Leíró és matematikai statisztika

Matematika alapszak, matematikai elemző specializáció

Zempléni András

Valószínűségelméleti és Statisztika Tanszék
Matematikai Intézet
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Honlap: `zempleni.elte.hu`

E-mail: `andras.zempleni@ttk.elte.hu`

Szoba: D 3-310

10. előadás

A χ^2-próba

Legyen $A_1, \ldots, A_r$ teljes eseményrendszer.
Végezzünk n darab független megfigyelést, jelölje az i-edik esemény bekövetkezési gyakoriságát N_i $(i = 1, \ldots, r)$. A megfigyelések egyes eredményei segítségével definiálható az X_i valószínűségi változó, ami vegyen fel olyan értéket, amelyik számú esemény a teljes eseményrendszerből bekövetkezett. Ezáltal formálisan

$$N_i = \sum_{j=1}^{n} I(X_j = i) \text{ és } \sum_{i=1}^{r} N_i = n$$

H_0: $P(A_i) = p_i$, $i = 1, \ldots, r$ $\rightsquigarrow$ tfh. $p_i > 0 \ \forall i$, $p_1 + \ldots + p_r = 1$
H_1: a nullhipotézis tagadása

Próbastatisztika: $T_n(\mathbf{X}) := \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{r-1}$ eloszlásban

Kritikus tartomány: $\mathcal{X}_k = \left\{\mathbf{x} : T_n(\mathbf{X}) > \chi^2_{r-1;1-\alpha}\right\}$

A χ^2-próba II

Alkalmazásai:

- illeszkedésvizsgálat: egy minta adott eloszlást követ-e
- homogenitásvizsgálat: két minta eloszlása megegyezik-e
- függetlenségvizsgálat: két szempont, ismérv, tulajdonság független-e egymástól

Megjegyzések:

- a χ^2-próba **aszimptotikus** próba, ami azt jelenti, hogy "nagy" mintaelemszámra lehet végrehajtani. "Kicsi" minták esetén a kritikus érték nem használható, azt szimulálni kell a konkrét minta alapján.
- Mikor elég "nagy" már egy minta – hüvelykujjszabály: ha legalább 100 elemű. Egyébként H_0-tól függ, hogy legalább mekkora n-re van szükség, hogy kritikus értéknek a χ^2-eloszlás kvantiliseit lehessen használni.
- Végrehajtásának további feltétele, hogy minden osztályban "elegendő" mennyiségű gyakoriság legyen (szokásos feltétel: $N_i \geq 4$).
- A próbastatisztikában lévő összeg tagjai $\frac{(O-E)^2}{E}$ alakúak, ahol E: elméleti gyakoriságok, O: tapasztalati gyakoriságok

Illeszkedésvizsgálat

H_0: a minta egy adott eloszlásból származik
H_1: a minta nem ilyen eloszlású

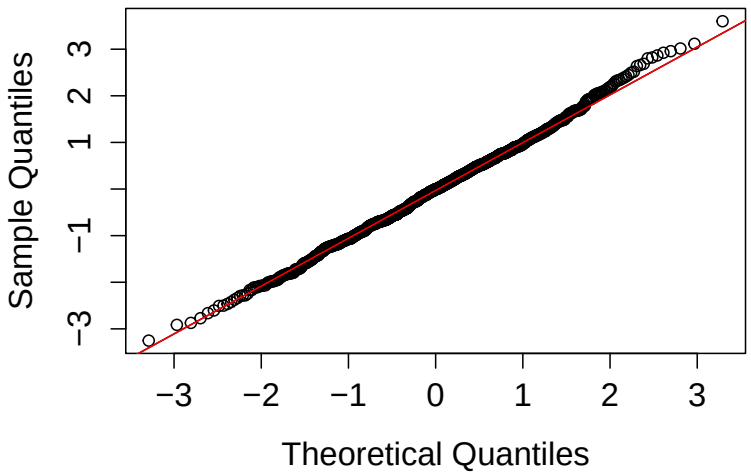
Végrehajtása:

