Segédanyag az *Idősorok és többdimenziós statisztika gyakorlat* (matematika BSc, matematikai elemző szakirány) tantárgyhoz.
2018. december 2.

Riemann-Stieltjes integrál és várható érték

A Riemann-Stieltjes (vagy néha csak Stieltjes-) integrál a Riemann-integrál kiterjesztése, amikor egy függvény határozott integrálját egy másik függvény megváltozása szerint számítjuk ki. Felépítése nagyon hasonlóan történhet, mint a Riemann-integrálé, és a tulajdonságaik is nagyon hasonlóak. Segítségével általánosabb definíciót lehet adni a valószínűségi változók várható értékére.

Definíció. **Függvény teljes variációja.**
Legyen $f : [a;b] \to \mathbb{R}$, $\mathcal{F}_n = \{x_0, x_1, \dots, x_n\}$ az $[a;b]$ intervallum egy tetszőleges feloszlása. Jelölje a felosztás finomságát $\delta(\mathcal{F}_n) := \max_{1\le i\le n}(x_i - x_{i-1})$.

f teljes variációja $[a;b]$-n: $V_a^b(f) = \lim_{\delta(\mathcal{F}_n)\to 0} \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|$.

Definíció. **Korlátos változású függvény.**
Az $f : [a;b] \to \mathbb{R}$ függvény korlátos változású, ha $V_a^b(f) < \infty$.
Jel.: $f \in BV[a;b]$

Tétel. **Korlátos változású függvények reprezentációja.**
$f \in BV[a;b] \iff \exists g, h : [a;b] \to \mathbb{R}$ monoton növő függvények, hogy $f = g - h$

Jel.: f Riemann-integrálható $[a;b]$-n $\iff f \in R[a;b]$

Definíció. **Riemann-Stieltjes integrál.**
Legyenek $f, g : [a;b] \to \mathbb{R}$, $\mathcal{F}_n = \{x_0, x_1, \dots, x_n\}$ az $[a;b]$ intervallum egy tetszőleges feloszlása, $\mu_k \in [x_{k-1}, x_k]$ tetszőleges pontok, $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_n\}$.

Riemann-Stieltjes integrálközelítő összeg: $R(f, g, \boldsymbol{\mu}) := \sum_{i=1}^{n} f(\mu_i)[g(x_i) - g(x_{i-1})]$

f Riemann-Stieltjes integrálható g-re nézve, ha $R(f, g, \boldsymbol{\mu})$ minden $\delta(\mathcal{F}_n) \to 0$ felosztássorozat esetén konvergens.

Jel.: $\int_a^b f\,dg = \int_a^b f(x)\,dg(x) = \int_{[a;b]} f\,dg$

Jel.: f Riemann-Stieltjes integrálható $[a;b]$-n $\iff f \in RS[a;b]$

Megjegyzés. Az f neve integrandus, a g neve integrátor

Megjegyzés. A Riemann-Stieltjes integrál hasonlóan kiterjeszthető improprius értelemben, mint a Riemann-integrál.

Tétel. Ha $f \in C[a;b]$ és $g \in BV[a;b]$, akkor $f \in RS([a;b], g)$.

Állítás. A R-S integrál nem létezik, ha f és g ugyanabban a pontban ugranak.

Állítás. **A R-S integrál tulajdonságai.**
Tegyük fel, hogy $f_i \in RS([a;b], g_i)$, $i = 1, 2$, $a < b < c \in \mathbb{R}$. Ekkor

a.) $\int_a^b (cf_1 + f_2)\,dg_1 = c\int_a^b f_1\,dg_1 + \int_a^b f_2\,dg_1$

b.) $\int_a^b f_1\,d(cg_1 + g_2) = c\int_a^b f_1\,dg_1 + \int_a^b f_1\,dg_2$

c.) $\int_a^c f\,dg = \int_a^b f\,dg + \int_b^c f\,dg$, amennyiben az itt szereplő mindhárom integrál létezik.

A következő két tétel a Riemann-Stieltjes integrál kiszámítását abban a két szélsőséges esetben mutatja be, amikor a g integrátor tiszta ugrófüggvény, illetve folytonosan differenciálható. Mivel minden (nem szinguláris) $g(x)$ valós függvény felbontható egy g_1 tiszta ugrófüggvény és egy g_2 folytonosan differenciálható függvény $g = g_1 + g_2$ összegére, így a g szerinti integrálása a fenti állítás b.) része alapján szétbontással könnyedén elvégezhető.

Tétel. Legyen $f \in R[a;b]$. Ha

a.) g tiszta ugrófüggvény az a_i pontokban c_i nagyságút ugrik, azaz $g(x) = \sum_{i=1}^{n} c_i I(x \le a_i)$, akkor $\int_a^b f\,dg = \sum_{i=1}^{n} f(a_i)\cdot c_i$

b.) $g \in C^1[a;b]$, akkor $\int_a^b f\,dg = \int_a^b f(x)g'(x)\,dx$

Végül definiálhatjuk a valószínűségi változók várható értékét.
Definíció. Legyen X valószínűségi változó F eloszlásfüggvénnyel. Ekkor X várható értéke $EX = \int_{-\infty}^{\infty} x\,dF(x)$.

Állítás. Legyen X valószínűségi változó, $g : \mathbb{R} \to \mathbb{R}$ "szép" függvény ("szép"=Borel-mérhető). Ekkor $E(g(X)) = \int_{-\infty}^{\infty} g(x)\,dF(x)$.

Megjegyzés. A standard valószínűségelméletben a várható érték legáltalánosabb definíciója az alábbi: $EX = \int_\Omega X\,dP = \int_{\omega\in\Omega} X(\omega)\,dP(\omega)$, ahol X valószínűségi változó az $(\Omega, \mathcal{A}, P)$ Kolmogorov-féle valószínűségi téren van értelmezve, a P a valószínűségi mérték. Ebben az értelemben tehát a várható érték a P valószínűségi mérték szerinti integrál. A mérték szerinti integrál felépítését és fontosabb tulajdonságait mértékelmélet tankönyvekben lehet megtalálni.

