērs.*
Pieņemsim, ka 10 studentu grupa, kārtojot eksāmenu, ir ieguvusi šādas atzīmes (desmit ballu sistēmā): 7;5;6;9;7;8;3;4;4;7.
Studentu sadalījums pēc atzīmēm:

x_i	3	4	5	6	7	8	9
n_i	1	2	1	1	3	1	1

2.tabula



3.zīm. **A**tzīmju biežumu poligons.



4.zīm. Atzīmju sadalījumu empīriskā sadalījuma funkcija.

Variācijas rindas raksturotāji

Statistiskie raksturotāji jāatspoguļo statistiskā objekta (parādības) objektīvās īpašības. Statistikā izmanto rādītājus, kuri īsi un koncentrēti raksturo galvenās sadalījuma īpašības. Šos rādītājus aprēķina tieši pēc sākotnējiem datiem. Datus apstrādā ar datortehniku.
Pieņemsim, ka dota izlase $(x_1, x_2, ..., x_n)$.

Empīriskais (aritmētiskais) vidējais -- $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

svērtais (dati sagrupēti) -- $\overline{x} = \dfrac{\sum_{i=1}^{k} x_i^* n_i}{n}$.

Empīriskais l-tās kārtas sākuma moments-- $\alpha_l := \frac{\sum_{i=1}^{n} x_i^l}{n},\ l \in N$

svērtais -- $\alpha_l := \frac{\sum_{i=1}^{k} (x_i^*)^l n_i}{n},\ l \in N$

Empīriskais l-tās kārtas centrālais

moments -- $\beta_l := \frac{\sum_{i=1}^{n} (x_i - \overline{x})^l}{n},\ l \in N$

svērtais -- $\beta_l := \frac{\sum_{i=1}^{k} (x_i^* - \overline{x})^l n_i}{n},\ l \in N$

Empīriskā dispersija --

$$s^2 := Var(x) = \beta_2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2 = \alpha_2 - \alpha_1^2$$

Ja doti kāda divdimensiju gadījuma lieluma (ξ, η) n novērojumi, $\{(x_i, y_i), i = 1,2,...,n\}$ tad

$\overline{xy} := \frac{\sum_{i=1}^{n} x_i y_i}{n}$ -- ***empīriskais 1-ās kārtas sākuma moments***

$\frac{\sum_{i=1}^{n} (x_i y_i)^k}{n}$, $k \in N$ ***-- empīriskais k-tās kārtas sākuma moments.***

3.*Piemērs.* Aprēķināt vidējo mēneša darba algu 8 cilvēku brigādei, ja atsevišķiem brigādes strādniekiem šajā mēnesī ir izmaksāts: 215,27; 230,91; 190,15; 250,31; 201,50; 222,93; 219,10; 205,129(lati).

Mūsu rīcībā ir tieši, nesagrupēti dati par katra strādnieka algu. Tādēļ ir jālieto empīriskā (nesvērtā) vidējā formula.

Izdarot aprēķinus

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{8} = 216,91(lati)$$

Strādnieka vidējā mēneša alga pēc noapaļošanas ir 217lati.

$$Var(x) = \frac{\sum_{i=1}^{8} x_i^2}{8} - \bar{x}^2 = 306,\ S = \sqrt{Var(x)} = 17.5$$

Mediāna variācijas rindas vidū esošā variants.
Ja vienību skaits variācijas rindā ir nepāra skaitlis $2m+1$, tad mediāna ir x_{m+1}. Ja vienību skaits ir pāra skaitlis $2m$, tad mediāna ir variantu x_m un x_{m+1} aritmētiskais vidējais.

$Me=x_{m+1}$, ja $n=2m+1$; $Me=\frac{x_m + x_{m+1}}{2}$, ja $n=2m$.

Lai atrastu mediānu intervālu variācijas rindai

-- jāatrod ***mediānas*** i***ntervāls*** (x_l, x_{l+1})**.** Mediānas intervāls ir tas, kurā uzkrātie absolūtie biežumi pirmo reizi pārsniedz pusi no kopas vienību skaita vai kurā uzkrātie relatīvie biežumi pirmo reizi pārsniedz 50%

-- pieņemot, ka mediānas intervāla ietvaros varianti sadalās vienmērīgi, izmanto interpolācijas formulu

$$Me = x_l + \frac{(0.5 - \sum_{i=1}^{l-1} \frac{n_i}{n})(x_{l+1} - x_l)}{n_l / n}$$

Piemēram, 1-ajā piemērā no 1-ās tabulas atrodam mediānas intervālu. Tas ir (22,26), jo tajā uzkrātais relatīvais biežums pirmo reizi pārsniedz 0.5 (50%), (piemērā tieši 50%). Me=22
Par ***modu*** *Mo* sauc variantu, kurš sadalījuma rindā ir sastopams visbiežāk.

Diskrētā variācijas rinda gadījumā moda nolasāma tieši kā variants ar vislielāko absolūto vai relatīvo biežumu.

Nepārtrauktā gadījumā variācijas rindu vispirms attēlo ar histogrammu. Pēc tam, lai atrastu modu intervālu variācijas rindai:

-- jānosaka ***modas*** jeb ***modālais intervāls*** (x_l, x_{l+1}). Tas ir intervāls ar vislielāko biežumu.

-- pieņemot hipotēzi, ka sadalījums intervālu ietvaros ir vienmērīgs, modu aprēķina pēc ***interpolācijas formulas*** :

$$Mo = x_l + \frac{(x_{l+1} - x_l)(n_l - n_{l-1})}{2n_l - n_{l+1} - n_{l-1}}$$

Tā, 1-ajā piemērā modālais intervāls ir (18, 22), Mo = 21,5.

Nosakot mediānu, ņem vērā visus rindas locekļus, bet modu nosaka galvenokārt modālais intervāls; interpolācijas gadījumā bez tam ņem vērā vēl pirms modālo un pēc modālo intervālus.

Modu lieto galvenokārt tad, ja ir svarīgi izdalīt tieši to variantu, kurš sastopams visbiežāk. Piemēram, preču pieprasījuma statistikā moda var būt visbiežāk pieprasītais apavu izmērs, gatavo apģērbu izmērs utt. Šajā gadījumā aritmētiskie vidējie ir maznozīmīgi, jo nav vajadzīgs un nav paredzēts ražot, piemēram, apavus, kuru izmērs tieši atbilstu aritmētiskajam vidējam (izmēram).

Parametru novērtējumu īpašības [1],[2]

Pieņemsim, ka gadījuma lieluma ξ sadalījuma funkcija $F_\xi(x, \theta)$ satur nezināmu parametru θ. Pieņemsim, ka dota izlase x=($x_1, x_2, \ldots x_n$).

Par parametra θ ***novērtējumu*** sauc funkciju $T(x_1, x_2, ...x_n)$, kas atkarīga tikai no novērotājām vērtībām $x_1, x_2, ...x_n$, kura kaut kādā nozīmē tuva novērtējamam parametram θ.

- Statistiku $T(x_1, x_2, ...x_n)$ sauc par parametra θ ***nenovirzītu (nenobīdītu) novērtējumu,*** ja jebkuram *n* $ET(x_1, x_2, ...x_n) = \theta$. Pretēja gadījumā novērtējums ir novirzīts.
- Statistiku $T(x_1, x_2, ...x_n)$ sauc par parametra θ ***asimptotiski nenovirzītu*** novērtējumu, ja $\lim_{n\to\infty} ET(x_1, x_2, ...x_n) = \theta$.
- Statistiku $T(x_1, x_2, ...x_n)$ sauc par parametra θ ***būtisku novērtējumu***, ja

$$\forall \varepsilon, \delta > 0 \quad \exists n_0 : n > n_0 \Rightarrow P(|T(x_1, x_2, ..., x_n) - \theta| > \varepsilon) < \delta,$$

kas nozīmē, ka statistika $T(x_1, x_2, ...x_n)$ konverģē pēc varbūtības uz skaitli θ.

- Statistiku $T^*(x_1, x_2, ...x_n)$ sauc par parametra θ ***efektīvu novērtējumu,*** ja $\min_{T \in \{T: ET = \theta\}} DT(x_1, x_2, ...x_n) = DT^*$.

Tas nozīmē, ka pie dota izlases apjoma *n* starp nenovirzītiem novērtējumiem statistikai $T^*(x_1, x_2, ...x_n)$ ir vismazākā dispersija.

Piemēram, $E\overline{x} = E\xi$, $ES^2 = \frac{n-1}{n} D\xi$, kas nozīmē, ka empīriskais vidējais ir nenovirzīts novērtējums matemātiskajai cerībai, bet empīriskā dispersija ir asimptotiski nenovirzīts novērtējums teorētiskai dispersijai.

Savukārt, statistika $s^2 := \frac{n}{n-1} S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$ būs nenovirzīts novērtējums teorētiskai dispersijai, jo izpildās $Es^2 = D\xi$.

Empīriskie raksturlielumi ir nozīmīgi, jo pie lieliem *n* tie ir tuvi, konverģences pēc varbūtības nozīmē, atbilstošiem teorētiskiem lielumiem.

Teorēma [1],[8].

Ja $n \to \infty$, spēkā sekojošas sakarības (visur konverģence pēc varbūtības)

- $F_n(x) \underset{n\to\infty}{\to} F_\xi(x)$ jebkuram x, $x, -\infty < x < \infty$
- $\alpha_k = \dfrac{\sum_{i=1}^{n} x_i^k}{n} \underset{n\to\infty}{\to} E\xi^k$, ja $E\xi^k < \infty$
- $S^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \underset{n\to\infty}{\to} D\xi$, ja $E\xi^2 < \infty$.

Atzīmēsim, ka dažreiz būtisks, bet novirzīts novērtējums dod labāku rezultātu.
Kā likums, ar tiešiem aprēķiniem var pārbaudīt vai parametra θ novērtējums ir novirzīts, vai nenovirzīts. Lai pārbaudītu vai novērtējums ir būtisks, vai nebūtisks jālieto LSL un sekojošu lemmu.

Lemma.

Ja gadījuma lielumu virkne $(\xi_n, n \geq 1)$ konverģē pēc varbūtības uz skaitli *a,* gadījuma lielumu virkne $(\eta_n, n \geq 1)$ konverģē pēc varbūtības uz skaitli *b,* divu mainīgo *x* un *y* funkcija $f(x, y)$ nepārtraukta punktā $x = a, y = b$, tad
$f(\xi_n, \eta_n) \underset{n\to\infty}{\to} f(a, b)$ pēc varbūtības.

Lai pārbaudītu vai novērtējums ir efektīvs, vai nav jālieto ***Rao-Kramera nevienādību.***

Teorēma[1],[9].

Pieņemsim, ka $p(x,\theta), x=(x_1,x_2,...x_n)$ ir gadījuma vektora $X=(X_1,X_2,...X_n)$ varbūtību sadalījuma blīvuma funkcija, parametrs $\theta \in \mathrm{R}$, un spēkā nosacījumi:

a) kopa $G=\left\{x \in \mathrm{R}^n : p(x,\theta)>0\right\}$ nav atkarīga no parametra θ (*regulārais nosacījums)*

b) blīvuma funkcijas logaritms diferencējams pēc θ $\forall x \in G$ un

$$I_n := \int_G \left[\frac{\partial \ln p(x,\theta)}{\partial \theta}\right]^2 p(x,\theta)dx = E\left[\frac{\partial \ln p(X,\theta)}{\partial \theta}\right]^2 < \infty,$$

tad jebkuram parametra θ nenovirzītam novērtējumam $\hat{\theta}=T_n(X_1,X_2,...,X_n)$ izpildās ***Rao-Kramera nevienādība***

$$E(\hat{\theta}-\theta)^2 = \mathrm{var}(\hat{\theta}) \geq \frac{1}{I_n}.$$

I_n -(Fišera koeficients) ir informācijas daudzums par parametru θ izlasē $X=(X_1,X_2,...X_n)$.

Saskaņā ar mūsu modeli, izlases visas komponentes ir neatkarīgas un vienādi sadalītas, tātad $I_n = nI_1$, kur

$I_1 = E\left(\frac{\partial \ln p(X_1,\theta)}{\partial \theta}\right)^2$ ir informācijas daudzums par parametru θ vienā komponentē X_1.

Rao-Kramera nevienādība nosaka parametra θ novērtējuma dispersijas zemāko robežu. Tātad, ja kaut kāda parametra θ nenovirzīta statistika to sasniedz, tad varam apgalvot, ka statistika ir parametra θ efektīvs novērtējums.

Piemēri.

1. . Pieņemsim, ka $\xi \sim N(\theta,\sigma^2)$, kur $x_1,x_2,...,x_n$ ir gadījuma lieluma ξ vērtību izlase, ko izmantosim, lai iegūtu

parametra θ efektīvu novērtējumu. Šim nolūkam uz *n* novērojumiem ($x_1, x_2, ..., x_n$) skatāmies kā uz *n*-dimensiju gadījuma lieluma $(X_1, X_2, ..., X_n)$ vienu vērtību. Gadījuma lielumi $X_i, i = 1, 2, ..., n$ kopumā neatkarīgi un vienādi sadalīti, $F_{X_i}(x) = F_{\xi}(x)$.

$$\ln p(x_1, \theta) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{(x_1 - \theta)^2}{2\sigma^2} \Rightarrow \frac{\partial \ln p(x_1, \theta)}{\partial \theta} =$$

$$= \frac{x - \theta}{\sigma^2} \Rightarrow E\left(\frac{X_1 - \theta}{\sigma^2}\right)^2 = \frac{1}{\sigma^2} \Rightarrow I_n = \frac{n}{\sigma^2}.$$

No šejienes varam secināt, ka statistika $\bar{x}$, kurai $E\bar{x} = \theta$ un $D\bar{x} = \frac{\sigma^2}{n}$, ir parametra θ efektīvs novērtējums.

2. Lai gadījuma lielums η sadalīts pēc eksponenciālā likuma ar nezināmo parametru $\lambda > 0$. Ievērojot, ka $E\eta = \frac{1}{\lambda}, D\eta = \frac{1}{\lambda^2}$ un ka $E\bar{x} = E\eta,\ D\bar{x} = \frac{\sigma_\eta^2}{n}$, ievedīsim jaunu parametru $\theta := \frac{1}{\lambda}$ un tam meklēsim efektīvu novērtējumu.

$$p(x, \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}},\ x > 0 \Rightarrow \ln p(x_1, \theta) = -\ln\theta - \frac{x}{\theta} \Rightarrow \frac{\partial \ln p(x_1, \theta)}{\partial \theta} =$$

$$= -\frac{1}{\theta} + \frac{x}{\theta^2} \quad \Rightarrow \quad I_1 = \int_0^\infty \left(-\frac{1}{\theta} + \frac{x}{\theta^2}\right) p(x, \theta)\, dx = \frac{1}{\theta^2} \Rightarrow I_n = \frac{n}{\theta^2}.$$

Līdz ar to statistika $\bar{x}$ ir nenovirzīts un efektīvs novērtējums parametram $\theta := \frac{1}{\lambda}$.

Parametru novērtējumu atrašanas metodes [1],[9]

I.***Vislielākās ticamības metode***

Uzskata, ka eksperimenta rezultātā esam ieguvuši visticamākās vērtības, tas nozīmē:

a) diskrētiem gadījuma lielumiem -

$$P_\theta(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) =: L(x_1, x_2, ..., x_n, \theta),$$

$$\max_\theta L(x_1, x_2, ..., x_n, \theta) = L(x_1, x_2, ..., x_n, \tilde{\theta}).$$

Par parametra θ novērtējumu ņemsim $\tilde{\theta}$, kas dod ***ticamības funkcijas*** $L(x_1, x_2, ..., x_n, \theta)$ maksimumu.

Ja ξ ir diskrēts gadījuma lielums, kuram n-vērtību izlasē ir *r* novērotas vērtības $x_1, x_2, ..., x_r$ ar atbilstošiem biežumiem

$n_1, n_2, ...n_r$, $\sum_{i=1}^{r} n_i = n$, tad,

ņemot vērā $X_i, i = 1, 2, ..., n$ neatkarību, ticamības funkciju varam pārrakstīt formā

$$L(x_1, x_2, ..., x_r, \theta) = \prod_{i=1}^{r} P_\theta^{n_i}(X_i = x_i).$$

b) nepārtrauktiem gadījuma lielumiem -

Pieņemsim, ka *n* neatkarīgu novērojumu rezultātā notiek varbūtīgais notikums, t.i.,

$$P\left(X_1 \in [x_1, x_1 + dx_1], X_2 \in [x_2, x_2 + dx_2], ..., X_n \in [x_n, x_n + dx_n], \theta\right) =$$

$$= \max_\theta$$

Par parametra θ novērtējumu, tāpat kā diskrētā gadījumā, ņem $\tilde{\theta}$, kas dod ***ticamības funkcijas***

$L(x_1, x_2, ..., x_n, \theta) = \prod_{i=1}^{n} p(x_i, \theta)$ maksimumu, kur $p(x, \theta)$ ir gadījuma lieluma ξ sadalījuma blīvuma funkcija.
Tātad, jāatrisina vienādojums

$$\frac{dL(x_1, x_2, ..., x_n, \theta)}{d\theta} = 0.$$

Atrisinājums, kas dod ticamības funkcijas maksimumu, būs parametra θ novērtējums. Iepriekšējā vienādojuma vietā reizēm ērtāk apskatīt šādu vienādojumu

$$\frac{d \ln L(x_1, x_2, ..., x_n, \theta)}{d\theta} = \frac{dL(x_1, x_2, ..., x_n, \theta)}{d\theta} \frac{1}{L} = 0.$$

Divu nezināmo parametru θ_1 un θ_2 gadījumā jāatrisina vienādojumu sistēma

$$\begin{cases} \frac{\partial L(x_1, x_2, ..., x_n, \theta_1, \theta_2)}{\partial \theta_1} = 0 \\ \frac{\partial L(x_1, x_2, ..., x_n, \theta_1, \theta_2)}{\partial \theta_2} = 0 \end{cases} \text{ vai } \begin{cases} \frac{\partial \ln L(x_1, x_2, ..., x_n, \theta_1, \theta_2)}{\partial \theta_1} = 0 \\ \frac{\partial \ln L(x_1, x_2, ..., x_n, \theta_1, \theta_2)}{\partial \theta_1} = 0 \end{cases}$$

Piemēri.

1. Novērtēsim notikuma A varbūtību *p*, ja notikums A *n* neatkarīgos
   mēģinājumos parādījies *m* reizes.

Varbūtību *p* uzskatīsim par parametru, kas ietipst diskrētā gadījuma lieluma ξ sadalījuma funkcijā. Gadījuma lielums ξ pieņem tikai divas vērtības: $x_1 = 1, x_2 = 0$ atkarībā no tā, vai notikums A dotajā mēģinājumā notiek vai nē. Tātad, ticamības funkcija būs

$$L(x_1, x_2, ..., x_n, \theta) = p^m (1 - p)^{n-m},$$

no kurienes

$$\frac{d \ln L(x_1, x_2, ..., x_n, p)}{dp} = \frac{m}{p} - \frac{n-m}{1-p} = 0.$$

Līdz ar to $p = \frac{m}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Tātad, varbūtības *p* novērtējums ir notikuma A parādīšanās relatīvais biežums $\tilde{p} = \frac{m}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{x}.$

Šis novērtējums būtisks (lielo skaitļu likums) un nenovirzīts, jo

$$E\tilde{p} = E\frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{np}{n} = p,$$

kā arī asimptotiski normāls (centrālā robežteorēma).

2. Pieņemsim, ka ξ ir gadījuma lielums, sadalīts pēc Puasona likuma ar nezināmo parametru $\mu > 0$, t.i., gadījuma lielums, kas pieņem nenegatīvas veselu skaitļu vērtības ar varbūtībām $P(\xi = i) = \frac{\mu^i}{i!} e^{-\mu}, \quad \mu > 0, \ i = 0,1,2,...$.

Pieņemsim, ka *n* neatkarīgos mēģinājumus novērotas gadījuma lieluma ξ vērtības $x_1, x_2, ..., x_k$ ar atbilstošiem biežumiem $n_1, n_2, ..., n_k$, $\sum_{i=1}^{k} n_i = n$.

Tad

$$L(x_1, x_2, ..., x_k, \mu) = \prod_{i=1}^{k} P^{n_i}(X_i = x_i) = \prod_{i=1}^{k} \left(\frac{\mu^{x_i}}{x_i!} e^{-\mu}\right)^{n_i},$$

$$\ln L = \sum_{i=1}^{k} n_i [x_i \ln \mu - \ln x_i! - \mu], \quad \frac{d \ln L}{d\mu} = \sum_{i=1}^{k} n_i \left[\frac{x_i}{\mu} - 1\right] = 0,$$

$$\mu = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i} = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \tilde{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{x}.$$

3. Pieņemsim, ka $\xi \sim N(\mu, \sigma^2)$, t.i., gadījuma lielums ξ ir sadalīts pēc normālā sadalījuma likuma ar nezināmiem parametriem μ un σ^2.

Pieņemsim, ka $x_1, x_2, ..., x_n$ gadījuma lieluma ξ vērtību izlase, kuru izmantosim, lai iegūtu parametru μ un σ^2 novērtējumu. Tad no sakarības

$$L(x_1, x_2, ..x_n, .., \mu, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\}$$

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\}.$$

Vienādojumi μ un σ^2 novērtējumu iegūšanai ir sekojoši:

$$\begin{cases} \frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \\ \frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases}.$$

Atrodam

$$\tilde{\mu} = \overline{x},\ \tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{x})^2}{n} = S^2.$$

Pirmais novērtējums-nenovirzīts, būtisks un efektīvs, otrais-asimptotiski nenovirzīts un būtisks.

4. Pieņemsim, ka gadījuma lielums $\xi \sim v.s.[a,b]$, t.i., varbūtību sadalījuma blīvuma funkcija gadījuma lieluma ξ ir

$$p(x,a,b)=\begin{cases}\dfrac{1}{b-a}, & x\in[a,b]\\ 0, & x\notin[a,b]\end{cases}.$$

Pieņemsim, ka $x_1, x_2, ..., x_n$ gadījuma lieluma ξ vērtību izlase, kuru izmantosim, lai iegūtu parametru *a* un *b* novērtējumu.

Ticamības funkcija $L(x_1, x_2, ..., x_n, a, b)=\dfrac{1}{(b-a)^n}$ ir nezināmo parametru *a* un *b* funkcija un maksimumu sasniedz pie

$$\tilde{a}=\min_i X_i,\ \tilde{b}=\max_i X_i.$$

Vislielākās ticamības metode
-dod būtiskus, kaut arī reizēm novirzītus novērtējumus, kas asimptotiski
normāli sadalīti;
- kaut kāda nozīmē vislabāk izmanto visu novērojumu izlases doto informāciju par nezināmo parametru.

II. ***Momentu metode***

Šī metode dod vienkāršākus un ērtākus raksturojumus: izlases sākuma momenti kalpo par novērtējumiem gadījuma lieluma ξ sadalījuma atbilstošiem teorētiskiem sākuma momentiem, kas satur nezināmus parametrus. Savukārt novērtējamie parametri izsakās kā teorētisko momentu zināmas funkcijas. Aizvietojot tajās teorētiskos momentus ar to novērtējumiem, iegūstam parametru novērtējumus.

Piemēri.

a) Pieņemsim, ka gadījuma lielums $\xi \sim v.s.[a,b]$.

$x_1, x_2, ..., x_n$ ir gadījuma lieluma ξ vērtību izlase, kuru izmantosim, lai iegūtu parametru *a* un *b* novērtējumu ar momentu metodes palīdzību. Ir zināms, ka

$$E\xi = \frac{b+a}{2}, \quad D\xi = \frac{(b-a)^2}{12}.$$

Tad, $\begin{cases} \overline{x} = \dfrac{b+a}{2} \\ S^2 = \dfrac{(b-a)^2}{12} \end{cases} \Rightarrow \begin{cases} \tilde{b} = \overline{x} + \sqrt{3}S \\ \tilde{a} = \overline{x} - \sqrt{3}S \end{cases}.$

b) Aplūkosim kā iegūt korelācijas koeficienta novērtējumu izmantojot momentu metodi.

$$\rho(\xi,\eta) := \frac{cor(\xi,\eta)}{\sigma_\xi \sigma_\eta} \overset{def}{=} \frac{1}{\sigma_\xi \sigma_\eta} E(\xi - E\xi)(\eta - E\eta) =$$

$$= \frac{1}{\sigma_\xi \sigma_\eta}(E\xi\eta - E\xi E\eta).$$

Aizvietojot teorētiskos momentus ar to novērtējumiem, iegūstam sistēmu

$$\begin{cases} \overline{x} = E\xi \\ \overline{y} = E\eta \\ \dfrac{\sum_{i=1}^{n} x_i y_i}{n} = E\xi\eta \\ s_x^2 = \sigma_\xi^2 \\ s_y^2 = \sigma_\eta^2 \end{cases} \Rightarrow r_{x,y} = \frac{1}{s_x s_y}(\overline{xy} - \overline{x}\,\overline{y})$$

Tātad,

$$r_{(x,y)} = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{Cov(x,y)}{s_x s_y} = \frac{1}{s_x s_y}(\overline{xy} - \bar{x}\,\bar{y})$$

ir korelācijas koeficienta $\rho(\xi,\eta)$ novērtējums.

