**Matematikai statisztika tétel-kidolgozás**

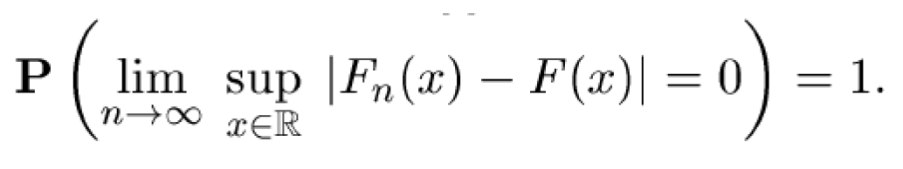
# 1. tétel

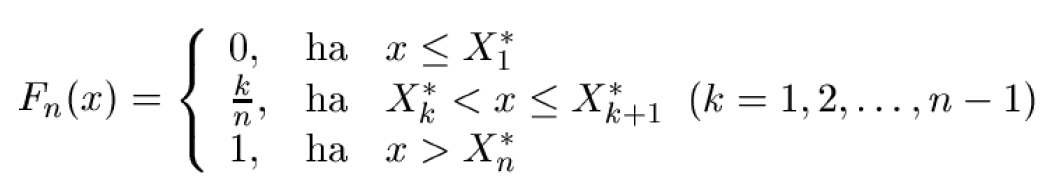
**Minta**: A populáció egy kis elemszámú részhalmazára vonatkozó megfigyelések adatai. A minta úgy kell, hogy tükrözze a populáció tulajdonságait, ahogy a cseppben látjuk a tengert. Azaz a minta reprezentatív kell, hogy legyen.   
Az X valószínűségi változóval azonos eloszlású, egymással teljesen független X1, X2,…, Xn valószínűségi változók együttesét **statisztikai mintának** nevezzük.

**Statisztika**: A vizsgált populációra vonatkozólag előre megtervezett módon, matematikai elvek figyelembe vételével beszerzett adatokkal, a minta feldolgozásával állítja elő a sokaságra vonatkozó hasznos következtetéseket. Vagy: A minta realizáció adataiból adott képlettel számolt adat a statisztika számított értéke.

**Paraméterek**: A minta eloszlásfüggvényét egy adott **v** paraméter konkretizálja, ha ennek ismerjük az értékét akkor pontosan meg tudjuk adni az eloszlásfüggvényt.

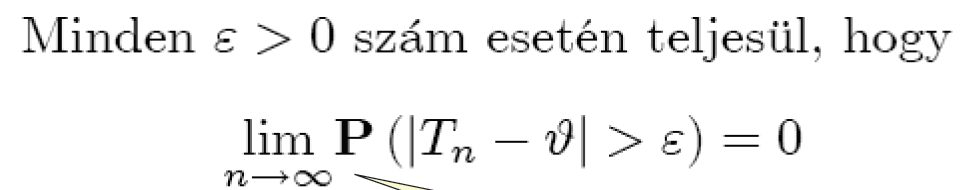
**A matematikai statisztika alaptétele (Glivenkó-Cantelli):**

****

****

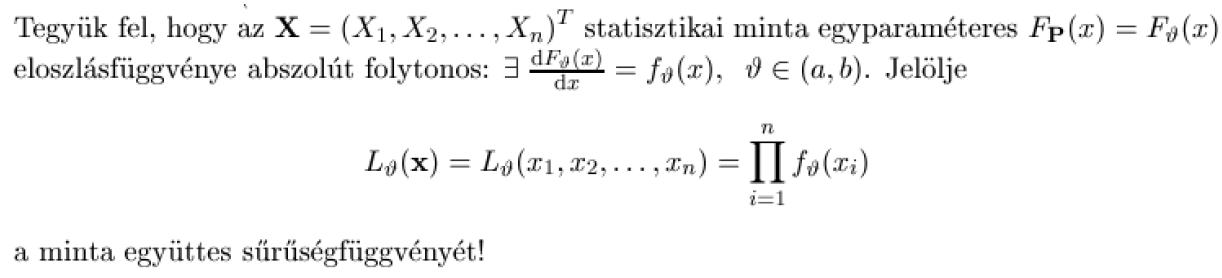
Az empirikus eloszlásfüggvény (Fn(x) ) 1 valószínűséggel egyenletesen konvergál az eloszlásfüggvényhez.

**Becslések**: Az eloszlás paraméterek meghatározásához sokszor paraméterbecslést alkalmazunk. A becsléshez valamilyen statisztikát alkalmazunk ahol paramétere a becsült paraméter.

* **Torzítatlan**: statisztika a paraméter torzítatlan becslése, ha . Lehet határértékben is értelmezni a torzítatlanságot, ekkor csak az esetre igaz az állítás.
* **Konzisztens**: a minta elemszám növekedésével növekszik a becslés pontosságának valószínűsége  
  .
* **Hatásos**: Két torzítatlan becslés közül nyilván a kisebb varianciájú a jobb, hiszen kisebb mértékben ingadozik a paraméter körül! Egy torzítatlan becslés akkor lesz hatásos, ha varianciája minden más torzítatlan becslés varianciájánál kisebb!
* **Elégséges**: a statisztika a minta eloszlásának paraméterére vonatkozóan minden információt magába sűrít, egymaga képes helyettesíteni a mintát. A paraméterbecsléshez elégséges, ha az elégséges statisztikák függvényei közt keresünk.

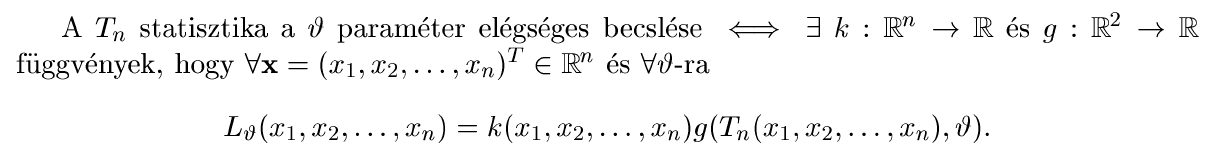
**Cramer-Rao-egyenlőtlenség:**

A Cramer-Rao-egyenlőtlenség elvi alsó korlátot ad a torzítatlan becslések szórásnégyzeteire. Ha egy statisztikára belátjuk, hogy szórásnégyzete éppen az alsó korláttal egyenlő, akkor az biztosan hatásos, sőt az egyetlen hatásos becslés!



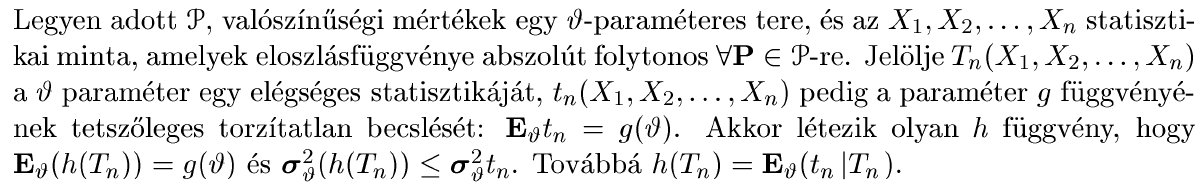
Példák:

* Az átlagstatisztika hatásos normális esetben
* Az átlagstatisztika hatásos exponenciális esetben
* Az átlagstatisztika hatásos Poisson esetben
* Az a korrigált empirikus szórásnégyzet hatásos normális esetben

**Neyman-Fisher faktorizációs tétel: **

Ennek a tételnek a segítségével, lehet ellenőrizni, hogy egy statisztika elégséges-e.

**Rao-Blackwell-Kolmogorov-tétel:** Ha létezik hatásos (legjobb torzítatlan) becslés, akkor elég azt az elégséges becslés függvényei között keresni!

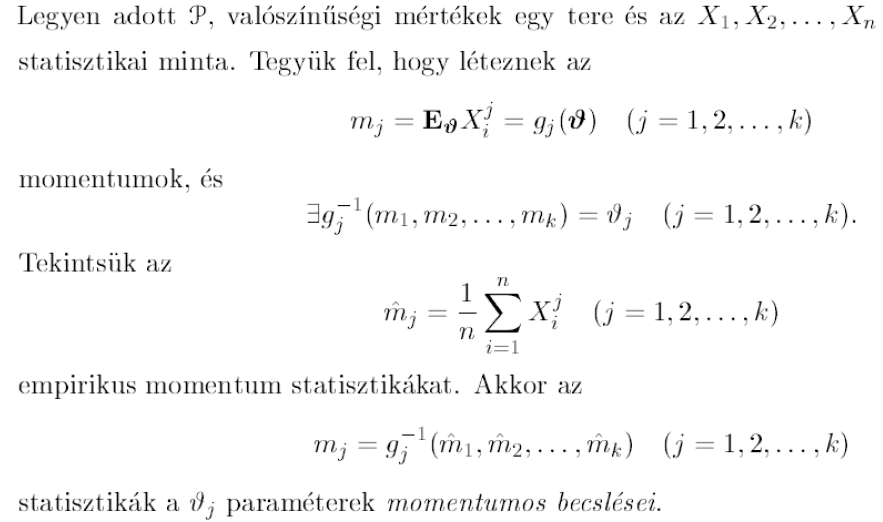


# 2. tétel

**Maximum likelihood-módszer:**

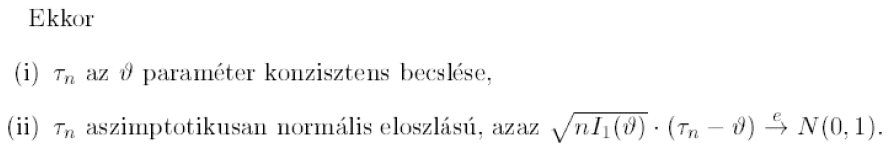
* A mintánk eloszlásfüggvénye a paramétertől függ
* Ha egy kísérletnél több esemény is bekövetkezhet, legtöbbször a legnagyobb valószínűségű eseményt fogjuk megfigyelni.
* A sokaságra vett mintavételezés során kaptunk egy realizációt. Feltételezzük, hogy azért éppen ezt a realizációt kaptuk, és nem mást, mert az összes realizációk közül ennek volt a legnagyobb a bekövetkezési valószínűsége.
* Vegyük tehát, az összes lehetséges paraméter közül azt, amelynél éppen kapott realizáció bekövetkezése a maximális.