Ajánlott irodalom: Richardson: Advanced Calculus, 7. fejezet

Definíció. X **val.változó eloszlásfüggvénye:** $F_X(x) = P(X < x)$.

Állítás. Az eloszlásfüggvény tulajdonságai:
- $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$;
- balról folytonos;
- monoton növő.

Nevezetes diszkrét eloszlások:

Eloszlás neve	Jelölése	Eloszlása	EX	$\mathrm{D}^2\mathrm{X}$
Karakterisztikus (indikátorvált.)	$\mathrm{Ind}(p)$	$P(X=1)=p$ $P(X=0)=1-p$	p	$p(1-p)$
Geometriai (Pascal)	$\mathrm{Geo}(p)$	$P(X=k)=p(1-p)^{k-1}$ $k=1,2,\dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hipergeometriai	$\mathrm{Hipgeo}(N,M,n)$	$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ $k=0,1,\dots,n$	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n-1}{N-1}\right)$
Binomiális	$\mathrm{Bin}(n,p)$	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k=0,1,\dots,n$	np	$np(1-p)$
Negatív binomiális	$\mathrm{NegBin}(n,p)$	$P(X=k)=\binom{k-1}{n-1}p^n(1-p)^{k-n}$ $k=n,n+1,\dots$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poisson	$\mathrm{Poi}(\lambda)$	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$ $k=0,1,\dots$	λ	λ

Nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	$\mathrm{D}^2\mathrm{X}$
Egyenletes	$\mathrm{E}(a,b)$	$\begin{cases}0 & \text{ha } x\le a\\ \frac{x-a}{b-a} & \text{ha } a<x\le b\\ 1 & \text{ha } b<x\end{cases}$	$\begin{cases}\frac{1}{b-a} & \text{ha } a<x\le b\\ 0 & \text{különben}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponenciális	$\mathrm{Exp}(\lambda)$	$\begin{cases}1-e^{-\lambda x} & \text{ha } x\ge 0\\ 0 & \text{különben}\end{cases}$	$\begin{cases}\lambda e^{-\lambda x} & \text{ha } x\ge 0\\ 0 & \text{különben}\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Standard norm.	$\mathrm{N}(0,1^2)$	$\Phi(x)=\dots$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\quad x\in\mathbb{R}$	0	1
Normális	$\mathrm{N}(m,\sigma^2)$	$\dots$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}\quad x\in\mathbb{R}$	m	σ^2

További nevezetes abszolút folytonos eloszlások:

Eloszlás neve	Jelölése	Eloszlásfüggvény	Sűrűségfüggvény	EX	$\mathrm{D}^2\mathrm{X}$
Cauchy	$Cauchy(a,b)$ $a\in\mathbb{R}, b>0$	$\frac{1}{\pi}\mathrm{arctg}\left(\frac{x-a}{b}\right)+\frac{1}{2}$	$\frac{1}{\pi b\left[1+\left(\frac{x-a}{b}\right)^2\right]}\quad x\in\mathbb{R}$	$\nexists$	$\nexists$
Pareto*	$Pareto(\alpha,\beta)$ $a,b>0$	$\begin{cases}1-\left(\frac{\beta}{x}\right)^{\alpha} & \text{ha } x\ge\beta\\ 0 & \text{ha } x<\beta\end{cases}$	$\begin{cases}\frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1} & \text{ha } x\ge\beta\\ 0 & \text{ha } x<\beta\end{cases}$	$\frac{\alpha\beta}{\alpha-1}$	$\frac{\beta^2\alpha}{(\alpha-1)^2(\alpha-2)}$

* A Pareto-eloszlásnak akkor van véges várható értéke a képletnek megfelelően, ha $\alpha>1$, szórásnégyzete pedig akkor, ha $\alpha>2$.

Eloszlás neve	Jelölése	Eloszlás-függvény	Sűrűségfüggvény	EX	$\mathrm{D}^2\mathrm{X}$
Khí-négyzet	$\chi^2_k\quad k\in\mathbb{N}$	$\dots$	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}\quad x\in\mathbb{R}$	k	$2k$
Gamma	$\Gamma(\alpha,\lambda)\quad\alpha,\lambda>0$	$\dots$	$\begin{cases}\frac{1}{\Gamma(\alpha)}\lambda^{\alpha}e^{-\lambda x}x^{\alpha-1} & \text{ha } x\ge 0\\ 0 & \text{ha } x<0\end{cases}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Béta	$Beta(\alpha,\beta)$ $\alpha,\beta>0$	$\dots$	$\begin{cases}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} & x\in[0;1]\\ 0 & \text{különben}\end{cases}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Lognormális	$\mathrm{LN}(m,\sigma^2)$ $m\in\mathbb{R},\sigma>0$	$\dots$	$\begin{cases}\frac{1}{x\sqrt{2\pi}\sigma}e^{-\frac{(\log x-m)^2}{2\sigma^2}} & \text{ha } x\le 0\\ 0 & \text{ha} x<0\end{cases}$	$e^{m+\sigma^2/2}$	$(e^{\sigma^2-1})e^{2m+\sigma^2}$

Többváltozós valószínűségszámítás

Definíció. **Valószínűségi vektorváltozó:** $\mathbf{X}\colon \Omega\to\mathbb{R}^d$ (Borel-)mérhető függvény, azaz amire $\{\omega:\mathbf{X}(\omega)\in B\}\in\mathcal{A}$ minden $B\subseteq\mathbb{R}^d$ nyílt (Borel-)halmazra.

Definíció. **$\mathbf{X}$ valószínűségi vektorváltozó eloszlásfüggvénye:**
$F_{\mathbf{X}}(\mathbf{x})=P(\mathbf{X}<\mathbf{x})=P(X_1<x_1,\dots,X_d<x_d)$.

