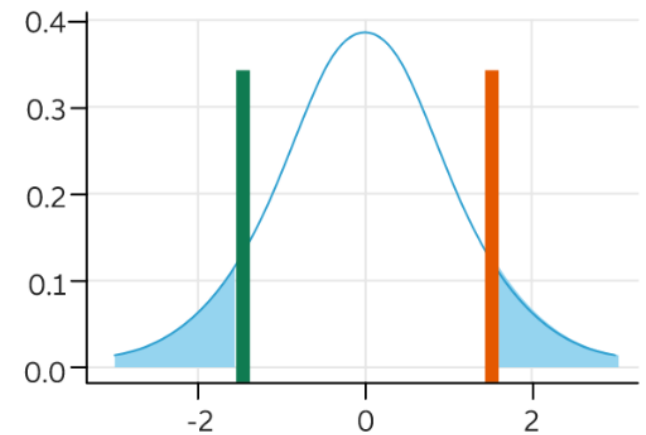
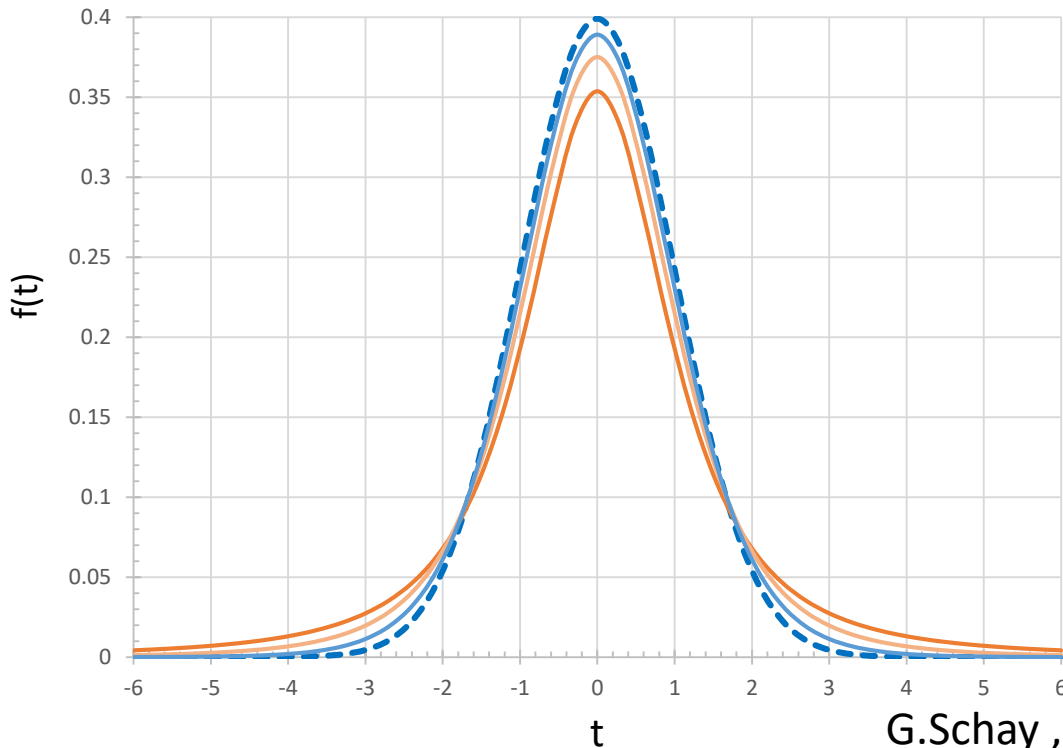
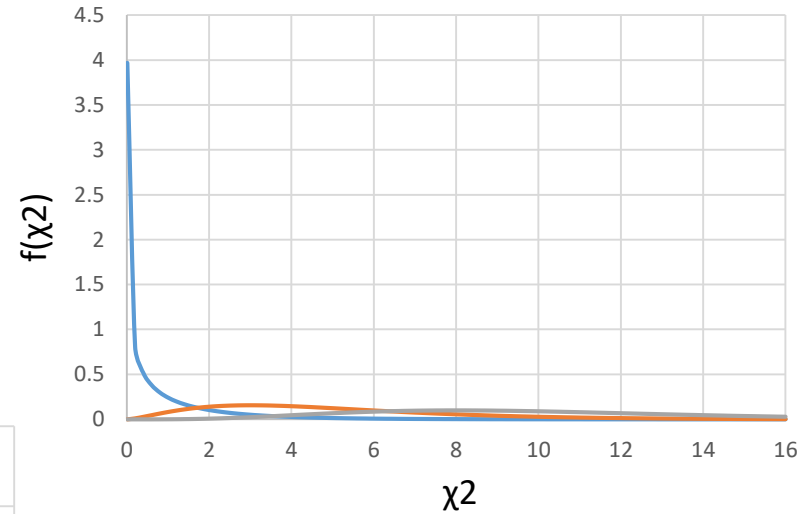
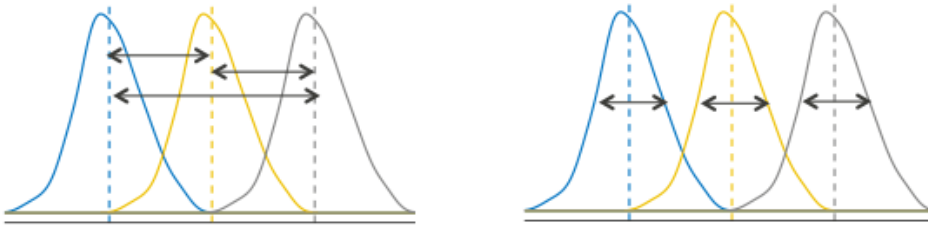