**Momentumok módszere:**

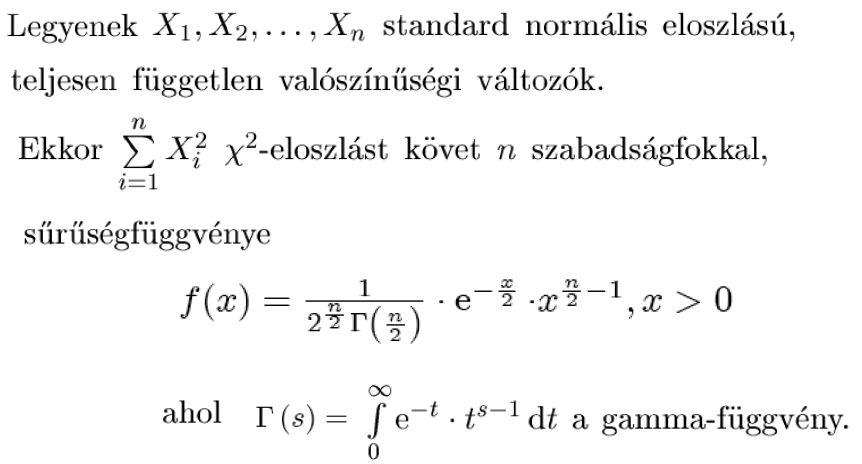


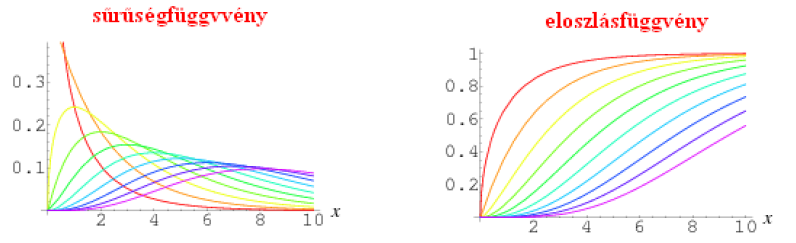
**Cramer-Dugue tétel:**



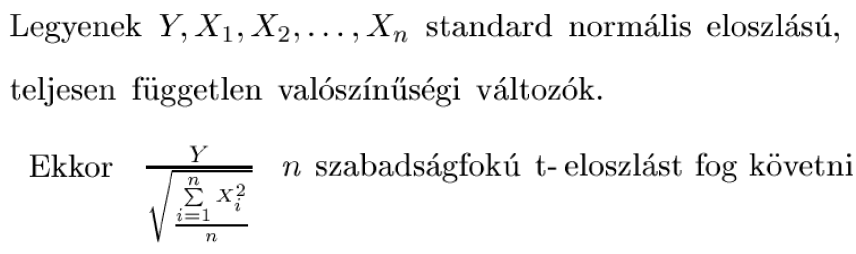


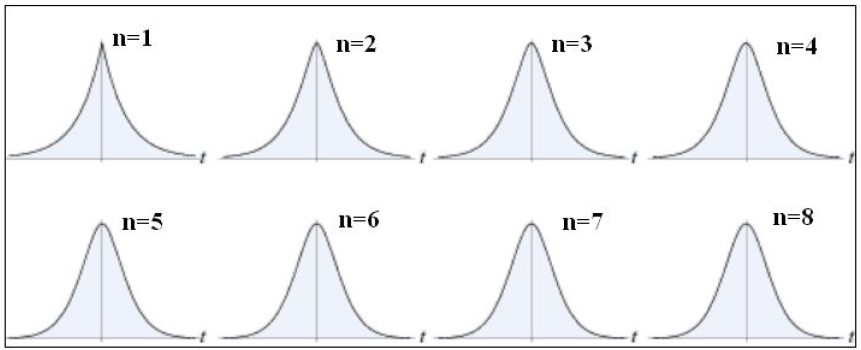
**Chi-négyzet eloszlás:**

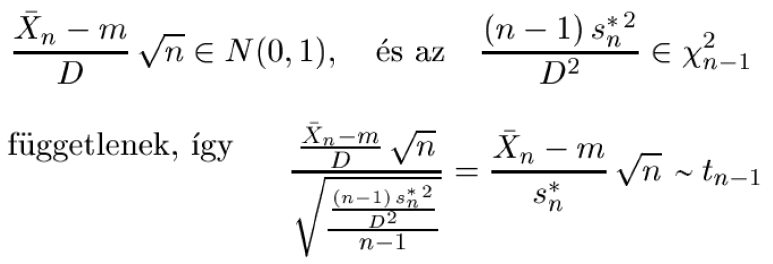




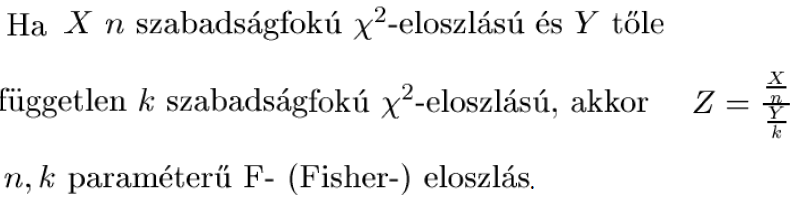
**Student-t eloszlás:**

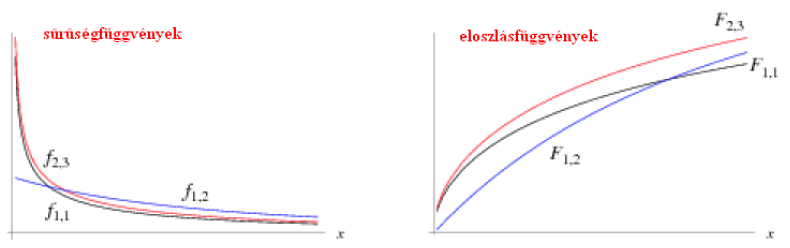






**Fisher eloszlás:**

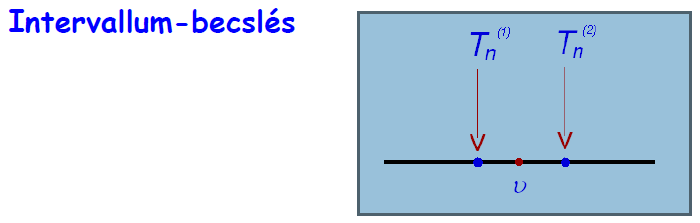
****

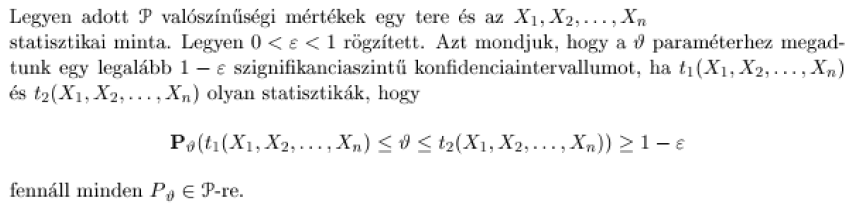
****

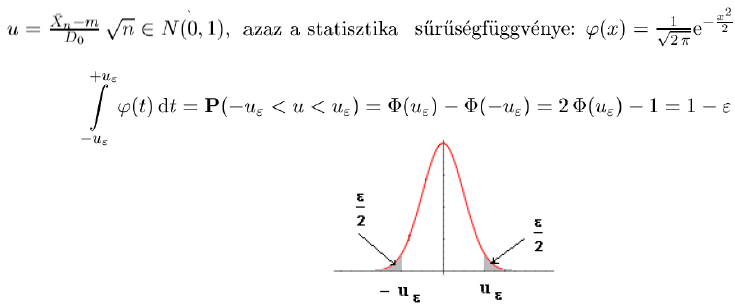
**Lukács tétel:**



**Konfidencia-intervallumok, példák:**

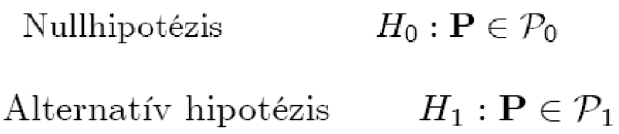
****

****

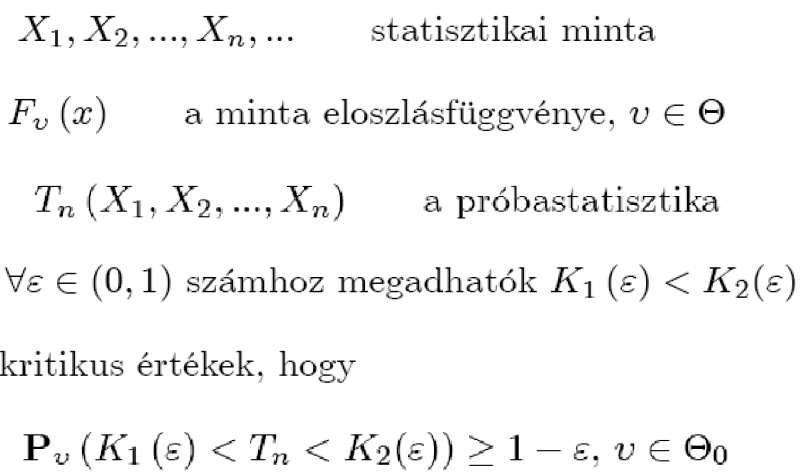
****

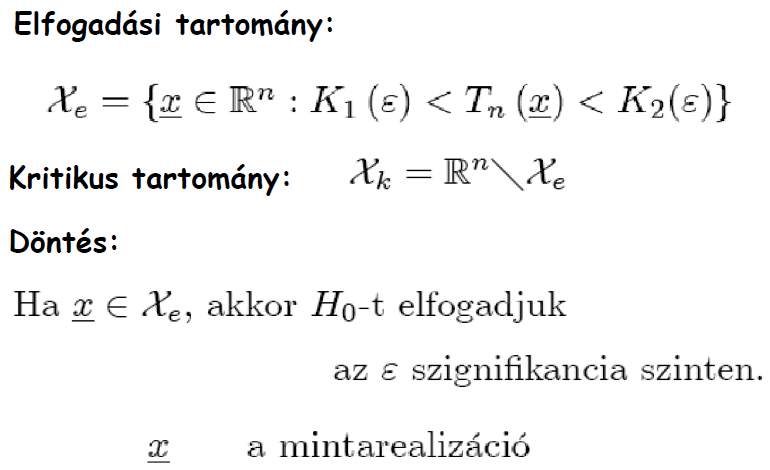
# 3. tétel

**Statisztikai hipotézisvizsgálat**: Döntési eljárást dolgozunk ki annak eldöntésére, hogy a nullhipotézis igaz-e. Ha úgy kell döntenünk, hogy a nullhipotézis nem igaz, automatikusan az alternatív hipotézist fogjuk elfogadni. A döntésünkhöz szignifikancia szintet fogunk rendelni, amivel jellemezzük, hogy a nullhipotézisünk melletti döntés milyen erős.



**Próbastatisztika és kritikus tartomány:**

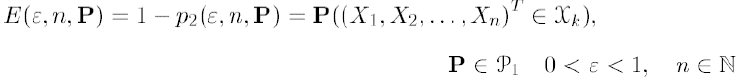


****

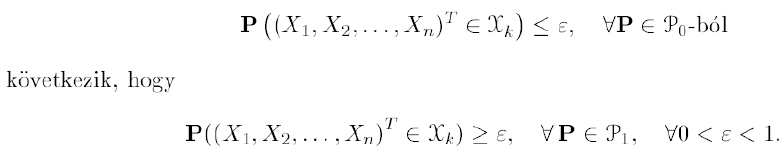
**Első - és másodfajú hibavalószínűség:**

* **Elsőfajú hibavalószínűség**: Akkor követjük el, ha igaz a nullhipotézis, de a mintarealizáció mégis a kritikus tartományba esik, és a döntésünk elutasító! Az elsőfajú hibavalószínűség az éppen az , amit mi állítunk be!
* **Másodfajú hibavalószínűség**: Akkor követjük el, ha elfogadjuk a nullhipotézist, holott valójában nem igaz. Értéke nehezebben állapítható meg.

**Erőfüggvény:**

****

**A próba konzisztenciája, torzítatlansága és ereje:**

* **konzisztencia:   
  **
* **torzítatlanság:  
  **
* **ereje:  
  **

Egy próba egyenletesen legjobb próba, ha az adott elsőfajú hibával rendelkező próbák között a legkisebb a másodfajú hibája.

**Neyman-Pearson**: A tétel arról szól, hogyan lehet adott n minta elemszámú minta esetén rögzített elsőfajú hibavalószínűséghez a lehető legkisebb másodfajú hibavalószínűségű próbát megkonstruálni.

**Stein-lemma: Na ne.**

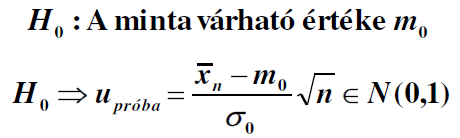
# 4. tétel

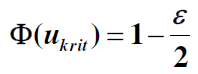
**Paraméteres próbák:** A próbákban az a közös, hogy az elemzett minta eloszlása normálist követ. A nullhipotézist éppen a normális eloszlás paramétereivel kapcsolatosan fogalmazzuk meg.