Definíció. **$\mathbf{X}$ valószínűségi vektorváltozó abszolút folytonos** , ha létezik olyan $f_{\mathbf{X}}(x_1,\dots,x_d)$ függvény, amelyre
$F_{\mathbf{X}}(x_1,\dots,x_d)=\int_{-\infty}^{x_1}\dots\int_{-\infty}^{x_d}f_{\mathbf{X}}(t_1,\dots,t_d)\,dt_1\dots dt_d$.
Ilyenkor $f_{\mathbf{X}}(\mathbf{x})$-et **sűrűségfüggvény**nek hívjuk.

$d=2$ esetén vezessük be a következő jelöléseket és elnevezéseket:

- $F_{X,Y}(x,y)=P(X<x,Y<y)\quad\rightsquigarrow$ együttes eloszlásfüggvény
- $F_X(x)=P(X<x)$, $F_Y(y)=P(Y<y)\quad\rightsquigarrow$ peremeloszlásfüggvények
- $f_{X,Y}(x,y)\quad\rightsquigarrow$ együttes sűrűségfüggvény
- $f_X(x), f_Y(y)\quad\rightsquigarrow$ peremsűrűségfüggvények
  - $F_X(x)=\lim\limits_{y\to\infty}F_{X,Y}(x,y)\quad$ és $\quad F_Y(y)=\lim\limits_{x\to\infty}F_{X,Y}(x,y)$

Állítás.
- $F_{X,Y}(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x}f_{X,Y}(u,v)\,dudv$
- $f_{X,Y}(x,y)=\partial_y\partial_x F_{X,Y}(x,y)=\partial_x\partial_y F_{X,Y}(x,y)$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f_{X,Y}(x,y)\,dxdy=1$
- $f_X(x)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)\,dy\quad$ és $\quad f_Y(y)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)\,dx$

Állítás. Legyen (X,Y) abszolút folytonos, $A\subseteq\mathbb{R}$, $B\subseteq\mathbb{R}^2$ mérhető halmazok.
- $P(X\in A)=\int_{x\in A}dF_X(x)=\int_{x\in A}f_X(x)\,dx$
- $P((X,Y)\in B)=\iint_{(x,y)\in A}dF_{X,Y}(x,y)=\iint_{(x,y)\in A}f_{X,Y}(x,y)\,d(x,y)$

Állítás.
- X,Y függetlenek $\quad\Leftrightarrow\quad F_{X,Y}(x,y)=F_X(x)\cdot F_Y(y)$
- X,Y függetlenek $\quad\Leftrightarrow\quad f_{X,Y}(x,y)=f_X(x)\cdot f_Y(y)$
- X,Y függetlenek $\quad\Leftrightarrow\quad P(X=x,Y=y)=P(X=x)\cdot P(Y=y)$
- X,Y függetlenek $\quad\Rightarrow\quad E(XY)=EX\cdot EY$

Definíció. **X és Y kovarianciája:** $\mathrm{Cov}(X,Y)=E\left[(X-EX)(Y-EY)\right]$.
Köv.: $\mathrm{Cov}(X,Y)=E(XY)-EXEY$.
Elnevezés: ha $\mathrm{Cov}(X,Y)=0$, akkor azt mondjuk, hogy X és Y **korrelálatlanok**.

Állítás.
- X és Y függetlenek $\implies X$ és Y korrelálatlanok
- X és Y korrelálatlanok $\not\Longrightarrow X$ és Y függetlenek !!!!!

Definíció. X **és** Y **lineáris korrelációja:** $Cor(X,Y)=\begin{cases}\frac{\mathrm{Cov}(X,Y)}{DX\,DY} & \text{ha } DX, DY>0\\ 0 & \text{ha } DX=0 \text{ v. } DY=0\end{cases}$

Ez a Pearson-féle lineáris korreláció két valószínűségi változó közti *lineáris* kapcsolat irányát és erősségét méri.

Definíció. **Kovarianciamátrix.** Legyen $\mathbf{X}$ valószínűségi vektorváltozó. Ekkor $\Sigma(\mathbf{X}) := E(\mathbf{X}\cdot\mathbf{X}^T) - E(\mathbf{X})E(\mathbf{X})^T$

A többdimenziós adatelemzés lényeges eszköze a korrelációs mátrix, aminek (i,j)-edik eleme az $R(X_i,X_j)$ lineáris korrelációs együttható. A korrelációs mátrix átlójában csupa 1-ek szerepelnek.

A többdimenziós normális és az egyenletes a gyakorlatban legtöbbször előforduló abszolút folytonos többdimenziós valószínűségi változók.

Ha $\mathbf{X}$ d **dimenziós** nem-elfajuló **normális eloszlás**t követ $\mathbf{m}$ várható érték vektorral és $\mathbf{\Sigma} > 0$ kovarianciamátrixszal (jel.: $\mathbf{X} \sim N_d(\mathbf{m},\mathbf{\Sigma})$), akkor sűrűségfüggvénye:

$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{d}{2}}|\det(\mathbf{\Sigma})|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right\}$.

$\mathbf{X}$ **egyenletes eloszlás**t követ a d dimenziós tér $B\subseteq\mathbb{R}^d$ mérhető részhalmazán (jel.: $\mathbf{X}\sim E(B)$), ha sűrűségfüggvénye:

$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases}\frac{1}{t(B)} & \text{ha } \mathbf{x}\in B\\ 0 & \text{egyébként}\end{cases}$ ahol a $t(\cdot)$ függvény a d dimenziós térfogatot jelöli

Tétel. **Valószínűségi vektorváltozó transzformáltjának sűrűségfüggvénye.** Legyen $\mathbf{X}=(X_1,\dots,X_n)$ abszolút folytonos valószínűségi vektorváltozó $f_{\mathbf{X}}$ sűrűségfüggvénnyel, $A\subseteq\mathbb{R}^n$ összefüggő és nyílt halmaz. Legyen $\mathbf{g}: A\to A$ függvény, amely invertálható és inverze folytonosan differenciálható. Legyen $\mathbf{Y}=\mathbf{g}(\mathbf{X})$, $\mathbf{J}=\partial_{\mathbf{y}}\mathbf{g}^{-1}(\mathbf{y})$ a Jacobi-mátrix. Ekkor

$f_{\mathbf{g}(\mathbf{X})}(\mathbf{y}) = |\det(\mathbf{J})|\cdot f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}))$

Ajánlott irodalom: Márkus L. előadásfóliái a többdimenziós normális eloszlásról: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/Tobbdim_normalis_elo.pdf`

Feltételes várható érték

Legyen a valószínűségi mező a szokásos $(\Omega,\mathcal{A},P)$ hármas, $\mathcal{F}\subset\mathcal{A}$ σ-algebra.