Visbiežāk sastopamās sadalījuma funkcijas[6],[8]

Vienmērīgais sadalījums $[a,b]$, $\xi \sim v.s.[a,b]$

Varbūtību blīvuma funkcija ir

$$p(x) = \begin{cases} 0, x \le a \\ \dfrac{1}{b-a}, a < x \le b \\ 1, x > b \end{cases}, E\xi = \frac{b+a}{2},\ D\xi = \frac{(b-a)^2}{12}$$

Normālais sadalījums, $\xi \sim N(a,\sigma^2)$

Normālā sadalījuma blīvuma funkcija ir

$$p(x,a,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}, a \in R,\ \sigma > 0$$

$$E\xi = a,\ D\xi = \sigma^2,\ Mo(\xi) = Me(\xi) = a.$$

$$P(a - 3\sigma < \xi < a + 3\xi) = 0.9986$$

Ja σ=1 un *a*=0, tad sadalījumu sauc par ***standartu (normētu) normālo sadalījumu.***

Daudzdimensiju normālais sadalījums, $\xi \sim N(a,M)$

$\xi = (\xi_1, \xi_2, ..., \xi_k), a = (a_1, a_2, ..., a_k), a \in \mathrm{R}^k$,
M -simetriskā, pozitīvi definīta matrica.

Daudzdimensiju sadalījumu sauc par normālu, ja tā blīvuma funkcija ir

$$p(x, a, A) = (2\pi)^{-k/2} \mid A \mid^{1/2} \exp\{-\tfrac{1}{2}(A(x-a),(x-a)\},$$

kur $x \in R^k$, $\mid A \mid$- matricas A determinants, $(\cdot,\cdot)$-skalārais reizinājums, $M := A^{-1} = \| \operatorname{cov}(\xi_i, \xi_j) \|$, $i, j = 1, 2, ..., k$ - kovariācijas matrica.

Ja A^{-1} ir diagonālmatrica, tad normālā vektora ξ komponentes neatkarīgas un

$$p(x, a, A) = \left(\sqrt{2\pi}\right)^{-n} \prod_{i=1}^{k} \sigma_i^{-1} \exp\left(-\frac{(x_i - a_i)^2}{2\sigma_i^2}\right).$$

Spēkā arī pretējs apgalvojums, ja normālā vektora komponentes neatkarīgas, tad atbilstošā kovariācijas matrica diagonāla.
Divdimensiju normālā sadalījuma blīvuma funkcija ir šāda:
$p(x, y) =$

$$= \frac{1}{2\pi\sigma_\xi\sigma_\eta\sqrt{1-\rho}} \exp\left(-\frac{1}{2(1-\rho)}\left(\frac{(x-a)^2}{\sigma_\xi^2} - 2\rho\frac{(x-a)(y-b)}{\sigma_\xi\sigma_\eta} + \frac{(y-b)^2}{\sigma_\eta^2}\right)\right)$$

kur $\rho(\xi, \eta)$ -korelācijas koeficients, $E\xi = a, E\eta = b.$

□

Lemma [11]

$\xi \sim N(a, M),\ \eta = C\xi \Rightarrow \eta \sim N(Ca, CMC^*)$

□

Pīrsona jeb χ^2 − sadalījums.

Pieņemsim, ka $\xi_i \sim N(0,1), i = 1, 2, ..., n,$ neatkarīgi

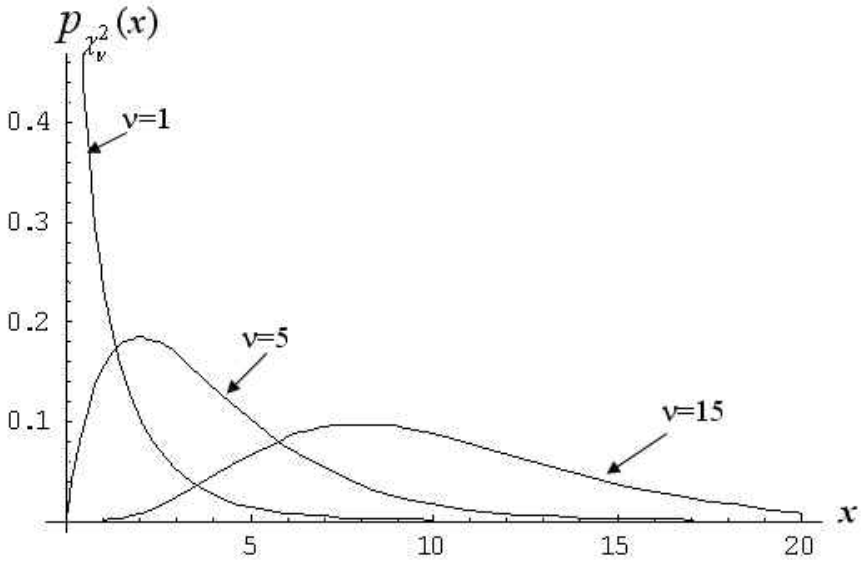
Gadījuma lieluma $\chi^2_{\mathrm{n}} = \sum_{i=1}^{n} \xi_i^2$ sadalījumu sauc par Pirsona jeb χ^2 – sadalījumu ar *n* brīvības pakāpēm (hī- kvadrāts ar n brīvības pakāpēm).

χ^2 – sadalījuma blīvuma funkcija ir

$$p_{\chi_n^2}(x) = \begin{cases} \dfrac{1}{(2)^{\frac{n}{2}}\,\Gamma\left(\dfrac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x > 0, \\ 0, \qquad\qquad x \le 0, \end{cases}$$

kur *n* ir brīvības pakāpes.

$E\chi_n^2 = n,\ \ D\chi_n^2 = 2n,\ \dfrac{\chi_n^2 - n}{\sqrt{2n}} \underset{n\to\infty}{\to} \tau$ pēc sadalījuma, kur $\tau \sim N(0,1)$.



5.zīm. χ^2 – sadalījumu blīvuma funkcijas.

□

Stjūdenta jeb t-sadalījums.

Pieņemsim, ka $\xi \sim N(0,1)$, $\chi_\nu = \sqrt{\chi_\nu^2}$ ir neatkarīgi gadījuma lielumi.

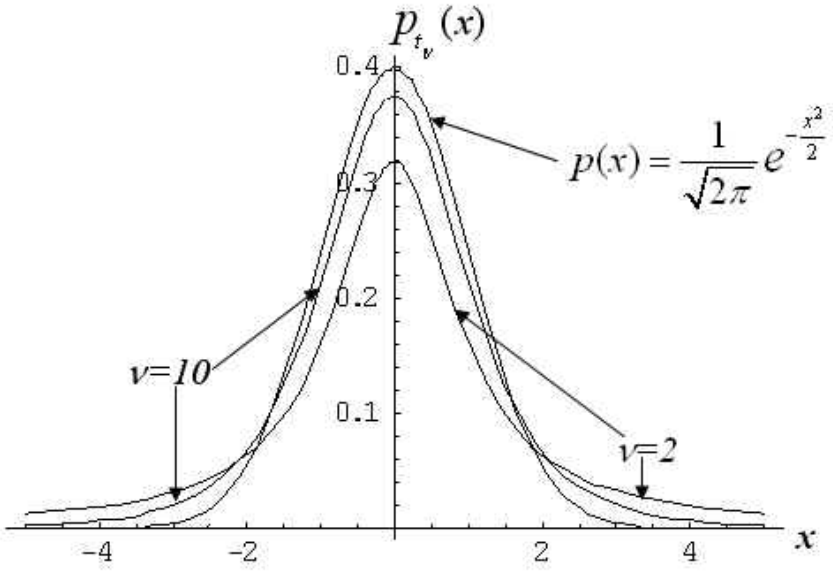
Gadījuma lieluma $t_\nu = \dfrac{\xi}{\chi_\nu}\sqrt{\nu}$ sadalījumu sauc par Stjūdenta jeb *t*-sadalījumu ar ν brīvības pakāpēm.

t_ν sadalījuma blīvuma funkcija ir

$$p_{t_\nu}(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\dfrac{\nu+1}{2}\right)}{\Gamma\left(\dfrac{\nu}{2}\right)} \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in (-\infty, \infty).$$

Ja ν=1, tad iegūstam Koši sadalījumu.

$t_\nu \underset{\nu\to\infty}{\to} \tau$ pēc sadalījuma, kur $\tau \sim N(0,1)$



6.zīm. t-sadalījumu blīvuma funkcijas.

F-*sadalījums.*

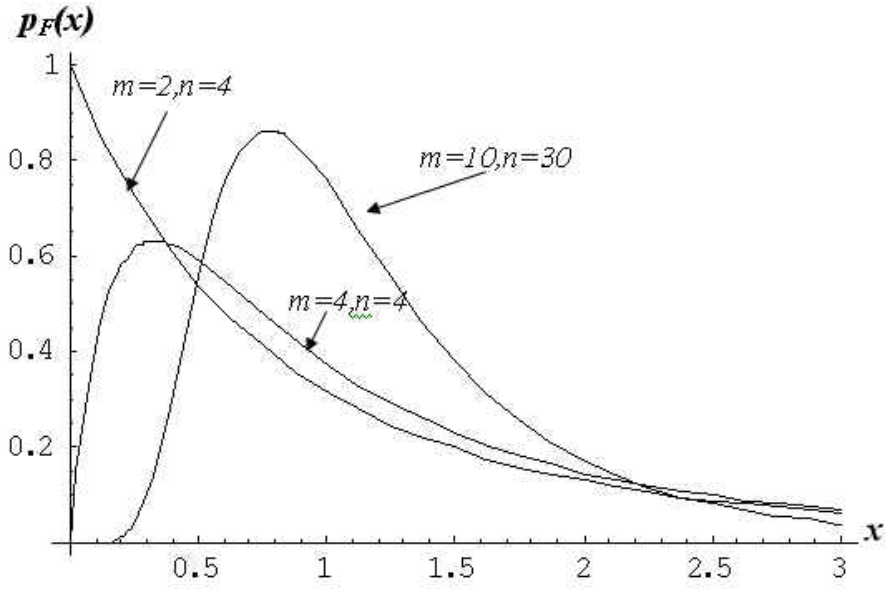
Pieņemsim, ka χ_m^2, χ_n^2 - neatkarīgi gadījuma lielumi.

Gadījuma lieluma $\frac{\chi_m^2}{\chi_n^2}\frac{n}{m} = F_{m,n}$ sadalījumu sauc par Snedekora jeb F- sadalījumu ar *m* un *n* brīvības pakāpēm.

$F_{m,n}$ sadalījuma blīvuma funkcija ir

$$p_{F_{m,n}}(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{m+n}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right)\Gamma\left(\dfrac{m}{2}\right)}\left(\dfrac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1}\left(1+\dfrac{m}{n}x\right)^{-\frac{m+n}{2}}, & x>0, \\ 0, & x \le 0. \end{cases}$$

Ja *m*=1, tad $\sqrt{F_{1,n}} \sim t_n$.



7.zīm. F-sadalījumu blīvuma funkcijas.

Gadījuma lieluma $z = \frac{1}{2}\ln \mathrm{F}_{m,n}$ sadalījumu sauc par Fišera sadalījumu ar *m* un *n* brīvības pakāpēm.

□

Pieņemsim, ka ξ – nepārtraukti sadalīts gadījuma lielums ar sadalījuma funkciju $F_\xi(x)$

Atradīsim gadījuma lieluma $\eta = F_\xi(\xi)$ sadalījuma funkciju.

No tā, ka $F_\xi(x) \in [0,1]$ izriet

$$F_\eta(x) = P(\eta < x) = P(F_\xi(\xi) < x) = \begin{cases} 0, & x \le 0, \\ 1, & x > 1. \end{cases}.$$

Ja $x \in (0,1)$ iegūstam

$$F_\eta(x) = P(\eta < x) = P(F_\xi(\xi) < x) =$$

$$= P(\{\omega : \xi < F_\xi^{-1}(x)\}) = F_\xi(F_\xi^{-1}(x)) = x, \quad 0 < x \le 1.$$

Līdz ar to

$$F_\eta(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x \le 1 \\ 1, & x > 1 \end{cases}.$$

Tādējādi, gadījuma lielums η ir vienmērīgi sadalīts segmentā [0,1]. Izmantojot formulu

- $\xi = F_\xi^{-1}(\eta)$ varam modelēt nepārtraukti sadalītus gadījuma lielumus (η - vienmērīgi sadalīts [0,1] gadījuma lielums) .

Statistikas, saistītas ar normālo izlasi

Kokrena teorēma **[12]**

Ja gadījuma vektora $\xi = (\xi_1, \xi_2, ..., \xi_n)$ komponentes ir neatkarīgas un sadalītas pēc standarta normālā sadalījuma likuma un eksistē k kvadrātiskas formas $(V_j\xi, \xi),\ j = 1, 2, ..., k$

ar atbilstošiem rangiem $n_1, n_2, .., n_k$, kurām

$$\sum_{i=1}^{n} \xi_i^2 = \sum_{j=1}^{k} (V_j\xi, \xi),$$

tad ir spēkā šādi apgalvojumi:

1. no vienādības $\sum_{j=1}^{k} n_j = n$ seko, ka katrai kvadrātiskai formai $(V_j\xi, \xi),\ j = 1, 2, ..., k$ ir $\chi^2_{n_j}$ sadalījums un tās visas ir neatkarīgas;
2. ja katrai kvadrātiskai formai $(V_j\xi, \xi),\ j = 1, 2, ..., k$ ir $\chi^2_{m_j}$ sadalījums, tad katra šī sadalījuma brīvības pakāpju skaits m_j vienāds ar atbilstošo kvadrātiskās formas rangu $n_j, j = 1, 2, ..., k$ un ir spēkā $\sum_{j=1}^{k} n_j = n$;
3. ja visas kvadrātiskās formas $(V_j\xi, \xi),\ j = 1, 2, ..., k$ ir neatkarīgas kopumā, tad katra šī gadījuma lieluma sadalījums ir $\chi^2_{n_j}\ j = 1, 2, ..., k$ un ir spēkā $\sum_{j=1}^{k} n_j = n$.

Teorēmas rezultātus bieži izmanto matemātiskajā statistika, veicot pētījumos, kas saistīti ar normālo sadalījumu.

Saskaņā ar mūsu modeli, no Kokrenas teorēmas seko:

a) ja $\xi \sim N(0,1)$, $y = (y_1, y_2, \ldots y_n)$ ir gadījuma ξ izlase,

tad $\begin{cases} s_y{}^2 n \sim \chi^2_{n-1} \\ \sqrt{n}\,\overline{y} \sim N(0,1) \end{cases}$ un neatkarīgi.

Pierādījums.

$$\frac{\sum_{i=1}^{n} y_i^2}{n} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n} + \overline{y}^2$$

pārrakstīsim formā $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 + n\overline{y}^2$.

Kvadrātiskās formas $\sum_{i=1}^{n} (y_i - \overline{y})^2$ rangs vienāds ar *n*-1, savukārt kvadrātiskās formas $n\overline{y}^2$ rangs vienāds ar 1. Tā kā izpildās Kokrena teorēmas pirmā apgalvojuma nosacījumi varam secināt, ka $s_y{}^2 n \sim \chi^2_{n-1}$, $n\overline{y}^2 \sim \chi^2_1$ un tās ir neatkarīgas. Ievērojot gadījuma lieluma $\chi^2_1 = \tau^2$, kur $\tau \sim N(0,1)$, definīciju iegūstam: $s_y{}^2 n \sim \chi^2_{n-1}$, $\sqrt{n}\,\overline{y} \sim N(0,1)$ neatkarīgi.

b) ja

$$\xi \sim N(a, \sigma^2),$$

$$p(x, a, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}, a \in R,\ \sigma^2 \in R_+$$

$x = (x_1, x_2, \ldots x_n)$ - ir gadījuma ξ izlase, tad

- $\begin{cases} \dfrac{\overline{x}-a}{\sigma/\sqrt{n}} \sim N(0,1) \\ \dfrac{s_x{}^2 n}{\sigma^2} = \chi^2_{n-1} \end{cases}$ neatkarīgi

Pierādījums.

Saskaņā ar mūsu modeli, gadījuma vektora $y = (y_1, y_2, ...y_n)$, kur $y_i := \dfrac{x_i - a}{\sigma}$, $i = 1, 2, ..., n$, komponentes neatkarīgas kopumā un sadalītas pēc standarta normāla sadalījuma likuma. Tādējādi, no punkta (a) seko, ka

$$s_y{}^2 n = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 = \frac{s_x^2 n}{\sigma^2} \sim \chi^2_{n-1},$$

$\sqrt{n}\,\overline{y} = \dfrac{\overline{x}-a}{\sigma}\sqrt{n} \sim N(0,1)$, un tās ir neatkarīgas, kas arī bija jāpierada.

- $\dfrac{\overline{x}-a}{s}\sqrt{n-1} \sim t_{n-1}$ (pēc t-sadalījuma definīcijas)
- $\dfrac{s\sqrt{n}}{\sigma} = \chi_{n-1}$ (pēc χ-sadalījuma definīcijas)

c)

Pieņemsim, ka $x = (x_1, x_2, ..., x_n), y = (y_1, y_2, ...y_m)$ – divas neatkarīgas izlases no sadalījumiem:

$$\xi_i \sim N(a_i, \sigma_i{}^2), \;\; p(x, a_i, \sigma_i{}^2) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-a_i)^2}{2\sigma_i{}^2}},$$

$$a_i \in R, \;\; \sigma_i{}^2 \in R_+, i = 1, 2.$$

Ja ξ_1, ξ_2 -neatkarīgi, tad

$$\bullet \frac{s_x^2 n(m-1)\sigma_1^2}{s_y^2 m(n-1)\sigma_2^2} = \frac{s_{\text{izl},x}^2 \sigma_1^2}{s_{\text{izl},y}^2 \sigma_2^2} \sim F(n-1, m-1),$$

kur

$$s_{\text{izl},x}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1},\ s_{\text{izl},y}^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$$

(seko no F-sadalījuma definīcijas).

d)

Ja $\xi -$ gadījuma lielums ar sadalījuma funkciju $F_\xi(x)$,

$E\xi = a, D\xi = \sigma^2$, $x = (x_1, x_2, \ldots x_n)$ - ir gadījuma ξ izlase, un izlases apjoms *n* ir liels, tad no Centrālās Robeža teorēmas (CRT),

$$\bullet \frac{\overline{x} - a}{\sigma / \sqrt{n}} \sim N(0,1)$$

e)

$\xi -$ gadījuma lielums ar sadalījuma funkciju $F_\xi(x)$, variācijas rindas variantus apvieno k intervālos: $(x_{i+1}, x_i), i = 1, 2, \ldots, k,$

$x_1 := \min\{x_i, i = 1, \ldots, n\}, x_{k+1} := \max\{x_i, i = 1, \ldots, n\}$, n_i - novēroto vērtību skaits katrā i-tajā intervālā, $i = 1, 2, \ldots, k$, $\sum_{i=1}^{k} n_i = n$,

izlases apjoms *n* ir liels, tad

$$\bullet \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi_{k-1-l}^2,$$

$p_i = P(\xi \in (x_i, x_{i+1})) = F_\xi(x_{i+1}) - F_\xi(x_i), i = 1, 2, \ldots, k,$

$x_1 = -\infty,\ x_{k+1} = \infty$, *l*- novērtējumu skaits, kas ir nepieciešams lai aprēķinātu $p_i,\ i = 1, 2, \ldots, k.$

Ticamības intervāla konstruēšana

Pieņemsim, ka $F_{\xi}(x,\theta)$ ir gadījuma lieluma ξ sadalījuma funkcija, funkcija zināma, bet satur nezināmu parametru θ, $\theta \in$ R. Pieņemsim, ka $x=(x_1,x_2,...x_n)$-gadījuma lieluma ξ vērtību izlase, ar kuru palīdzību tika iegūts parametra θ novērtējums $\gamma(x_1,x_2,...,x_n)$. Nepieciešams vēl novērtēt novērtējuma precizitāti un drošību, t.i., jānoskaidro, kādas kļūdas rada nezināmā parametra aizvietošana ar tā novērtējumu. Tas sevišķi svarīgi, ja novērojumu skaits ir maz. Pieņemsim, ka iegūts parametra θ nenovirzīts novērtējums $\gamma(x_1,x_2,...,x_n)$.

Jo mazāks $|\theta-\gamma|$, jo precīzāk novērtējums $\gamma(x_1,x_2,...,x_n)$ nosaka parametru θ. Tātad, ja $|\theta-\gamma|<\delta$, tad skaitlis δ raksturo novērtējuma precizitāti.

Statistiskās metodes ļauj spriest tikai par varbūtību β, ar kādu realizējas pēdēja nevienādība.

Par parametra θ novērtējuma $\gamma(x_1,x_2,...,x_n)$ ***drošību*** vai ***ticamības varbūtību*** sauc varbūtību β, ar kādu realizējas nevienādība $|\theta-\gamma|<\delta$.

Ņemsim, piemēram, varbūtību β pietiekoši lielu, tādu, lai notikumu ar varbūtību β varētu uzskaitīt par ***praktiski drošu,*** t.i., $\beta \in [0.95;0.999]$.

Tad

$$P(|\theta-\gamma|<\delta) \geq \beta,$$

vai arī

$$P(\gamma-\delta<\theta<\gamma+\delta) \geq \beta,$$

t.i., aizvietojot parametru θ ar $\gamma(x_1,x_2,...,x_n)$, kļūda būs diapazonā δ, vai ar varbūtību β nezināmā parametra θ vērtība būs intervālā $J_{\theta,\beta}=(\gamma-\delta,\gamma+\delta)$. Kļūda, pēc absolūtās vērtības lielāka nekā δ,

parādīsies ar varbūtību, kas mazāka par $\alpha = 1 - \beta$. Šādā gadījumā intervālu

$$J_{\theta,\beta} = (\gamma - \delta, \gamma + \delta)$$

sauc par ***parametra θ ticamības intervālu*** ar drošības koeficientu β.

Ja varbūtību teorijas kursā apskatīja varbūtību, ar kādu gadījuma lielums pieņem vērtības no dotā intervālā, tad šeit citādi, šeit θ ir konstante, bet $J_{\theta,\beta} = (\gamma - \delta, \gamma + \delta)$ ir gadījuma intervāls.

β - nav varbūtība tam, ka θ būs dotajā intervālā, bet varbūtība tam, ka gadījuma intervāls $J_{\theta,\beta} = (\gamma - \delta, \gamma + \delta)$ pārklās punktu θ.

Ticamības intervālu var uzskatīt par tādu θ vērtību intervālu, kas saskan ar novērojumu rezultātiem, nav ar tiem pretrunā. Tās θ vērtības, kam $|\theta - \gamma| > \delta$, var uzskatīt par pretrunīgām novērojumiem, jo tāda notikuma varbūtība mazāka par $\alpha = 1 - \beta$, t.i., praktiski neiespējams gadījums.

Teorētiski ticamības intervālu ar ticamības varbūtību, piemēram, 0.5, var ilustrēt sekojoši: ja izdarītu *n* novērojumu sērijas, tad puse no ticamības intervāliem, kas konstruēti pēc katras sērijas novērojumiem, segtu novērtējamo parametra θ.

Pieņemsim, ka $F_\xi(x,\theta)$ ir gadījuma lieluma ξ sadalījuma funkcija, θ, $\theta \in \Xi$ - parametrs. $x = (x_1, x_2, ...x_n)$ -gadījuma lieluma ξ vērtību izlase,

$\gamma_1(x), \gamma_2(x) -$ statistikas, $\alpha \in \{0.05; 0.01; 0.001\}$ - ***nozīmības līmenis***, $\beta = 1 - \alpha$ ***- drošība,***

$T(x,\theta)$ - gadījuma funkcija no novērotājām vērtībām $x_1, x_2, ...x_n$ un parametra θ:

$$T : \prod_{i=1}^{n} \mathrm{X}_i \times \Xi \to \mathrm{R}^k,\ k \in \mathrm{N}$$

$J_{\theta,1-\alpha} = (\gamma_1(x), \gamma_2(x))$ sauc par parametra θ ***ticamības intervālu,*** ja izpildās

$$P((\gamma_1(x) < \theta < \gamma_2(x)) = 1 - \alpha\ ,$$

t.i., ar varbūtību $\beta = 1-\alpha$ intervāls $(\gamma_1(x), \gamma_2(x))$ ar gadījuma galiem $\gamma_1(x)$ un $\gamma_2(x)$ pārklāj parametra θ patieso vērtību.

