# Villamosmérnök A4 — 11. gyakorlat (2013. 04. 25-26.)
# Kétdimenziós normális eloszlás, regressziók

## Kétdimenziós normális összefoglalás

Egy kétdimenziós $(X,Y)$ valószínűségi változó **kovariancia mátrixa**:

$$\Sigma = \begin{pmatrix} \mathbb{V}\mathrm{ar}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathbb{V}\mathrm{ar}(Y) \end{pmatrix}$$

**Korrelációs együttható**: $\mathrm{r}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathbb{SD}(X)\mathbb{SD}(Y)}$.

**Kétdimenziós normális eloszlás**: Standard kétdimenziós normális eloszlású egy $(U,V)$ pár, ha sűrűségfüggvénye

$$\varphi(u,v) := \frac{1}{2\pi}\exp\left(-\frac{u^2+v^2}{2}\right) \quad (u,v\in\mathbb{R}).$$

$(U,V)$ zérus várható érték vektorú, egységmátrix kovarianciájú pár. Azt mondjuk, hogy az $(X,Y) = (U,V)\mathbf{A}+\mu$ pár kétdimenziós normális eloszlású $\mu\in\mathbb{R}^2$ várható érték vektorral és $\Sigma\in\mathbb{R}^{2\times 2}$ (invertálható) kovariancia mátrixszal, ha $(U,V)$ standard normális pár, és $\mathbf{A}^\top\mathbf{A} = \Sigma$, ahol $\mathbf{A}\in\mathbb{R}^{2\times 2}$.
Egy kétdimenziós normális $(X,Y)$ pár sűrűségfüggvénye:

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

ahol $\mathbb{E}X=\mu_X$, $\mathbb{E}Y=\mu_Y$, $\mathbb{SD}(X)=\sigma_X$, $\mathbb{SD}(Y)=\sigma_Y$, és $r=r(X,Y)$ $X$ és $Y$ korrelációs egyyütthatója. Kétdimenziós normális eloszlás esetében a perem- illetve feltételes eloszlások is normálisok: a marginálisok eloszlása $N(\mu_X,\sigma_X)$ és $N(\mu_Y,\sigma_Y)$, a feltételes eloszlások pedig $N(\mu_{X|Y=y},\sigma_{X|Y=y})$, ahol

$$\mu_{X|Y=y} = \mu_X + (y-\mu_Y)\,r\,\frac{\sigma_X}{\sigma_Y} \qquad \text{és} \qquad \sigma_{X|Y=y} = \sqrt{1-r^2}\,\sigma_X\,.$$

## Regressziók összefoglalás

Adott egy $X$ valószínűségi változó, mely $c$ érték(ek)re lesz a $h_1(c) := \mathbb{E}|c-X|$ **hiba minimális**? A válasz: az $m(X)$ **mediánra** (amiből lehet több is, diszkrét változó esetén). És mely $c$ értékre lesz $h_2(c) := \mathbb{E}\big[(c-X)^2\big]$ minimális? A válasz: az $\mathbb{E}X$ **várható értékre** (Steiner-tétel, tanultuk.)

Hasonlóképpen, ha az $X_1,\dots,X_n$ független kísérleteket látjuk egy ismeretlen eloszlású valószínűségi változóra, és ezek alapján akarjuk becsülni a változót, akkor:

(a) az a $c$, amire $h_1(c) := \sum_{i=1}^n |c-X_i|$ a minimális, az a **minta** $m(X_1,\dots,X_n)$ **mediánja**, azaz a nagyságra középső érték a kísérletekből (páros $n$ esetén a két középső érték között bármi), és persze nagy $n$ esetén ez egy jó becslés lesz a valószínűségi változó mediánjára; illetve

(b) az a $c$, amire $h_2(c) := \sum_{i=1}^n (c-X_i)^2$ a minimális, az a **mintaátlag** $\mu(X_1,\dots,X_n) := (X_1+\dots+X_n)/n$, és ez egy jó becslés lesz a valószínűségi változó várható értékére.