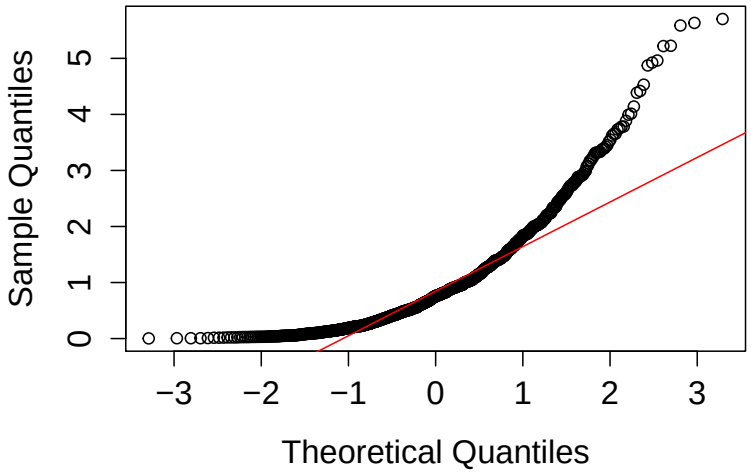
- grafikus módszerek ("szemmel" jónak tűnik-e az illeszkedés):
  - Q-Q plot
  - P-P plot
  - hisztogram/magfüggvényes sűrűségfüggvény-becslés, valamint az illesztett sűrűségfüggvény egy ábrában
- statisztikai próbák:
  - diszkrét eloszlás esetén χ^2-próba
  - folytonos eloszlás esetén több statisztikai próba közül lehet választani
    - diszkretizálás (mesterséges osztályok létrehozása) révén χ^2-próba
    - Kolmogorov-Szmirnov próba
    - Cramér-von Mises próba
    - Anderson-Darling próba
    - Shapiro-Wilk próba: kizárólag normalitásvizsgálatra, amire ez a legjobb

Illeszkedésvizsgálat grafikusan

Q-Q plot (kvantilis-kvantilis ábra)

- Az illesztett eloszlás kvantiliseit vetjük össze a tapasztalati kvantilisekkel, azaz a következő pontokat ábrázoljuk:
$\left(F^{-1}\left(\frac{k}{n+1}\right), x_k^*\right) \quad k = 1, \ldots, n$
ahol
  - F: az illesztett eloszlás eloszlásfüggvénye
  - x_k^* a k. rendezett mintaelem
- Be szokták húzni a 45 fokos egyenest és minél jobban rásimulnak a pontok az egyenesre, annál jobbnak tekinthető az illeszkedés.
- Felnagyítja az eloszlás szélein az eltéréseket, ezért szinte mindig előnyben részesítik a P-P plot-tal szemben.





Illeszkedésvizsgálat grafikusan II

P-P plot (percentilis-percentilis ábra)

- Az illesztett eloszlás egyes valószínűségeit vetjük össze a tapasztalati valószínűségekkel, azaz a következő pontokat ábrázoljuk:
  $\left(\frac{k}{n+1}, F\left(x_k^*\right)\right) \quad k = 1, \ldots, n \qquad$ ahol
  - F: az illesztett eloszlás eloszlásfüggvénye
  - x_k^* a k. rendezett mintaelem
- Be szokták húzni a 45 fokos egyenest és minél jobban rásimulnak a pontok az egyenesre, annál jobbnak tekinthető az illeszkedés.
- Felnagyítja az eloszlás közepén az eltéréseket

A Q-Q plot és P-P plot nem helyettesíti a formális tesztelést, inkább kiegészíti azt!

Illeszkedésvizsgálat χ^2-próbával

Osztályok	1	2	...	r	Összesen
Valószínűségek	p_1	p_2	...	p_r	1
Gyakoriságok	N_1	N_2	...	N_r	n

H_0 : a valószínűségek: $\mathbf{p}=(p_1,\ldots,p_r)$

H_1: nem ezek a valószínűségek

Próbastatisztika: $T_n(\mathbf{X}) = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1}$ elo.-ban, ha $n \to \infty$

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : T_n(\mathbf{x}) > \chi^2_{r-1,1-\alpha}\}$

Becsléses illeszkedésvizsgálat: csak annyit "sejtünk", hogy a minta valamilyen eloszlású, viszont a paramétereiről nincs sejtésünk. Ilyenkor amennyiben ML-módszerrel becsüljük meg az s darab ismeretlen paramétert, akkor a próbastatisztika: $T_n(\mathbf{X}) \overset{H_0 \text{ esetén}}{\longrightarrow} \chi^2_{r-1-s}$ eloszlásban, ha $n \to \infty$.