1.Teorēma[9]

Ja $\exists\, T(x,\theta)$:

- $F_{T(x,\theta)}(t), t \in R$ nav atkarīga no θ, $\forall \theta \in \Xi$
- $\forall x \Rightarrow T(x,\theta) \Uparrow \vee \Downarrow$ pēc θ neatkarīgi no x
- $\exists$ viens vienīgs atrisinājums $T(x,\theta) = t_0$ attiecība pret θ

tad, $\forall$x

$\exists T_1(x,\beta), T_2(x,\beta) : P_\theta(x : T_1(x,\beta) < \theta < T_2(x,\beta)) = \beta,$

$\beta = 1-\alpha$ – drošība

Piemēram, gadījumā, kad $T(x,\theta) \Uparrow$ visiem x un $\beta = p_2 - p_1$ iegūstam

$$\begin{cases} P_\theta(T(x,\theta) < t_1) = F_{T(x,\theta)}(t_1) = p_1 \Rightarrow t_1(p_1) \\ P_\theta(T(x,\theta) < t_2) = F_{T(x,\theta)}(t_2) = p_2 \Rightarrow t_2(p_2) \end{cases}.$$

Tātad

$$\beta = P(t_1(\beta) < T(x,\theta) < t_2(\beta)) = P(T^{-1}(x,t_1) < \theta < T^{-1}(x,t_2))$$

Nezināmā parametra θ ticamības intervāls būs

$J_{\theta,\beta} = (T^{-1}(x,\beta), T^{-1}(x,\beta))$, x - izlase, β-drošība.

$J_{\theta,\beta} = (T^{-1}(x,\beta), T^{-1}(x,\beta))$ ir tādu θ vērtību intervāls, kas saskan ar novērojumu rezultātiem, t.i., nav ar to pretrunā.

Vēlams, lai intervāla garums, kas ir gadījuma lielums, būtu mazāks. Tomēr, visbiežāk intervāla robežas t_1, t_2 izvēlās tā, lai

$$P_\theta(T(x,\theta) > t_2) = P_\theta(T(x,\theta) < t_1) = \frac{\alpha}{2}.$$

Piemēri.

1. Ar uzdoto drošību β noteikt ticamības intervālu normāli sadalīta gadījuma lieluma ξ parametram $\theta := E\xi$, ja zināma $D\xi = \sigma^2$.

$T(x,\theta) = \dfrac{\bar{x} - \theta}{\sigma}\sqrt{n}$ apmierina teorēmas nosacījumus. Līdz ar to:

- $T(x,\theta) = \dfrac{\bar{x} - \theta}{\sigma}\sqrt{n} \sim N(0,1)$;
- ar Laplasa funkcijas tabulām, pie uzdota β atrodam *t* no vienādojumā

$$\beta = P\left(-t < \frac{\bar{x} - \theta}{\sigma}\sqrt{n} < t\right),$$

no šejienes iegūstam

$$\beta = P\left(\bar{x} - \frac{t\sigma}{\sqrt{n}} < \theta < \bar{x} + \frac{t\sigma}{\sqrt{n}}\right);$$

- tātad, matemātiskās cerības ticamības intervāls normāli sadalītam gadījuma lielumam ξ, gadījumā, kad $D\xi = \sigma^2$ zināma, ir

$$J_{\theta,\beta} = \left(\bar{x} - \frac{t\sigma}{\sqrt{n}}, < \bar{x} + \frac{t\sigma}{\sqrt{n}}\right).$$

2. Gadījuma lielums ξ normāli sadalīts ar nezināmiem parametriem θ_0, θ_1. Analizējot izlasi $x = (x_1, x_2, \ldots x_n)$, konstruēt ar uzdotu drošību β ticamības intervālus parametriem θ_0 un θ_1.

a) $T(x,\theta_0) = \dfrac{\bar{x} - \theta_0}{s}\sqrt{n-1}$ apmierina teorēmas nosacījumus.

- $T(x,\theta_0) = \dfrac{\bar{x} - \theta_0}{s}\sqrt{n-1} \sim t_{n-1}$ (Stjūdenta sadalījums ar *n*-1 brīvības pakāpēm);
- ņemot vērā, ka gadījuma lieluma t_ν ar ν brīvības pakāpēm sadalījuma blīvuma funkcija ir pāra funkcija, secinām, ka

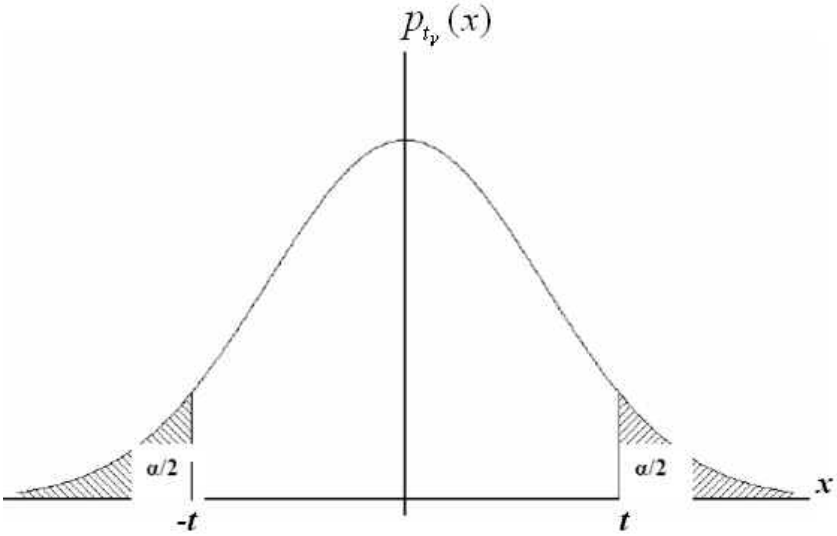
$t = -t_1 = t_2$

- ievērojot iepriekšējo, pēc Stjūdenta sadalījuma funkcijas tabulām, atrodam t no vienādojuma

$$\beta = P\left(t_1 < \frac{\bar{x} - \theta_0}{s}\sqrt{n-1} < t_2 \right).$$

- Ievietojot $t = -t_1 = t_2$ formulā, iegūstam

$$\beta = P\left(\bar{x} - \frac{ts}{\sqrt{n-1}} < \theta_0 < \bar{x} + \frac{ts}{\sqrt{n-1}} \right)$$



8.zīm.Simetrisks intervāls zem Stjūdenta sadalījuma līknes.

- līdz ar to secinām, ka pie uzdota β matemātiskās cerības ticamības intervāls normāli sadalītam gadījuma lielumam ξ, gadījumā, kad dispersija ξ nav zināmā, ir

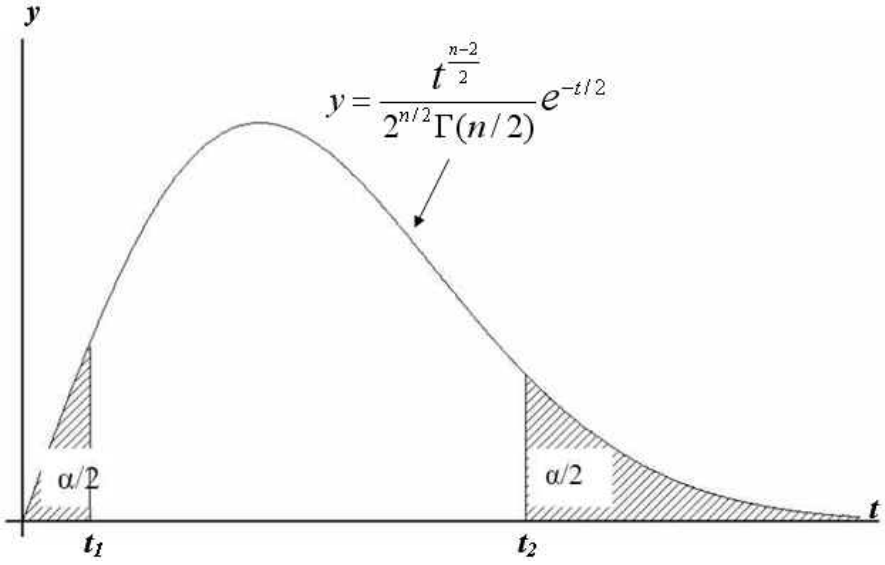
$$J_{\theta_0,\beta} = \left(\bar{x} - \frac{ts}{\sqrt{n-1}}, < \bar{x} + \frac{ts}{\sqrt{n-1}} \right).$$

b) $T(x,\theta_1)=\dfrac{s^2 n}{\theta_1}\sim \chi^2_{n-1}$ un apmierina teorēmas nosacījumus. Līdz ar to uzdotam β var atrast bezgalīgi daudz skaitļu pāru (t_1,t_2) tādu, ka

$$P(t_1<\chi^2_{n-1}<t_2)=\beta\,.$$

Skaitās lietderīgi intervāla robežas t_1 un t_2 izvēlēties tā, lai izpildās

$$P_{\theta_1}(T(x,\theta_1)>t_2)=P_{\theta}(T(x,\theta_1)<t_1)=\frac{1-\beta}{2}=\frac{\alpha}{2}\,.$$



9.zīm. Intervāls zem χ^2_ν sadalījumu blīvuma funkcijas.

- t_1 un t_2 atrod no tabulām un ievieto formulā

$$\beta=P(t_1<\frac{s^2 n}{\theta_1}<t_2)=P\left(\frac{s^2 n}{t_2}<\theta_1<\frac{s^2 n}{t_1}\right);$$

- no šejienes secinām, ka pie uzdota β dispersijas ticamības intervāls normāli sadalītam gadījuma lielumam ξ ir

$$J_{\theta_1,\beta}=\left(\frac{s^2 n}{t_2},\frac{s^2 n}{t_1}\right).$$

1.piezīme.

Dispersijas ticamības intervāla konstruēšanai mēs lietojam gadījuma lieluma χ^2_ν sadalījuma tabulas, kurās parasti brīvības pakāpes $\nu\le 30$.

Izmantosim faktu, ka $E\chi_\nu^2 = \nu,\ D\chi_\nu^2 = 2\nu$ un to, ka gadījuma lieluma $\frac{\chi_\nu^2 - \nu}{\sqrt{2\nu}}$ sadalījuma funkcija, ja $\nu \to \infty$, tiecas uz standartu normālu sadalījumu. Tad, pie lielām ν vērtībām $\chi_\nu^2 \sim N(\nu, 2\nu)$. Tātad, ja ν>30,

$$\beta = P(t_1 < \frac{s^2 n}{\theta_1} < t_2) = P\left(\frac{t_1 - (n-1)}{\sqrt{2(n-1)}} < \frac{s^2 n - (n-1)}{\sqrt{2(n-1)}} < \frac{t_2 - (n-1)}{\sqrt{2(n-1)}} \right) =$$

$$= F_{N(0,1)}\left(\frac{t_2 - (n-1)}{\sqrt{2(n-1)}} \right) - F_{N(0,1)}\left(\frac{t_1 - (n-1)}{\sqrt{2(n-1)}} \right).$$

Ievērojot, ka

$$P_{\theta_1}\left(\frac{s^2 n}{\sigma^2} > t_2 \right) = P_\theta\left(\frac{s^2 n}{\sigma^2} < t_1 \right) = \frac{1-\beta}{2},$$

iegūstam

$$F_{N(0,1)}\left(\frac{t_1 - (n-1)}{\sqrt{2(n-1)}} \right) = \frac{1-\beta}{2},\ F_{N(0,1)}\left(\frac{t_2 - (n-1)}{\sqrt{2(n-1)}} \right) = \frac{1+\beta}{2},$$

kā arī to, ka $\frac{t_2 - (n-1)}{\sqrt{2(n-1)}} = -\frac{t_1 - (n-1)}{\sqrt{2(n-1)}}.$

2.piezīme.

Ja novērojamā normāli sadalīta gadījuma lieluma matemātiskā cerība *a* ir zināma, tad, lai konstruētu dispersijas σ^2 ticamības intervālu, var izmantot statistiku

$$T(x, \sigma^2) = \frac{\overline{x} - a}{\sigma}\sqrt{n} \sim N(0,1).$$

3.piezīme.

Pieņemsim, ka $F(x, \theta)$ ir gadījuma lieluma ξ sadalījuma funkcija ar vienu nezināmu parametru $\theta \in \mathrm{R}$. Atzīmēsim, ka šajā gadījumā gan $E\xi$, gan $D\xi$ ir funkcijas no parametra θ.

Pieņemsim, ka $x_1, x_2, ..., x_n$ ir gadījuma lieluma ξ izlase. Tā kā $\{x_i, i = 1, 2, ..., n\}$ kopumā neatkarīgi un vienādi sadalīti, tad pie lieliem *n* pēc centrālās robežteorēmas izpildās

$$\frac{\sum_{i=1}^{n}(x_i - Ex_i)}{\sqrt{\sum_{i=1}^{n} Dx_i}} \sim N(0,1),$$

kas ļauj konstruēt ticamības intervālu nezināmam parametram $\theta \in \mathrm{R}$.

3. Ar uzdotu drošību β noteikt ticamības intervālu eksponenciāla sadalīta gadījuma lieluma ξ parametram $\theta := \lambda$, gadījumā, kad izlases apjoms *n* ir liels.

Blīvuma funkcija eksponenciālajam sadalījumam ir

$$p_\xi(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0,\ \lambda > 0. \end{cases}$$

$E\xi = \frac{1}{\lambda}$, $D\xi = \frac{1}{\lambda^2}$ un pie lieliem *n* pēc CRT funkcija

$$T(x, \lambda) = \frac{\bar{x} - \frac{1}{\lambda}}{\frac{1}{\lambda}} \sqrt{n} = \lambda(\bar{x} - 1)\sqrt{n} \sim N(0,1).$$

Tātad, pie uzdotas drošības β no

$$\beta = P\left(-t < (\lambda\bar{x} - 1)\sqrt{n} < t\right)$$

atrodam *t*. Līdz ar to, eksponenciāli sadalīta gadījuma lieluma ξ parametra λ ticamības intervāls ir

$$J_{\lambda,\beta} = \left(\frac{1}{\bar{x}}\left(1 - \frac{t}{\sqrt{n}}\right), \frac{1}{\bar{x}}\left(1 + \frac{t}{\sqrt{n}}\right) \right).$$

4.Piezīme.
Pie lieliem *n* līdzīgi konstruē ticamības intervālu nezināmam parametram, piemēram, tādiem gadījuma lielumiem:

$\circ\ \xi \sim b(1,P),\ T(x,p)=\dfrac{\bar{x}-p}{\sqrt{p(1-p)}}\sqrt{n} \sim N(0,1),\ n-\text{liels},$

$\circ\ \xi \sim P(\lambda),\ T(x,\lambda)=\dfrac{\bar{x}-\lambda}{\sqrt{\lambda}}\sqrt{n} \sim N(0,1),\ n-\text{liels},$

$\circ\ \chi^2_\nu,\ T(x,\nu)=\dfrac{\bar{x}-\nu}{\sqrt{2\nu}}\sqrt{n} \sim N(0,1),\ \ n-\text{liels}.$

5.Piezīme.
Lai aprēķinātu korelācijas koeficienta $\rho(\xi,\eta)$ ticamības intervālu, var izmantot statistikas

$$T(x,y)=\frac{1}{2}\ln\frac{1+r_{x,y}}{1-r_{x,y}}$$

sadalījuma funkciju, kura ļoti tuva normālajam sadalījumam ar matemātisko cerību $\frac{1}{2}\ln\frac{1+\rho(\xi,\eta)}{1-\rho(\xi,\eta)}$ un dispersiju $\frac{1}{\sqrt{n-3}}$ (R.Fišers).

Statistiskā hipotēžu pārbaude[1],[2],[5]

Hipotēzes par varbūtību statistiskā pārbaude

Pieņemsim, ka zināmu apsvērumu rezultātā mums ir pamats uzskatīt, ka dotā notikuma varbūtība ir *p*. Zinot, ka n neatkarīgos mēģinājumos notikums parādījies *m* reizes, gribām pārbaudīt mūsu hipotēzi par notikuma varbūtību *p*. Piemēram, metot monētu, uzskatām, ka varbūtība vienā metienā uzkrist ģerbonim ir 0.5. Pieņemsim, ka 100 reizes metot monētu, ģerbonis ir uzkritis 40 reizes. Mūsu piemērā uzkritušo ģerboņu skaits μ - gadījuma lielums, kas sadalīts pēc binomiālā sadalījuma likuma ar matemātisko cerību $E\mu = np = 50$ un dispersiju $D\mu = npq = 25.$ Mūs interesē, vai eksperimentā uzkritušo ģerboņu skaitu 40 var uzskatīt par pietiekoši tuvu teorētiskai normai 50, kas atbilst hipotēzei $p=\frac{1}{2}$.

Vispirms izvēlēsimies robežas gadījuma lieluma μ pieļaujamām novirzēm, ja spēkā mūsu hipotēze, t.i., uzrādīsim tādu „kritisko novirzi", kuras pārsniegšana pie mūsu hipotēzes ir iespējama ar tik mazu varbūtību, ka to praktiski var skaitīt par neiespējamu. Tāpēc, ja šī pārsniegšana tomēr notiek, mūsu hipotēze nav savienojama ar novērojumiem, novērojumi neapstiprina hipotēzi. Un citādi, ja faktiskās novirzes mazākas par kritisko, mēs varam uzskatīt, ka eksperimenta rezultāts nav pretrunā ar izvirzīto hipotēzi un novērotajām novirzēm ir gadījuma raksturs. Parasti par praktiski neiespējamām novirzēm pieņem tādas, kuru varbūtība nepārsniedz 0.05 vai 0.01. Šo varbūtību sauc par nozīmības līmeni α, tai atbilstošo lielo noviržu apgabalu par kritisko apgabalu, bet pašu pārbaudes likumu- par nozīmes kritēriju (pārbaudes procedūru, hipotēzes kritēriju). Princips, pēc kura notikumi ar mazu varbūtību skaitās neiespējami, bet notikumi ar varbūtību tuvu vieniniekam – gandrīz droši, ir gandrīz visas matemātiskās statistikas pamatā. Skaitli α izvēlas aiz praktiskiem apsvērumiem. Dažos gadījumos var ņemt vērā notikumus, ja to varbūtība <0.05, bet, ja jānoskaidro, piemēram, lidmašīnas avārijas iespēja, jāņem vērā pat gadījumi, kuru varbūtība ir 0.001 un mazāka.

Pieņemsim, ka α=0.05. Ņemot vērā, ka pie pietiekoši lieliem *n* gadījuma lielums μ ir tuvu normāli sadalītam gadījuma lielumam, lielums

$$\eta = \frac{\mu - np}{\sqrt{npq}}$$

tuvu normāli sadalītam gadījuma lielumam ar parametriem $E\eta = 0,\ D\eta = 1.$ No šejienes atradīsim kritiskā apgabala robežu *t*

$$P(|\eta| > t) = \alpha \Rightarrow t = 1.96.$$

Tātad, gadījuma lieluma η vērtības, kas pēc absolūtās vērtības lielākas par *t,* iespējamas ar varbūtību 0.05, tik maza, ka pie pieņemtās hipotēzes dotā eksperimentā, tas ir neiespējams notikums. Jānoskaidro gadījuma lieluma η eksperimentā realizējusies vērtība

$$\eta^* = \frac{40 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.25}} = -5.$$

Tā kā $|\eta^*| > 1.96$, hipotēzi nevar pieņemt par pareizu, jo eksperimentā realizējies notikums ar tik mazu varbūtību, ka mēs to pieņemam par neiespējamu.

Princips:
notikumi ar mazu varbūtību tiek uzskatīti par neiespējamiem, notikumi ar varbūtību tuvu 1 gandrīz drošiem.

Ja η^* būtu bijis iespējamo vērtību robežas, tad hipotēzes būtu jāpārbauda tālāk, jo eksperimentā iegūtais rezultāts nebūtu ar to pretrunā, bet tas vēl nebūtu hipotēzes pareizības pierādījums.

Hipotēžu pārbaudes vispārīgais uzdevums

Pieņemsim, ka runājot par kādas parādības veidu, hipotēze H_0, zināmu apsvērumu dēļ, ir pieņemta kā galvenā (nulles), nošķirot to no visas alternatīvo hipotēžu $\{H_i, i=1,2,...\}$ kopas. Ir eksperiments, kura rezultāts x ir kādas kopas X (ģenerālās kopas) elements. Pēc izlases vērtības x jānosaka, kurai no hipotēzēm dodama priekšroka.
Hipotēzes H_0 pārbaudes process formāli ir šāds:
- izvēlas kādu kopu S, ko sauc par hipotēzes H_0 kritisko kopu
- izdara eksperimentu
- ja eksperimenta rezultāts $x \in S$, tad hipotēzi H_0 noraida.

Vēlams, lai S apmierinātu sekojošas prasības:
a) $P(x \in S \mid H_0) = 0$, jo tādā gadījumā mēs nekad nenoraidītu pareizu hipotēzi H_0
b) $P(x \in S \mid H_i) = 1$, $i \neq 0$, jo tādā gadījumā mēs vienmēr noraidītu hipotēzi H_0, ja patiesībā pareiza būtu jebkura no hipotēzēm $H_i, i \neq 0$.
Taču praktiski interesantos gadījumos, lai izpildās $P(x \in S \mid H_0) = 0$, kopai S jābūt tukšai. Bet tad arī $P(x \in S \mid H_i) = 0$ visiem i, un pārbaudes process ir bijis veltīgs. Tāpēc jāpieļauj no 0 atšķirīgas

vērtības varbūtībai $P(x \in S \mid H_0)$, izvēloties vispirms „nozīmības līmeni", t.i., izvēloties skaitli $\alpha > 0$ un prasot, lai

$$P(x \in S \mid H_0) \le \alpha .$$

Ha hipotēze H_0 ļoti nozīmīga, α jābūt mazam. Parasti izvēlas vienu no skaitļiem 0.05, 0.01, 0.001, kas ir visu atzīti un kuriem tāpēc ir sastādītas atbilstošas statistiskas tabulas. Tātad, vispirms izvēlas α, tad izvēlas *S*, kas apmierina

$$P(x \in S \mid H_0) \le \alpha$$

un izdara eksperimentu.

Acīmredzot $P(x \in S \mid H_0)$ ir varbūtība noraidīt hipotēzi H_0, ja tā pareiza. Šādu kļūdu sauc par ***pirmā veida kļūdu.*** No $P(x \in S \mid H_0) \le \alpha$ seko, ka pirmā veida kļūdas varbūtība nepārsniedz nozīmības līmeni α.

Ja $x \in S$, kur *S* apmierina $P(x \in S \mid H_0) \le \alpha$, tad „hipotēze H_0 tiek noraidīta nozīmības līmeni α".

Ja $x \notin S$, tad var sacīt, ka „hipotēze H_0 netiek noraidīta nozīmības līmeni α". Apgalvot, ka „hipotēze H_0 tiek pieņemta", nedrīkst , jo iespējams, ka citā nozīmības līmenī tā tiks noraidīta. Protams, ka mūs galvenokārt interesē gadījumi, kad hipotēze H_0 tiek vai netiek noraidīta visos saprātīgos nozīmības līmeņos. Tādi gadījumi sastopami bieži.

Bez pirmā veida kļūdas iespējama ***otrā veida kļūda***- hipotēze H_0 netiek noraidīta, lai gan patiesībā pareiza ir nevis H_0, bet gan kāda no hipotēzēm $H_i, i \neq 0$. Šis kļūdas varbūtība ir

$$\gamma(i) := P(x \notin S \mid H_i) = 1 - P(x \in S \mid H_i) .$$

Funkciju

$$\beta(i) = P(x \in S \mid H_i),$$

kas vienāda ar varbūtību noraidīt hipotēzi H_0, jo pareiza hipotēze H_i $i \neq 0$, sauc par statistiskā kritērija *S* ***jaudu***. Lai otrā veida kļūda būtu

vismazākā starp visiem *S'*, kas apmierina $P(x \in S' \mid H_0) \le \alpha$,
jāizvēlas tas *S*, kam spēkā

$$P(x \in S \mid H_i) = \sup_{S'} P(x \in S' \mid H_i),\ i \neq 0.$$

Vispār *S* atkarīgs no *i*. Ja eksistē *S*, kas nav atkarīgs no *i*, tad tas ir vislabākais kritērijs, kas diemžēl diezgan reti eksistē. Otrā veida kļūdu regulē ar kritiskā apgabala izvēli. Kritiskais apgabals jāizvēlas tā, lai gadījumā, kad noraida $H_i \ \ i \neq 0$, varbūtība nokļūt apgabalā S, kas šajā gadījumā ir nepareizas hipotēzes H_0 noraidīšanas varbūtība, būtu vislielākā.

Pieņemsim, ka eksperimenta rezultātā novēro statistiku *T(x)*.
Tad, gadījumā *i*=2, hipotēzes H_0 pārbaudes procesu formulē šādi:

Pieņemsim, ka $F(x,\theta)$ gadījuma lieluma ξ sadalījuma funkcija, funkcija zināma, bet satur nezināmu parametru

$$\theta \in \Xi,\ \Xi = \Xi_0 \bigcup \Xi_1,\ \Xi_0 \bigcap \Xi_1 = \varnothing,$$

$x = (x_1, x_2, ..., x_n)$ -gadījuma lieluma ξ vērtību izlase,
$\alpha \in (0.05; 0.01; 0.001)$ – nozīmības līmenis.