Izgalmasabb, amikor egy kétdimenziós $(X,Y)$ valószínűségi változóból látunk független kísérleteket, és ezek alapján értenénk meg, hogyan függ $Y$ az $X$-től; pontosabban, $X$ **milyen függvényével tudnánk $Y$-t a legjobban becsülni**? A válasz függ attól, hogyan mérjük, milyen jó a becslésünk:

(a) Akkor lesz a $h_1(f) := \mathbb{E}|f(X)-Y|$ hiba minimális, ha $\mathbb{E}\Big[|f(x)-Y|\,\Big|\,X=x\Big]$-et minimalizáljuk minden rögzített $x$-re, azaz $f(x)$ az $Y$ **feltételes mediánja** az $X=x$ feltétel mellett.

(b) Akkor lesz a $h_2(f) := \mathbb{E}\big[(f(X)-Y)^2\big]$ hiba minimális, ha $\mathbb{E}\Big[(f(x)-Y)^2\,\Big|\,X=x\Big]$-et minimalizáljuk, azaz $f(x)$ az $Y$ **feltételes várható értéke** az $X=x$ feltétel mellett.

Ha csak **lineáris** $f$ függvényeket engedünk meg, akkor a $h_2(f)$ négyzetes hibát az **első regressziós egyenes** minimalizálja: $f(x) = \mu_2 + r\,(x-\mu_1)\,\sigma_2/\sigma_1$, ahol $\mu_1$ és $\sigma_1$ az $X$ várható értéke és szórása, $\mu_2$ és $\sigma_2$ az $Y$-éi, $r$ pedig $X$ és $Y$ korrelációs együtthatója. A **második regressziós egyenes** pedig azon $g$ lineáris függvény, mely az $\mathbb{E}\big[(g(Y)-X)^2\big]$ hibát minimalizálja: $g(y) = \mu_1 + r\,(y-\mu_2)\,\sigma_1/\sigma_2$.

Ezek nem csak azért fontosak, mert a lineáris összefüggéseket fogadja be a legkönnyebben az értelmünk, hanem mert a $\mu_i,\sigma_i,r$ értékeket természetes módon becsülhetjük egy $(X_1,Y_1),\dots,(X_n,Y_n)$ adathalmazból. Mégpedig: a $\mu(X)$ mintaátlagot már feljebb definiáltuk, a **minta varianciája** pedig $\sum_{i=1}^n (X_i-\mu(X))^2/(n-1)$; kovariancia hasonlóan. Az $n-1$-gyel osztás $n$ helyett nem nyomdahiba, hanem így lesz $\mathbb{E}\sum_{i=1}^n (X_i-\mu(X))^2/(n-1) = \mathbb{V}\mathrm{ar}(X)$, ha utánaszámolunk.

Láttuk fent, hogy **kétdimenziós normális** eloszlásokra a feltételes eloszlás normális, így mediánja és várható értéke megegyezik, ráadásul lineáris függvénye a feltételnek, így megegyezik a regressziós egyenessel.

**Kétdimenziós normális feladatok**

1. Tegyük fel, hogy egy jólmenő étterem heti összbevétele normális eloszlást követ 1 millió forint várható haszonnal, és 700000 forint szórással. Mi annak a valószínűsége, hogy kevesebb, mint 1.5 millió forint a bevétele két, egymást követő héten? Itt tegyünk még fel függetlenséget! Majd nézzük meg, hogyan változik annak a valószínűsége, hogy a második héten több mint 2 millió forint a bevétel feltéve, hogy az első héten 1.5 millió forint volt a bevétel és a korreláció $-0.5$! Mi a második hét várható bevétele ugyanezen feltétel mellett? Mennyi a két hét várható összbevétele? Mi a szórás?