Definíció. $\mathcal{F}$**-mérhetőség.**
Az $X\colon\Omega\to\mathbb{R}$ valószínűségi változó $\mathcal{F}$-mérhető, ha minden $B\subseteq\mathbb{R}$ Borel-halmazra $X^{-1}(B)\in\mathcal{F}$.

Definíció. X **feltételes várható értéke** $\mathcal{F}$**-re nézve.**
Legyen X integrálható. Az $Y:=E(X|\mathcal{F})$ az a valószínűségi változó, amelyre egyrészt Y $\mathcal{F}$-mérhető, másrészt $\forall B\in\mathcal{F}$ halmazra $\int\limits_B X\,dP = \int\limits_B Y\,dP$.

Speciálisan, ha $\mathcal{F}=\sigma(Y)$, azaz $\mathcal{F}$-et az Y valószínűségi változó generálja, akkor $E(X|\mathcal{F})$ helyett $E(X|Y)$-t írunk.

Tehát $E(X|Y)$-ra úgy gondolunk, mint egy valószínűségi változóra, konkrétabban az Y valószínűségi változó egy mérhető $h(Y)$ függvényére; és ha Y egy adott értéket vesz fel, azaz ha $E(X|Y=y)$, akkor mint konkrét számra.

Abszolút folytonos eloszlások esetén a következő képlettel számítható:

$$\mathrm{E}(g(X)|Y) = \int\limits_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)dx\Big|_{y=Y}$$

ahol $f_{X|Y}(x|y) = \begin{cases}\frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{ha } f_Y(y)>0\\ 0 & \text{különben}\end{cases}$ a feltételes sűrűségfüggvény.

Definíció. σ**-algebrától való függetlenség.**
X valószínűségi változó független az $\mathcal{F}$ σ-algebrától, ha $\forall A\in\sigma(X)$ és $\forall B\in\mathcal{F}$ eseményekre $P(A\cap B)=P(A)P(B)$.

Állítás. Tulajdonságok. Legyen g $\mathcal{F}$-mérhető függvény.

- $E(X|\mathcal{F})$ 1 valószínűséggel egyértelműen létezik
- $E(E(X|\mathcal{F})) = EX$ $\rightsquigarrow$ teljes várható érték tétel (TVÉT)
- X $\mathcal{F}$-mérhető $\Rightarrow$ $E[g(X)|\mathcal{F}] = g(X)$
- X független $\mathcal{F}$-től $\Rightarrow$ $E(X|\mathcal{F}) = EX$
- X $\mathcal{F}$-mérhető $\Rightarrow$ $E(XY|\mathcal{F}) = XE(Y|\mathcal{F})$

Állítás. Ha X független $\mathcal{F}$-től, Y mérhető $\mathcal{F}$-re nézve, $g(X,Y)$ integrálható, akkor $E(g(X,Y)|\mathcal{F}) = E(g(X,y))|_{y=Y}$.

Állítás. **Teljes valószínűség tétele folytonos esetben.**
Legyen A tetszőleges esemény, Y abszolút folytonos valószínűségi változó. Ekkor $P(A)=\int\limits_{\infty}^{\infty} P(A|Y=y)f_Y(y)\,dy$.

Feladat: Y-t szeretnénk közelíteni X tetszőleges függvénye segítségével:

$$E[Y-h(X)]^2 \longrightarrow \min_h \qquad \rightsquigarrow \text{Megoldása: } h_{opt}=E(Y|X)$$

Ajánlott irodalom: Márkus L. előadásfóliái a feltételes várható értékről: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/FeltVarhErt.pdf`

Sztochasztikus folyamatok alapjai

Definíció. **Sztochasztikus folyamat:** $(X_t)_{t\in T}$, ahol T a paramétertér és minden t-re X_t valószínűségi változó.

A sztochasztikus folyamat **diszkrét paraméterű** (vagy diszkrét idejű), ha T számossága legfeljebb megszámlálhatóan végtelen, tipikusan $T=\mathbb{Z}$ vagy $T=\mathbb{Z}_+$. A sztochasztikus folyamat **folytonos paraméterű** (vagy folytonos idejű), ha T kontinuum számosságú, jellemzően $T=[0;1]$, $T=\mathbb{R}$ vagy $T=\mathbb{R}_+$. A félév során előforduló sztochasztikus folyamatok:

- Poisson-folyamat: folytonos paraméterű
- Markov-folyamat: diszkrét vagy folytonos idejű, a gyakorlaton csak a diszkrét idejűekkel foglalkozunk
- Wiener-folyamat: folytonos paraméterű
- idősormodellek (autoregresszív, mozgóátlag folyamatok): diszkrét idejű

Definíció. **Gauss-folyamat:** olyan sztochasztikus folyamat, melynek bármely véges számú peremeloszlása együttesen normális eloszlású, azaz minden $n \in \mathbf{Z}_+$, $t_1 \in T, \ldots, t_n \in T$ esetén $(X_{t_1}, \ldots, X_{t_n})$ együttesen normális eloszlású.

Idősor: Olyan sztochasztikus folyamat, amikor a T paramétertartományra '*idő*'-ként gondolunk.

Definíció. **Erős stacionaritás.** $(X_t)_{t\in T}$ erősen stacionárius, ha minden $n \in \mathbb{Z}_+$, $t_1 \in T, \ldots, t_n \in T$ és $h \in T$ esetén $(X_{t_1}, \ldots, X_{t_n})$ együttesen ugyanolyan eloszlású, mint a h-val való eltoltja, $(X_{t_1+h}, \ldots, X_{t_n+h})$.