* Várható érték a paraméter
* Szórás a paraméter

**u-próba**: Akkor lehet használni, ha a normális eloszlású minta(ák) szórása előzetesen ismert.

* egymintás, és kétmintás is van







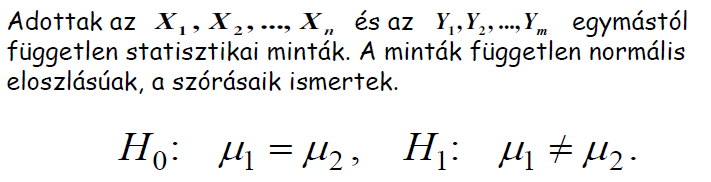
Az elsőfajú hibavalószínűség éppen az szignifikancia szint.

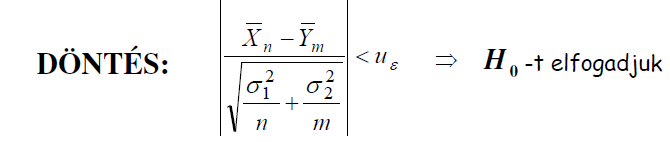
u-próba tulajdonságai:

* a próba torzítatlan és konzisztens
* egyenletesen legjobb próba is

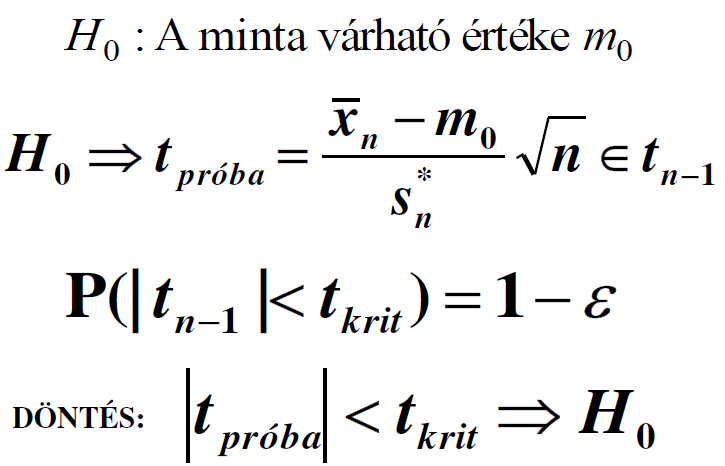
A gyakorlatban akkor is alkalmazzák az u-próbát, amikor a minta nem normális eloszlású, de a minta elemszám „nagy”. Az alkalmazás jogosságát a centrális határeloszlás-tétellel lehet indokolni. Ugyanis a próbastatisztika normális eloszlású lesz aszimptotikusan, mivel a CHT szerint a mintaátlag már közel normális eloszlású!

kétmintás u-próba:

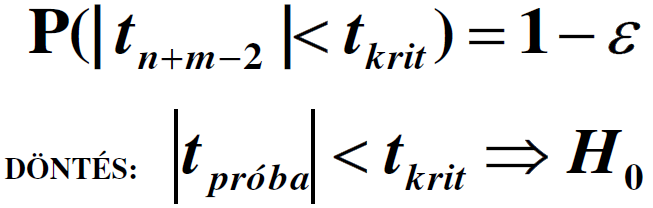




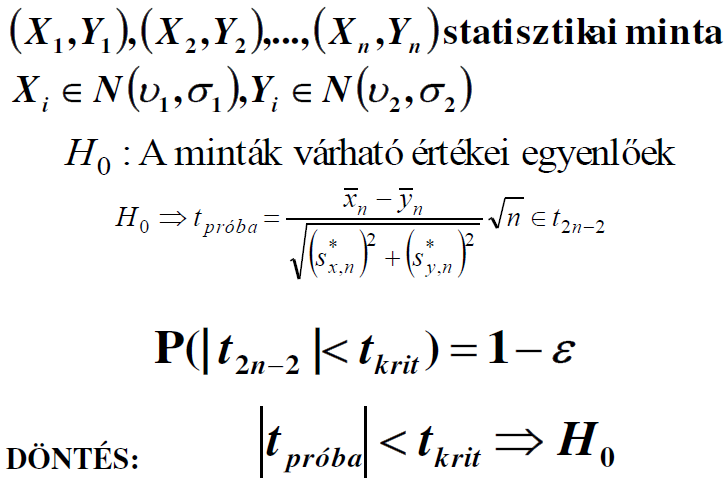
**t-próba:** Akkor lehet használni, ha a normális eloszlású minta(ák) szórása egyenlőnek tekinthető, de nem ismert (ennek ellenőrzése F-próbával lehetséges).



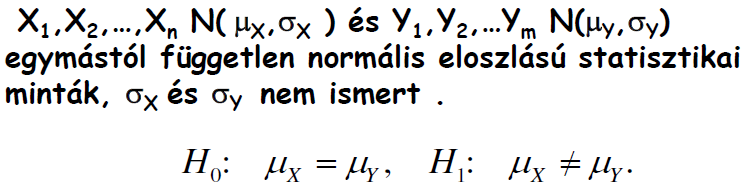
két független mintás t-próba:



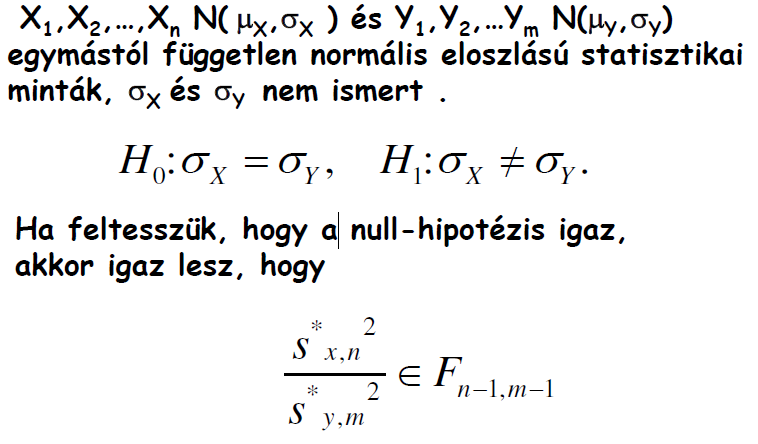
két összetartozó mintás t-próba:



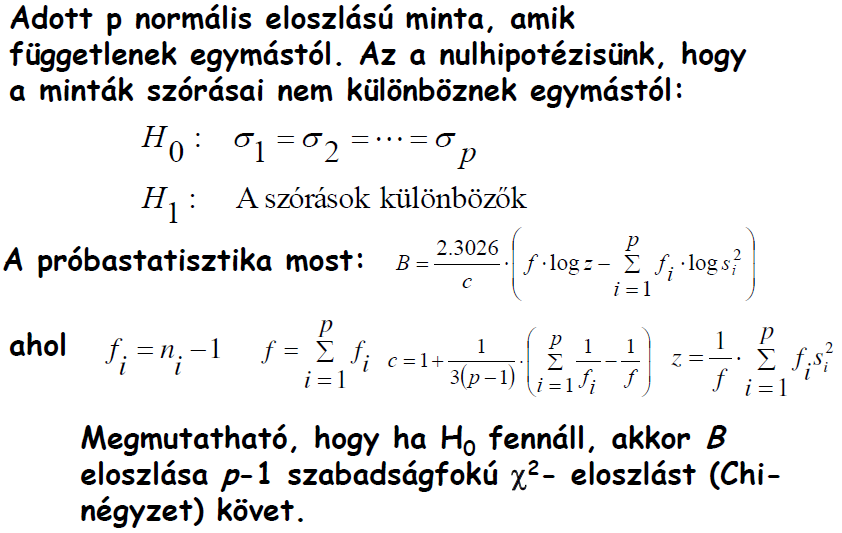
**Welch-próba:** Ha az F-próbát el kell vetni, nem alkalmazható a két független mintás t-próba. Ehelyett alkalmazzuk a Welch-tesztet.



**F-próba:**



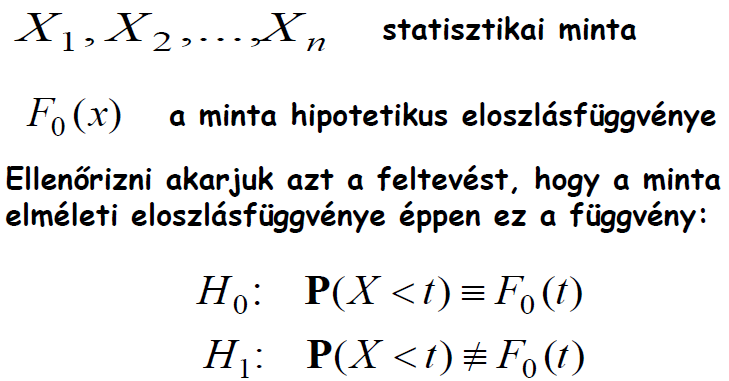
**Bartlett-próba:**

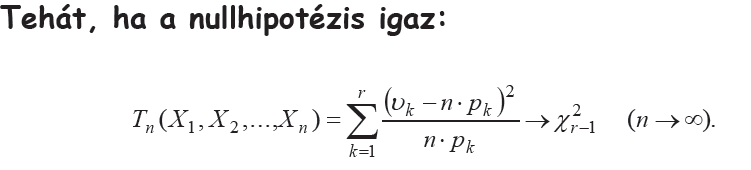


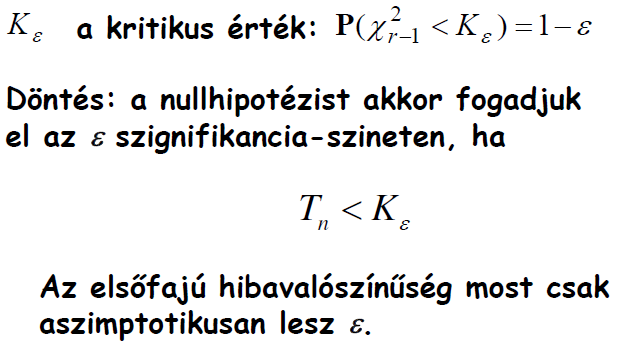
**Nemparaméteres próba:**

Ha az alapsokaság (a statisztikai minta) eloszlását nem tekintjük eleve ismertnek, akkor nemparaméteres próbákról beszélünk. Ilyenkor tehát az előzetes feltevéseink nagyon általánosak, de természetesek; pl. feltesszük, hogy a minta eloszlása folytonos, vagy feltesszük, hogy a szórás véges, stb. Mivel kevesebb feltételt követelünk meg kiinduláskor (a priori feltevések), a következtetéseink levonásához nagyobb elemszámú mintákra lesz szükségünk, mint a paraméteres próbák esetén. A próbastatisztikák eloszlását csak aszimptotikusan ismerjük.