A χ^2-próba végrehajtásának feltételei (hüvelykujjszabály): $N_i \geq 4$ és $np_i \geq 4$ minden i-re. Ha ezek nem teljesülnek, akkor osztályokat kell összevonni.

Illeszkedésvizsgálat Kolmogorov-Szmirnov próbával

$H_0 : F_{X_1}(x) = F(x) \ \ \forall x \in \mathbb{R}$ ahol F egy adott eloszlás elofv.-e

H_1: a nullhipotézis tagadása

Próbastatisztika: $D_n(\mathbf{X}) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$

A próbastatisztika $\sqrt{n}$-szeresének eloszlása H_0 esetén az ún. Kolmogorov-eloszláshoz tart ($n \to \infty$). Jelöljük K_α-val a Kolmogorov-eloszlás α-kvantilisét.

Kritikus tartomány: $\mathcal{X}_k = \{\mathbf{x} : \sqrt{n} D_n(\mathbf{x}) > K_{1-\alpha}\}$

Megjegyzések:

- D_n kiszámításához elég csak a mintapontokban tekinteni az eltérést.
- Nem lehet használni a határeloszlást, ha paramétereket kell becsülnünk! Ilyen esetben a kritikus értéket szimulációval kaphatjuk meg.
- A Kolmogorov-eloszlás eloszlásfüggvénye: $1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}$

Feladatok

E**44.)** Egy gyártó megfigyelte, hogy 100, általa előállított SSD merevlemezen 5 év használat után hány hibás szektort talál az ezek felkutatására készített szoftver:

Hibás szektorok száma	0	1	2	3	4	5	7	Összesen
Gyakoriságok	45	35	12	5	1	1	1	100

Vizsgáljuk meg, hogy a szektorhibák száma Poisson-eloszlást követ-e!

E**45.)** Nézzük meg P-P plot-tal és Q-Q plot-tal, majd diszkretizálás után χ^2-próbával, valamint Kolmogorov-Szmirnov próbával, hogy a következő minta:

4,3	2,0	5,6	8,1	3,2	0,6	5,4	8,9	7,5	9,3
9,6	6,7	4,4	2,9	1,0	6,5	4,0	6,6	4,2	1,9

származhat-e az alábbi eloszlásokból:

a.) $E(0; 10)$;

b.) $N\left(5; \left(\frac{5}{\sqrt{3}}\right)^2\right)$.

Homogenitásvizsgálat

Adott két független minta, mindkettő egy közös szempont szerint r osztály egyikébe sorolva.

	Osztályok	1	2	...	r	Összesen
1. minta	Valószínűségek	p_1	p_2	...	p_r	1
	Gyakoriságok	N_1	N_2	...	N_r	n
2. minta	Valószínűségek	q_1	q_2	...	q_r	1
	Gyakoriságok	M_1	M_2	...	M_r	m

H_0: a két minta azonos eloszlású, azaz $(p_1, \ldots, p_r) = (q_1, \ldots, q_r)$
H_1: a nullhipotézis tagadása

Próbastatisztika: $T_{n,m}(\mathbf{X}, \mathbf{Y}) = nm \sum_{i=1}^{r} \frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i} \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{r-1}$ eloszlásban

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{X}, \mathbf{Y}) : T_{n,m}(\mathbf{X}, \mathbf{Y}) > \chi^2_{r-1,1-\alpha}\}$

Függetlenségvizsgálat

Feladat: van egy minta, két ismérv szerint csoportosítva. Azt kell eldönteni, hogy a két szempont független-e egymástól.