2. Budapesten májusban az átlagos hőmérséklet 25 °C, 7 °C szórással, valamint az átlagnyomás $10^5$ Pa, $2\times10^4$ Pa szórással. A hőmérséklet/nyomás változása szoros összhangban van, köztük lévő korreláció 0.7. Írjuk fel a kovariancia mátrixot majd határozzuk meg a következőket:

   (a) Mi a valószínűsége annak, hogy egy nap melegebb lesz, mint 40 °C? És, hogy alacsonyabb a nyomás $6\times10^4$ Pa-nál?

   (b) Egy nap 20 °C-ot mértünk. Mi annak a valószínűsége, hogy a légnyomás $1.2\times10^5$ Pa fölött járt? Átlagosan mekkora volt a légnyomás? Mekkora a szórás?

   (c) Feltéve, hogy egy nap $10^5$ Pa volt a légnyomás, mi annak a valószínűsége, hogy melegebb volt, mint 35 °C? Átlagosan hány fok volt aznap? Mekkora a szórás?

3. Magyarországon a felnőtt férfiak testmagassága átlagosan 178 cm, 9 cm szórással, míg testsúlyuk 85 kg, 10 kg szórással. A korrelációs együttható 0.7, azaz minél magasabb valaki, annál súlyosabb is. Írjuk fel a kovariancia mátrixot!

   (a) Mi a valószínűsége annak, hogy egy férfi magasabb 2 méternél? És, hogy nehezebb 100 kg-nál?

   (b) Feltéve, hogy egy férfi 80 kg, mi annak a valószínűsége, hogy magasabb, mint 180 cm? Várhatóan hány cm magas egy ilyen férfi? Mekkora a szórás?

   (c) Átlagosan mekkora súlyú egy 190 cm magas férfi?

   (d) Átlagosan milyen magas egy 94.3 kg-os férfi?

   (e) Hasonlítsuk össze az utolsó két eredményt.

4. Az előadásbeli példát folytatva, az $X$ áramerősség normális $N(230, 6)$ eloszlású, a mérőkészülék $Z$ hibája ettől független $N(0, 8)$ eloszlású, mi az $Y = X + Z$ értéket mérjük. Mi a valószínűsége, hogy $Z > X/20$?

5. Legyen a $(X, Y)$ pár kétdimenziós normális eloszlású, $r$ korrelációval. Mi az eloszlása $U = X + Y$-nak és $V = X - Y$-nak? Független-e $U$ a $V$-től? Számoljuk ki a várható értékeket és szórásokat is!

6. Legyen $U$ és $Z$ független valószínűségi változók, előbbi egyenletes a $(0, 2\pi)$ intervallumban, utóbbi exponenciális, 1 várható értékkel. Mi az eloszlása az $X = \sqrt{2Z}\cos(U)$, illetve $Y = \sqrt{2Z}\sin(U)$ valószínűségi változónak? Független-e $X$ $Y$-tól?

7. Hogyan generálna le kétdimenziós normális eloszlású véletlen pontokat a síkon, melyek várható értéke $(\mu_1, \mu_2)$, szórása $(\sigma_1, \sigma_2)$, korrelációs együtthatója pedig $r$. Függetlenek-e a koordináták, ha $r = 0$?

**Regressziós feladatok**

8. Vegyük a $4, 6, 1, 4, 13, 5$ adathalmazt (más néven mintát).

   (a) Határozzuk meg a $h_1(c)$ hibafüggvényt és a minta mediánjait!

   (b) Határozzuk meg a $h_2(c)$ hibafüggvényt és a mintaátlagot!

9. Egy kétdimenziós háromelemű mintánk első koordinátái $-1, 0, 1$, második koordinátái $3, 4, 5$, valamilyen sorrendben. Világos, hogy $3! = 6$-féleképpen lehet összepárosítani a koordinátákat. A koordinátákkénti minta-mediánok, -átlagok, és -szórások persze nem függenek a párosítástól. Mik ezek a koordinátákkénti értékek? És mi a korrelációs együttható a 6 lehetséges párosításban?