***Hipotēzes H_0 pārbaudes process*:**

- izvirzās
  $H_0 = \{\theta \in \Xi_0\}$ -nulles hipotēze
  $H_i = \{\theta \in \Xi_1\}$ - alternatīva hipotēze

- izvēlas α

- izvēlas statistiku $T(x) \in \mathrm{X}$,

- izvēlas hipotēzes H_0 kritisko kopu $S \subset \mathrm{X}$:

$$\begin{cases} C = \{S' : P(T(x) \in S' \mid H_0) = \alpha\} \\ \sup\limits_{S' \in C} P(T(x) \in S' \mid H_1) = \\ \qquad = P(T(x) \in S \mid H_1) \end{cases}$$

- veic eksperimentu
- analizē eksperimenta rezultātu

$T(x) \in S \Rightarrow$ hipotēze H_0 tiek noraidīta nozīmības līmenī α,

$T(x) \notin S \Rightarrow$ hipotēze H_0 netiek noraidīta nozīmības līmenī α.

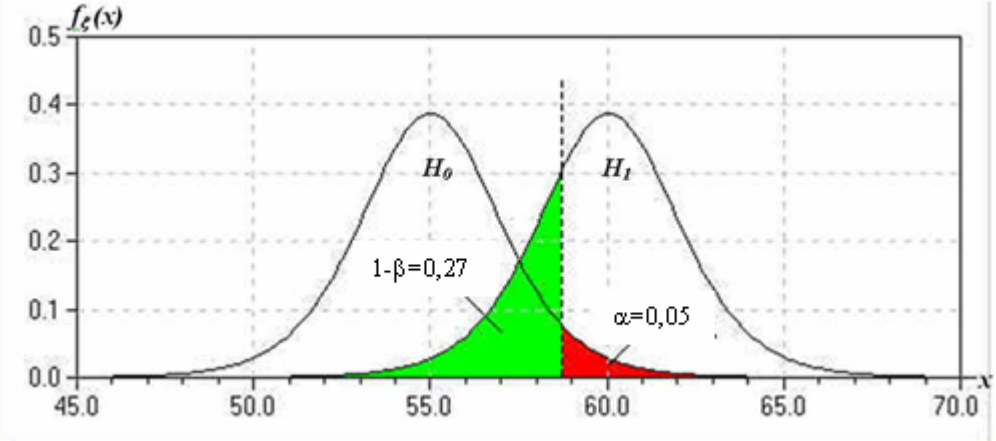
$\alpha = P(T(x) \in S \mid H_0) - pirmā\ veida\ kļūda$

$1 - \beta = P(T(x) \notin S \mid H_1) - otrā\ veida\ kļūda$

$\beta = P(T(x) \in S \mid H_1) - kritērija\ jauda$

Kritiskais apgabals

Pieņemsim, ka par gadījuma lieluma ξ sadalījumu izvirzās 2 hipotēzes



10.zīm.

$H_0 = (\theta = \theta_0), H_1 = (\theta = \theta_1)$

$f_\xi(x)$ – gadījuma lieluma ξ sadalījuma blīvuma funkcja

Pie šādām hipotēzēm, kas dotas zīmējumā, otrā veida kļūda minimāla, ja par kritisko apgabalu izvēlas *lielu pozitīvo noviržu apgabalu.*

Ja hipotēze H_1 būtu novirzīta pa kreisi attiecība pret hipotēzi H_0 par kritisko apgabalu būtu jāizvēlas *lielu negatīvo noviržu apgabalu.*

Ja nav zināmas konkurējošās hipotēzes novirzes tendences, tad par kritisko apgabalu izvēlas *pēc absolūtās vērtības lielu noviržu apgabalu.*

Vidējo salīdzināšana.

Praksē bieži vien eksperimentu veic vairākas reizes, pie tam eksperimentos iegūtie vidējie lielumi jūtami atšķiras. Kā piemēru varam minēt produkcijas kvalitātes izlases pārbaudi.
Jānoskaidro, vai šo atšķirību dod eksperimentu gadījuma kļūdas, izlases ierobežotība vai arī būtiskas atšķirības izlasēs.
Piemēram: produkcijas kvalitāte būtiski atšķiras atkarībā no partijas, jo netiek ievērots ražošanas tehnoloģijas process .
Pieņemsim, ka katrā sērijā tiek mērīts viens un tas pats lielums a.
Mērīšanas kļūdas pirmajā un otrajā eksperimentā sadalītas pēc normālā sadalījuma likuma attiecīgi ar parametriem $(0,\sigma_1^2)$ un $(0,\sigma_2^2)$. Jānoskaidro, vai mērāmo lielumu patiesā vērtība ir a, kaut gan $\overline{x}$ un $\overline{y}$ atšķīrās.

Atrisinājuma etapi:

1. Izvirzām nulles hipotēzi H_0:
   „starp salīdzināmo izlašu parametriem būtisku atšķirtību nav”
   un alternatīvas hipotēzi H_1:
   „salīdzināmo izlašu parametri būtiski atšķiras”.

2.Atrodam gadījuma lieluma $T(x,y)=\bar{x}-\bar{y}$, vai kādas citas statistikas $T(x,y)=f(x,y)$ sadalījuma funkciju.

3.Uzdodot $\beta=1-\alpha$ konstruējam $T(x,y)$ statistikas „kritisko apgabalu" *S*.

4.Izdarot eksperimentu izrēķinām statistikas iegūto vērtību *T**.

5. Ja iegūtā vērtība *T** ir kritiskā apgabalā, tad izvirzīto nulles hipotēzi noraidām. Ja nē, tad varam uzskatīt, ka atšķirību izsaukušas eksperimenta gadījuma kļūdas. Taču mūsu nulles hipotēze pierādīta nav, nevaram apgalvot, ka abas izlases ir iegūtas viena un tā paša lieluma mērījumu rezultātā.

a)***Vidējo salīdzināšana, ja zināma mērījumu precizitāte***

Pieņemsim, ka dotas divas izlases $x_1,x_2,...,x_n;\ y_1,y_2,...,y_m$, un ka abās mērījumu sērijās, abos eksperimentos, zināmi lielumi σ_1 un σ_2. Tad, ņemot vērā, ka

$x_i \sim N(a_1,\sigma_1^2),\ i=1,2,...,n,\ \ y_i \sim N(a_2,\sigma_2^2),\ i=1,2,...,m,$

iegūstam, ka gadījuma lielumi

$$\bar{x} \sim N(a_1,\frac{\sigma_1^2}{n}),\ \ \bar{y} \sim N(a_2,\frac{\sigma_2^2}{m})$$

un ir neatkarīgi. No tas seko, ka

$$\bar{x}-\bar{y} \sim N\left(a_1-a_2,\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}\right).$$

No šejienes

$$T(x,y)=\frac{\bar{x}-\bar{y}-(a_1-a_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \sim N(0,1)\ .$$

Tālāk pēc dota α noteiksim kritisko apgabalu

$$P(|T(x,y)|>t\,|\,H_0)=\alpha \text{ jeb } P\left(\left|\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\right|>t\right)=\alpha .$$

No šejienes pēc Laplasa integrāļa tabulām atrodam *t*.
Tālāk, ja

$$T^*(x,y)=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \quad \text{kur,} \quad \bar{x},\ \bar{y},\ n,\ m,\ \sigma_1^2,\ \sigma_2^2$$

konkrēti, dotie vai eksperimentā iegūti skaitļi , pēc moduļa lielāki par *t,* tad hipotēzi H_0*:* ka starp salīdzināmo izlašu parametriem būtiski atšķirtību nav, noraidām kā nepareizu. Tātad atšķirtība ir būtiska. Pretējā gadījumā atšķirības var būt gadījuma rakstura, nenozīmīgas.

b)***Vidējo salīdzināšana, ja mērījumu precizitāte nav zināma***

Pieņemsim, ka σ_1^2 un σ_2^2 nav zināmi, bet abās izmēģinājumu sērijās vienādi (piemēram, abās sērijās mērījumi tiek veikti ar vienu instrumentu).
Ievērojot, ka

$$\left(\frac{s_1\sqrt{n}}{\sigma_1}\right)^2 \sim \chi^2_{n-1}, \left(\frac{s_2\sqrt{m}}{\sigma_2}\right)^2 \sim \chi^2_{m-1},$$

pieņēmumu $\sigma_1^2=\sigma_2^2=:\sigma$, kā arī faktu, ka neatkarīgiem χ^2_{n-1} un χ^2_{m-1} spēkā $\chi^2_{\text{n-1}}+\chi^2_{\text{m-1}}=\chi^2_{\text{n+m-2}}$, iegūstam

$$\frac{s_1^2 n+s_1^2 m}{\sigma^2}\sim\chi^2_{n+m-2}.$$

Ņemot vērā, ka gadījuma lielumi

$$\frac{\bar{x}-\bar{y}-(a_1-a_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1), \quad \frac{s_1^2 n + s_1^2 m}{\sigma^2} \sim \chi^2_{n+m-2}$$

ir neatkarīgi, konstruējam statistiku sadalītu pēc *t*-sadalījuma likuma

$$T(x,y) = \frac{\bar{x}-\bar{y}-(a_1-a_2)}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{s_1^2 n + s_2^2 m}}\sqrt{n+m-2} \sim t_{n+m-2.}$$

Pēc dota α, no t-sadalījuma tabulām atrodam *t* no

$$P(|\,T(x,y)\,| > t \,|\, H_0) = \alpha,$$

t.i.,

$$\mathrm{P}\left(\left|\frac{\bar{x}-\bar{y}}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{s_1^2 n + s_2^2 m}}\sqrt{n+m-2}\right| > t\right) = \alpha.$$

Tālāk, pēc eksperimentā iegūtajiem skaitļiem, jāaprēķina

$$T^*(x,y) = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{s_1^2 n + s_2^2 m}}\sqrt{n+m-2}$$

un jāanalizē, vai *T*(x ,y)* vērtība pieder kritiskajam apgabalam vai nē.

Piezīme.
Šajā izlases salīdzināšanas uzdevumā var tikt mērīts viens un tas pats normāli sadalīts gadījuma lielums. Tad gadījumā a) σ_1 un σ_2, kas tiek uzdoti, vairs nav tikai mērījumu kļūdu rezultāts, gadījumā b) spriedumi paliek tie paši.

Dispersiju hipotēžu pārbaude

Dispersiju hipotēzēm pielietojumos ir svarīga nozīme, jo dispersija raksturo tādus konstruktīvus un tehnoloģiskus rādītājus kā mašīnu un instrumentu precizitāte, tehnoloģisko procesu precīza izpilde utt. Pieņemsim, ka gribam pārbaudīt, vai divas normālu izlašu dispersijas ir vienādas. Izvirzām hipotēzi, ka dispersijas ir vienādas. Kā mēs redzējām, šāda veida hipotēžu pārbaudē jāatrod statistika, kuras, ja H_0 pareiza, sadalījums nav atkarīgs no novērojumu sadalījuma likuma parametriem, jo tādam ne no kā neatkarīgam sadalījumam var viegli sastādīt tabulas. Apskatīsim attiecību

$$T(x,y)=\frac{s_x^2 n(m-1)}{s_y^2 m(n-1)}=\frac{m-1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2},$$

kur $x_1, x_2, ..., x_n;\ y_1, y_2, ..., y_m$ -divas neatkarīgas izlases no sadalījumiem $\xi \sim N(a_1, \sigma_1^2),\ \eta \sim N(a_2, \sigma_2^2)$, turklāt skaitītajam sadalījums ir $\chi^2_{n-1}\frac{\sigma_1^2}{n-1}$, saucējam - $\chi^2_{m-1}\frac{\sigma_2^2}{m-1}$. Ja hipotēze H_0 pareiza, tad attiecībai ir *F*- (Snedekora) sadalījums ar atbilstošajām brīvības pakāpēm *n*-1 un *m*-1, $T(x,y) \sim F_{n-1,m-1}$.

Ievērojot *F*- sadalījuma blīvuma funkcijas formu, kritisko apgabalu izvēlas tā, lai pie nozīmības līmeņa izpildās α

$$P(F_{n-1,m-1} > t_2) = P(F_{n-1,m-1} < t_1) = \frac{\alpha}{2}.$$

Tāda kritiskā apgabala izvēle dod lielāku pārbaudes kritērija jaudu.

Acīmredzot, ka $\frac{1}{F_{n-1,m-1}} \sim F_{m-1,n-1}$. Tad

$$P(F_{n-1,m-1} < t_1) = P\left(\frac{1}{F_{n-1,m-1}} > \frac{1}{t_1}\right) = P(F_{m-1,n-1} > \frac{1}{t_1}).$$

Tātad, $F_{n-1,m-1}$ - sadalījuma kreisajam kritiskajam punktam atbilst $F_{m-1,n-1}$ sadalījuma labais kritiskais punkts. Tāpēc pietiek ar vienpusīgām ticamības robežu tabulām, lai noteiktu kritisko apgabalu.

Korelācijas koeficienta hipotēzes pārbaude

Lai pārbaudītu hipotēzi par korelācijas koeficientu, var izmantot ticamības intervāla metodi. Šai pārbaudei ir divi soļi:

1. izmantojot statistiku

$$T(x,y)=\frac{1}{2}\ln\frac{1+r_{x,y}}{1-r_{x,y}}$$

nosaka ar drošību $\beta=1-\alpha$ ticamības intervālu jeb statistikas pieļaujamo vērtību apgabalu;

2. izdara secinājumus:
   - ja nulles hipotēzes vērtība pieder ticamības intervālam, tad nevaram noraidīt nulles hipotēzi $H_0:\rho(\xi,\eta)=\rho_0$;
   - ja vērtība atrodas ārpus ticamības intervāla, nulles hipotēzi noraida un uzskata, ka izpildās $H_1:\rho(\xi,\eta)\neq\rho_0$.

Tādējādi ticamības intervālu varam uzskatīt par nulles hipotēzes pieļaujamo apgabalu, bet intervālu ārpus pieļaujamā apgabala – par nulles hipotēzes noraidīšanas apgabalu.

Sadalījuma likumu hipotēžu pārbaude

Pieņemsim, ka jāpārbauda hipotēze H_0, kas apgalvo, ka gadījuma lieluma ξ sadalījuma funkcija ir $F_\xi(x)$. Kā zināms, šajā gadījumā, tiek konstruēts kritiskais apgabals S, un hipotēze tiek noraidīta, ja izlases vērtība $x\in S$ un netiek noraidīta, ja $x\notin S$. Kritisko apgabalu konstruē, izmantojot statistiku $T(x,y)\in Z$, kuras sadalījuma likums $F_{T(x,y)}(t)$, $t\in\mathrm{R}$ zināms. Tad, kritiskais apgabals var būt, piemēram, šāda veida $\{z:T(x,y)>\gamma_\alpha\}$, kur

$P\{z:T(x,y)>\gamma_\alpha \mid H_0\}\le\alpha$, α -iepriekš uzdots nozīmības līmenis. Konstruētā kritērija pirmā veida kļūdas varbūtība nepārsniedz α . Taču α ir pārāk mazs, lai kritērijs mūs pilnīgā apmierinātu. Vajag, lai tam būtu pietiekoša jauda, t.i., lai varbūtība $P\{z:T(x,y)>\gamma_\alpha \mid H_1\}$ būtu pietiekoši tuva vieniniekam. Pie uzdotās alternatīvas H_1 vairumam statistiku šīs īpašība neizpildās.
Tāpēc īpaši interesantas ir universālās statistikas, kuru atbilstošajiem kritērijiem ir pietiekoša jauda plašā gadījumu klasē.

χ^2 *statistika un kritērijs (Pirsona kritērijs)*

Dota izlase ar apjomu *n.* Jāpārbauda, vai eksperimentālie dati nav pretrunā ar hipotēzi, ka gadījuma lielums ξ sadalīts pēc sadalījuma likuma $F_\xi(x)$.
Sadalīsim skaitļu taisni *r* savstarpēji nešķeļošos nogriežņos: $(-\infty,a_1),[a_1,a_2),...,[a_{r-1},\infty)$. Ar $p_i, i=1,2,...,r$ apzīmēsim varbūtību izlases vērtībai iekļūt *i*-tajā nogrieznī. Acīmredzot,

$$p_i = F_\xi(a_i) - F_\xi(a_{i-1}).$$

Definēsim statistiku

$$\chi^2 := \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i},$$

kur *n* –izlases apjoms, n_i- *i*-tajā intervālā esošo novēroto vērtību skaits, *i*=1,2,...,r; $a_0=-\infty$, $a_r=\infty$. Kritēriju, kura pamatā ir χ^2 ***statistika,*** sauc par χ^2 ***kritēriju***. Ja hipotēze H_0 pareiza, tad $\frac{n_i}{n} \underset{n\to\infty}{\to} p_i$, $i=1,2,...,r$ gandrīz droši, un statistika

$$\chi^2 := \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{r} \left(\frac{n_i}{n} - p_i\right)^2 \frac{n}{p_i}$$

pie pietiekoši lieliem *n* ar lielu varbūtību būs ārpus kritiskā apgabala $S = \{\chi^2 > \gamma_\alpha\}$. Ja pareiza H_1, t.i., ģenerālās kopas sadalījuma funkcija $F_\xi'(x) \neq F_\xi(x)$, tad eksistē tāds *i*, ka

$\left(\frac{n_i}{n} - p_i\right)^2 > 0$. Tātad $\chi^2 \underset{n\to\infty}{\to} \infty$. Šie spriedumi rāda , ka χ^2 kritērijam ir pietiekoša jauda. To apstiprina gan precīzi aprēķini, gan šī kritērija sekmīga pielietošana praktisku uzdevumu risināšanā.

Teorēma[1],[6].

Ja pareiza hipotēze $H_0 : F_\xi(x) = F(x)$ un izlases apjoms tiecās uz bezgalību, tad statistikas

$$\chi^2 := \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

sadalījums tiecas uz χ^2_{r-1-l} sadalījumu ar $r-1-l$ brīvības pakāpēm, kur *n* –izlases apjoms; $(-\infty, a_1), [a_1, a_2), ..., [a_{r-1}, \infty)$ - skaitļu taisnes sadalījums savstarpēji nešķeļošos intervālos, $a_0 = -\infty$, $a_r = \infty$; n_i- *i*-tajā intervālā esošo novēroto vērtību skaits, *i*=1,2,...,r; $p_i = F(a_i) - F(a_{i-1})$, *i*=1,2,...,r; *l*- novērtējumu skaits, kas izmantots, aprēķinot teorētisko sadalījumu. Teorēma ļauj par kritisko apgabalu ņemt $S = (\gamma, \infty)$, kur γ ir vienādojuma

$$1 - F_{\chi^2_{r-1-l}}(\gamma) = \alpha$$

atrisinājums.

Piezīme.

Galvenie uzdevumi, kurus risina, izmantojot χ^2 kritēriju, ir šādi:

1. pārbauda, vai divi gadījuma lielumi ir neatkarīgi viens no otra;
2. pārbauda, vai dotais empīriskais sadalījums atbilst izvēlētajam teorētiskajam sadalījumam;
3. pārbauda, vai divi vai vairāki empīriskie sadalījumi ir vienāda veida, nenoskaidrojot, kādam teorētiskam sadalījumam tie atbilst.

χ^2 *kā empīriskā un teorētiskā sadalījuma atbilstības pārbaudes kritērijs*

Pieņemsim, ka dota izlase ar apjomu *n.* Ar drošību $\beta = 1 - \alpha$ jāpārbauda, vai eksperimentālie dati nav pretrunā ar hipotēzi, ka gadījuma lielums ξ sadalīts pēc sadalījuma likuma $F_\xi(x)$.Tātad,

$$\begin{cases} H_0 : F_\xi(x) = F(x) \\ H_1 : F_\xi(x) \neq F(x) \end{cases}$$

Sadalīsim skaitļu taisni *r* savstarpēji nešķeļošos nogriežņos:

$$(-\infty, a_1), [a_1, a_2), ..., [a_{r-1}, \infty).$$

Pieņemot, ka nulles hipotēze ir spēkā, aprēķināsim varbūtību, ar kādu izlases vērtība var iekļūt *i*-tajā nogrieznī . t.i., $p_i, i = 1, 2, ..., r$.
Ja sadalījuma funkcija satur nezināmus parametrus, jāizmanto parametru atbilstošie novērtējumi.
Definēsim statistiku

$$T(x) = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i},$$

kur *n* –izlases apjoms, n_i- *i*-tajā intervālā esošo novēroto vērtību skaits, *i*=1,2,...,r; $a_0 = -\infty$, $a_r = \infty$.

Ja pareiza H_0, pēc teorēmas, $T(x) \underset{n\to\infty}{\to} \chi^2_{r-1-l}$ pēc sadalījuma, kur *l*-novērtējumu skaits, kas izmantots, aprēķinot $p_i, i = 1, 2, ..., r$.

Ja spēkā H_1, tad vismaz pie viena *i*

$$\lim_{n\to\infty} \left(\frac{n_i}{n} - p_i \right)^2 > 0 \text{ , un } T(x) \underset{n\to\infty}{\to} \infty .$$

Līdz ar to $S = (\gamma, \infty)$, kur γ atrodam no vienādojumā

$$P(T(x) > \gamma \mid H_0) = \alpha .$$

Izrēķinām statistikas eksperimentā iegūto vērtību $T^*(x)$.

Secinām:

ja $T^*(x) < \gamma$, tad hipotēzi jāpārbauda tālāk, jo eksperimentā iegūtais rezultāts nav ar to pretrunā,
ja $T^*(x) > \gamma$ izvirzīto nulles hipotēzi noraidām.

Kolmogorova kritērijs **[10],[4]**

Lai noteiktu eksperimentālo datu saskaņotības pakāpi ar hipotēzi, kas apgalvo, ka nepārtraukta gadījuma lieluma ξ sadalījuma likums ir $F(x)$, var lietot Kolmogorova kritēriju. Izvēlas sekojošu statistiku

$$\sqrt{n}D_n = \sqrt{n}\sup_x |F_n(x) - F(x)|,$$

kur $F_n(x)$ ir empīriskā sadalījuma funkcija , $F(x)$ ir teorētiskā sadalījuma funkcija. A. Kolmogorovs ir pierādījis, ka

$$P(\sqrt{n}D_n < x) \underset{n\to\infty}{\to} K(x),$$

kur

$$K(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2} .$$

Kritiskā apgabala konstruēšanai var izmantot Kolmogorova sadalījuma tabulas, kas saista *x* un *K(x)*
Ja gadījuma lieluma ξ sadalījuma funkcija ir $F_1(x) \neq F(x)$,
t.i., pareiza alternatīvā hipotēze, tad pēc Gļivenko teorēmas ar varbūtību vienādu ar 1 ir spēkā

$$\lim_{n\to\infty}\sup_x |F_n(x) - F_1(x)| = 0$$

un, tātad

$$\lim_{n\to\infty} D_n = \lim_{n\to\infty}\sup_x |F_n(x) - F(x)| > 0 .$$

No šejienes ar varbūtību 1

$$\lim_{n\to\infty}\sqrt{n}D_n = \infty.$$

Tas nozīmē, ka pie jebkura nozīmības līmeņa α, izlases apjomam neierobežoti palielinoties, Kolmogorova kritērija jauda tiecas uz vienīnieku.

Kolmogorova kritērija pielietošanas shēma ir šādā:

a)tiek konstruēta empīriskā sadalījuma funkcija $F_n(x)$ un izvirzīta hipotēzi par teorētisko sadalījuma funkciju $F(x)$; no eksperimenta rezultātiem tiek noteikta $D_n\sqrt{n} = x^*$;
b)pēc tabulām tiek atrasts lielums $1 - K(x^*)$, t.i., varbūtība, ka gadījuma cēloņu dēļ $F_n(x)$ un $F(x)$ atšķirība, ja pareiza nulles hipotēze, nebūs mazāka par faktiski novēroto x^*;
c)ja $1 - K(x^*) < \alpha$, tad nulles hipotēzi vajag noraidīt kā nepareizu, pretējā gadījumā to var uzskatīt par savienojumu ar prakses datiem.

Kolmogorova kritērijs no Pirsona kritērija atšķiras ar savu vienkāršību. Taču, lai to lietotu, precīzi jāzina $F(x)$, t.i., jāzina ne tikai funkcijas veidu, bet arī visi tajā ietilpstošie parametri, kas praksē reti ir iespējams.
Parasti no teorētiskiem apsvērumiem zināms ir tikai funkcijas $F(x)$ vispārīgais veids, tajā ieejošos parametrus nosaka pēc statistiskā materiāla. Pirsona kritērijā tas tiek ievērots, samazinot sadalījuma hī-kvadrāta brīvības pakāpju skaitu par nosakāmo parametru skaitu. Lietot Kolmogorova kritēriju, ja parametri nav zināmi, nav ieteicams.