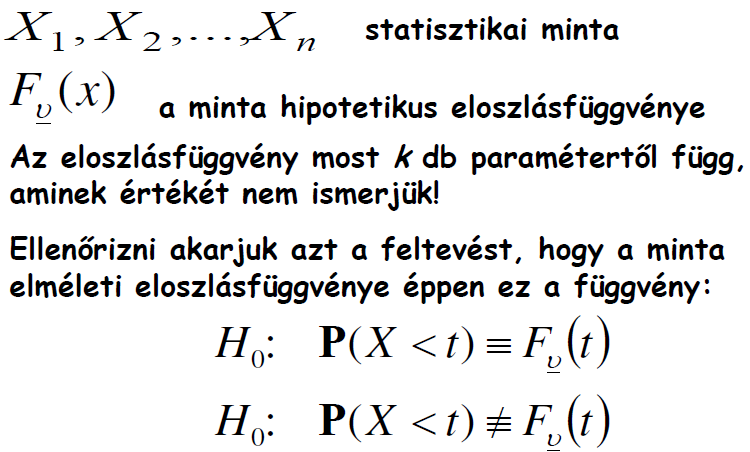
**Tiszta illeszkedésvizsgálat:**

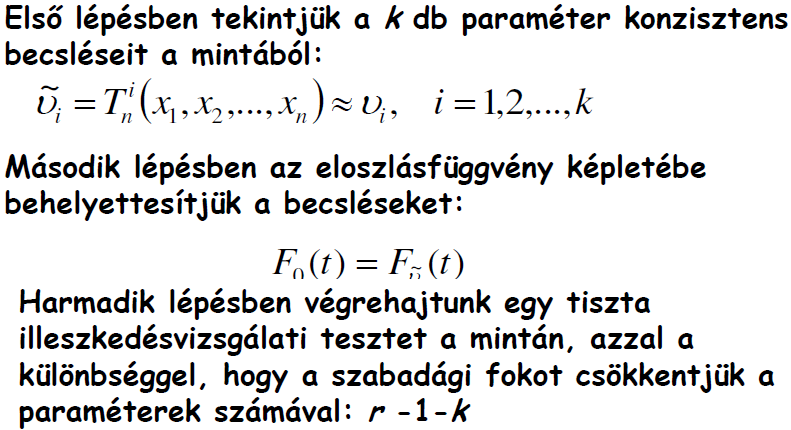






**Paraméteres illeszkedésvizsgálat chi-négyzet próbával:**

****

****

# 5. tétel

**Szórásanalízis**:   
A szórásanalízis (ANOVA=analysis of variance) modellek rugalmas statisztikai eszközök valamely kvantitatív (numerikus vagy intervallum skálájú) változónak (függő változó) egy vagy több nem feltétlenül kvantitatív változóval (független változók) való kapcsolata kielemzésére. Arra vagyunk itt kíváncsiak, hogy van-e hatása a független változóknak a függő változóra illetve, hogy ez a hatás egyforma vagy különböző. A hatás, apcsolat függvényszerű feltárása akkor sem cél, ha a független változók kvantitatívek .

* A vizsgált független változók kvalitatívek is lehetnek (pl. név, lakcím stb.)
* Nem cél a független és a függő változó közti függvénykapcsolat feltárása

Érdemes a regresszió analízis előtt végrehajtani az ANOVA-t.   
Azt vizsgáljuk, hogy egy bizonyos faktornak (körülménynek) van-e hatása a kimeneti változó (válasz) várható értékére. **A nullhipotézis ebben a feladatban, hogy a faktornak nincs hatása, azaz a várható értékek egyenlőek.**

**Alapfogalmak**:

* **faktor (factor)**: Faktornak nevezzük a vizsgálatba bevont független változókat.
* **faktor szint**: A faktor értékkészletének eleme, ezen beállítások mellett figyeljük meg a függőváltozót.
* **interakció**: Az egyes faktorok közt feltételezett kapcsolat. Pl. a dolgozó neme, és fizetési kategóriája között feltételezhető kapcsolat.
* **egyfaktoros**, **és** **több** **faktoros** **elemzések**: A modelleket a faktorok száma szerint csoportosítjuk, így beszélünk egy-, két-, háromfaktoros modellekről stb. Bizonyos kérdéseket csak többfaktoros modellekben tehetünk fel, pl. a faktorok között fellépő interakció kérdése ilyen.
* **véletlen** **és** **beállított** **faktorok**: véletlen faktor pl. a csapadék mennyisége, beállított faktor: egy gyártó berendezés hibájának szórását egy adott beállítás mellett teszteljük
* **kvalitatív és kvantitatív faktorok:** Ha a faktorszintek nem numerikusak, vagy intervallum skálájúak kvalitatív, ellenkező esetben kvantitatív faktorokról beszélünk.
* **kezelések (cellák)**: Egyfaktoros esetben a kezelések megfelelnek a faktor szintjeinek, többfaktoros esetben a figyelembe vett faktorok szintjeiből előálló kombinációk lesznek a kezelések. Pl. amikor a két faktor egy boríték színei és a címzésnél használt betűtípus, akkor a kezelések a (szín, betűtípus) párok összes lehetséges kombinációiból állnak.

**Kísérleti elrendezések**: Több faktor hatásának egyidejű vizsgálatakor a módszereket három csoportba sorolhatjuk.

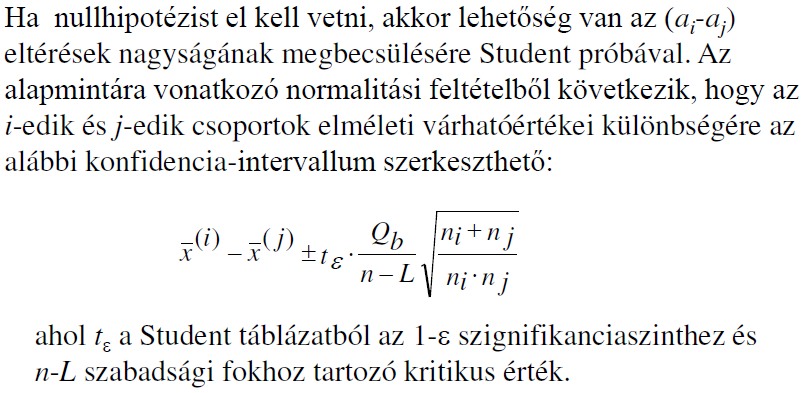
* **hierarchikus** **osztályozás**: A faktorok hierarchiában vannak és egy faktor összes szintje a felette álló faktor csak egy szintjéhez kapcsolódik.
* **keresztosztályozás**: Az A és B faktor szintjeinek minden (i,j) párosításához (kezelés, cella) veszünk egy- vagy többelemű mintát:
* **nem teljes kísérleti elrendezések**: Olyankor alkalmazandó, amikor egy vizsgálandó faktor mellett más, nem kívánt de számon tartott hatás is fellép, és azokat ki akarjuk küszöbölni. (egy faktor hatásának elimináláshoz egy másik faktorhoz tartozó elemeket véletlenszerűen permutáljuk, műtrágyás példa), latin négyezetek módszere, kiegyensúlyozott nem teljes blokk (4 állat, 7 alom, 7 gyógyszer)

**Alkalmazási feltételek:**

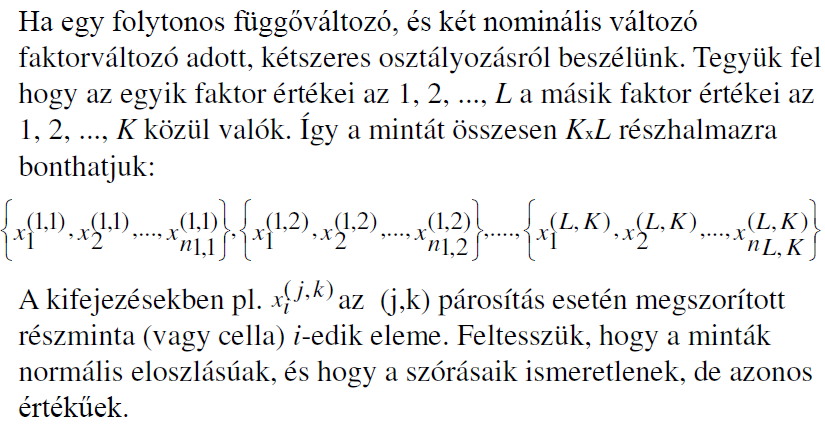
* A függőváltozó eloszlásának normálisnak kell lennie. Tehát tetszőleges kezeléshez tartozó mintának követnie kell a normális eloszlást.
* minták szórásnégyzeteinek meg kell egyezniük. Ez azon múlik, hogy a kezelések eredményét azonos módon mérik-e.
* Az egyes kezelésekhez tartozó mintáknak függetleneknek kell lenniük.

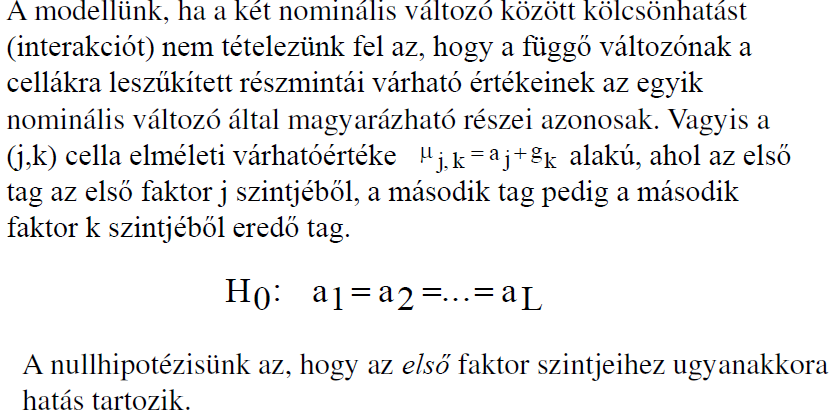
**Egyszeres- és kétszeres osztályozás**:

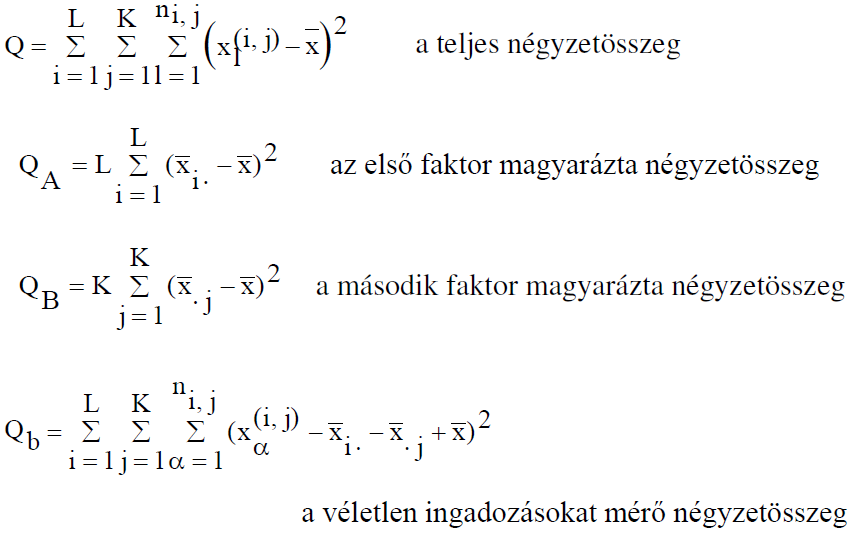
Egyszeres osztályozás (One-way ANOVA)