Definíció. **Autokovariancia függvény:** $R(t,s) = \mathrm{cov}(X_t, X_s)$

Definíció. **Gyenge stacionaritás.** $(X_t)_{t\in T}$ gyengén stacionárius, ha EX_t nem függ t-től (azaz konstans), illetve az autokovariancia függvény $R(t,s)$ értéke csak a $t-s$ eltéréstől függ.
Gyengén stacionárius sztochasztikus folyamat autokovariancia függvénye tehát tulajdonképpen egyváltozós, ezt az egyváltozós függvényt is R-rel fogjuk jelölni: $R(t,s) = R(t-s)$. Tehát gyengén stacionárius sztochasztikus folyamat autokovariancia függvénye $R(h) = \mathrm{cov}(X_{t+h}, X_t)$ módon számolható.

Megjegyzés. A gyenge stacionaritásból nem következik az erős, de az erős stacionaritásból se a gyenge (nem biztos, hogy léteznek momentumai).

Megjegyzés. A *stacionaritás* szó bizonyos szempontból időbeli állandóságot, stabilitást jelent. Szeretjük, ha egy idősor stacionárius, és igyekszünk adatainkat úgy transzformálni, hogy azok "közel", illetve "látszólag" stacionáriusak legyenek.

Állítás. $R(0) = D^2X_t$ minden t-re.

Definíció. **Autokorreláció függvény (ACF):** $r(h) = \mathrm{Cor}(X_t, X_{t+h})$, $h \in T$.

Állítás. $r(h) = \frac{R(h)}{R(0)}$

Definíció. **Független értékű zaj folyamat:** $\varepsilon_t \sim i.i.d.(0, \sigma^2)$, ha
$E\varepsilon_t = 0$, $D^2\varepsilon_t = \sigma^2$, valamint ε_t és ε_s minden $t \neq s$ esetén független egymástól.

Definíció. **Fehér zaj folyamat (white noise):**
$\varepsilon_t \sim WN(0, \sigma^2)$, ha $E\varepsilon_t = 0$, $D^2\varepsilon_t = \sigma^2$ és $\mathrm{cov}(\varepsilon_t, \varepsilon_s) = 0$, ha $t \neq s$.

Megjegyzés. gyakran kényelmes feltenni a fehér zajról, hogy Gauss-folyamat, ilyenkor Gauss-féle fehér zajról beszélünk (GWN).

Ajánlott irodalom: Márkus L. előadásfóliái idősorelméletről: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/Idosorok_1_Slides_2017_12_07.pdf`

A Poisson-folyamat

Definíció. Az X_t sztochasztikus folyamat **független növekményű**, ha tetszőleges $t_1 < t_2 \leq t_3 < t_4$ esetén $X_{t_2} - X_{t_1}$ és $X_{t_4} - X_{t_3}$ növekmények függetlenek egymástól.

Definíció. Az X_t sztochasztikus folyamat **stacionárius növekményű**, ha tetszőleges $t_1 < t_2$ és h esetén $X_{t_2} - X_{t_1} \sim X_{t_2+h} - X_{t_1+h}$, azaz tetszőleges növekmények tetszőleges eltoltjai ugyanolyan eloszlásúak.

Megjegyzés. Az előző két definíció során feltesszük, a tetszőlegesen választott időpontok olyanok, hogy azok nem vezetnek ki a T paramétertartományból.

Definíció. **Poisson-folyamat.**
$X_t, t \geq 0$ Poisson-folyamat $\lambda > 0$ intenzitással, ha
- $X_0 = 0$,
- független növekményű,
- stacionárius növekményű,
- $P(X_t = 1) = \lambda t + o(t)$ és $P(X_t \geq 2) = o(t)$, ha $t \to 0$

Tétel. **A Poisson-folyamat tulajdonságai.**
Legyen X_t Poisson-folyamat λ intenzitással. Jelöje τ_i az $i.$ esemény bekövetkezésének időpontját, $i = 1, 2, \ldots$ Ekkor
a.) $X_t \sim \mathrm{Poi}(\lambda t)$
b.) az autokovariancia függvénye $\mathrm{Cov}(X_t, X_s) = \lambda \cdot \min(t, s)$;
c.) $\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \ldots$ függetlenek és azonos, $\mathrm{Exp}(\lambda)$ eloszlásúak
d.) ha $t > s$, akkor $X_s | X_t \sim \mathrm{Bin}\big(X_t, \frac{s}{t}\big)$
e.) ha $t < s$, akkor $X_s | X_t \sim \ X_t + \mathrm{Poi}(\lambda(s-t))$

Következmény. $\tau_n \sim \Gamma(n, \lambda), \ \ n = 1, 2, \ldots$

Állítás. **Poisson-folyamatok egyesítése.**
Legyenek $X_t^1, \ldots, X_t^m$ független Poisson-folyamatok $\lambda_1, \ldots, \lambda_m$ intenzitásokkal. Ekkor $X_t = X_t^1 + \ldots + X_t^m$ is Poisson-folyamat, $\lambda_1 + \ldots + \lambda_m$ intenzitással.

Tétel. **Poisson-folyamat ritkítása.**
Legyen X_t Poisson-folyamatok λ intenzitással. Minden egyes esemény bekövetkezésekor egymástól függetlenül feldobunk egy érmét, a fejdobás valószínűsége $p \in [0; 1]$. Legyen X_t^1 a t időpontig kapott fejek száma, X_t^2 pedig a t időpontig kapott írások száma.
Ekkor X_t^1 és X_t^2 egymástól független Poisson-folyamatok $\lambda \cdot p$ és $\lambda \cdot (1-p)$ intenzitásokkal.

Ajánlott irodalom: Márkus L. előadásfóliái a Poisson-folyamatról: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/Egyszeru_poisson2.pdf`

Diszkrét idejű Markov-folyamatok (Markov-láncok)

Legyen S megszámlálható halmaz, neve: állapottér; $X_0, X_1, X_2, \ldots$ valószínűségi változó sorozat, $P(X_i \in S) = 1$ minden i-re. Az állapottér elemeire sokszor egyszerűbb az $1, 2, \ldots$ számokként gondolni.