ω^2 *pārbaudes kritērijs*

Pieņemsim, ka hipotēze H_0 apgalvo, ka nepārtraukta gadījuma lieluma ξ sadalījuma funkcija ir *F(x)*. Lietosim statistiku

$$\omega^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \, dF(x),$$

kur

$$F_n(x) = \begin{cases} 0, & x \le x_1, \\ \dfrac{k}{n}, & x_k < x \le x_{k+1}, \quad k = 1, 2, \ldots, n \\ 1, & x > x_n \end{cases}$$

($x_1,x_2,...,x_n$)-izlases vērtības, dotas variāciju rindas veidā.

$$\omega^2 = \int_{-\infty}^{x_1} F^2(x)dF(x) + \sum_{k=1}^{n-1} \int_{x_k}^{x_{k+1}} \left[\frac{k}{n} - F(x)\right]^2 dF(x) +$$

$$+ \int_{x_n}^{\infty} \left[1 - F(x)\right]^2 dF(x) = \frac{F^3(x)}{3} + \sum_{k=1}^{n-1} \frac{\left[F(x_{k+1}) - \frac{k}{n}\right]^3}{3} -$$

$$- \sum_{k=1}^{n-1} \frac{\left[F(x_k) - \frac{k}{n}\right]^3}{3} - \frac{\left[1 - F(x_n)\right]^3}{3} = \frac{1}{12n^2} + \frac{1}{n}\sum_{k=1}^{n} \left[F(x_k) - \frac{2k-1}{2n}\right]^2.$$

Ir zināms, ka statistikas $n\omega^2$ sadalījums pie pietiekoši lieliem *n* tuvs dotam tabulētam robežsadalījumam. Pie $n \to \infty$
ω^2 - pārbaudes kritērija jauda, tāpat kā Pirsona un Kolmogorova kritērija jaudas, tiecas uz vienīnieku.

χ^2 *kā neatkarības pārbaudes kritērijs*

Pieņemsim, ka jāpārbauda $H_0 : F_{(\xi,\eta)}(x,y) = F_\xi(x)F_\eta(y)$,
kas nozīmē, ka gadījuma lielumi ξ un η ir neatkarīgi, pret alternatīvo hipotēzi *H_1*: „gadījuma lielumi ξ un η ir atkarīgi”.
Aplūkosim diskrētu gadījumu:

$\xi \in \{a_1, a_2, ..., a_s\}$; $\eta \in \{b_1, b_2, ..., b_k\}$.

Apkoposim novērojumu rezultātus $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ ***darba tabulā***:

ξ \ η	b_1	b_2 . . .	b_k	Σ
a_1	ν_{11}	ν_{12} . . .	ν_{1k}	$\nu_{1\cdot}$
a_2 . . .	ν_{21} . . .	ν_{22} . . .	ν_{2k}	$\nu_{2\cdot}$
a_s	ν_{s1}	ν_{s2} . . .	ν_{sk}	$\nu_{s\cdot}$
Σ	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$ . . .	$\nu_{\cdot k}$	n

Kad nulles hipotēze ir pareiza, izpildās

$$p_{ij} := P(\xi = a_i, \eta = b_j) = P(\xi = a_i)P(\eta = b_j) =: p_{i\cdot}p_{\cdot j},$$
$$i = 1,2,...s, \quad j = 1,2,...,k.$$

Mūsu apzīmējumos, tas nozīmē, ka

$$p_{ij} = p_{i\cdot} \times p_{\cdot j}, \ i = 1,2,...s, \ j = 1,2,...,k, \ \sum_{i=1}^{s} p_{i\cdot} = 1, \ \sum_{j=1}^{k} p_{\cdot j} = 1$$

Lai aprēķinātu $p_{ij}, i = 1,...,s, \ j = 1,...,k$, jāizmanto s-1 novērtējumus $\tilde{p}_{i\cdot} = \frac{\nu_{i\cdot}}{n}$ un k-1 novērtējumus $\tilde{p}_{j\cdot} = \frac{\nu_{\cdot j}}{n}$. Tātad, χ^2 pārbaudes kritērija statistiku var pierakstīt

$$\sum_{i=1}^{s}\sum_{j=1}^{k} \frac{\left(\nu_{ij} - n\tilde{p}_{ij}\right)^2}{n\tilde{p}_{ij}} = \sum_{i=1}^{s}\sum_{j=1}^{k} \frac{\left(\nu_{ij} - n\frac{\nu_{i\cdot}}{n}\cdot\frac{\nu_{\cdot j}}{n}\right)^2}{\frac{\nu_{i\cdot}\cdot\nu_{\cdot j}}{n}} =$$

$$= n\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\nu_{ij}\right)^2}{\nu_{i\cdot}\cdot\nu_{\cdot j}} - 2\sum_{i=1}^{s}\sum_{j=1}^{k}\nu_{i\cdot}\cdot\nu_{\cdot j} + \sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\nu_{i\cdot}\cdot\nu_{\cdot j}}{n} =$$

$$= n\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\nu_{ij}\right)^2}{\nu_{i\cdot}\cdot\nu_{\cdot j}} - n = n\left(\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\nu_{ij}\right)^2}{\nu_{i\cdot}\cdot\nu_{\cdot j}} - 1\right).$$

Saskaņā ar hī- kvadrāta kritēriju, brīvības pakāpju skaitu šajā gadījumā nosaka, atskaitot no kopējā klašu skaita savstarpēji neatkarīgo lineāro nosacījumu skaitu, kas saista aplūkojamos datus, t.i., $sk-(s-1)-(k-1)=(s-1)(k-1)$.

Tātad, pie pietiekoši lieliem n

$$\chi^2 = n\left(\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\nu_{ij}\right)^2}{\nu_{i\cdot}\cdot\nu_{\cdot j}} - 1\right) \sim \chi^2_{(s-1)(n-1)}\,.$$

Atliek, pēc uzdotā α,

- atrast atbilstošo kritisko vērtību,
- aprēķināt empīrisko hī – kvadrātu un
- izdarīt vajadzīgie secinājumi.

Piezīme.

Gadījumā, kad $(\xi,\ \eta)$ – nepārtraukti sadalīts vektors (jeb viena no komponentēm ir nepārtraukti sadalīts gadījuma lielums), jākonstruē statistiskā rinda, jāizvēlas intervālu pārstāvjus un eksperimenta rezultātus jāpkopo darba tabulā.

Divu izlašu līdzības kritērijs (*Smirnova -Kolmogorova kritērijs)* **[10]**

Smirnova -Kolmogorova kritēriju lieto, lai pārbaudītu hipotēzi H_0, kas apgalvo, ka neatkarīgas izlases

$$x = (x_1, x_2, , , x_n) \text{ un } y = (y_1, y_2, , , y_m)$$

ir no ģenerālām kopām, kuru sadalījuma likumi ir vienādi. Tā pamatā ir statistika

$$T(x, y) = D_{nm}\sqrt{\frac{nm}{n+m}},$$

kur $D_{nm} = \sup_x |F_n(x) - F_m(x)|$, $F_n(x)$, $F_m(x)$ -izlašu empīriskās sadalījuma funkcijas. Pie $n \to \infty$ statistikas *T(x,y)* sadalījuma funkcija tiecas uz, tā saucamo, Kolmogorova sadalījuma funkciju *K(x)*.

χ^2 kā izlašu līdzības pārbaudes kritērijs **[1],[2]**

Pieņemsim, ka dotas ir k neatkarīgas izlases

$x_1 = (x_{11}, x_{12}, ..., x_{1n_1}), x_2 = (x_{21}, x_{22}, ..., x_{2n_2}), ..., x_k = (x_{k1}, x_{k2}, ..., x_{kn_k})$.

Lai ar χ^2 kritēriju pārbaudītu hipotēzi, ka visas izlases ir no ģenerālām kopām, kuru sadalījuma likumi ir vienādi, vispirms jākonstruē darba tabula:

- visu izlašu vērtību kopu sadala *s* savstarpēji nešķeļošos nogriežņos: $x_0 - x_1, x_1 - x_2, ..., x_{s-1} - x_s$, $x_0 = \min_{i,j} x_{ij}$, $x_s = \max_{i,j} x_{ij}$,
- katrai izlasei atsevišķi konstruē statistisko rindu, apzīmējot ar ν_{ij} - *i*-tā izlase novērojumu vērtību skaits *j*-tajā intervālā, t.i.,

$$\sum_{l=1}^{s} \nu_{il} = n_i =: \nu_{i\cdot}, i = 1,2,...,k; \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \nu_{i\cdot} = n;$$

$$\sum_{l=1}^{k} \nu_{lj} =: \nu_{\cdot j}, \quad j = 1,2,...,s, \quad \sum_{i=1}^{s} \nu_{\cdot j} = n;$$

x_j-x_{j+1} / ξ_i	x_o-x_1	x_1-x_2 . . .	x_{s-1}-x_s	Σ
ξ_1	ν_{11}	ν_{12} . . .	ν_{1k}	$\nu_{1\cdot}=n_1$
ξ_2 . . .	ν_{21} . . .	ν_{22} . . .	ν_{2k}	$\nu_{2\cdot}=n_2$
ξ_k	ν_{k1}	ν_{k2} . . .	ν_{ks}	$\nu_{k\cdot}=n_k$
Σ	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$ . . .	$\nu_{\cdot k}$	n

- definē statistiku

$$\sum_{l=1}^{s} \frac{\left(\nu_{1l} - n_1 \frac{\nu_{\cdot l}}{n}\right)^2}{n_1 \frac{\nu_{\cdot l}}{n}} + \sum_{l=1}^{s} \frac{\left(\nu_{2l} - n_2 \frac{\nu_{\cdot l}}{n}\right)^2}{n_2 \frac{\nu_{\cdot l}}{n}} + ... + \sum_{l=1}^{s} \frac{\left(\nu_{kl} - n_k \frac{\nu_{\cdot l}}{n}\right)^2}{n_k \frac{\nu_{\cdot l}}{n}} =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{s} \frac{\left(\nu_{ij} - n_i \cdot \frac{\nu_{\cdot j}}{n}\right)^2}{n \frac{\nu_{\cdot j}}{n}} \sim \chi_{k(s-1)-l},$$

kur, ņemot vērā, ka $\sum_{j=1}^{s} \nu_{\cdot j} = n$ un to, ka visiem i spēkā

$\tilde{p}_{ij} = \frac{\nu_{\cdot j}}{n}, j = 1,2,...s$, novērtējumu skaits l ir vienāds ar $s-1$.

Acīm redzami, ka statistika, ar kuru palīdzību pārbaudām nulles hipotēzi, pārrakstāma formā

$$\chi^2 = n\left(\sum_{i=1}^{k}\sum_{j=1}^{s}\frac{\left(\nu_{ij}\right)^2}{\nu_{i\cdot}\cdot\nu_{\cdot j}} - 1\right)$$

un sadalīta pēc $\chi^2_{(s-1)(n-1)}$ likuma.

Divu izlašu līdzības pārbaudes kritērijs (Vilkoksona kritērijs) **[10]**

Pieņemsim, ka dotas divas neatkarīgas izlases $x_1, x_2, ..., x_n$ un $y_1, y_2, ..., y_m$. Pieņemsim, ka

H_0: „izlašu vērtības ir no ģenerālām kopām, kuru sadalījuma likumi ir vienādi"

H_1: „izlašu vērtības ir no ģenerālām kopām, kuru sadalījuma likumi ir dažādi"

Jānoskaidro, vai pie uzdotā nozīmības līmeņa α nulles hipotēzi jānorāda, uzskatot H_1 par pareizu, vai tomēr nulles hipotēzi H_0 noraidīt nevar.

Abu izlašu vērtību kopu uzrakstīsim augošā secībā. Piemēram,

$$y_1' \le x_1' \le x_2' \le y_2' \le y_3' \le y_4' \le x_3' \le \cdots \le x_n'.$$

Ja, piemēram, y' atrodas pirms x' teiksim, ka pāris (x', y') veido vienu „inversiju", jeb x' uzdod vienu „inversiju".

Tā, mūsu piemērā: x_1' veido vienu inversiju. Tā kā pirms x_2' atrodas tikai y_1', tad attiecīgi x_2' uzdod arī vienu inversiju. x_3' veido četras inversijas, jo pirms x_3' atrodas y_1', y_2', y_3', y_4', un atbilstoši x_n' uzdod m inversijas.

Spēkā apgalvojums:

ja H_0 ir pareiza, n un m pietiekoši lieli, tad inversiju skaits u ir normāli sadalīts gadījuma lielums ar parametriem

$$Eu = \frac{mn}{2},\ Du = \frac{mn}{12}(m+n+1),\ (n>10,\ m>10).$$

Tad, ja nulles hipotēze ir pareiza, statistika

$$\frac{u-\frac{mn}{2}}{\sqrt{\frac{mn}{12}(m+n+1)}} \sim N(0,1).$$

Līdz ar to konstruējam statistikas kritisko apgabalu

$$P\left(\left|\frac{u-\frac{mn}{2}}{\sqrt{\frac{mn}{12}(m+n+1)}}\right| > \varepsilon \mid H_0\right) = \alpha \;\Rightarrow\; S = (-\infty,-\varepsilon)\cup(\varepsilon,\infty).$$

Tātad secinām:
ja statistikas iegūtā vērtība pēc absolūtas vērtības lielāka par ε, tad izvirzīto nulles hipotēzi noraidām; pretējā gadījumā nulles hipotēzi noraidīt nevar.

Novērojumu rupjo kļūdu kritērijs.

Pieņemsim, ka tika izdarīti *n* mērījumi un rezultātā iegūtas gadījuma lieluma ξ, ar sadalījuma funkciju $F_{\xi}(x)$, *n* vērtības $x_1, x_2, ...x_n$.
Pieņemsim, ka starp iegūtiem skaitļiem ir tāds, kas būtiski atšķīrās no citiem. Tas var rasties dažādu iemeslu dēļ : rupjas mērīšanās kļūdas, operatora kļūdas rezultātā utt.

Konstruēsim kritēriju, kas ļauj atmest no izlases šī vērtības.
Pārbaudām, piemēram, hipotēzi:
$H_0 : x_{\max} \in \mathrm{X}$, X - ξ ģenerālā kopa,
$H_1 : x_{\max} \notin \mathrm{X}$
Aplūkosim statistiku $T(x) = \max\limits_{1\le i\le n} x_i$.
Pie dotā α, ņemot vērā
$F_{x_{\max}}(x) = P(x_{\max} < x) = P(x_1 < x, x_2 < x, ..., x_n < x) =$

$$= \prod_{i=1}^{n} P(x_i < x) = \left(F_\xi(x)\right)^n ,$$

un to, ka kritiskais apgabals ir lielu pozitīvo noviržu apgabals $S = (\varepsilon, \infty)$, atrodam ε no vienādojuma

$$1 - \left(F_\xi(\varepsilon)\right)^n = \alpha .$$

Ja gadījuma lieluma $x_{\max}$ empīriskā vērtība ir mazāka par kritisko vērtību ε, nulles hipotēzi noraidīt nevar.
Pretējā gadījumā nulles hipotēzi noraidām, tas nozīmē, ka secinām: gadījuma lieluma $x_{\max}$ empīriskā vērtība nepieder gadījuma lieluma ξ ģenerālai kopai.

Piezīme.
Jāatzīmē, ka nulles hipotēzes $H_0 : \theta = \theta_0$ pārbaudei pie alternatīvas $H_1 : \theta \neq \theta_0$ katram nozīmības līmenim α var atrast kritēriju $T(x)$ tādu, ka kritērija jauda būs maksimālā (***Neimana-Pirsona lemma***).
Nepārtraukta sadalījumu gadījumā kritērija forma ir sekojošā:

$$\int_X T(x) p(x, \theta_0) dx = \alpha,$$

$$T(x) = \begin{cases} 1, & p_\xi(x, \theta_1) < \gamma\, p_\xi(x, \theta_0), \\ 0, & p_\xi(x, \theta_1) < \gamma\, p_\xi(x, \theta_0) \end{cases}$$

pie kaut kādas γ vērtības.

Kritērijs hipotēzes pārbaudei, ja hipotēze apgalvo, kā dotā izlašu kopa ir no vienās un tās pašas ģenerālās kopas **[6],[1]**

Pieņemsim, ka dotas ir *r* izlases. Apzīmēsim ar n_i, i-tās izlases apjomu, $i = 1, 2, ..., r.$ Apzīmēsim $n = \sum_{i=1}^{r} n_i.$ Acīmredzot n- doto izlašu kopīgais izlases vērtību skaits. Pieņemsim, ka x_{ij} – i-tās izlases *j*-tā

vērtība, $\overline{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, $\overline{x} = \frac{1}{n}\sum_{i=}^{r}\sum_{j=1}^{n_i} x_{ij}$. Zināms, ka *i*-tā izlase ir no ģenerālās kopas, kuras sadalījums $N(a_i,\sigma^2)$ tāds, ka dispersija σ^2 nemainās pa izlasēm, bet tās vērtība nav zināma, tapāt nav zināmas matemātiskās cerības a_i. Nulles hipotēze apgalvo, ka visi a_i ir vienādi. Tāda veida uzdevumi sastopami dabaszinātnēs, kad jānoskaidro kāda faktora ietekme uz pētāmo parādību. To sauc par ***faktoranalīzi jeb dispersiju analīzi.***

Apskatīsim statistiku

$$S_1 := \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 ,$$

kas raksturo izlases vērtību atšķirības atsevišķu izlašu iekšienē, un statistiku

$$S_2 := \sum_{i=1}^{r} n_i(\overline{x}_i - \overline{x})^2 ,$$

kas raksturo atšķirību starp izlasēm.

Viegli pierādīt, ka tad, ja nulles hipotēze ir pareiza, $\frac{S_1}{\sigma^2}$ un $\frac{S_2}{\sigma^2}$ ir neatkarīgi gadījuma lielumi, sadalīti pēc χ^2_ν sadalījuma likuma attiecīgi ar brīvības pakāpēm $n-r$ un $r-1$. Līdz ar to kritērija konstruēšanai ieteicams lietot statistiku $\frac{S_1}{S_2}$, kurai ir F(*r-1,n-r*) (Snedekora) sadalījums.

Korelācijas teorijas pamati

1.§. Regresijas līknes. Nosacītās dispersijas [2],[3]

Pieņensim, ka divi gadījuma lielumi ir stohastiski saistīti, t.i., viens no tiem uz otra izmaiņām reaģē ar sava sadalījuma likuma iznaiņu. Stohastiskā saite parasti eksistē, ja ir kopēji gadījuma faktori, kas ietekmē abus gadījuma lielumus, darbojoties kopā ar citiem gadījuma faktoriem, kas ir dažādi katram gadījuma lielumam. Praksē bieži ir jāzina apskatāmo gadījumu lielumu sadalījuma likums, kā arī varbūtības, ar kādām iespējamas tās vai citas vērtību kombinācijas. Pieņemsim, ka X un Y ir gadījuma lielumi, t.i., apskatīsim novērojumus kā divdimensiju gadījuma vektorus (x_i, y_i). Nepieciešams precīzi definēt kopu, no kuras iegūti novērojumi.

Par Y ***pa*** X ***regresijas līkni*** sauksim gadījuma lieluma Y nosacīto matemātisko cerību pie atbilstošajām dotajām $x \in X$ vērtībām. Apzīmēsim to $\mathbb{E}(Y|x) = \bar{y}(x)$.
Analoģiski, X pa Y regresija — $\mathbb{E}(X|y) = \bar{x}(y)$. Ja atkarība ir funkcionāla, tad mainīgais Y pie dotās $x \in X$ vērtības var pieņemt tikai vienu noteiktu vērtību $\bar{y}(x)$, t.i., izkliede ap $\bar{y}(x)$ ir vienāda ar nulli. Šo gadījumu var uzskatīt par stohastiskās atkarības robežgagījumu. Parasti pie $x \in X$ novērojama vairāk vai mazāk izteikta Y izkliede ap centru $\bar{y}(x)$. Par šīs izkliedes meru var būt Y ***nosacītā dispersija pie dotā*** x

$$\sigma^2_{Y|x} = \mathbb{D}(Y|x) = \mathbb{E}(Y - \bar{y}(x)|\ x)^2 = \int\limits_{-\infty}^{\infty} [y - \bar{y}(x)]^2 dF_{Y|x}(y),$$

kur $F_{Y|x}(y)$ — Y nosacītā sadalījuma funkcija pie dotā x. Tātad, kopā ar regresijas līkni $\bar{y}(x)$ vēl ir jāapskata līnija $\sigma^2_{Y|x}$ un divi analoģiski raksturlielumi $\bar{x}(y)$ un $\sigma^2_{X|y}$. Ja regresijas līkni $\bar{y}(x)$ izmanto, lai prognozētu lielumu Y pēc novērotās X vērtības x, tad $\sigma^2_{Y|x}$ var uzskatīt par lieluma prognozes vidējo kvadrātisko kļūdu. Lielums $\sigma^2_{Y|x}$ un līdz ar to prognozes precizitāte ir atkarīgi no lieluma x. Ja vēlamies iegūt priekštatu par Y pēc x prognozes precizitāti visā X izmaiņu diapazonā, tad, acīmredzot, jāņem nosacīto

dispersiju vidējais normētais lielums

$$\sigma^2_{Y|x} = \mathbb{E}\sigma^2_{Y|x} = \int\limits_{-\infty}^{\infty} \sigma^2_{Y|x} dF_X(x) = \mathbb{E}(Y - \bar{y}(X))^2.$$

Korelācijas teorijā visbiežāk sastopami divi sekojoši uzdevumi:

1) noteikt regresijas funkcijas veidu, t.i, atkarības formu starp gadījuma lielumiem, līdz ar to kļūst iespējama viena gadījuma lieluma vērtību prognoze pie uzdotām otra gadījuma lieluma vērtībām;

2) novērtēt gadījuma lieluma atkarības pakāpi un prognozes kļūdu.

Ievērosim, ka no visām funkcijām $u(x)$ lieluma $\mathbb{E}(Y - u(x))^2$ minimumu dod funkcija $\bar{y}(x)$, t.i.,Y pa X regresijas līkne. Tas redzams no tā, ka

$$\mathbb{E}(\xi - c)^2 = \mathbb{E}(\xi - \mathbb{E}\xi + \mathbb{E}\xi - c)^2 = D\xi + (c - \mathbb{E}\xi)^2,$$

no kurienes

$$\min_c \mathbb{E}(\xi - c)^2 = \mathbb{E}(\xi - \mathbb{E}\xi)^2.$$

•Tātad, $\bar{y}(x)$ — funkcija, kas minimizē lieluma Y pa X prognozes vidējo kvadrātisko kļūdu. Analoģiskas īpašības ir arī funkcijai $\bar{x}(y)$, kas raksturo X pa Y regresiju.

2.§. Lineārā regresija[4]

Pieņemsim, ka gadījuma lieluma X vērtības x nosaka eksperimentators, bet mainīgais Y rāda novērojuma rezultātus eksperimentā, atbilstošus mainīgā x vērtībām. Pieņemsim, ka tā saucamais neatkarīgais mainīgais x tiek dots bez kļūdām. Tātad, apskatīsim visvienkāršāko regresijas analīzes gadījumu, pieņenot, ka mūsu rīcībā ir sekojoši novērojumi $(x_1, y_1), \ldots, (x_k, y_k)$. Turpmāk uzskatīsim, ka visi novērojumi $Y_1, \ldots, Y_k$ neatkarīgi un normāli sadalīti ar parametriem a_i un σ^2, bez tam regresijas līnija $\bar{y}(x)$ ir šāda veida x lineāra funkcija

$$\bar{y}(x) = \alpha + \beta(x - \bar{x}), \tag{1}$$

kur

$$\bar{x} = \frac{\sum\limits_{i=1}^{k} x_i}{k},$$

β -***regresijas koeficients.*** Ja precīzās regresijas līnijas ir taisnes, tad korelāciju sauc par lineāru. Mūsu uzdevums - pēc novērojumiem atrast parametru α, β, σ^2 atbilstošos novērtējumus a, b, S^2. Tad vienādojums

$$\tilde{y}(x) = a + b(x - \bar{x}) \tag{2}$$

definē taisni, kas ir regresijas taisnes novērtējums.