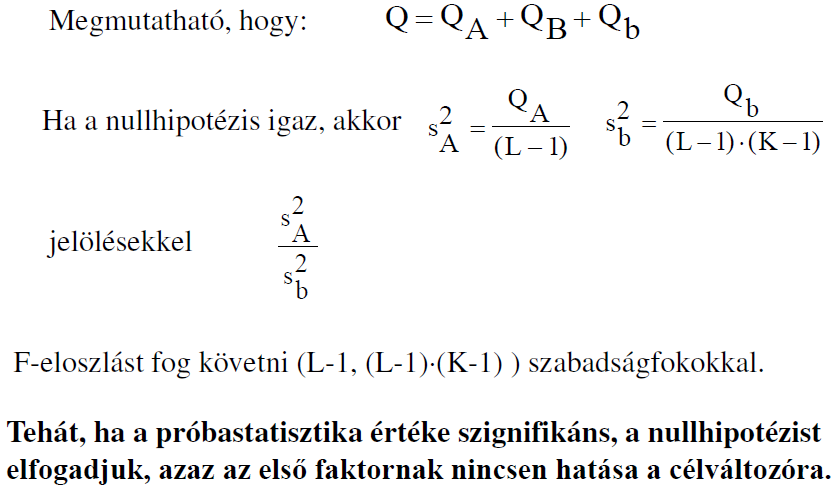
****

**Kétszeres osztályozás (interakció nélkül):**

****

****

****

****

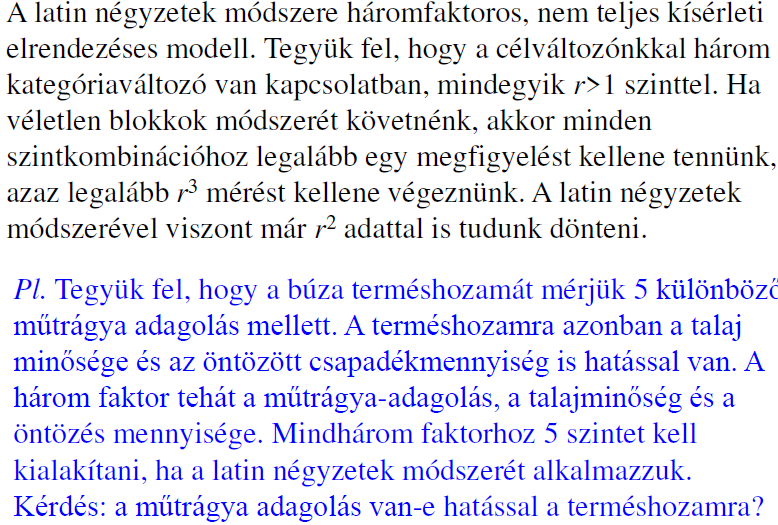
**Az interakció figyelembevétele:**

Ha figyelembe vesszük az interakciót két nominális faktor közt akkor a párosításnál megjelenik egy új tag, amely azt fejezi ki, hogy a két párosított faktor gyengíti vagy erősíti egymás hatását. Így egyszerre három hipotézist is ellenőrizhetünk:

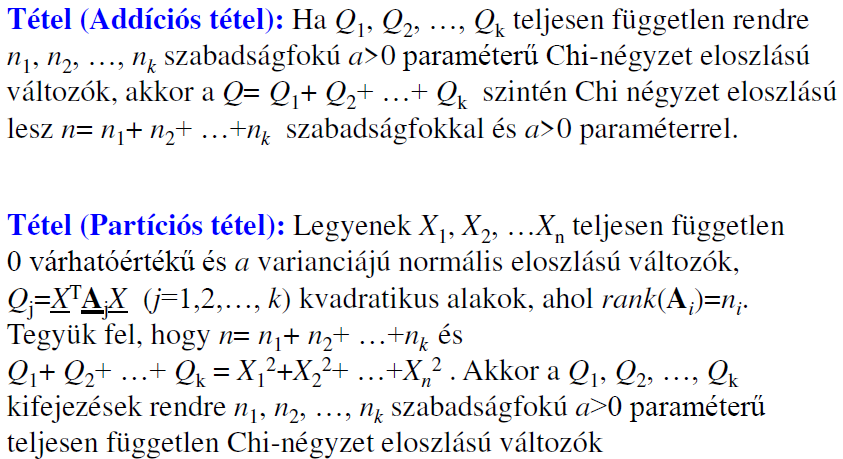
* H1:
* H2:
* H1,2 : interakció hatása zérus (c,ij = 0)

Először az interakciót vizsgáljuk meg (H1,2 -t) (F-eloszlást kell kövessen), ha ezt elfogadtuk H2 és H1-et nézzük meg.

**A latin négyzetek módszere**:



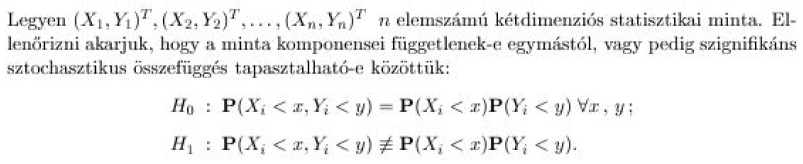
**Fisher-Cohran tételek.**



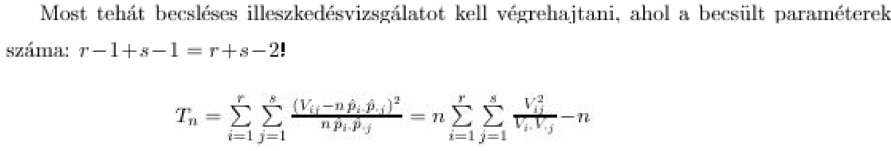
# 6. tétel

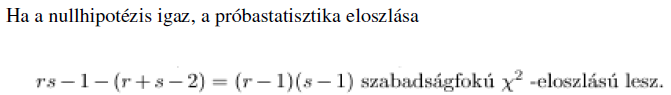
Nemparaméteres próbák

**Függetlenség vizsgálat:**



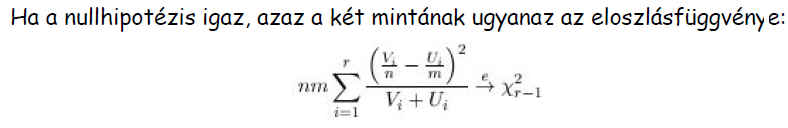
**A H0 hipotézis az, hogy a minták függetlenek!**





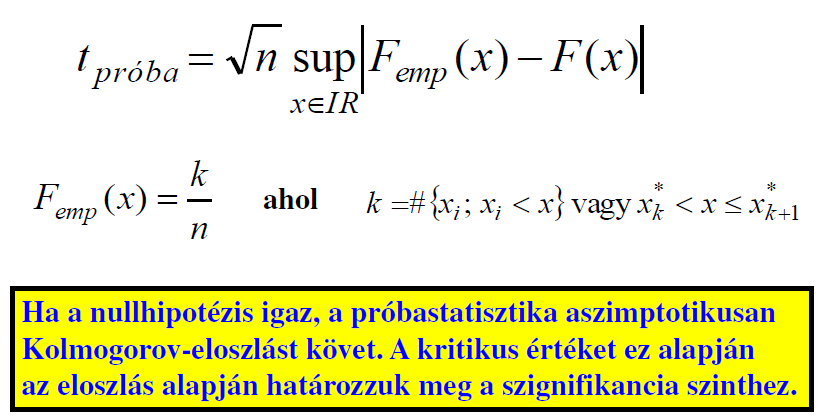
**Homogenitás vizsgálat:**

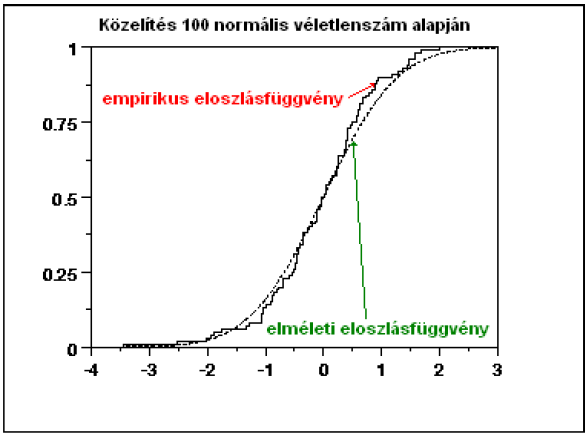
A homogenitás vizsgálat annak a kérdésnek a megválaszolására szolgál, hogy két valószínűségi változó azonos eloszlású-e. Adott két sorozat független minta, és a feladat, hogy eldöntsük, hogy a két minta eloszlása azonos-e.   
**A nullhipotézis itt az, hogy a minták azonos eloszlásúak!**

****

**Kolmogorov-Szmirnov**:

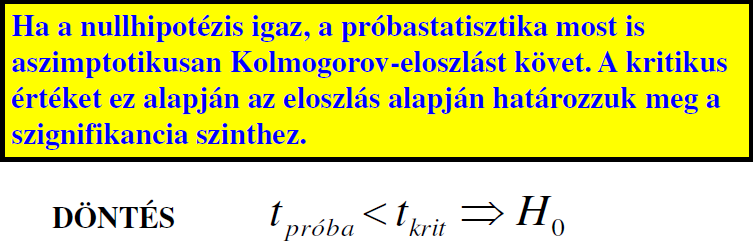
A nullhipotézis ezesetben az, hogy a minta eloszlásfüggvénye F(x).





**Kétmintás K-S:**

A nullhipotézis, hogy a két eloszlás azonos! Ez egy homogenitás vizsgálat ebben az esetben!



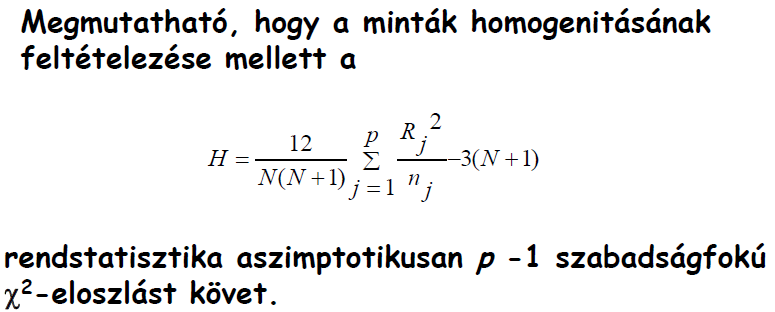
Kolmogorov- és Gnyegyenko-Koroljuk-tétel

**Mann-Whitney próba**: (két független minta homogenitásának vizsgálata)

Egy X minta adatait két részre osztjuk egy Y csoportképző változó segítségével. Megvizsgáljuk, hogy a két minta azonos eloszlásfüggvényhez tartozik-e. Rendezett mintát képezünk és rangszámokat adunk (Rx, Ry) Ha n és m nagy, akkor Rx és Ry normális lesz aszimptotikusan. Kis minták esetén a M-W táblázatot alkalmazzuk.

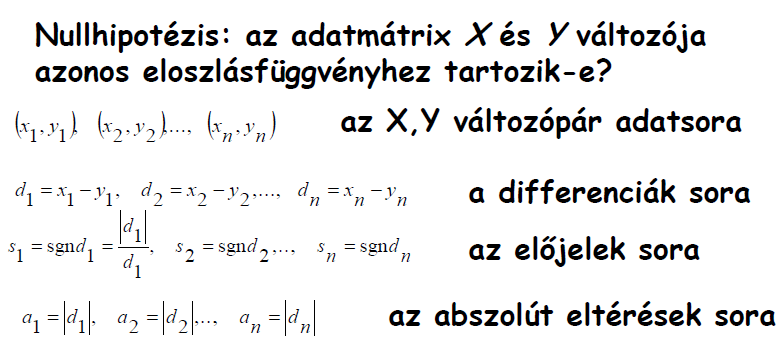
**Kruskal-Wallis próba**

Ellenőrizni szeretnénk azt a nullhipotézist, hogy p független minta ugyanabból az eloszlásból származik-e, vagyis a mintáknak közös-e az eloszlásfüggvényük.  
Pl.: a gdp eloszlása azonos-e az egyes földrészeken?



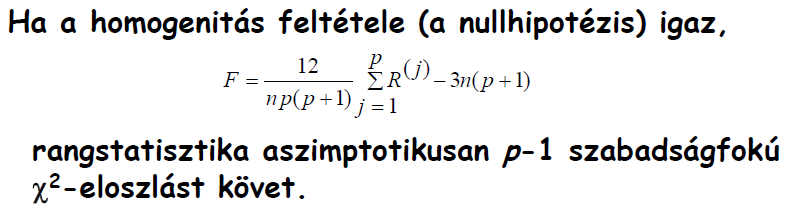
**Wilcoxon próba**

Két összetartozó minta homogenitásának ellenőrzésére szolgál.

****

**Friedman próba:**

Összesen p változó azonos eloszláshoz tartozását ellenőrizzük. Elő kell állítani a rangszámokat soronként!