Definíció. **Markov-tulajdonság.** Minden $0 \leq n \in \mathbb{Z}$ és $i_j \in S,\ j = 1, 2, \ldots$ esetén $P(X_{n+1} = i_{n+1} | X_n = i_n, \ldots, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n)$.

Definíció. **Markov-folyamat/lánc.** A Markov-tulajdonsággal rendelkező diszkrét idejű sztochasztikus folyamatokat Markov-folyamatoknak vagy Markov-láncoknak hívjuk.

Definíció. **Homogén Markov-lánc.** A Markov-lánc homogén (vagy stacionárius), ha a $P(X_{n+1} = j | X_n = i)$ feltételes valószínűség nem függ n-től, azaz minden n-re ugyanaz.
Jel. a kezdeti eloszlást $\mathbf{q}^T = (q_1, q_2, \ldots)^T$, ahol $q_k = P(X_0 = i),\ \ i \in S$.
Jel. homogén Markov-lánc esetén $p_{i,j} := P(X_{n+1} = j | X_n = i),\ \ i, j \in S$. Ezek neve: egylépéses átmenetvalószínűségek vagy röviden átmenetvalószínűségek.
Jel. $\mathbf{P} = (p_{i,j})_{i,j \in S}$, neve: **átmenetvalószínűség mátrix**

Jel. homogén Markov-lánc esetén $p_{i,j}^{(n)} := P(X_n = j | X_0 = i),\ \ i, j \in S$. Ezek neve: n lépéses átmenetvalószínűségek.
Jel. $\mathbf{P}^{(n)} = \left(p_{i,j}^{(n)}\right)_{i,j \in S}$, neve: n lépéses átmenetvalószínűség mátrix

Tétel. **Chapman-Kolmogorov egyenletek.** Ha $m < k < n$, akkor

$$P(X_n = i_n | X_m = i_m) = \sum_{i_k \in S} P(X_n = i_n | X_k = i_k) \cdot P(X_k = i_k | X_m = i_m).$$

Következmény. $\mathbf{P}^{(n)} = \mathbf{P}^n$
Állítás. A $\mathbf{P}$ mátrix minden sorösszege 1.

Állítás. $P(X_n = j) = \sum_{i \in S} q_i \cdot (\mathbf{P}^n)_{i,j} = \mathbf{q}^T \mathbf{P}^n,\ \ j \in S$

Definíció. i állapotból j állapot **elérhető**, ha $\exists n \geq 1$: $p_{i,j}^{(n)} > 0$. Jel.: $i \to j$
Definíció. i és j **érintkeznek**, ha vagy $i = j$, vagy ($i \to j$ és $j \to i$). Jel.: $i \leftrightarrow j$

Állítás. Az elérhetőség tranzitív, az érintkezés pedig ekvivalenciareláció.

Definíció. **Pontosan n lépésben visszatérés valószínűsége.**
$f_n(i,i) := P(X_n = i, X_{n-1} \neq i, \ldots, X_1 \neq i | X_0 = i),\ \ i \in S$
Definíció. **Visszatérés valószínűsége.** $f(i,i) := \sum_{n=1}^{\infty} f_n(i,i),\ \ i \in S$

Állítás. $p^{(n)}(i,i) = f_n(i,i) + \sum_{k=1}^{n-1} f_k(i,i) p^{(n-k)}(i,i)$

Definíció. **Visszatérő állapot.** Az i állapot visszatérő, ha $\forall j \in S$-re $i \to j \ \Rightarrow j \to i$.
Megjegyzés. Ekvivalens elnevezések: visszatérő=lényeges=perzisztens állapot.

Definíció. **Átmeneti állapot.** Az i állapot átmeneti, ha nem visszatérő.
Megjegyzés. Ekvivalens elnevezések: átmeneti=lényegtelen=**tranziens** állapot

Tétel. Az i állapot visszatérő $\iff f(i,i) = 1 \iff \sum_{n=0}^{\infty} p^{(n)}(i,i) = \infty$

Következmény. Az i állapot átmeneti $\iff f(i,i) < 1 \iff \sum_{n=0}^{\infty} p^{(n)}(i,i) < \infty$

Definíció. **Elnyelő állapot.** Az i állapot elnyelő, ha $p_{i,i} = 1$.

Definíció. Az i állapot periódusa: $d(i) := \text{lnko}\{n \geq 0 : p^{(n)}(i,i) > 0\}$

Definíció. **Periodikus állapot.** Az i állapot periodikus, ha $d(i) > 1$.

Definíció. **Aperiodikus állapot.** Az i állapot aperiodikus, ha $d(i) = 1$.

Definíció. **Irreducibilis ML:** az állapotai érintkeznek egymással.

Megjegyzés. Az irreducibilis Markov-lánc gráfja összefüggő.

Definíció. **Ergodikus ML:** irreducibilis, minden állapota visszatérő és aperiodikus.

Tétel. Legyen $\mathbf{P}$ egy ergodikus Markov-lánc átmenetvalószínűség mátrixa.
Ekkor $\forall i, j \in S$-re $p^{(n)}(i,j) \underset{n \to \infty}{\longrightarrow} \frac{1}{\sum_{n=1}^{\infty} n f_n(j,j)}$.

Vegyük észre, hogy az előző tétel alapján amihez konvergál, már nem függ a kiinduló i állapottól. Jelölje $\pi_j := \frac{1}{\sum_{n=1}^{\infty} n f_n(j,j)}\ \ j = 1, 2, \ldots$, ezzel $\lim_{n \to \infty} \mathbf{P}^n = \begin{pmatrix} \pi_1 & \pi_2 & \cdots \\ \vdots & \vdots & \ldots \\ \pi_1 & \pi_2 & \cdots \end{pmatrix}$.