Standartmetode, ko lieto regresiju analīzē novērtējumu iegūšanai, ir mazāko kvadrātu metode. Šīs metodes pamatā ir tādu novērtējumu a un b izvēle, kuri minimizētu novēroto vērtību Y_i un prognozēto vērtību $\tilde{y}(x_i)$ noviržu kvadrātu summu. Tātad, pec mazāko kvadrātu metodes a un b novērtējumu tiek atrasti, minimizējot noviržu kvadrātu summu.

$$Q = \sum_{i=1}^{k}(Y_i - \tilde{y}(x_i))^2 = \sum_{i=1}^{k}(Y_i - a - b(x_i - \bar{x}))^2.$$

Lai atrastu a un b, kas minimizē R, diferencēsim pēdējo izteiksmi pec a un b un pielīdzināsim atvasinājumus nullei:

$$\frac{\partial Q}{\partial a} = -2\sum_{i=1}^{k}[Y_i - a - b(x_i - \bar{x})] = 0,$$

$$\frac{\partial Q}{\partial a} = -2\sum_{i=1}^{k}[Y_i - a - b(x_i - \bar{x})](x_i - \bar{x}) = 0. \tag{3}$$

Pārrakstīsim iegūtos vienādojumus

$$ka + b\sum_{i=1}^{k}(x_i - \bar{x}) = \sum_{i=1}^{k} y_i, \tag{4}$$

$$a\sum_{i=1}^{k}(x_i - \bar{x}) + b\sum_{i=1}^{k}(x_i - \bar{x})^2 = \sum_{i=1}^{k}(x_i - \bar{x})Y_i. \tag{5}$$

Tā kā $\sum\limits_{i=1}^{k}(x_i - \bar{x}) = 0$, tad α un β novērtējumu ir sekojoši

$$a = \frac{b\sum\limits_{i=1}^{k} Y_i}{k} = \bar{Y}$$

$$b = \frac{\sum\limits_{i=1}^{k}(x_i - \bar{x})Y_i}{\sum\limits_{i=1}^{k}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{k}(x_i - \bar{x})Y_i - \bar{Y}\sum\limits_{i=1}^{k}(x_i - \bar{x})}{\sum\limits_{i=1}^{k}(x_i - \bar{x})^2} =$$

$$= \frac{\sum_{i=1}^{k}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{k}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{s_x^2} = r_{x,y}\frac{s_x}{s_y}. \qquad (6)$$

Parādīsim, ka b ir parametra β nenovirzīts novērtējums. Pie $x_i \in X$ sagaidāmā $\mathbb{E}Y_i$ vērtība $\alpha + \beta(x_i - \bar{x})$, tāpēc

$$\mathbb{E}b = \frac{\sum_{i=1}^{k}(x_i - \bar{x})\mathbb{E}Y_i}{\sum_{i=1}^{k}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})[\alpha + \beta(x_i - \bar{x})]}{\sum_{i=1}^{k}(x_i - \bar{x})^2} = \beta.$$

Acīmredzot, $\mathbb{E}a = \alpha$. Viegli var izrēķināt arī novērtējumu a un a dispersijas

$$\mathbb{D}a = \mathbb{D}\left(\frac{\sum_{i=1}^{k} Y_i}{k}\right) = \frac{\sigma^2}{k},$$

$$\mathbb{D}b = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2\mathbb{D}(Y_i)}{\left[\sum_{i=1}^{k}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{k}(x_i - \bar{x})^2}. \qquad (7)$$

Parametru novērtēšanas metode neizmanto nosacījumu par normālo sadalījumu, tas kļūst nepieciešams nosakot ticamības intervālu un veicot hipotēžu, kas izvirzītas par parametriem α un β, pārbaudi. Var parādīt, ka mazāko kvadrātu metodes dotie novērtējumi ir ar vismazāko dispersiju visu lineāro nenovirzīto novērtējumu klasē.

Statistikas saistītas ar lineāro regresiju[12]

Pieņemsim, ka mūsu rīcība ir sekojoši novērojumi $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$, kur, tā saucamais, neatkarīgais mainīgais x tiek dots bez kļūdām, visi novērojumi $Y_i,\ i = 1, ..., n$ ir neatkarīgi un normāli sadalīti gadījuma lielumi, ar parametriem $\mathbb{E}Y_i = \overline{y}(x_i)$ un $DY_i = \sigma^2$, regresijas līnija ir lineārā funkcija

$$\overline{y}(x) = \alpha + \beta(x - \overline{x}).$$

Starpību starp novērojumiem Y_i un vērtībām $\overline{y}(x_i)$ var izteikt kā summas

$$Y_i - \overline{y}(x_i) = (Y_i - \widetilde{y}(x_i)) + (\widetilde{y}(x_i) - \overline{y}(x_i)) =$$

$$= (Y_i - \widetilde{y}(x_i)) + [a + b(x_i - \overline{x}) - \alpha - \beta(x_i - \overline{x})] =$$

$$= (Y_i - \widetilde{y}(x_i)) + (a - \alpha) + (b - \beta)(x_i - \overline{x}). \tag{8}$$

Pēdējās sakarības abas puses kāpināsim kvadrātā un summēsim pa i. Ievērojot, ka

$$\begin{cases} \sum\limits_{i=1}^{n}(Y_i - \widetilde{y}(x_i)) = 0 \\ \sum\limits_{i=1}^{n}(Y_i - \widetilde{y}(x_i))(x_i - \overline{x}) = 0, \end{cases} \tag{9}$$

iegūsim

$$\sum_{i=1}^{n}(Y_i - \overline{y}(x_i))^2 = \sum_{i=1}^{n}(Y_i - \widetilde{y}(x_i))^2 + (b-\beta)^2\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(a-\alpha)^2. \tag{10}$$

Kvadrātu summa $\sum\limits_{i=1}^{n}(Y_i - \overline{y}(x_i))^2$ sadalīta kā $\sigma^2\chi^2_n$.
Šī summa sastāv no trim kvadrātiskajām formām, no kurām pēdējās divas ir attiecīgi atkarīgas tikai no b un a. Tātad,
- otrajai un trešajai formai brīvības pakāpju skaits ir vienāds ar vieninieku, kas ir atbilstošo formu rangi;
- pirmais loceklis satur n starpības $Y_i - \widetilde{y}(x_i),\ i = 1, ..., n$ ar diviem nosacījumiem (9), tātad, tam ir $n - 2$ brīvības pakāpes.

Tad, pēc Kokrena teorēmas [12] katrs loceklis labajā pusē ir sadalīts kā $\sigma^2\chi^2_\nu$ ar attiecīgo ν brīvības pakāpju skaitu un šie locekļi savā starpā ir neatkarīgi.
No iepriekš teiktā izriet, ka ir pareizas šādas svarīgas vienādības:

$$\sum_{i=1}^{n}\left(\frac{Y_i - \widetilde{y}(x_i)}{\sigma}\right)^2 \sim \chi^2_{n-2} \quad \searrow$$

$$(b-\beta)^2\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^2 \sim \chi^2_1 \quad \longrightarrow \quad \text{neatkarīgi}$$

$$n\left(\frac{a-\alpha}{\sigma}\right)^2 \sim \chi^2_1 \quad \nearrow$$

No šejienes, izmantojot standarta normālā gadījuma lieluma τ sadalījuma funkcijas īpašības, iegūstam

$$\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2\sim\sigma^2\chi^2_{n-2}\quad\searrow$$

$$(b-\beta)\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sim\sigma\tau\quad\longrightarrow\quad\text{neatkarīgi.}$$

$$\sqrt{n}(a-\alpha)\sim\sigma\tau\quad\nearrow$$

Līdz ar to statistika

$$S^2:=\frac{\sum\limits_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2}{n-2}\sim\sigma^2\chi^2_{n-2}\frac{1}{n-2}$$

nav atkarīga no gadījuma lielumiem a un b, un $\mathbb{E}S^2=\sigma^2$. Tādējādi, varam secināt

▶ statistika S^2 ir parametra σ^2 nenovirzīts novērtējums.

Savukārt, no tā, ka gadījuma lielums $n(a-\alpha)^2$ sadalīts kā $\sigma^2\chi^2_1$, seko, ka $\mathbb{E}[n(a-\alpha)^2]=\sigma^2$, tātad $Da=E(a-\alpha)^2=\sigma^2/n$.
Aizvietojot formulā parametru σ^2 ar tā novērtējumu S^2 iegūsim statistiku

$$s.k.(a):=\sqrt{S^2_a}:=\sqrt{\frac{S^2}{n}}=\sqrt{\frac{\sum\limits_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2}{n(n-2)}},$$

kas tiek saukta par ***statistikas*** a ***standartkļūdu***.
Analoģiski iegūstam, ka

$$Db=\mathbb{E}(b-\beta)^2=\mathbb{E}\left(\frac{\sigma^2\chi^2_1}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right)=\frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}.$$

No tā, ***statistikas*** b ***standartkļūda*** būs

$$s.k.(b):=\sqrt{S^2_b}:=\sqrt{\frac{S^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}=\sqrt{\frac{\sum\limits_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2}{(n-2)\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}.$$

Tā kā

$$\overline{Y-\widetilde{y}(x)}=\frac{1}{n}\sum_{n=1}^{n}(Y_i-\widetilde{y}(x_i))=\overline{y}-\frac{1}{n}\sum_{n=1}^{n}(a+b(x_i-\overline{x}))=\overline{y}-\overline{y}=0,$$

$s.k.(a)$ un $s.k.(b)$ var pierakstīt formā

$$s.k.(a)=\sqrt{\frac{Var(Y-\widetilde{y}(x))}{n-2}},\ \ s.k.(b)=\sqrt{\frac{Var(Y-\widetilde{y}(x))}{(n-2)Var(x)}},$$

kur

$$Var(Y-\widetilde{y}(x))=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i)-\overline{Y-\widetilde{y}(x)})^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2.\ (11)$$

Atzīmēsim, ka, saskaņā ar iegūtajām formulām, $\mathbb{E}S_a^2=Da$ un $\mathbb{E}S_b^2=Db$, kas nozīmē, ka S_a^2 ir nenovirzīts novērtējums parametram Da, un S_b^2 ir nenovirzīts novērtējums parametram Db. Jāpiezīmē, jo lielāka dispersijas σ^2 vērtība, jo lineāras regresijas koeficientu vidējo kvadrātisko noviržu (regresijas koeficientu standartkļūdas) ir nozīmīgākās. Tagad konstruēsim statistikas, kas ir saistītas ar hipotēžu pārbaudes un lineāras regresijas parametru ticamības intervālu konstruēšanas uzdevumiem.

Statistikas, saistītas ar hipotēžu pārbaudēm un ticamības intervālu konstruēšanām lineārās regresijās

Ievērojot, ka $Y_i, i=1,...,n$ ir neatkarīgi un normāli sadalīti gadījuma lielumi, iegūstam, ka regresijas līnijas novērtējuma $\widetilde{y}(x)$ parametri a u b, kā gadījuma lielumu $Y_i, i=1,...,n$ lineārās kombinācijas, arī ir neatkarīgi un normāli sadalīti gadījuma lielumi. Bez tam, regresijas līknes novērtējums $\widetilde{y}(x)$, katrai $x\in X$ vērtībai, arī ir normāli sadalīts gadījuma lielums ar

$$\mathbb{E}\widetilde{y}(x)=\overline{y}(x),\ D\widetilde{y}(x)=Da+(x-\overline{x})^2Db=\sigma^2\left(\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right).\ (12)$$

Tad, ir spēkā:

$$\frac{a-\alpha}{s.k.(a)}=t_{n-2}.\qquad\qquad (I)$$

Pierādījums. No $a \sim N(\alpha, \sigma^2/n)$ seko, ka $n(a-\alpha)/\sigma \sim N(0,1)$. Ievērojot gadījuma lielumu a un $\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2)/\sigma^2$ neatkarību, konstruēsim statistiku, kas būs sadalīta pēc Stjūdenta likuma ar $n-2$ brīvības pakāpēm :

$$\frac{n(a-\alpha)}{\sqrt{\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2}}\sqrt{n-2} = \frac{a-\alpha}{\sqrt{\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2}}\sqrt{n(n-2)} = \frac{a-\alpha}{s.k.(a)} = t_{n-2}, \blacktriangle$$

$$\frac{b-\beta}{s.k.(b)} = t_{n-2}. \qquad \text{(II)}$$

Pierādījums. $b \sim N(\beta, \sigma^2/\sum_{i=1}^{n}(x_i-\overline{x})^2)$. Tad, ņemot vērā, ka

$$s.k.(b) = \frac{\sqrt{\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i)^2)}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2(n-2)}},$$

un gadījuma lielumu b un $\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2)/\sigma^2$ neatkarību, iegūstam

$$\frac{(b-\beta)\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}{\sqrt{\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2}}\sqrt{n-2} = \frac{b-\beta}{s.k.(b)} = t_{n-2}, \blacktriangle$$

$$\frac{\widetilde{y}(x)-\overline{y}(x)}{s.k.(\widetilde{y}(x))} = t_{n-2}. \qquad \text{(III)}$$

Pierādījums. Katram $x \in X$ gadījuma lielumi

$$\widetilde{y}(x) \sim N\left(\overline{y}(x),\ \frac{\sigma^2}{n} + \frac{\sigma^2(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right) \text{ un } \frac{\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2)}{\sigma^2} \sim \chi^2_{n-2}$$

ir neatkarīgi. Līdz ar to

$$\frac{\widetilde{y}(x)-\overline{y}(x)}{\sigma\sqrt{\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \sim N(0,1),$$

$$\frac{\widetilde{y}(x)-\overline{y}(x)}{\sqrt{\sum\limits_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2}\sqrt{\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}}\sqrt{n-2}=t_{n-2}.$$

Ievērojot, ka, lai iegūtu $[s.k.(\widetilde{y}(x))]^2$, $D\widetilde{y}(x)$ izteiksmē parametra σ^2 vietā jāievieto statistika S^2, t.i.,

$$[s.k.(\widetilde{y}(x))]^2=\left(\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right)\sum_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2\frac{1}{n-2},$$

pēdējo formulu parrakstām formā III,▲

$$\frac{(b-\beta)^2}{[s.k.(b)]^2}=\frac{(b-\beta)^2\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}{\sum\limits_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2}(n-2)=F(1,n-2). \qquad \text{(IV)}$$

Pierādījums. Tā kā gadījuma lielumi

$$\sum_{i=1}^{n}\left(\frac{Y_i-\widetilde{y}(x_i)}{\sigma}\right)^2\sim\chi^2_{n-2},\ \ (b-\beta)^2\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma}\right)^2\sim\chi^2_1$$

ir neatkarīgi un, ņemot vērā Snedekora sadalījuma definīciju, iegūstam apgalvojumu.▲

Katram $x\in X$, kas $x\notin\{x_1,x_2,...,x_n\}$, var ar uzdotu drošību prognozēt gadījuma lielumam Y_x vērtību intervālu, izmantojot gadījuma lielumu

$$\frac{Y_x-\widetilde{y}(x)}{\sqrt{\sum\limits_{i=1}^{n}(Y_i-\widetilde{y}(x_i))^2)}\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}}\sqrt{n-2}=t_{n-2}. \qquad \text{(V)}$$

Pierādījums. Gadījuma lielums Y_x nav atkarīgs no

$$\widetilde{y}(x)=T(x_1,x_2,...,x_n,Y_1,Y_2,...,Y_n),\ Y_x\sim N \text{ ar } \mathbb{E}Y_x=\overline{y}(x),\ DY_x=\sigma^2.$$

Tad

$$\mathbb{E}(Y_x - \widetilde{y}(x)) = \overline{y}(x) - \overline{y}(x) = 0,$$

$$D(Y_x - \widetilde{y}(x)) = DY_x + D\widetilde{y}(x) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} \right).$$

Līdz ar to gadījuma lielums

$$\frac{Y_x - \widetilde{y}(x)}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}} \sim N(0,1)$$

un nav atkarīgs no gadījuma lieluma

$$\sum_{i=1}^{n} \left(\frac{Y_i - \widetilde{y}(x_i)}{\sigma} \right)^2 \sim \chi^2_{n-2}.$$

Ņēmot vērā Stjudenta sadalījuma definīciju, esam pierādījuši apgalvojumu.▲

1.Piezīme. Regresijas taisni $\overline{y}(x) = \alpha + \beta(x - \overline{x})$ var pierakstīt šādi $\overline{y}(x) = \alpha^{'} + \beta x$, kur $\alpha^{'} = \alpha - \beta\overline{x}$, $\alpha^{'}$ - regresijas līknes brīvais loceklis, β - regresijas koeficients, jeb šīs taisnes virziena koeficients, t.i., tā leņķa tangenss, ko regresijas taisne veido ar x-asi.
Vienādojuma brīvais loceklis raksturo ordinātu ass nogriežņa lielumu no koordinātu sistēmas sākuma līdz punktiem, kurā regresijas taisne krusto ordinātu asi. Ievērojot, ka statistikas a un b ir sadalītas pēc normālā likuma, redzams, ka statistika

$$c := a - b\overline{x}$$

arī normāli sadalīta ar parametriem

$$\mathbb{E}c = \alpha - \beta\overline{x} = \alpha^{'}, \quad Dc = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{nVarx} \right),$$

$$s.k.^2(c) = \left(\frac{1}{n} + \frac{\overline{x}^2}{nVarx} \right) \frac{\sum\limits_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i) \right)^2}{n-2},$$

un ir parametra $\alpha^{'}$ nenovirzīts un būtisks novērtējums. Bez tam, ir zināms, ka

$$\frac{c - \mathbb{E}c}{\sqrt{Dc}} \sim N(0,1), \quad \sum_{i=1}^{n} \left(\frac{Y_i - \widetilde{y}(x_i)}{\sigma} \right)^2 \sim \chi^2_{n-2}$$

un ir neatkarīgi. Līdz ar to

$$\frac{c - \mathbb{E}c}{s.k.(c)} = t_{n-2.}$$

2.Piezīme. Regresijas vienādojuma brīvo locekli dažkārt interpretē kā atkarīgā mainīgā vidējo lielumu ar nosacījumu, ka neatkarīgā mainīgā vērtība ir nulle. Tāda interpretācija ir pieļaujama tikai atsevišķos gadījumos, jo regresijas vienādojums ir spēkā tikai noteiktā eksistences jeb definīcijas apgabalā. Eksistences apgabalu parasti uzdod ar nevienādībām, kas parāda neatkarīgā mainīgā lieluma vērtību, kuru robežās vienādojums pareizi atspoguļo reālās sakarības. Šīs robežas visbiežāk sakrīt ar vismazāko un vislielāko neatkarīgā mainīgā vērtību, kāds sastopams sākotnējos datos. Pēc analīzes apsvērumiem eksistences apgabals varētu būt plašāks. Taču, paplašinot vienādojuma eksistences apgabalu ārpus faktisko datu variācijas apgabala, var rasties kļūdas. Tā, piemēram, sakarību forma var būt lineāra izpētītajā apgabalā, bet ārpus tā var būt nelineāra. Tādēļ, regresijas vienādojuma brīvo locekli var interpretēt kā atkarīgā mainīgā vidējo lielumu, kur neatkarīgā mainīgā vērtības ir nulle, tikai tad, ja neatkarīgā mainīgā nulles vērtība ietilpst vienādojuma eksistences apgabalā.

3.§. Tuvinātās regresijas taisnes [2]

Pieņemsim, ka X un Y ir gadījuma lielumi ar dotu sadalījuma likumu. Apskatīsim, kā konstruēt tā saucamas tuvinātās regresijas "taisnes", kas rada priekštatu par regresijas līnijas formu tajos gadījumos, kad tās pietiekošu tuvinājumu iespējams aprakstīt ar taisnēm. Tātad, uzdevums ir noteikt taisni $y = ax + b$ plaknē (X, Y), kas minimizētu vidējo kvadrātisko kļūdu Y pa X, t.i., atrast tādus parametrus a un b, ka $\Delta(a, b) = \mathbb{E}(y - ax - b)^2 = \min$.

Tad pēc gadījuma lieluma X vērtības varam prognozēt gadījuma lieluma Y vērtību, izmantojot taisni $\tilde{y} = ax + b$, kur aizvietojam y ar $\tilde{y}$. Šajā gadījumā prognozes kļūda ir $y - \tilde{y}$. Turklāt parametru a un b noteikšanai izmantojam sakarības

$$\frac{\partial \Delta(a, b)}{\partial a} = 0, \quad \frac{\partial \Delta(a, b)}{\partial b} = 0. \tag{13}$$

No šīm sakarībām iegūstam

$$\begin{cases} \mathbb{E}yx - a\mathbb{E}x^2 - b\mathbb{E}x = 0 \\ \mathbb{E}y - a\mathbb{E}x - b = 0. \end{cases}$$

Tātad

$$b = \mathbb{E}y - a\mathbb{E}x,$$

$$a = \frac{\mathbb{E}yx - \mathbb{E}y\mathbb{E}x}{\mathbb{E}x^2 - (\mathbb{E}x)^2} = \rho_{x,y}\frac{\sigma_y}{\sigma_x},$$

kur $\rho_{x,y}$ - gadījuma lielumu X un Y korelācijas koeficients.

Līdz ar to tuvinātās Y pa X regresijas taisne ir šāda:

$$\frac{\tilde{y} - \mathbb{E}y}{\sigma_y} = \rho_{x,y}\frac{x - \mathbb{E}x}{\sigma_x}.$$

Analoģiski

$$\frac{\tilde{x} - \mathbb{E}x}{\sigma_x} = \rho_{x,y}\frac{y - \mathbb{E}y}{\sigma_y} \text{ — tuvinātās } X \text{ pa } Y \text{ regresijas taisne.}$$

Viegli aprēķināt prognozes kļūdu

$$\Delta_{\min}(a,b) = \mathbb{E}(y - \tilde{y})^2 =$$

$$= \mathbb{E}(y - \mathbb{E}y)^2 - 2\mathbb{E}(y - \mathbb{E}y)(x - \mathbb{E}x)\rho_{x,y}\frac{\sigma_y}{\sigma_x} + \rho_{x,y}^2\frac{\sigma_y^2}{\sigma_x^2}\mathbb{E}(x - \mathbb{E}x)^2 =$$

$$= \sigma_y^2(1 - \rho_{x,y}^2). \tag{14}$$

Iegūtais lielums ļauj spriest par Y pa X prognozes precizitāti, izmantojot vislabāko lineāro tuvinājumu. Analoģiski,

$$\Delta'(a,b) = \sigma_x^2(1 - \rho_{x,y}^2) \tag{15}$$

ļauj spriest par X pa Y prognozes precizitāti.
Lielumus

$$\beta_{Y|x} = \rho_{x,y}\frac{\sigma_y}{\sigma_x}, \;\; \beta_{X|y} = \rho_{x,y}\frac{\sigma_x}{\sigma_y}$$

sauksim par ***regresijas koeficientiem***. Acīmredzot šo regresijas koeficientu reizinājums vienāds ar ρ^2. Tātad tuvinātās regresijas taisnes ir taisnes, kas novilktas pēc mazāko kvadrātu metodes.

4.§. Normālā korelācija[11]

Pieņemsim, ka X, Y - gadījuma lielumi ar normālo kopējo sadalījuma likumu, t.i., sadalījuma blīvums ir

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot e^{-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)}},$$

kur

$$u=\frac{x-a}{\sigma_x},\quad v=\frac{y-b}{\sigma_y},\quad \rho=\mathbb{E}\left(\frac{x-\mathbb{E}x}{\sigma_x}\cdot\frac{y-\mathbb{E}y}{\sigma_y}\right),\quad a=\mathbb{E}X,\ b=\mathbb{E}Y.$$

Parādīsim, ka šajā gadījumā starp gadījuma lielumiem pastāv lineāra korelācija.
Zinot kopējo gadījuma lielumu (X,Y) blīvumu, viegli izrēķināt komponentes sadalījuma blīvumu. Tātad

$$p_X(x)=\int\limits_{-\infty}^{\infty}p(x,y)dy=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\int\limits_{-\infty}^{\infty}e^{\frac{-u^2+2\rho uv-v^2}{2(1-\rho^2)}}dy=$$

$$=\frac{e^{-\frac{(x-a)^2}{2\sigma_x^2}}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\int\limits_{-\infty}^{\infty}e^{-\frac{\left[\frac{y-b}{\sigma_y}-\rho\cdot\frac{x-a}{\sigma_x}\right]^2}{2(1-\rho^2)}}dy=\frac{1}{\sqrt{2\pi}\sigma_x}\cdot e^{-\frac{(x-a)^2}{2\sigma_x^2}},$$

t.i., viendimensiju gadījuma lielums X, kas ir (X,Y) sastāvdaļa, arī ir normāli sadalīts.

Aprēķināsim gadījuma lieluma Y nosacīto sadalījuma blīvumu pie $X=x$.

$$p(y|x)=\frac{p(x,y)}{p(x)}=\frac{\sqrt{2\pi}\sigma_x}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\cdot e^{-\frac{u^2-2\rho uv+v^2}{2(1-\rho^2)}+\frac{u^2}{2}}=$$

$$=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\cdot e^{-\frac{(u\rho-v)^2}{2(1-\rho^2)}}=\frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}}\cdot e^{-\frac{\left[y-b-\rho\frac{\sigma_y}{\sigma_x}(x-a)\right]^2}{2\sigma_y^2(1-\rho^2)}}.$$

No iegūtās izteiksmes redzams, ka gadījuma lielums Y pie noteikuma, ka $X=x$, ir normāli sadalīts ar parametriem:

$$\mathbb{E}(Y|x)=b+\rho\frac{\sigma_y}{\sigma_x}(x-a),\ \mathbb{D}(Y|x)=\sigma_y^2(1-\rho^2),\ \text{t.i., }\bar{y}(x)=\mathbb{E}y+\rho\frac{\sigma_y}{\sigma_x}(x-\mathbb{E}x).$$

Analoģiski, gadījuma lieluma X nosacītais sadalījums pie $Y=y$ ir normālais ar centru

$$\mathbb{E}(X|y)=\mathbb{E}X+\rho\frac{\sigma_x}{\sigma_y}(y-\mathbb{E}y)\ \text{ un dispersiju }\sigma_x^2(1-\rho^2).$$

Tātad abas regresijas funkcijas ir lineāras, un līdz ar to ir lineārā korekācija.