$p_{i,j} = P$(egy megfigyelés az (i,j) osztályba kerül)
$N_{i,j} =$ennyi megfigyelés kerül az (i,j) osztályba

		2. szempont					
		1	...	j	...	s	Összesen
1. szempont	1	N_{11}	...	N_{1j}	...	N_{1s}	$N_{1\bullet}$
	⋮	⋮		⋮		⋮	⋮
	i	N_{i1}	...	N_{ij}	...	N_{is}	$N_{i\bullet}$
	⋮	⋮		⋮		⋮	⋮
	r	N_{r1}	...	N_{rj}	...	N_{rs}	$N_{r\bullet}$
Összesen		$N_{\bullet 1}$	...	$N_{\bullet j}$	...	$N_{\bullet s}$	n

ahol $N_{i\bullet} = \sum_{j=1}^{s} N_{ij}$ és $N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$

Függetlenségvizsgálat II

Itt formálisan a mintánk két dimenziós: a megfigyelések az $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ párok, ahol az X-ek r, az Y-ok pedig s különböző értéket vehetnek fel nemnulla valószínűséggel: $p_{i,j} = P(X_1 = x_i, Y_1 = y_j)$, ahol $i = 1, \ldots, r$ és $j = 1, \ldots, s$.

Továbbá $N_{i,j} = \sum_{k=1}^{r} \sum_{l=1}^{s} I(X_k = x_i, Y_l = y_j)$.

H_0 : az ismérvek függetlenek, azaz $p_{i,j} = p_{i\bullet} \cdot p_{\bullet j} \quad \forall i, j$-re

H_1: az ismérvek nem függetlenek

Próbast.: $T_n(\mathbf{X}, \mathbf{Y}) = \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{i,j} - N_{i\bullet} N_{\bullet j}/n)^2}{N_{i\bullet} N_{\bullet j}/n} \right) \overset{H_0 \text{ esetén}}{\underset{n\to\infty}{\longrightarrow}} \chi^2_{(r-1)(s-1)}$

elo.-ban

Kritikus tartomány: $\mathcal{X}_k = \{(\mathbf{X}, \mathbf{Y}) : T_n(\mathbf{X}, \mathbf{Y}) > \chi^2_{(r-1)(s-1),1-\alpha}\}$

Ha $r = s = 2$, akkor a próbastatisztika $T_n = n \cdot \frac{(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1\bullet}N_{2\bullet}N_{\bullet 1}N_{\bullet 2}}$-re egyszerűsödik, az aszimptotikus eloszlás pedig 1 szabadságfokú χ^2.

Feladat

E**46.)** Egy webtervező azt gyanítja, hogy az általa létrehozott internetes vásárlás honlapján a vásárlások mértéke összefügg azzal, hogy milyen nap van a héten. Ennek a sejtésnek az ellenőrzésére egy héten keresztül adatokat gyűjt – összesen 3758 látogatót számlált meg:

Vásárlás	H	K	Sz	Cs	P	Sz	V	Össz.
Nem vásárolt	399	261	284	263	393	531	502	2633
1 vásárlás	119	72	97	51	143	145	150	777
Több vásárlás	39	50	20	15	41	97	86	348
Összesen	557	383	401	329	577	773	738	3758

Alkalmas statisztikai próbával döntsünk arról, hogy helyes-e a webtervező sejtése!

Regresszióelemzés, lineáris modell – motiváció

- Gyakorlati szempontból a félév egyik legfontosabb témája!
- Az 1. órai kérdőíves felmérés alapján mennyire magyarázható jól
  - a hallgatók testmagassága a súlyuk segítségével?
  - a hallgatók testmagassága a súlyuk és a cipőméretük segítségével?
  - a hallgatók statisztika érdemjegye a testmagasságuk segítségével?
- Egy részvény holnapi árfolyamát hogyan jelezzük előre a tegnapi, tegnapelőtti, stb. árfolyamadatok segítségével?
- Egy gazda földvásárlási dilemmája – egy bizonyos földterületen a várható termésátlag mennyire jelezhető előre a földterület fontosabb jellemzői alapján (a talaj kémhatása, a $CaCO_3$ megjelenési mélysége, a humusztartalom, topográfiai helyzet)?
- Meg lehet-e becsülni annak az esélyét, hogy valaki élete során megbetegszik tüdőrákban? Hogyan modellezzük ezt? Például: megbetegedés esélye $\longleftarrow$ dohányzik-e, hány éven át dohányzott élete során, van-e tüdőrákos a közeli rokonságban, van-e egyéb tüdőbetegsége, poros/füstös helyen dolgozik-e?