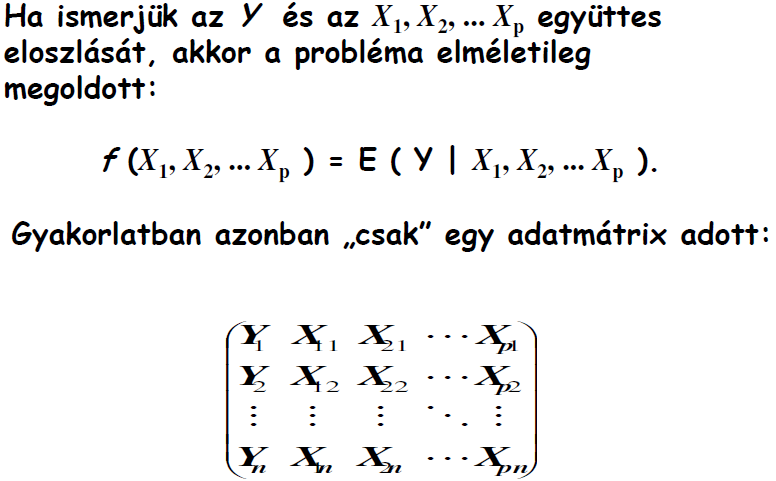


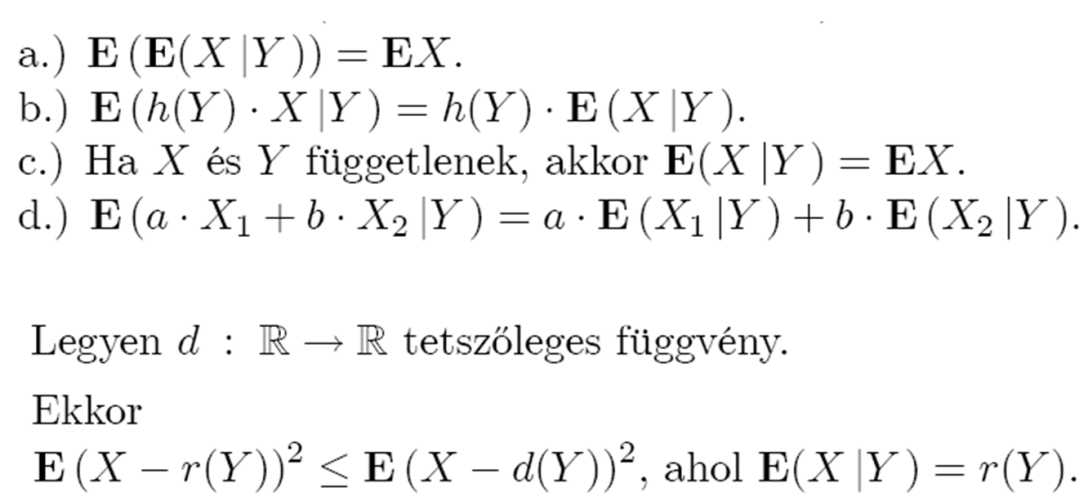
Ha az n minta elemszám kicsi, akkor a Friedman-táblázatot használjuk. Abban az esetben, ha a homogenitást el kellett vetni, akkor az összes (i,j) párokra vonatkozó kétdimenziós mintákon egyenként ellenőrizzük a homogenitás fennállását, pl. Wilcoxon próbával.

# 7. tétel

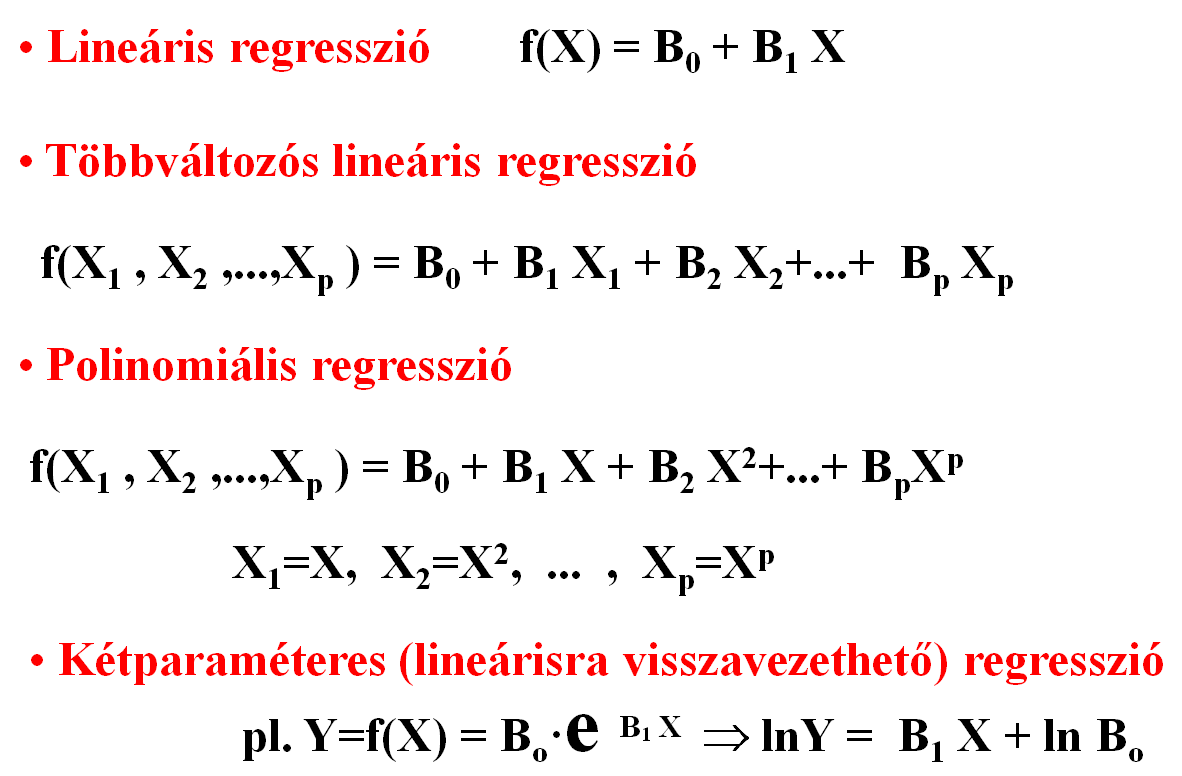
Regresszióanalízis I. - Regressziószámításkor egy változót egy (vagy több) másik változóval becslünk.