5.§. Korelācijas koeficients[6]

Aplūkotās tuvinātās regresijas taisnes ļauj iegūt jaunus secinājumus par korelācijas koeficientu. Apzīmēsim $z := Y - \tilde{y} = Y - (aX + b)$. Ņemot vērā, ka

$$a = \rho_{x,y}\frac{\sigma_y}{\sigma_x}, \quad b = \mathbb{E}Y - a\mathbb{E}X,$$

varam izrēķināt

$$\mathbb{E}z = \mathbb{E}(Y - aX - b) = \mathbb{E}\left(Y - aX - \mathbb{E}Y + \rho_{x,y}\frac{\sigma_y}{\sigma_x}\mathbb{E}X\right) =$$

$$= \mathbb{E}[Y - \mathbb{E}Y - a(X - \mathbb{E}X)] = 0,$$

$$\operatorname{cov}(z, X) = \mathbb{E}(Y - aX - b)(X - \mathbb{E}X) =$$

$$= \mathbb{E}\left[(Y - \mathbb{E}Y) - \rho_{x,y}\frac{\sigma_y}{\sigma_x}(X - \mathbb{E}X)\right](X - \mathbb{E}X) =$$

$$= \mathbb{E}(Y - \mathbb{E}Y)(X - \mathbb{E}X) - \rho_{x,y}\frac{\sigma_y}{\sigma_x}\mathbb{E}(X - \mathbb{E}X)^2 =$$

$$= \operatorname{cov}(Y, X) - \operatorname{cov}(Y, X) = 0.$$

No šejienes secinām, ka z un X ir nekorelēti gadījuma lielumi. Tālāk

$$\mathbb{D}z = \mathbb{E}(Y - aX - b)^2 = \mathbb{E}\left[(Y - \mathbb{E}Y) - \rho_{x,y}\frac{\sigma_y}{\sigma_x}(X - \mathbb{E}X)\right]^2 =$$

$$= \sigma_y^2 + \rho_{x,y}^2\frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2 - 2\rho_{x,y}\frac{\sigma_y}{\sigma_x}\operatorname{cov}(Y, X) = \sigma_y^2 + \rho_{x,y}^2\sigma_y^2 - 2\rho_{x,y}^2\sigma_y^2 =$$

$$= \sigma_y^2(1 - \rho_{x,y}^2).$$

Tālāk Y izsakām summas veidā

$$Y = Y - \tilde{y}(x) + \tilde{y}(x) = z + \tilde{y}(x) = z + aX + b,$$

kur gadījuma lielumi z un X ir nekorelēti, bet summa $aX + b$ ir lineāra X funkcija. Nekorelētu gadījuma lielumu summas dispersiju varam uzrakstīt šādā veidā

$$\mathbb{D}Y = \mathbb{D}z + \mathbb{D}(aX + b) = \sigma_y^2(1 - \rho_{x,y}^2) + a^2\sigma_x^2 =$$

$$= \sigma_y^2(1 - \rho_{x,y}^2) + \rho_{x,y}^2\sigma_y^2.$$

No šejienes redzams, ka ρ^2 izsaka to lieluma Y dispersijas daļu, kuru dod lineāri prognozējamā komponente pie katras X vērtības, kamēr ar X nekorelētā komponente dod dispersijas daļu $1-\rho^2$. No formulām (14) un (15) seko, ka jo tuvāk ρ^2 vieniniekam, jo blīvāk sadalījums koncentrēts pie katras no taisnēm. Robežgadījumā pie $\rho = \pm 1$ abas regresijas taisnes sakrīt.

•Tātad, ρ^2 ir X un Y saites lineārās pakāpes mērs.

6.§. Korelācijas attiecība [3]

Ja regresijas līnijas nav taisnes, tad korelācijas koeficientu tikai tuvināti var uzskatīt par mainīgo X un Y atkarības pakāpes rādītāju. Nelineāras sakarības gadījumā eksistē lielumi, kas raksturo sadalījuma koncentrāciju ap taisnēm $\bar{y}(x)$ un $\bar{x}(y)$. Šie lielumi ir korelācijas attiecības $\eta^2_{Y|x}$ un $\eta^2_{X|y}$. Lai saprastu $\eta^2_{Y|x}$ struktūru, apskatīsim vienādību

$$\mathbb{D}Y = \mathbb{E}(Y-\mathbb{E}Y)^2 = \mathbb{E}(Y-\bar{y}(x)+\bar{y}(x)-\mathbb{E}Y)^2 =$$

$$= \mathbb{E}(Y-\bar{y}(x))^2 + \mathbb{E}(\bar{y}(x)-\mathbb{E}Y)^2 + 2\mathbb{E}(Y-\bar{y}(x))(\bar{y}(x)-\mathbb{E}Y) =$$

$$= \bar{\sigma}^2_{Y|x} + \mathbb{E}[\bar{y}(x)-\mathbb{E}Y]^2 + 2\int\limits_{-\infty}^{\infty}[\bar{y}(x)-\mathbb{E}Y]\int\limits_{-\infty}^{\infty}(Y-\bar{y}(x))dF_{Y|x}(y)dF_X(x) =$$

$$= \bar{\sigma}^2_{Y|x} + \mathbb{E}[\bar{y}(x)-\mathbb{E}Y]^2. \tag{16}$$

Tagad definēsim rādītāju $\eta^2_{Y|x}$

$$\eta^2_{Y|x} = \frac{\mathbb{E}(\bar{y}(x)-\mathbb{E}Y)^2}{\sigma^2_y}. \tag{17}$$

Izmantojot (16), var rakstīt

$$\eta^2_{Y|x} = 1 - \frac{\bar{\sigma}^2_{Y|x}}{\sigma^2_y}. \tag{18}$$

No (16), (17) seko, ka $0 \le \eta^2_{Y|x} \le 1$, turklāt $\eta^2_{Y|x} = 1$ tad un tikai tad, ja $\sigma^2_{Y|x} = 0$, t.i., eksistē viennozīmīga funkcionāla Y atkarība no X. Tālāk, $\eta^2_{Y|x} = 0$ tad un tikai tad, ja $\bar{\sigma}^2_{Y|x} = \sigma^2_y$, t.i., Y nav korelēts ar X.

Līdzīgas īpašības ir arī otrajam sakarības rādītājam starp X un Y : $\eta_{X|y}$. Starp rādītājiem $\eta_{X|y}$ un $\eta_{Y|x}$ neeksistē vienkāršas sakarības, Y var būt nekorelēta ar X, tātad $\eta_{X|y} = 0$, kamēr $\eta_{Y|x} = 1$.

Atzīmēsim, ka vienmēr izpildās $\rho^2 < \eta^2_{Y|x}$ un $\rho^2 < \eta^2_{X|y}$ tātad, ja kaut viens no rādītājiem $\eta_{X|y}$, $\eta_{Y|x}$ vienāds ar nulli, tad $\rho = 0$.

7.§. Empīriskais determinācijas koeficients un korelācijas attiecība[3]

Lai konstruētu determinācijas koeficienta novērtējumu jāanalizē atkarīgā mainīga Y empīriskā dispersija $Var(Y)$. Dispersiju var pierakstīt šādā veidā:

$$Var(Y) = \frac{1}{n}\sum_{n=1}^{n}(Y_i - \overline{y})^2 = \frac{1}{n}\sum_{n=1}^{n}(Y_i - \widetilde{y}(x_i) + \widetilde{y}(x_i) - \overline{y})^2 =$$

$$= \frac{1}{n}\sum_{n=1}^{n}(Y_i - \widetilde{y}(x_i))^2 + \frac{1}{n}\sum_{n=1}^{n}(\widetilde{y}(x_i) - \overline{y})^2 + 2\frac{1}{n}\sum_{n=1}^{n}(Y_i - \widetilde{y}(x_i)(\widetilde{y}(x_i) - \overline{y}).$$

Ievērojot, ka

$$\frac{1}{n}\sum_{n=1}^{n}(Y_i - \widetilde{y}(x_i)) = \overline{y} - \frac{1}{n}\sum_{n=1}^{n}(a + b(x_i - \overline{x}) = \overline{y} - \overline{y} = 0,$$

$$\frac{1}{n}\sum_{n=1}^{n}\widetilde{y}(x_i) = a = \overline{y}, \quad \frac{1}{n}\sum_{n=1}^{n}(Y_i - \widetilde{y}(x_i)(\widetilde{y}(x_i) - \overline{y}) =$$

$$= \frac{1}{n}\sum_{n=1}^{n}(Y_i - \overline{y} - b(x_i - \overline{x})(\overline{y} - b(x_i - \overline{x}) - \overline{y}) =$$

$$= b\frac{1}{n}\sum_{n=1}^{n}(Y_i - \overline{y})(x_i - \overline{x}) - b^2\frac{1}{n}\sum_{n=1}^{n}(x_i - \overline{x})^2 =$$

$$= bCov(x,y) - b^2Var(x) = bCov(x,y) - bCov(x,y) = 0,$$

un ievietojot tos izteiksmē, iegūstam

$$Var(Y) = Var\widetilde{y}(x) + Var(Y - \widetilde{y}(x)).$$

Sakarība rāda, ka atkarīgā mainīgā Y no aritmētiskā vidējā noviržu kvadrātu summa sastādās no diviem locekļiem:
pirmais saskaitāmais raksturo Y empīriskās dispersijas daļu, ko var izskaidrot ar empīriskās regresijas līnijas palīdzību, otrais saskaitāmais - to Y empīrisko dispersiju daļu, kas nav saistīta ar mainīgo x.
Ar empīrisko regresiju izskaidrotās dispersijas daļas attiecību pret visu Y

empīrisko dispersiju sauc par ***empīrisko determinācijas attiecību*** un apzīmē ar burtu R^2:

$$R^2 = \frac{Var\widetilde{y}(x)}{Var(Y)}.$$

Kvadrātsakni no empīriskās determinācijas attiecības sauc par ***empīrisko korelācijas attiecību***.
Tā kā parasti ir zināma nevis izskaidrotā, bet neizskaidrotā dispersijas daļa, tad iegūstam, ka

$$R^2 = 1 - \frac{Var(Y - \widetilde{y}(x))}{Var(Y)}.$$

Determinācijas attiecība rāda, kādu daļu no atkarīgā mainīga dispersijas izskaidro aprēķinātais regresijas vienādojums.

Lemma. Lineārās regresijas gadījumā $R^2 = r^2_{(x,y)}$, kur $r_{(x,y)}$ ir empīriskās korelācijas koeficients starp X un Y.

Pierādījums.

$$R^2 = \frac{Var\widetilde{y}(x)}{Var(Y)} = \frac{\sum\limits_{i=1}^{n}(\widetilde{y}(x_i) - \overline{\widetilde{y}(x_i)})^2}{nVar(Y)} = \frac{\sum\limits_{i=1}^{n}(a + b(x - \overline{x}) - a)^2}{nVar(Y)} = \frac{b^2 Var(x)}{Var(Y)} =$$

$$= \frac{Cov^2(x,y)}{Var^2(x)}\frac{Var(x)}{Var(Y)} = \frac{Cov^2(x,y)}{Var(x)Var(Y)} = r^2_{(x,y)}.\blacktriangle$$

No lemmas seko $Var\widetilde{y}(x) = b^2 Var(x)$. Tātad,

$$s.k.(b) = \sqrt{\frac{Var(Y - \widetilde{y}(x))}{(n-2)Var(x)}} = b\sqrt{\frac{Var(Y - \widetilde{y}(x))}{(n-2)Var\widetilde{y}(x)}} =$$

$$= b\sqrt{\frac{1 - (1 - [Var(Y - \widetilde{y}(x))/Var(Y)])}{(n-2)[Var\widetilde{y}(x)/Var(Y)]}} = b\sqrt{\frac{1 - R^2}{(n-2)R^2}} = b\sqrt{\frac{1 - r^2_{(x,y)}}{(n-2)r^2_{(x,y)}}}.$$

8.§ Hipotēžu pārbaude lineārās regresijas analīzē[1],[5]

Pētot sakarības ekonomikā, vai citās zinātņu nozarēs, parasti samērā viegli var noteikt, kura no pētāmām pazīmēm ir neatkarīga un kura atkarīga. Atkarīgos un neatkarīgos mainīgos lielumus nosaka, vadoties no to kvalitatīvajām īpašībām un pētnieka zināšanām. Tā, piemēram, kopējas inflācijas temps ir atkarīgs no inflācijas tempa, kuru izraisa darba algas pieaugums;

pārtikas izdevumi vienam cilvēkam ir atkarīgi no tā ienākuma. Taču nereti nākas sastapties ar statistiskām pazīmēm, kuras atspoguļo tādas parādības, par kurām nevar pateikt, kura no tām ir cēlonis un kura ir sekas. Šādos gadījumos jāanalizē divas regresijas līnijas.
Pieņemsim tagad, ka ir zināms, ka starp gadījuma lielumiem X un Y eksistē sakarība un ka, pieaugot neatkarīgā mainīgā lieluma X vērtībām, lineāri pieaug atkarīgā mainīgā lieluma Y vidējās vērtības, kas nozīmē, regresijas līkni ir lineārā funkcija

$$\overline{y}(x) = \alpha + \beta(x - \overline{x}).$$

Nulles hipotēzi par regresijas koeficientiem β un α

Viena no svarīgākajām hipotēzēm regresijas analīzē ir:
$H_0 : \beta = \beta_0$ pret alternatīvo hipotēzi $H_1 : \beta \neq \beta_0$.
Hipotēzes pārbaudei jāizmanto statistika, kas sadalīta pēc Stjūdenta likuma ar brīvības pakāpju skaitu $\nu = n - 2$,

$$\frac{b - \beta}{s.k.(b)} = t_{n-2},$$

jeb statistiku sadalītu pec Snedekora likuma

$$\left(\frac{b - \beta}{s.k.(b)}\right)^2 \sim F(1, n-2),$$

kur gadījuma lielums b ir empīriskās regresijas virziena koeficients. Izvēloties nozīmības līmeni $\widehat{\alpha}$, konstruējam statistikas kritisko apgabalu
$\widehat{S} = (-\infty, -\varepsilon) \cup (\varepsilon, \infty)$. Kritisko robežu ε iegūstam no

$$\mathbb{P}\left(\left|\frac{b - \beta}{s.k.(b)}\right| > \varepsilon | H_0\right) = \widehat{\alpha}.$$

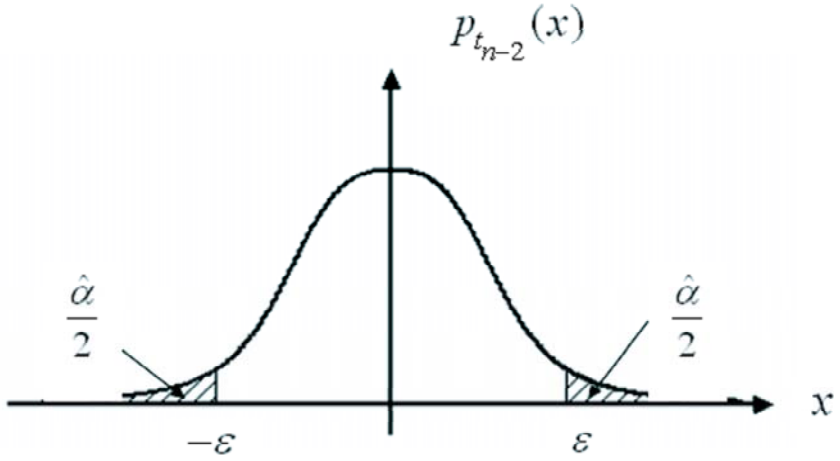
Ja pēc statistiskajiem datiem statistikas aprēķinātā vērtība

$\frac{b_0 - \beta_0}{s.k.(b)}$, kur b_0 – pēc datiem aprēķinātā b vērtība,

pieder kritiskajam apgabalam, tad izvirzīto nulles hipotēzi noraidām. Prētējā gadījumā nulles hipotēzi ir iespējama. Tātad, varbūtība noraidīt H_0, ja tā ir pareiza, t.i., pirmā veida kļūda, vienāda ar $\widehat{\alpha}$. Savukārt, varbūtība noraidīt pareizu alternatīvu H_1, t.i., otrā veida kļūda, ir

$$\mathbb{P}\left(\left|\frac{b - \beta}{s.k.(b)}\right| < \varepsilon | H_1\right) = 1 - \mathbb{P}\left(\frac{b - \beta}{s.k.(b)} \in \widehat{S} | H_1\right).$$

$$\widehat{\beta} := \mathbb{P}\left(\frac{b-\beta}{s.k.(b)} \in \widehat{S}|H_1\right) \text{ ir kritērija jauda.}$$



1.zīm. $H_0 : \beta = \beta_0$ pret alternatīvu $H_1 : \beta \neq \beta_0$.

Jāuzsver, ka nulles hipotēzes $H_0 : \rho_{(x,y)} = 0$ un $H_0 : \beta = 0$ ir identiskas un nozīme, ka nav korelācijas starp X un Y. Lineāras regresijas gadījumā tas nozīme, ka X un Y ir neatkarīgi gadījuma lielumi.
Lai noteiktu, vai regresijas modelis statistiski nozīmīgi izskaidro atkarīgā mainīgā vērtību izkliedi, jāpārbauda hipotēzi:

$$H_0 : \gamma^2_{Y|x} = 0; \; H_1 : \gamma^2_{Y|x} \neq 0.$$

Lineāras regresijas gadījumā, kad H_0 ir spēkā, var izmantot pēc Snedekora sadalītu statistiku

$$\frac{R^2}{1-R^2}(n-2) \sim F(1, n-2).$$

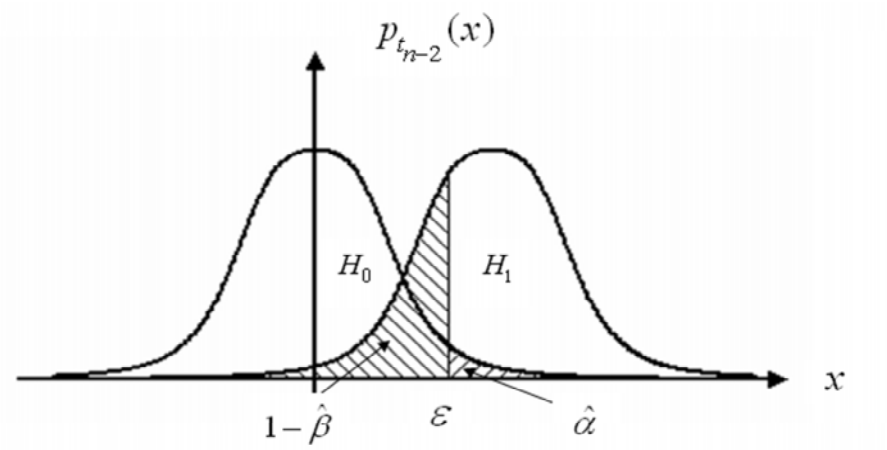
Tiešām, no $\gamma^2_{Y|x} = 0$ seko $\rho^2_{(x,y)} = 0$, kas ir identiskas, ka $\beta = 0$. Līdz ar to, ir spēkā

$$\frac{b^2 Var(x)}{Var(Y - \widetilde{y}(x))}(n-2) = \frac{b^2 Var(x)/Var(y)}{Var(Y - \widetilde{y}(x))/Var(y)}(n-2) =$$

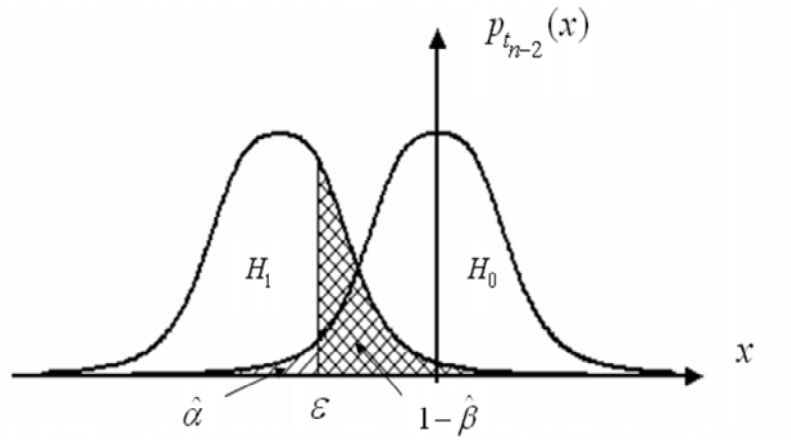
$$= \frac{R^2}{1-R^2}(n-2) = \frac{r^2_{(x,y)}}{1-r^2_{(x,y)}}(n-2) \sim F(1, n-2).$$

Vispārīgā gadījumā iepriekš minētās statistikas kritisko apgabalu, kas atbilst izvēlētajam nozīmības līmenim $\widehat{\alpha}$, jāizvēlas vienpusēji, vai divpusēji, atkarīgi no alternatīvās hipotēzes.

Nulles hipotēze H_0	Alternatīvā hipotēze H_1	Kritiskais apgabals, kurā ar nozīmības līmeni $\widehat{\alpha}$ noraida pareizu H_0
$\beta = \beta_0$	$\beta \neq \beta_0$	$\left\lvert \frac{b-\beta_0}{s.k.(b)} \right\rvert > \varepsilon$
$\beta = \beta_0$	$\beta > \beta_0$	$\frac{b-\beta_0}{s.k.(b)} > \varepsilon$
$\beta = \beta_0$	$\beta < \beta_0$	$\frac{b-\beta_0}{s.k.(b)} < \varepsilon$



2.zīm. $H_0 : \beta = \beta_0$ pret alternatīvo hipotēzi $H_1 : \beta > \beta_0$.



3.zīm. $H_0 : \beta = \beta_0$ pret alternatīvo hipotēzi $H_1 : \beta < \beta_0$.

Lai ar drošību $\widetilde{\beta}$ konstruētu regresijas taisnes virziena koeficieta ticamības intervālu, jāizmanto iepriekš minētā statistika. No vienādības

$$\widetilde{\beta} = \mathbb{P}\left(\left|\frac{b-\beta}{s.k.(b)}\right| < \varepsilon\right),$$

izmantojot Stjūdenta sadalījuma tabulas, iegūstam ε. Līdz ar to ar drošību $\widetilde{\beta}$ parametra β ticamības intervāls būs

$$J_{\widetilde{\beta},\beta} = (b - \varepsilon\, s.k.(b),\ b - \varepsilon\, s.k.(b)).$$

Regresijas vienādojuma brīvais loceklis a atspoguļo regresijas taisnes novietojumu plaknē un tiek aprēķināts pec izlases datiem. Tas nozıme, ka var izvirzīt jautājumu par brīvā locekļa novērtēšanu; brīvā locekļa kļūdas aprēķināšanu; ticamības intervāla konstruēšanu; hipotēžu izvirzīšanu.
Kā bija iepriekš minēts statistika

$$a = \frac{\sum\limits_{i=1}^{n} Y_i}{n} =: \overline{y}$$

ir parametra α nenovirzīts un būtisks novērtējums. Parametra a standartkļūda ir

$$s.k.(a) := \sqrt{S_a^2} := \sqrt{\frac{S^2}{n}} = \sqrt{\frac{\sum\limits_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right)^2}{n(n-2)}}.$$

Zinot statistikas

$$\frac{\sqrt{n}(a-\alpha)}{\sigma} \sim \tau,\ \text{kur}\ \tau \sim N(0,1),\ \ \frac{a-\alpha}{s.k.(a)} = t_{n-2}$$

sadalījuma likumi, varam secināt, ka ar drošību $\widetilde{\beta}$ parametra α ticamības intervāls ir

$$J_{\widetilde{\beta},\alpha} = (a - \varepsilon s.k.(a), a + \varepsilon s.k.(a)),$$

kur ε, izmantojot Stjūdenta sadalījuma tabulas, iegūstam no vienādības

$$\widetilde{\beta} = \mathbb{P}\left(\left|\frac{a-\alpha}{s.k.(a)}\right| < \varepsilon\right).$$

Var arı izvirzīt nulles hipotēzi par α koeficientu. Šajā gadījumā, atkarīgi na alternatīvas, jākonstruē vienpusējais vai divpusējais kritiskais apgabals $\widehat{S}$ tādā veidā, lai izpildās

$$\begin{cases} \mathbb{P}\left(\frac{a-\alpha}{s.k.(a)} \in \widehat{S} | H_0\right) & = \widehat{\alpha} \\ \mathbb{P}\left(\frac{a-\alpha}{s.k.(a)} \notin \widehat{S} | H_1\right) & = \min\limits_{\widehat{S}}. \end{cases}$$

Korelācijas koeficienta vērtēšana[2]

Lai aprēķinātu korelācijas koeficienta ticamības intervāla robežas, jāizmanto Fišera statistika

$$Z := \frac{1}{2}\ln\frac{1+r_{(x,y)}}{1-r_{(x,y)}}.$$

Tika pierādīts, ka gadījuma lieluma Z sadalījuma likums ir tuvs normalājām sadalījumam ar parametriem

$$\mathbb{E}Z = \frac{1}{2}\ln\frac{1+\rho_{(x,y)}}{1-\rho_{(x,y)}},\quad DZ = \frac{1}{n-3}.$$

Tātad, statistikas

$$\frac{\sqrt{n-3}}{2}\left[\ln\frac{1+r_{(x,y)}}{1-r_{(x,y)}} - \ln\frac{1+\rho_{(x,y)}}{1-\rho_{(x,y)}}\right] = \sqrt{n-3}\ln\frac{(1+r_{(x,y)})(1-\rho_{(x,y)})}{(1-r_{(x,y)})(1+\rho_{(x,y)})}$$

sadalījums tuvs $N(0,1)$. Līdz ar to, lai konstruētu parametra $\rho_{(x,y)}$ ticamības intervālu, jālieto speciālās gadījuma lieluma Z sadalījuma likuma tabulas, jeb ar uzrādītiem pārveidojumiem pāriet uz standarta normālo sadalījumu.
Lai pārbaudītu, vai starp neatkarīgo un atkarīgo mainīgo pastāv lineāra sakarība, izvirzām nulles hipotēzi:
$H_0 : \rho_{(x,y)} = 0$ attiecībā pret alternatīvo hipotēzi: $H_1 : \rho_{(x,y)} \neq 0$.
Ir zināms, ka, ja H_0 ir spēkā, statistika

$$\frac{r^2_{(x,y)}}{1-r^2_{(x,y)}}(n-2) \sim F(1,n-2)$$

sadalīta pec Snedekora likuma, savukārt statistika

$$\frac{r_{(x,y)}}{\sqrt{1-r^2_{(x,y)}}}\sqrt{(n-2)} \sim t_{n-2}$$

sadalīta pēc Stjūdenta likuma. Tālāk, izvēlamies nozīmības līmeni $\widehat{\alpha}$ un vienu no statistikām, piemēram pirmo, un konstruējam kritisko apgabalu $\widehat{S}$, lai izpildās

$$\mathbb{P}\left(\left|\frac{r^2_{(x,y)}}{1-r^2_{(x,y)}}(n-2)\right| > \varepsilon | H_0\right) = \widehat{\alpha}.$$

Pēc statistiskiem datiem aprēķinājām statistikas atbilstošo vērtību. Ja tā pieder kritiskajām apgabalam $\widehat{S}$ hipotēzi H_0 noraidām kā nepareizu un pieņemam alternatīvo hipotēzi H_1. Prētējā gadījumā hipotēze H_0 ir iespējama.