Regresszióelemzés – bevezetés

Legyenek $Y, X_1, \ldots, X_p$ véges szórású valószínűségi változók, amik egy véletlen jelenség egy-egy jellemzői.

A regresszióelemzés célja: a bennünket különösen érdeklő Y valószínűségi változó "minél jobb" közelítése az $X_1, \ldots, X_p$ valószínűségi változók segítségével.

Y elnevezései: eredményváltozó, függő változó, endogén változó
X_i-k elnevezései: magyarázó változók, független változók, exogén változók

Általában megfigyeléseink vannak, amik az $(Y, X_1, \ldots, X_p)^T$ valószínűségi vektorváltozó realizációinak tekinthetők:
$(y_i, x_{i,1}, \ldots, x_{i,p})^T \quad i = 1, 2, \ldots, n \qquad$ általában $n >> p$
Feltehetjük, hogy az y_i megfigyelések rendszerint mérési eredmények, amik sajnos pontatlanok. A mérési hibát ε_i-vel fogjuk jelölni, amiről természetes feltétel, hogy legyen 0 várható értékű és véges σ szórású valószínűségi változó.

Regresszióelemzés

Legyenek Y, X, $X_1, \ldots, X_p$ véges szórású valószínűségi változók, $c, a, b_1, \ldots, b_p$ valós számok.
Jelölje $\mathbf{X} = (X_1, \ldots, X_p)^T$, $\mathbf{b} = (b_1, \ldots, b_p)^T$ vektorokat.

	Feladat	Megoldás
a.)	$\min\limits_{c} E(Y-c)^2$	$\hat{c} = EY$
b.)	$\min\limits_{f:\mathbb{R}\to\mathbb{R} \text{ mérhető fv.}} E(Y-f(X))^2$	$\widehat{f}(X) = E(Y\|X)$
c.)	$\min\limits_{a,b} E(Y-(a+bX))^2$	$\hat{b} = \frac{cov(X,Y)}{D^2X}$, $\hat{a} = EY - \hat{b}EX$
d.)	$\min\limits_{f:\mathbb{R}^p\to\mathbb{R} \text{ mérhető fv.}} E(Y-f(X_1,\ldots,X_p))^2$	$\widehat{f}(X_1,\ldots,X_p) = E(Y\|X_1,\ldots,X_p)$
e.)	$\min\limits_{a,b_1,\ldots,b_p} E\left(Y-(a+\sum\limits_{i=1}^{p} b_iX_i)\right)^2$ [Többváltozós lineáris regresszió]	$\hat{\mathbf{b}} = (cov(\mathbf{X},\mathbf{X}))^{-1} cov(\mathbf{X}, Y)$ $\hat{a} = EY - \sum\limits_{i=1}^{p} \hat{b}_i EX_i$

$E(Y|X)$: feltételes várható érték

Eszköz: feltételes várható érték

- Adottak X, Y valószínűségi változók, amelyek között van összefüggés
- Szeretnénk Y-ból kinyerni minden X-re vonatkozó információt
- Az Y-ban lévő információt Y függvényei jelentik $\Leftrightarrow$ azt a $g(Y)$ valószínűségi változót szeretnénk meghatározni, amely leginkább hasonlít X-re, legközelebb van X-hez.
- A feladat általános megoldása egy alkalmas térben: nézzük a $g(Y)$-ok által kifeszített alteret és megkeressük az X merőleges vetületét erre az altérre $\Rightarrow$ ez lesz az X-hez legközelebbi $g(Y)$ alakú valószínűségi változó.
- Ez a vetület X-nek Y szerinti feltételes várható értéke: $g(Y) = E(X|Y)$