Jel. $\boldsymbol{\pi}^T = (\pi_1, \pi_2, \ldots)^T$, elnevezése: **stacionárius** vagy egyensúlyi **eloszlás**. A stacionárius eloszlás mutatja meg, hogy "hosszú idő után" a milyen valószínűséggel leszünk a Markov-lánc egyes állapotaiban.
A gyakorlatban a stacionárius eloszlást az alábbi egyenletrendszer megoldásával szokás kiszámolni: $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{P}$, ahol $\sum_i \pi_i = 1$. Ennek értelmében tehát a $\boldsymbol{\pi}$ vektor a $\mathbf{P}$ mátrix baloldali, 1-re normált sajátvektora.

Markov-láncoknál egy lényeges kérdés, hogy átlagosan mennyi időbe (lépésbe) telik, míg az egyik állapotból egy másik állapotba eljutunk. Jelölje $m_{i,j}$: ha jelenleg az i állapotban vagyunk, akkor várhatóan ennyi lépésre van szükség, hogy a j állapotba kerüljünk. Általánosan ezeket az értékeket nem lehet közvetlenül egyszerűen kiszámolni, de a teljes várható érték tétel alapján felírható rájuk a következő egyenlet: $m_{i,j} = p_{i,j} + \sum_{k \neq j} p_{i,j}(1 + m_{k,j})$, amit $m_{i,j} = 1 + \sum_{k \neq j} p_{i,j} m_{k,j}$-ra lehet egyszerűsíteni.

Állítás. Ergodikus Markov-lánc esetén $m_{i,i} = \frac{1}{\pi_i}$

Elnyelő Markov-láncok:

- van s tranziens állapot: $t_1, \ldots, t_s$

- van m elnyelő állapot: $a_1, \ldots, a_m$

Particionáljuk ezek alapján az átmenetvalószínűség mátrixot: $\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I}_m \end{pmatrix}$, ahol $\mathbf{Q} \in \mathbb{R}^{s \times s}$, $\mathbf{R} \in \mathbb{R}^{s \times m}$, $\mathbf{I}_m$ az m dimenziós egységmátrix.
Néhány lényeges mennyiség kiszámítása:

- ha t_i-ben vagyunk, akkor azon periódusok várható száma, amit t_j-ben töltünk, mielőtt egy elnyelő állapotba lépnénk: $\left((\mathbf{I}_s - \mathbf{Q})^{-1}\right)_{i,j}$
- ha t_i-ben vagyunk, akkor annak a valószínűsége, hogy a_j-be kerülünk: $\left((\mathbf{I}_s - \mathbf{Q})^{-1}\mathbf{R}\right)_{i,j}$

Elnyelő láncok esetén nem beszélhetünk olyan értelemben stacionaritásról, mint az ergodikus láncoknál, egyfajta egyensúly csak akkor érhető el, ha minden időszakban van(nak) új belépő(k) a rendszerbe. Tekintsük az n-edik időperiódust az $n-1$-edik és az n-edik időpont között eltelt időnek, $n = 1, 2, \ldots$
Vezessünk be jelöléseket:

- H_i: az egyes időperiódusok elején az i-edik állapotba belépő egyedek száma
- $N_i(n)$: az n-edik időperiódus elején az i-edik állapotban lévő egyedek száma
- $r_{i\bullet} = \sum_{j=1}^{m} r_{i,j}$, ami az i-edik állapotból egy elnyelő állapotba lépés valószínűsége
- $\widetilde{\mathbf{Q}} := \left(\begin{array}{c|c} \mathbf{Q} & \begin{matrix} r_{1\bullet} \\ \vdots \\ r_{s\bullet} \end{matrix} \end{array}\right)$

Kérdés, hogy léteznek-e a $\lim_{n\to\infty} N_i(n)$ határértékek. Ha léteznek, akkor jelöljük őket N_i-vel, amikből képezzük az $\mathbf{N} = (N_1, \ldots, N_s)^T$ egyensúlyi egyedszám vektort. Amennyiben létezik ilyen egyensúlyi helyzet, akkor minden időperiódusban az i-edik állapotba belépő egyedek számának (a lenti egyenletben a baloldal) meg kell egyeznie az onnan kilépő egyedek számával (jobboldal):

$$H_i + \sum_{k \neq i} N_k \cdot \widetilde{q}_{k,i} = N_i \cdot (1 - \widetilde{q}_{i,i}) \qquad i = 1, \ldots, s$$

Ajánlott irodalom: Wayne L. Winston: Operációkutatás, 17. fejezet

---

Lineáris modell (regressziószámítás)

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:

$$(y_i, x_{i,1}, \ldots, x_{i,p})^T \quad i = 1, 2, \ldots, n \qquad \text{általában } n >> p$$

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 egy véges σ szórású valószínűségi változó.
A **lineáris modell**: $\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$, ahol

- $\mathbf{y} = (y_1, \ldots, y_n)^T$
- $\mathbf{X} = \begin{bmatrix} x_{1,1} & \ldots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{bmatrix}$
- $\mathbf{b} = (b_1, \ldots, b_p)^T$
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$

Paraméterbecslés: $\widehat{\mathbf{b}} = (X^T X)^{-1} X^T \mathbf{y}$
Projekció az $F := \mathrm{Im}X$ altérre: $P_F = X(X^T X)^{-1} X^T$
Becsült értékek: $\widehat{\mathbf{y}} := X\widehat{\mathbf{b}}$
Reziduálisok: $\widehat{\varepsilon} = \mathbf{y} - \widehat{\mathbf{y}}$
Reziduális négyzetösszeg: $\mathrm{RNÖ} := \|\widehat{\varepsilon}\|^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$
Teljes négyzetösszeg: $\mathrm{NÖ} = \sum_{i=1}^{n} (y_i - \overline{y})^2$
Determinációs együttható: $R^2 = 1 - \frac{\mathrm{RNÖ}}{\mathrm{NÖ}} = \frac{\mathrm{NÖ} - \mathrm{RNÖ}}{\mathrm{NÖ}}$ $\rightsquigarrow$ az eredményváltozó változékonyságának hány %-át magyarázza a regressziós modell. Értéke 0 és 1 között lehet. Minél nagyobb, annál jobb.
Gyakori modellválasztási kritériumok:

- Korrigált determinációs együttható: $R^2_{\mathrm{adj}} = 1 - \frac{n-1}{n-r-1}\frac{SSR}{N}$ $\rightsquigarrow$ minél nagyobb, annál jobb
- **Akaike-féle információs kritérium**: $AIC = 2k - 2\log\widehat{L}$, ahol
  - k: a becsülendő paraméterek száma, a regressziós modellben $k = p + 1$
  - $\widehat{L}$ a likelihood-függvény értéke akkor, ha az ML-becslést használjuk (normális eloszlású hibáknál ez megegyezik a legkisebb négyzetes becsléssel)

  Minél kisebb, annál jobb.
- **Bayes-féle információs kritérium**: $BIC = \log n \cdot k - 2\log\widehat{L}$ $\rightsquigarrow$ minél kisebb, annál jobb

A regresszióelemzés lépései:

- az eredményváltozó(k) és a lehetséges magyarázóváltozók kiválasztása
- adatgyűjtés
- adattisztítás, adathibák korrekciója
- pontdiagrammal a potenciális modellek kiválasztása (lineáris, négyzetes, logisztikus stb.)
- paraméterbecslés
- modelldiagnosztika – az együtthatók szignifikanciája, a modell együttes jósága
- legjobb modell kiválasztása, "modellépítés" – több módszer/mutató közül választhatunk: korrigált R^2, cross-validation, AIC/BIC információs kritériumok stb.

- előrejelzés

Attól függően, hogy az eredmény-, illetve a magyarázóváltozó diszkrét-e vagy folytonos, az alábbi statisztikai módszerek használandók a kapcsolat vizsgálatára:

		Az eredményváltozó	
		diszkrét	abszolút folytonos
A magyarázóváltozó	diszkrét	asszociáció χ^2-próba	vegyes kapcsolat t-próba, ANOVA
	absz. folyt.	osztályozási eljárások, diszkriminancia analízis	korreláció regresszió

Ajánlott irodalom: Márkus L. előadásfóliái a regresszióról: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/RegressionSlides.pdf`

Szóráselemzés (ANOVA) / vegyes kapcsolat elemzése

A szóráselemzés a lineáris modell egyik legfontosabb alkalmazása. Az eljárásnak több elnevezése van: szóráselemzés = variancia-analízis = ANOVA (analysis of variance). Vegyes kapcsolat: egy diszkrét és egy folytonos ismérv közti kapcsolat.
A modell: $x_{i,j} = \mu_i + \varepsilon_{i,j}$, ahol

- $i = 1, \dots, k$ $\rightsquigarrow$ csoportok vagy osztályok száma
- $j = 1, \dots, n_i$ $\rightsquigarrow$ mintaelemszám egy osztályon belül
- $N = \sum_{i=1}^{k} n_i$ $\rightsquigarrow$ teljes mintaelemszám
- $\varepsilon_{i,j} \sim N(0, \tau^2)$ függetlenek, ahol $\tau > 0$

Feladat: annak eldöntése, hogy $\mu_1 = \dots = \mu_k$, azaz a csoporthoz tartozás nem befolyásolja az ismérv értékét.
Vezessünk be jelöléseket:

- $\overline{x}_i = \frac{1}{n_j} \sum_{j=1}^{n_i} x_{i,j}$ $\rightsquigarrow$ részátlagok vagy csoportátlagok
- $\overline{x} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{i,j} = \frac{1}{N} \sum_{i=1}^{k} n_i \overline{x}_i$ $\rightsquigarrow$ teljes átlag
- $\sigma_i = \sqrt{\frac{1}{n_j} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)^2}$ $\rightsquigarrow$ részszórások
- $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x})^2}$ $\rightsquigarrow$ teljes szórás
- $SS_w = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x}_i)^2 = \sum_{i=1}^{k} n_i \sigma_i^2$ $\rightsquigarrow$ csoportokon belüli eltérés-négyzetösszeg
- $SS_b = \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2$ $\rightsquigarrow$ csoportok közötti teljes eltérés-négyzetösszeg
- $SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x})^2$ $\rightsquigarrow$ teljes eltérés-négyzetösszeg

Állítás. $SS = SS_w + SS_b$

A kapcsolat erősségét mérő mutató a **szórásnégyzet-hányados:** $H^2 = 1 - \frac{SS_w}{SS} = \frac{SS_b}{SS}$
Tulajdonságai:

- $H^2 = 0$ esetén a két ismérv között nincs kapcsolat, DE (!!) ekkor nem feltétlen függetlenek egymástól (analógia: korrelálatlanságból nem következik a függetlenség)
- $H^2 = 1$ esetén a két ismérv között függvényszerű kapcsolat van
- $0 < H^2 < 1$ esetén a két ismérv között sztochasztikus kapcsolat van
- erős a kapcsolat, ha $H = \sqrt{H^2}$ közel van 1-hez és gyenge a kapcsolat, ha 0-hoz

Megjegyzés: ez nem más, mint a regressziónál az R^2.

Tekintsük az alábbi hipotézisvizsgálati feladatot: $H_0 : \mu_1 = \dots = \mu_k$
H_1 : nem igaz H_0
A hipotézisekről egy F-próbával döntünk, a F próbastatisztika kiszámításához az ún. **ANOVA táblázat**ot szokás elkészíteni:

Szóródás forrása	Szabadságfok	Négyzetösszegek	Tapasztalati szórásnégyzetek	
Külső (between group)	$k-1$	SS_b	$MS_b = \frac{SS_b}{k-1}$	$F = \frac{MS_b}{MS_w} = \frac{\frac{SS_b}{k-1}}{\frac{SS_w}{N-k}}$
Belső (within group)	$N-k$	SS_w	$MS_w = \frac{SS_w}{N-k}$	
Teljes	$N-1$	SS	–	

Tétel. Ha teljesülnek a modell feltételei és a H_0 nullhipotézis, akkor $F \sim F_{k-1,N-k}$, azaz a próbastatisztika F-eloszlást követ.

Ajánlott irodalom: Márkus L. előadásfóliái ANOVA-ról: `http://web.cs.elte.hu/probability/markus/ElemzoTS1/ANOVA_MANOVA_Sajat.pdf`