Piezīme. Lai pārbaudītu nulles hipotēzi $H_0 : \rho_{(x,y)} = 0$ attiecībā pret alternatīvo hipotēzi: $H_1 : \rho_{(x,y)} \neq 0$, kā redzams, var arī izmantot statistiku, kas sadalīta pec Stjūdenta likuma. Viegli pārliecināties, ka aprēķini dod vienādus rezultātus.

Atkarīgā mainīgā vērtību prognozēšana

Regresijas līkni var izmantot rezultatīvās pazīmes vērtību prognozei. Pieņemsim, ka neatkarīgā mainīga vērtība x neietilpst izlasē, bet ietilpst ģenerālajā kopā X. Šajā gadījumā statistika

$$\frac{Y_x - \widetilde{y}(x)}{s.k.(Y_x - \widetilde{y}(x))} = \frac{Y_x - \widetilde{y}(x)}{\sqrt{\sum\limits_{i=1}^{n}(Y_i - \widetilde{y}(x_i))^2)}\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}}\sqrt{n-2}$$

sadalīta pēc Stjūdenta likuma ar brīvības pakāpēm $\nu = n - 2$. Tālāk konstruēsim ticamības intervālu lielumam Y_x parastā secībā:

- izvēlamies ticamības varbūtību $\widetilde{\beta}$, t.i., ticamības intervāla drošību;
- aprēķina ε no vienādojuma

$$\mathbb{P}\left(\left|\frac{Y_x - \widetilde{y}(x)}{s.k.(Y_x - \widetilde{y}(x))}\right| > \varepsilon\right) = \widetilde{\beta};$$

- pec statistiskiem datiem aprēķina atbilstošās vērtības statistikām

$$\widetilde{y}\,(x),\ s.k.'(Y_x - \widetilde{y}(x));$$

- atrod ticamības intervālu

$$J_{(\widetilde{\beta},Y_x)} = (\widetilde{y}\,(x) - \varepsilon\, s.k.'(Y_x - \widetilde{y}(x)), \widetilde{y}\,(x) + \varepsilon s.k.'(Y_x - \widetilde{y}(x))).$$

Ar drošību $\widetilde{\beta}$ prognozēt regresijas taisnes vērtību konkrētai neatkarīgā mainīgā vērtībai $x = x_j,\ \ j = 1, 2, ..., n$, nozīmē konstruēt ticamības intervālu $\overline{y}(x_j)$. Vispirms izvēlēsimies statistiku

$$\frac{\widetilde{y}(x) - \overline{y}(x)}{s.k.(\widetilde{y}(x))} = \frac{\widetilde{y}(x) - \overline{y}(x)}{\sqrt{\sum\limits_{i=1}^{n}(Y_i - \widetilde{y}(x_i))^2}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}}\sqrt{n-2} = t_{n-2},$$

kas tiek sadalīta pēc Stjūdenta likuma ar $n - 2$ brīvības pakāpēm. Tālāk, ņemot vera iepriekš minētos soļus ticamības intervāla konstruēšanai, iegūstam

$$J_{(\widetilde{\beta},\overline{y}(x))} = (\widetilde{y}\,(x) - \varepsilon\, s.k.'(\widetilde{y}(x)), \widetilde{y}\,(x) + \varepsilon\, s.k.'(\widetilde{y}(x))),$$

kur ε atrodam no vienādojuma

$$\mathbb{P}\left(\left|\frac{\widetilde{y}(x)-\overline{y}(x)}{s.k.(\widetilde{y}(x))}\right|>\varepsilon\right)=\widetilde{\beta}.$$

1.Piemērs. Ir zināms, ka kopējā inflācija ir atkarīga no daudziem faktoriem un, viens no tiem ir inflācija, kuru izraisa darba algas pieaugums. Pieņemsim, ka starp kopējās inflācijas tempu p un inflācijas tempu w, kuru izraisa darba algas pieaugums, eksistē lineāra atkarība, t.i.,

$$p=\alpha+\beta w+u,$$

kur α,β - parametri, u - normāli sadalīta gadījuma komponente ar $\mathbb{E}u=0,\ Du=\sigma^2,\ \mathbb{E}uw=0$. Tātad, korelācija ir lineāra un regresijas līnija ir

$$\overline{p}(x)=\alpha+\beta x.$$

Pārbaudīsim nulles hipotēzi: $H_0:\beta=1$ pret alternatīvu $H_1:\beta\neq 1$, ja pēc n novērojumiem tika konstruēta empīriskā regresijas līnija

$$\widetilde{p}(x)=a+bx,$$

ar $a=-1.21, b=0.82$, un tieka aprēķinātas sekojošo statistiku vērtības:
$s.k.(a)=0.05,\ \ s.k.(b)=0.1.$
Vispirms, izvēlēsimies nozīmības līmeni, piemēram, 0.05. Tā kā statistika

$$\frac{b-\beta}{s.k.(b)}=t_{n-2}$$

sadalīta pēc Stjūdenta likuma ar $\nu=n-2$ brīvības pakāpēm, tad, ja, piemēram, izlases apjoms ir 20, ar drošību 0.95 t-statistikas iespējamo vērtību kopa būs $(-2.101, 2.101)$. Ja ir spēkā H_0, t-statistikas novērojamā vērtība vienāda ar

$$\frac{b-\beta}{s.k.(b)}=\frac{0.82-1}{0.1}=-1.8.$$

Tas nozīmē, ka nulles hipotēzi pie nozīmības līmeņa 0.05 (5% līmeņa) nevar noraidīt kā nepareizu. Acīmredzot, ka nevar noraidīt arı hipotēzi $\beta=0.8$, kā arī $\beta=0.9$ un t.t.. Bezjēdzīgi teikt, ka visas hipotēzes ir vienlaicīgi pareizas. Atzīmēsim, ka ar drošību 0.95 parametra β ticamības intervāls ir

$$J_{0.95,\beta}=(0.61,1.03),$$

negaidāmo parametra β vērtību kopa $\widehat{S}$, tā saucamais kritiskais apgabals nozīmības līmenim 0.05, ir $(-\infty,\ 0.61; 1.03,\ \infty)$.
Bez tam, pirmā veidā kļūda, varbūtība atmest pareizu H_0, ir vienāda ar nozīmības līmeni, mūsu piemēra tas ir 0.05. Savukārt, otrā veidā kļūda, varbūtība noraidīt pareizu alternatīvu H_1, ir minimālā.
Jāatzīmē, ka nulles hipotēzi varētu naraidīt tikai gadījumā, ja eksperimentu gaitā statistikas b novērojamā vērtība piederētu kritiskajam apgabalam $\widehat{S}$.

2.Piemērs. Pirms apmācīšanas ar testa palīdzību tika pārbaudīti 36 studentu spējas. Testa rezultāti novērtēti ar baļļu skaitu, apmācīšanas rezultāti ar: "1", ja students "pabeidz" un ar "0", ja "nepabeidz" apmācīšanu. Iegūtie rezultāti apkopoti tabulā.

Vai testu var uzskatīt par lietderīgu? Ja tā, tad kādai jābūt testa minimālajai

stud. nr.	testa rez.(x)	apmāc. rez.(y)	stud. nr.	testa rez.(x)	apmāc. rez.(y)	stud. nr.	testa rez.(x)	apmāc. rez.(y)
1	30	0	13	26	0	25	9	0
2	29	1	14	43	1	26	36	1
3	33	0	15	43	0	27	61	1
4	62	1	16	68	1	28	79	0
5	59	0	17	63	1	29	57	0
6	63	1	18	42	0	30	46	1
7	80	1	19	51	0	31	70	0
8	32	0	20	45	0	32	31	1
9	60	1	21	22	0	33	68	1
10	76	1	22	30	1	34	62	1
11	13	0	23	40	0	35	56	1
12	41	1	24	25	0	36	36	1

atzīme, lai skolas abiturientu varētu uzņemt mācību iestādē.
Tatād, uzdevums ir:
• aprēķināt lineāras regresijas $\overline{y}(x) = \alpha + \beta x$ koeficientu α un β novērtējumus
• ar drošību 0.95 konstruēt parametru ticamības intervālus
• uzrādīt minimālo baļļu skaitu, kad students ar varbūtību 0.95 beigs mācības programmu
• novērtēt prognozes kvalitāti
No tabulas iegūstam,

$$\sum_{i=1}^{36} x_i = 1688,\ \overline{x} = \frac{1688}{36} = 46.9,\ \sum_{i=1}^{36} y_i = 19,\ \overline{y} = 0.53,\ \sum_{i=1}^{36} x_i y_i = 1011,$$

$$\sum_{i=1}^{36} x_i^2 = 91256,\ \frac{\sum\limits_{i=1}^{36} x_i^2}{36} = 2534.9,\ Var(x) = \frac{\sum\limits_{i=1}^{36} x_i^2}{36} - \overline{x}^2 = 2534.9 - (46.9)^2 2 = 335.3,$$

$$b = \frac{Cov(x,y)}{Varx} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum\limits_{i=1}^{36} x_i^2 - n\overline{x}^2} = \frac{1011 - 36 \cdot 46.9 \cdot 0.53}{91256 - 36 \cdot (46.9)^2} = 0.0099,$$

$$a = \overline{y} - b\overline{x} = 0.53 - 0.0099 \cdot 46.9 = 0.0624,$$

$$\widetilde{y}(x) = 0.0624 + 0.0099x.$$

Ņemot vērā, ka

$$\sum_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right)^2 = \min_{a,b}$$

tikai tad, ja izpildās

$$\begin{cases} \sum\limits_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right) = 0 \\ \sum\limits_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right) x_i = 0 \end{cases},$$

iegūstam, ka

$$\sum_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right)^2 = \sum_{i=1}^{n} Y_i^2 - a\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} Y_i x_i = 7.7744,$$

$$\sum_{i=1}^{n} \left(\widetilde{y}(x_i) - \overline{y}\right)^2 = 1.1978,$$

$$\sum_{i=1}^{n} \left(Y_i - \overline{y}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right)^2 + \sum_{i=1}^{n} \left(\widetilde{y}(x_i) - \overline{y}\right)^2 = 8.9722.$$

Līdz ar to

$$Var(y) = \frac{\sum\limits_{i=1}^{n} \left(Y_i - \overline{y}\right)^2}{36} = \frac{8.9722}{36} = 0.25,$$

$$R^2 = \frac{\sum\limits_{i=1}^{n} \left(\widetilde{y}(x_i) - \overline{y}\right)^2}{\sum\limits_{i=1}^{n} \left(Y_i - \overline{y}\right)^2} = \frac{1.1978}{8.9722} = 0.1335.$$

$$S^2 = \frac{\sum\limits_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right)^2}{n-2} = \frac{7.7744}{34} = 0.2286,$$

$$s.k(b) = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right)^2}{(n-2)\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \sqrt{\frac{S^2}{nVarx}} = 0.0043,$$

$$s.k.(a) = \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{nVarx}\right)\frac{\sum_{i=1}^{n} \left(Y_i - \widetilde{y}(x_i)\right)^2}{n-2}} = S\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{nVarx}\right)} = 0.218.$$

Tā kā

$$\frac{b-\beta}{s.k(b)} = t_{n-2}, \quad \text{tad} \quad J_{0.95,\beta} = (0.0011,\ 0.0187).$$

Līdzīgi konstruējam ar drošību 0.95 ticamības intervālu parametram α. No

$$\frac{a-\alpha}{s.k(a)} = t_{n-2}, \ \text{seko} \ \ J_{0.95,\alpha} = (-0.38,\ 0.506).$$

Ja $\overline{y}(x)$ interpretējam, kā varbūtību studentam, ar testa rezultātu x, sekmīgi beigt apmācību, tad no vienādojuma

$$0.95 = 0.0624 + 0.0099x$$

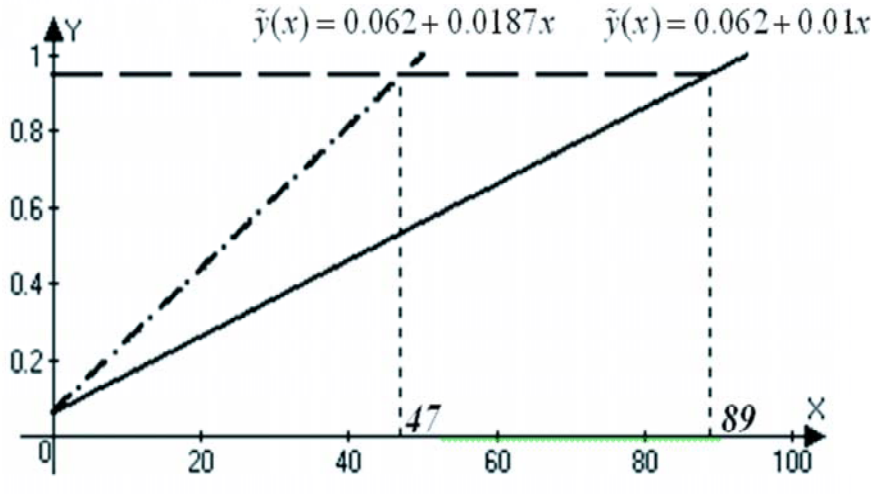
atrodam $x = 89$. Mūsu interpretacijā tas nozīme, ka studenti, kam baļļu skaits testā sasniedz var pārsniedz 89, beidz mācības programmu ar varbūtību, augstāku par 0.95. Analizējot statistiskos datus, ir redzams, ka iegūto rezultātu nevar uzskatīt par apmierinošu. Ja pētījums neprasa tik augstu ticamību, tad varam, izvēloties mazāku varbūtību, turpināt analīzi. Var izmantot arī faktu, ka ar varbūtību 0.95 pētāmā piemērā parametra β iespējamās vērtības intervāls ir $(0.0011, 0.0187)$. Līdz ar to no vienādojuma

$$0.95 = 0.0624 + 0.0187x$$

iegūstam testa rezultāta apakšējo robežu $x = 47$, atbilstošu parametra β augšējai robežai. Mūsu gadījumā

$$\widetilde{y}(47) = 0.529, \ \ s.k.(\widetilde{y}(47)) = 0.0797, \ \ J_{0.95,\overline{y}(47)} = (0.37,\ 0.69).$$

Tā kā $0.95 \notin J_{0.95,\overline{y}(47)}$ loģiski uzskatīt, ka minimālo baļļu skaits, kas atbilst varbūtībai 0.95, ir lielāk nekā 47.



4.zīm. Testa rezultāta ietekme uz studiju rezultātu.

Tā, mūsu interpretacijā, piemēram, ja $x = 63$, tad $\overline{y}(63)$ ir varbūtība studentam, kas testā ieguvis 63 balles, pabeigt mācību programmu. Tā kā $x = x_6 = 63$ pieder atkarīgā mainīgā ģenerālkopai $\mathbf{X}$, statistika

$$\frac{\widetilde{y}(x) - \overline{y}(x)}{s.k.(\widetilde{y}(x))} = t_{n-2},$$

$$[s.k.(\widetilde{y}(x))]^2 = \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \right) \sum_{i=1}^{n} (Y_i - \widetilde{y}(x_i))^2 \frac{1}{n-2},$$

sadalīta pec Stjūdenta likuma ar $\nu = 34$ brīvības pakāpēm. Tādēļ,

$$s.k.(\widetilde{y}(x)) = \sqrt{\frac{1}{36} + \frac{(63 - 46.9)^2}{36 \cdot 335.3} \cdot 0.2286} = 0.106,$$

un

$$J_{0.95,\overline{y}(63)} = (0.688 - 2.032 \cdot 0.106,\ 0.688 + 2.032 \cdot 0.106) = (0.47, 0.903).$$

Savukārt, gadījumā kad $x = x_{16} = 68$,

$$\widetilde{y}(68) = 0.7376,\ \ s.k.(\widetilde{y}(x)) = \sqrt{\frac{1}{36} + \frac{(68 - 46.9)^2}{36 \cdot 335.3} \cdot 0.2286} = 0.121$$

un $J_{0.95,\overline{y}(68)} = (0.49, 0.98)$. Tātad, ar varbūtību 0.95 secinām, ka studenti, kas testā ieguva vairāk kā 68 balles, beidz mācību programmu ar p varbūtību,

kur $p \in (0.49, 0.98)$.
Talak pārbaudīsim vai pētītā sakarība starp testa rezultātiem un apmācības rezultātiem ir statistiski nozīmīga. Šajā nolūkā izvirzām hipotēzi, saskaņā ar kuru regresijas koeficients ir nulles, un pārbaudām, vai pēc izlases datiem aprēķinātais rādītājs atbilst šai hipotēzei.
Tātad, jāanalizē $H_0 : \beta = 0$ pret alternatīvo hipotēzi $H_1 : \beta \neq 0$. Ir zināms, ka lineāras regresijas gadījumā, kad H_0 ir spēkā, var izmantot pēc Stjūdenta sadalītu statistiku

$$\frac{b}{s.k.(b)} = b\sqrt{\frac{Var(x)}{Var(Y - \widetilde{y}(x))}(n-2)} = t_{n-2},$$

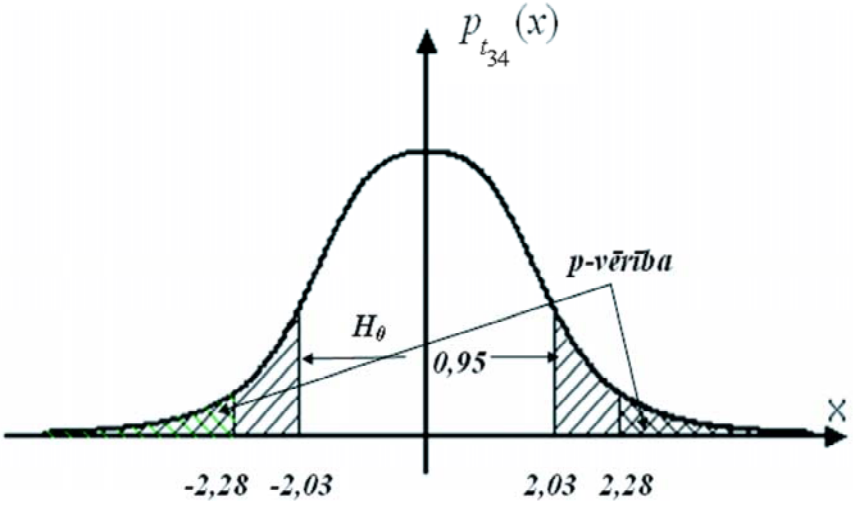
jeb pec Snedekora sadalītu statistiku

$$\frac{b^2 Var(x)}{Var(Y - \widetilde{y}(x))}(n-2) = \frac{R^2}{1-R^2}(n-2) = \frac{r^2_{(x,y)}}{1 - r^2_{(x,y)}}(n-2) \sim F(1, n-2).$$

Jāatzīmē, ka no $\beta = 0$ seko $\gamma^2_{Y|x} = 0$ un $\rho^2_{(x,y)} = 0$, kas ir identiskas.
Mūsu piemērā $\nu = 34$. Tam, pie nozīmības līmeņa 0.05, atbilst

$$\mathbb{P}\left(|t_{n-2}| < \varepsilon\right) = 0.05 \;\Rightarrow\; \varepsilon = 2.032.$$

Pēc statistiskajiem datiem t-statistikas aprēķinātā vērtība $2.288755 > 2.032$, no kā, ar varbūtību 0.95, izriet nulles hipotēzes noraidīšana.



5.zīm. $H_0 : \beta = 0$ pret alternatīvo hipotēzi $H_1 : \beta \neq 0$.

Tālāk, no

$$\mathbb{P}\left(|t_{34}| > 2.288755\right) = 0.028$$

iegūstam, ka p-vērtība vienāda ar 0.028. Tātad, secinām, ka nulles hipotēzi noraidām ar varbūtību 0.972.
Ja izvēlamies statistiku sadalītu pec Snedekora likuma, tad tabulā atrod statistikas $F(1,34)$ kritisko vērtību, kas atbilst nozīmības līmenim 0.05 un brīvības pakāpju skaitam 1 un 34. Tas ir 4.13. Līdz ar to var aprēķināt kritisko robežu determinacijas koeficienta R^2 vērtībām.

$$\frac{R^2_{kr}}{1-R^2_{kr}} = 4.13 \;\; \Rightarrow \;\; R^2_{kr} = 0.108.$$

Tā kā pec statistiskajiem datiem R^2 aprēķinātā vērtība $0.1335 > 0.108$ nulles hipotēzi noraidām ar varbūtību 0.95. Tātad, ar 5 % nozīmības līmeni nulles hipotēzi, ka lineārā regresija neuzlabo prognozes kvalitāti, salīdzinot ar triviālo prognozi $\overline{y}(x) = \overline{y}$, vajag noraidīt.

Literatūra

1. Боровков, А.А. *Математическая Статистика.* – Москва, Наука, 1984.
2. Ивченко, Г.И., Медведев, Ю.И. *Математическая Статистика.* – Москва, Высшая Школа. 1984.
3. Cox, D.R., Hinkley, D.V. *Mathematical Statistics.* - London, Chapman and Hall, 1974.
4. Дунин-Барковский, И.В., Смирнов, Н.В. *Курс Теории Вероятностей и Математической Статистики*. - Москва, Наука, 1969.
5. Lehmann, E.L. *Testing Statistical Hypotheses.* - NY, John Wiley&Sons, 1964.
6. Крамер, Г. *Математические Методы Статистики.* – Москва, МИР, 1975.
7. Kendell, M. G. And Stuart, A. *IThe Advanced Theory of Statistics. Vol.2. Inference and Relationship.* London, Charles Griffin & Co. Ltd., 1968.
8. Тутубалин, В.Н. *Теория Вероятностей.* Москва, МГУ, 1972.
9. Schmetterer, L. *Einführung in die Mathematische Statistik.* – NY , Springer-Verlag. 1966.
10. Большев, Л.И., Смирнов, Н.В. *Таблицы Математической статистики.* - Москва, Наука, 1983.
11. Carkova, V. *Matemātiskā statistika.* – Rīga, LU, 1979.
12. Carkovs, J. *Kočrena teorēma.* – Rīga, RTU, 1985.
13. Mardoch, J., Barnes, J.A. *Statistical Tables.* – London, MacMillan Press Ltd., 1998.
14. Krastiņš, O. *Varbūtību Teorija un Matemātiskā Statistika.* – Rīga, Zvaigzne, 1985.