10. Egy tízfős A4 csoportban, az $i$-edik diák első hét röpZH eredményének összegét jelölje $X_i$, első nagyZH-jának eredményét pedig $Y_i$. Az eredmények: $(21, 13)$, $(25, 28)$, $(19, 23)$, $(30.5, 26)$, $(28.5, 24)$, $(19, 15)$, $(27, 21)$, $(23, 27)$, $(33, 27.5)$, $(16.5, 17)$.

    (a) Határozzuk meg az $X$ és $Y$ minták átlagait, szórásait, mediánjait, és korrelációs együtthatójukat!

    (b) Írjuk föl a minta két regressziós egyenesét! Mennyire tűnik jónak az adatok alapján a lineáris közelítés, és mennyire gondoljuk, hogy elvileg lineárisnak kellene lennie az összefüggésnek?

    (c) Kiderül, hogy volt még egy láthatatlan diák is a csoportban, akinek a nagyZH-ja 25 pontos lett. Milyen röpZH összpontszámot tippelünk neki? És ha az derült volna ki, hogy a röpZH összpontszáma 26, akkor milyen nagyZH pontszámot tippelnénk?

11. (a) Kétszer dobtunk egy kockával, a dobások összege 10. Mi az első dobás feltételes várható értéke? És mit tippelünk az első dobásra?

    (b) Legyen $X$ két dobás összege, $Y$ pedig az első dobás. Határozzuk meg a regressziós egyenest!

    (c) Tízszer dobtunk egy kockával, a dobások összege 50. Most mi az első dobás feltételes várható értéke? És mit tippelünk az első dobásra? És mi a regressziós egyenes?

12. Legyenek $X_1, \dots, X_k$ független RAND() számok, minimumuk $X$, maximumuk $Y$. Határozzuk meg az $Y$ feltételes mediánját és várható értékét az $X = x$ feltétel mellett, és az első regressziós egyenest,

    (a) $k = 2$-re;

    (b) általános $k$-ra.

13. Legyen az $(X, Y)$ kétdimenziós valószínűségi változó együttes sűrűségfüggvénye:

    (a)
    $$f(x,y) = \begin{cases} 2\exp(-(x+2y)), & \text{ha } 0 \le x, y; \\ 0, & \text{egyébként.} \end{cases}$$

    (b)
    $$f(x,y) = \begin{cases} x+y, & \text{ha } 0 < x < 1;\ 0 < y < 1; \\ 0, & \text{egyébként.} \end{cases}$$

    (c)
    $$f(x,y) = \begin{cases} 24xy, & \text{ha } 0 \le x;\ 0 \le y \text{ és } 0 \le x+y \le 1; \\ 0, & \text{egyébként.} \end{cases}$$

    Határozzuk meg az $Y$ feltételes mediánját és várható értékét az $X = x$ feltétel mellett, és az első regressziós egyenest.

14. Egy hivatalban minden ügyfél kiszolgálása 10 perc várható értékű exponenciális valószínűségi változót vesz igénybe, egymástól függetlenül. Ha $k$ ügyfelet összesen $x$ idő alatt szolgálnak ki, akkor a legelső ügyfél kiszolgálásának mi a feltételes mediánja és várható értéke, mi a regressziós egyenes?

15. (a) Legyen $U$ egy egyenletes véletlen szám a $[0, 1]$ intervallumból, $X = U^2$ és $Y = U^3$. Mi $X$ és $Y$ korrelációs együtthatója? Mi az $\mathbb{E}[Y \mid X = x]$ feltételes várható érték és az $\sqrt{\mathbb{E}\big[(Y - \mathbb{E}[Y \mid X = x])^2 \,\big|\, X = x\big]}$ feltételes szórás? Határozzuk meg az első regressziós egyenest.

    (b) Most legyen $U$ egy egyenletes véletlen szám a $[0, 2]$ intervallumból, és, mint az előbb, $X = U^2$ és $Y = U^3$. Változott-e a korrelációs együttható?