Feltételes várható érték

Definíció: Ha X diszkrét valószínűségi változó és $P(B) > 0$, akkor X feltételes várható értéke a B feltétel mellett: $X \sim \begin{matrix} x_1, x_2, \dots \\ p_1, p_2, \dots \end{matrix}$

$$E(X|B) = \sum_{i=1}^{\infty} x_i \cdot P(X = x_i|B)$$

Valószínűségi változó szerinti feltételes várható érték X, Y diszkrét valószínűségi változó $X = \begin{matrix} x_1, x_2, \dots \\ p_1, p_2, \dots \end{matrix}$, $Y = \begin{matrix} y_1, y_2, \dots \\ q_1, q_2, \dots \end{matrix}$, $EX < \infty$
$Y = y_i$ esemény és $P(Y = y_i) > 0$, így $E(X|Y = y_i)$ definiált a fentiek szerint.
Definíció: $E(X|Y = y) = g(y)$ egy függvény az Y értékkészletén az $y_i \to E(X|Y = y_i)$ hozzárendelés szerint.

Valószínűségi változó szerinti feltételes várható érték

Definíció: Ha $g(y) = E(X|Y = y)$ az X feltételes várható értéke az $Y = y$ feltétel mellett, akkor a $g(Y) = E(X|Y)$ az X feltételes várható értéke az Y feltétel (vagy adott Y) mellett. $E(X|Y)$ valószínűségi változó.

Megjegyzés: Ha $Y(\omega) = y_i$, akkor $E(X|Y)(\omega) = g(y_i) = g(Y(\omega))$

Tulajdonságok:

- Lineáris : $E(c \cdot X_1 + X_2|Y) = c \cdot E(X_1|Y) + E(X_2|Y)$
- Ha $X \geq Z \Rightarrow E(X|Y) \geq E(Z|Y)$
- Ha X, Y függetlenek: $E(X|Y) = EX$
- Ha $X = h(Y)$, akkor $E(X|Y = y_j) = h(y_j)$ ezért $E(X|Y = y) = h(y) \Rightarrow E(X|Y) = h(Y) = X$
- $E(E(X|Y)|Y) = E(g(Y)|Y) = g(Y) = E(X|Y)$

A feltételes várható érték kiszámítása

Az abszolút folytonos eset: legyen X, Y együttes sfv.-e $f(x,y)$. X, ill. Y sfv.-e: $\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y = f_X(x)$, $\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x = f_Y(y)$

Definíció: A feltételes sűrűségfüggvény:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x}$$

Definíció: A feltételes eloszlásfüggvény: $F(x|y) = \int_{-\infty}^{x} f(t|y)\mathrm{d}t$

Definíció: A feltételes valószínűség y függvénye:

$$P(X \in A|\, Y = y) = \int_A f(t|y)\mathrm{d}t$$

Teljes valószínűség tétele

Állítás:

$$P(X \in A, Y \in B) == \int_A \int_B f(x|y) f_Y(y) \mathrm{d}y \mathrm{d}x =$$

$$\overset{\text{integrál csere}}{=} \int_B P(X \in A | Y = y) f_Y(y) \mathrm{d}y$$

Legyen $g(y) = \int_{-\infty}^{+\infty} x \cdot f(x|y) \mathrm{d}x$, akkor $g(Y) = E(X|Y)$

Következmény: Beírva $f(x|y)$ definícióját is kapjuk, hogy

$$g(y) = \int_{-\infty}^{+\infty} x \cdot \frac{f(x,y)}{f_Y(y)} \mathrm{d}x = \int_{-\infty}^{+\infty} \frac{x \cdot f(x,y)}{\int_{-\infty}^{+\infty} f(t,y) \mathrm{d}t} \mathrm{d}x$$

Példák

E47.) Legyen X és Y együttes sűrűségfüggvénye $h(x,y) = \exp(-y)$, ha $0 < x < y$, és 0 máshol. $E(X|Y) =?$

E48.) Legyen X és Y együttes sűrűségfüggvénye

$$h(x,y) = \frac{12}{5}(x+y) \text{ ha} 0 < \frac{x}{2} \leq y \leq 1 - \frac{x}{2}$$

és 0 különben. $E(X|Y) =?$