**Elméleti háttér: a feltételes várható érték.**

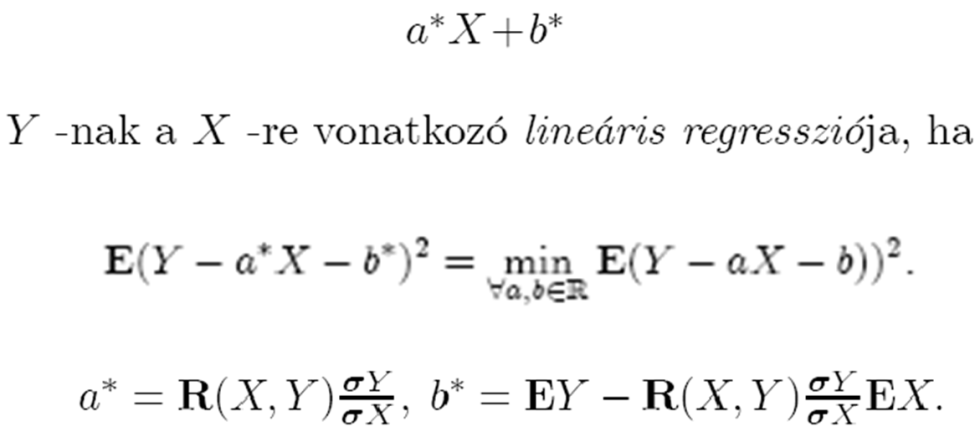




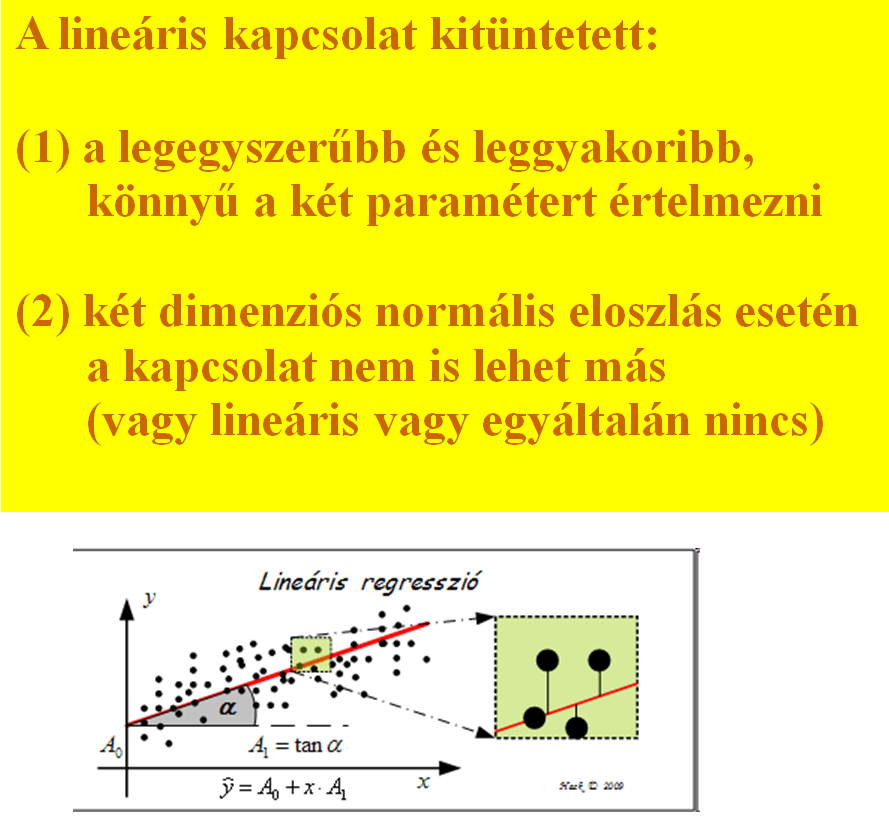
**A kétváltozós regresszió fajtái:**

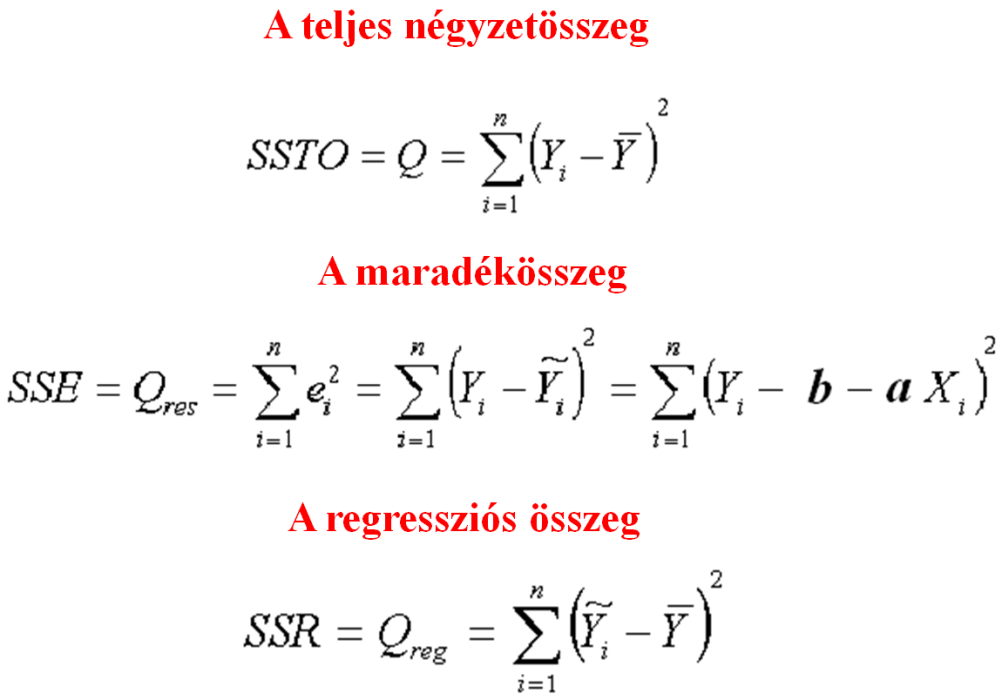
****

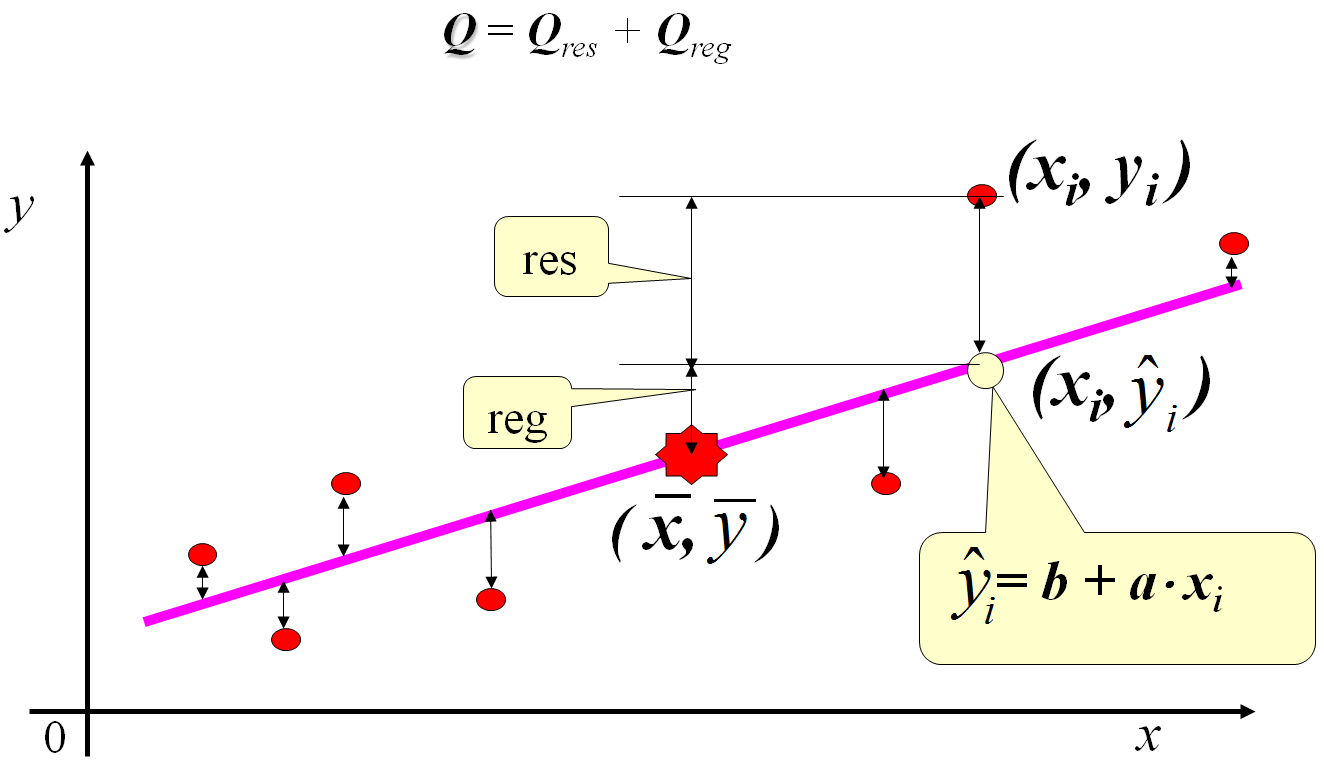
**Lineáris regressziók:**



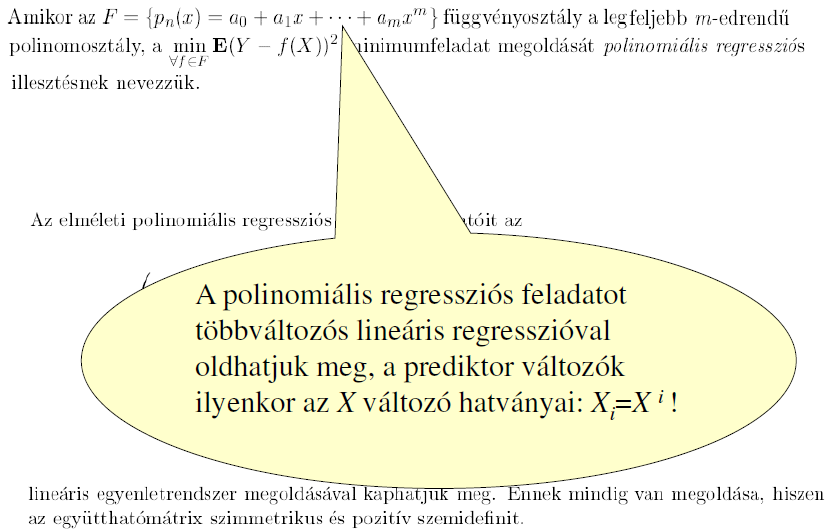
Ha X,Y együttes eloszlása normális, akkor a regresszió lineáris lesz.





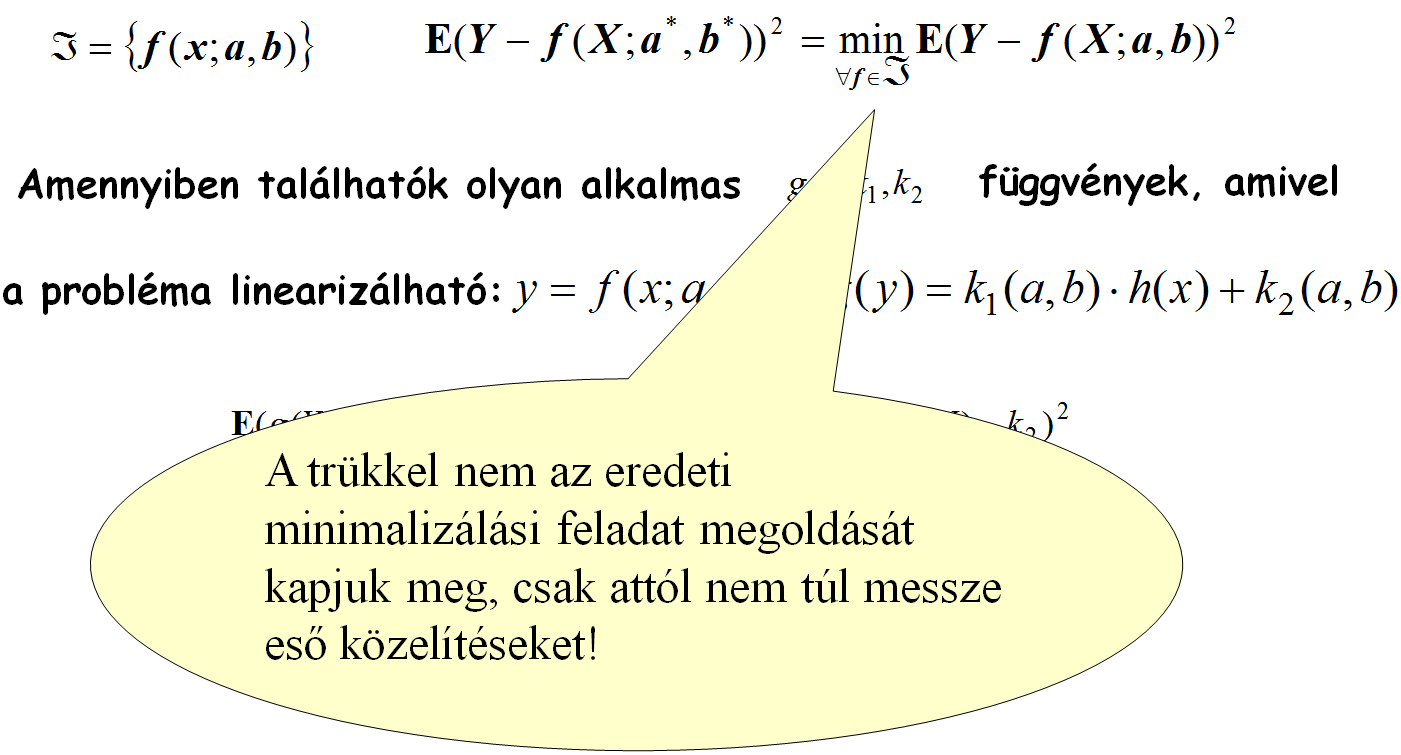


**Polinomiális regresszió:**



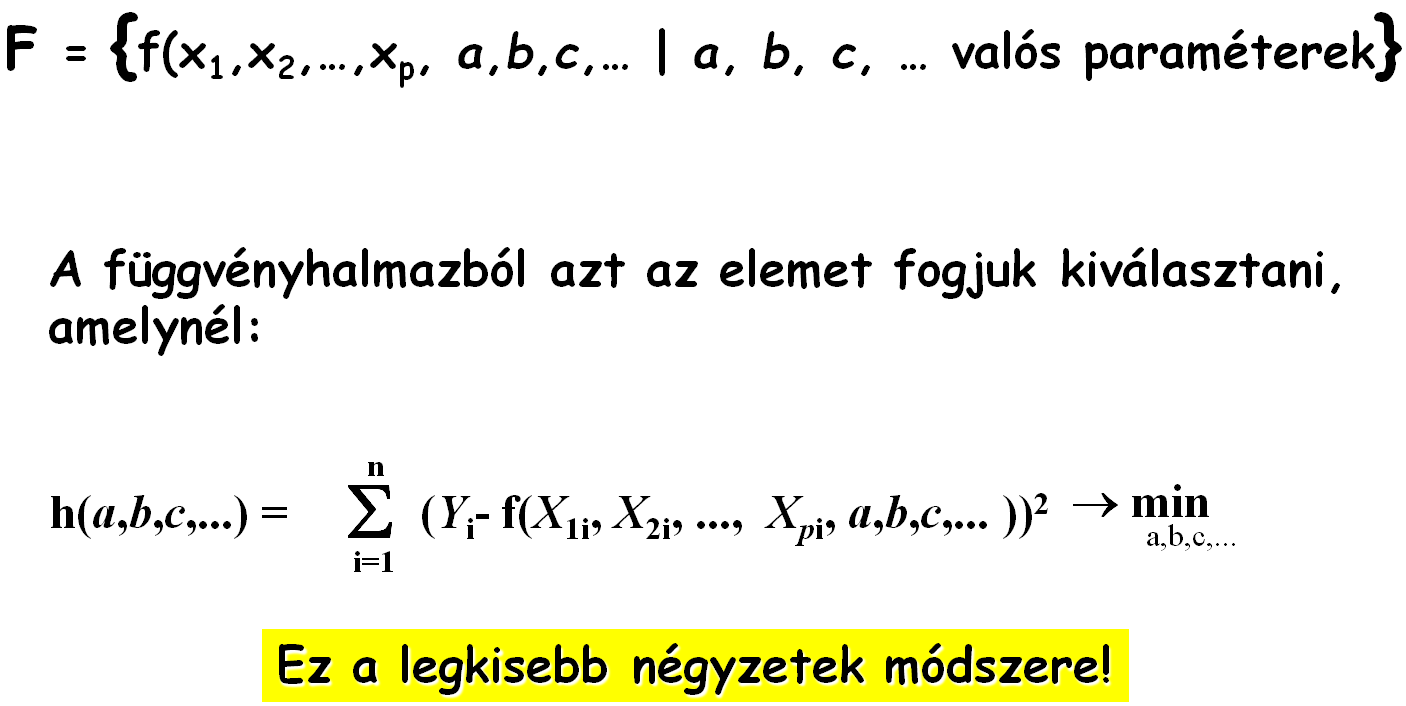
A polinomiális regresszió megoldását egy lineáris egyenletrendszer megoldásával kaphatjuk meg.

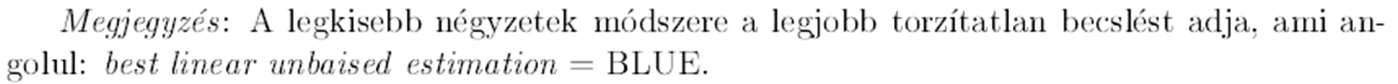
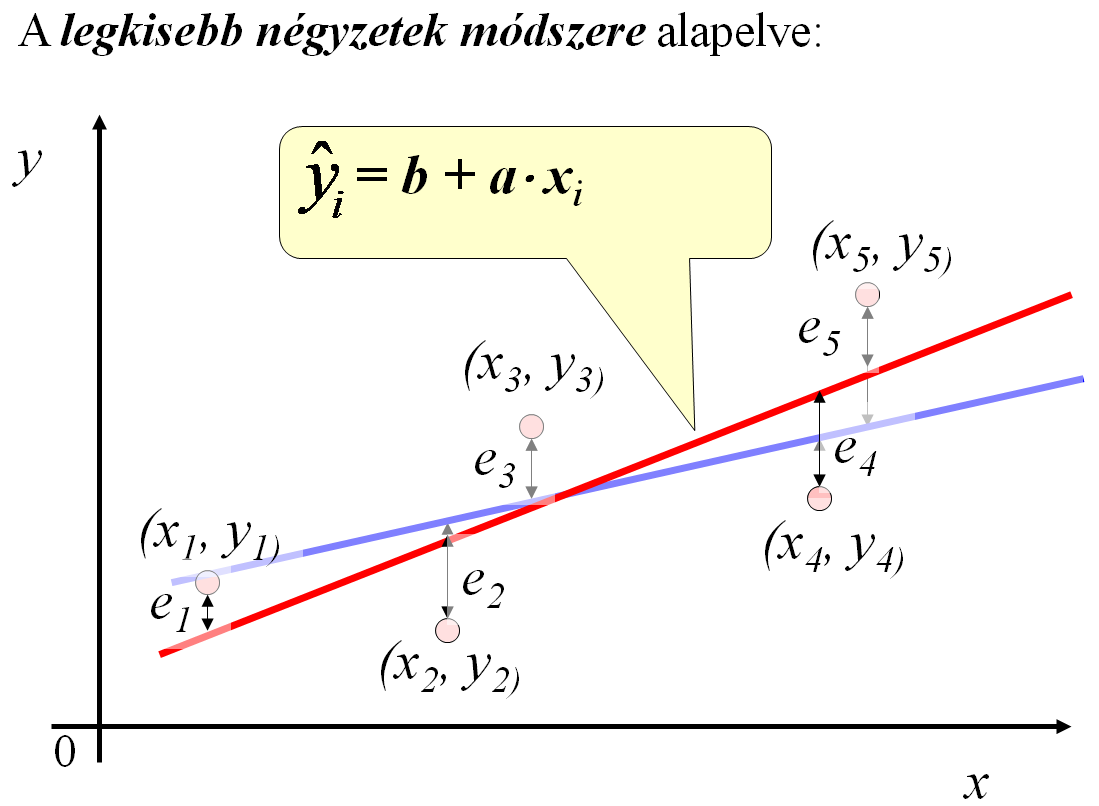
**Lineárisra visszavezethető kétparaméteres regressziók:**



* exponenciális kapcsolat:
* hatványfüggvény kapcsolat:
* Arrhenius:
* Reciprok:
* Racionális:
* homogén kvadratikus
* hiperbolikus
* logaritmikus

**A legkisebb négyzetek módszere:**

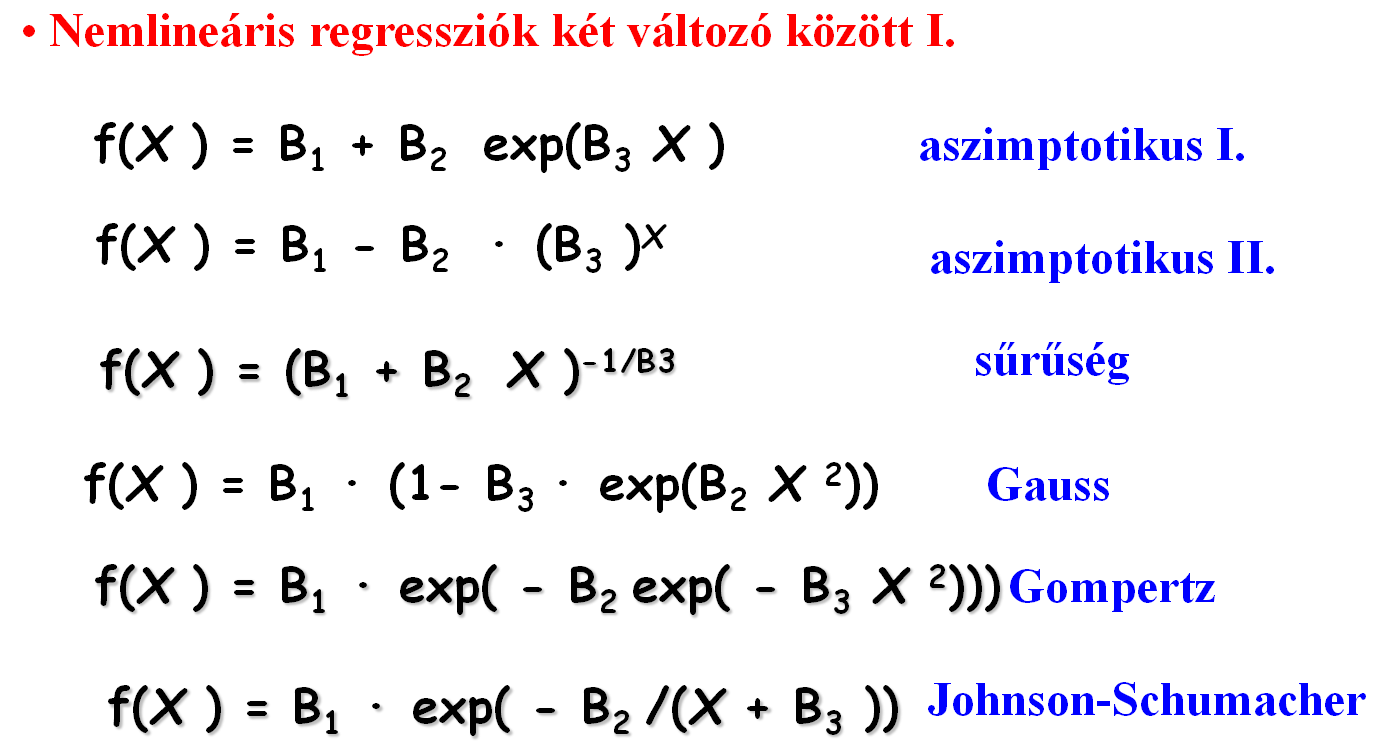




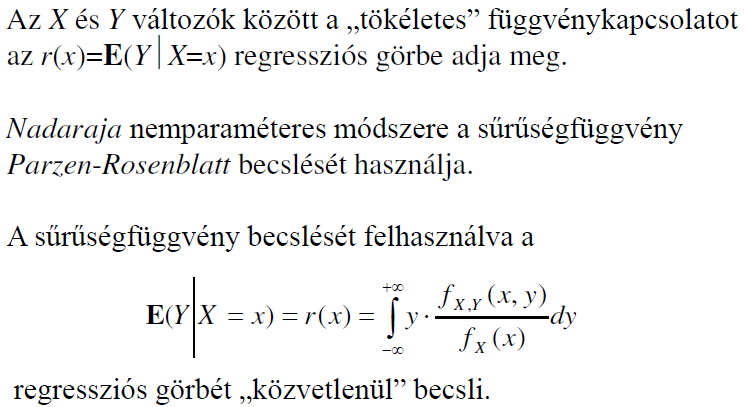
**Szórásanalízis (ANOVA) a modell érvényességének eldöntésére:**

**Meghatározottsági együttható (R-squared):**

**Nemlineáris regresszió:**



**Nadaraja-módszere:**



# 8. tétel

Regresszióanalízis II.

Többváltozós lineáris regresszió.

Modellépítési technikák.

Korrelációs együtthatók: totális-, többszörös-, parciális-.

A béta együtthatók.

Az adjusztált meghatározottsági együttható.

Multikollinearitás.

Heteroszkedaszticitás.

Outlier pontok detektálása, elemzése:

# 9. tétel

Faktor- és főkomponensanalízis.

A Kaiser-Meyer-Olkin statisztika és a minta-alkalmassági mérték (MSA).

A Bartlett-féle szfericitási-próba.

A k-faktoros modell és az átviteli mátrix.

Kummunalitás, a faktorok számának meghatározása.

Forgatások: varimax, equamax, quartimax.

A faktorok interpretálása. Watanabe-tétel.

# 10. tétel

Klaszteranalízis.

A metrikus tér, metrikafüggvények.

Dinamikus és hierarchikus módszer:

a k-közép módszer

agglomeratív klaszterezés.

McQueen-tétel.

Diszkriminanciaanalízis, osztályozási módszerek.

Tananyag, a tananyag előkészítése.

Tanulóalgoritmusok.

A legközelebbi társ módszere,

a legközelebbi k-társ módszere.

A legközelebbi társ gyors megkeresése.

# 11. tétel

Többdimenziós skálázás.

A skálák megbízhatósága.

A klasszikus MDS modellje.

Nemmetrikus módszerek,

a Shepard-Kruskal-algoritmus.

S-tress és stress mérőszámok,

RSQ.

Több kísérleti személy eredményeinek együttes kiértékelése.

Példák.

# 12. tétel

Kérdőíves felmérések módszertana.

Adatgyűjtési technikák.

Önkitöltős és kérdezőbiztosos adatfelvétel.

A kérdőív megszerkesztésének elvei.

Kérdések és állítások típusai.

Likert-skála, szimmetrikus differenciál, mátrix-kérdések.

Primer- és szekunder adatok.

Etikai vonatkozások.

# 13. tétel

**Alapfogalmak**:

**reprezentativitás**: A minta úgy kell, hogy tükrözze apopuláció tulajdonságait, ahogy a cseppben látjuk a tengert. Azaz a minta **reprezentatív** kell, hogy legyen.

**cenzus: népszámlálás**

**fókuszcsoport**:

**mintavételi keret**

**Mintavételezési technikák**: A populáció minden egyes elemének ugyanakkora esélyt kell biztosítani a mintába kerüléshez.

* **EVM**:
* **nem véletlen mintavételezés:**
* **rétegzett mintavételezés**: A populációt adott szempontok szerint csoportokba osztjuk, és a csoportok arányait a mintában is megtartjuk.
* **csoportos mintavételezés**: A mintába kerülő egyedeket sorsolással választjuk ki.
* **szekvenciális mintavételezés:**

**A szükséges minta elemszám meghatározása**: A minta elemszámának elég nagynak kell lennie ahhoz, hogy a következtetéseink átvihetők lehessenek a populációra is.

**A különböző egyenlőtlenségeken alapuló becslések a minimális mintaelem számra:**