Hipotézis vizsgálatok a gyakorlatban

t-próba, χ^2 -próba, többszörös összehasonlítás

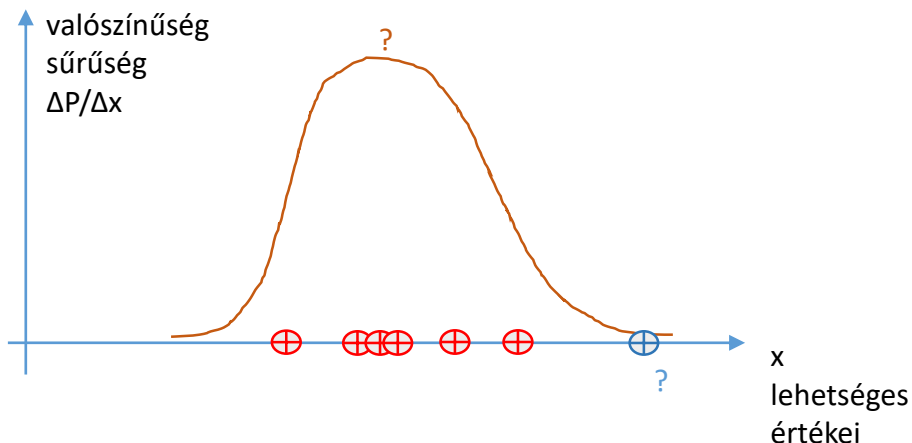


Bontsunk mindent I/N eldöntendő kérdés(ek) (sorozatára)

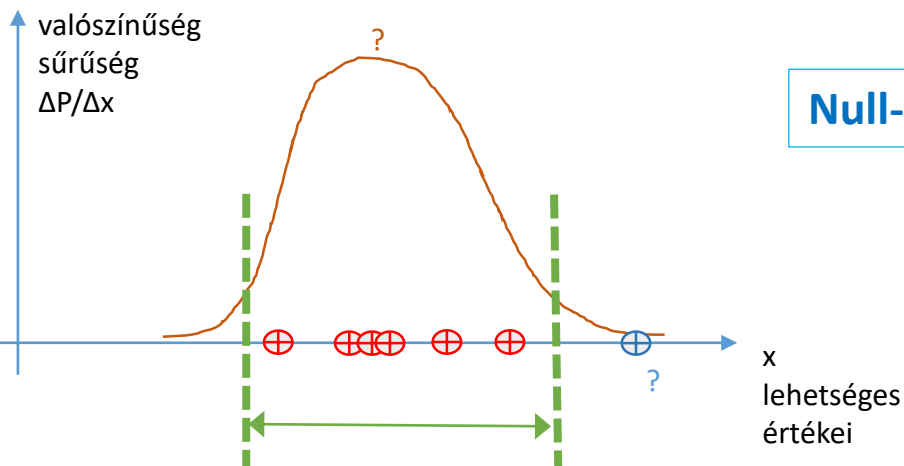
Hasonlítsuk az adott helyzetet egy **ismert, referencia** eloszláshoz!

A kérdés tehát úgy tehető fel, hogy „amit megfigyeltünk mint valószínűségi változót, vajon az általunk ismert, **előre feltételezett** eloszláshoz tartozik-e”

Null-hipotézis

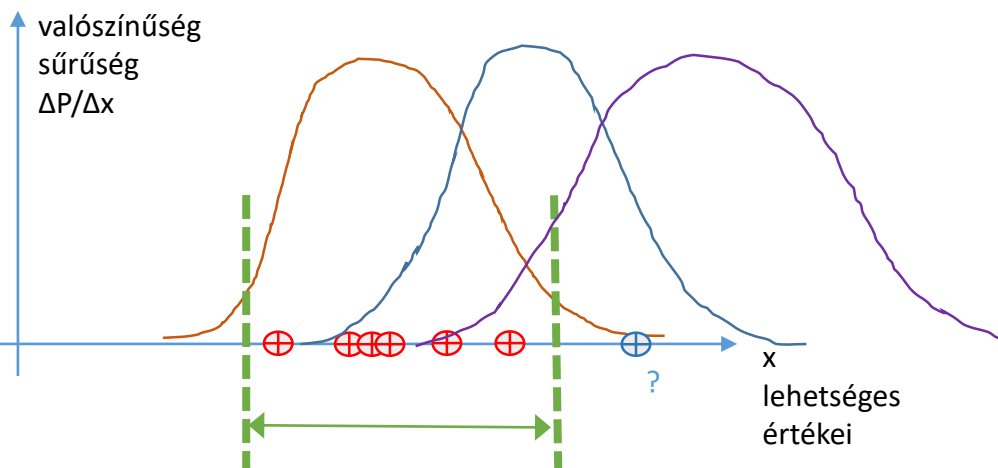


A kérdés tehát úgy tehető fel, hogy „amit megfigyeltünk mint valószínűségi változót, vajon az általunk ismert, **előre feltételezett** eloszláshoz tartozik-e”



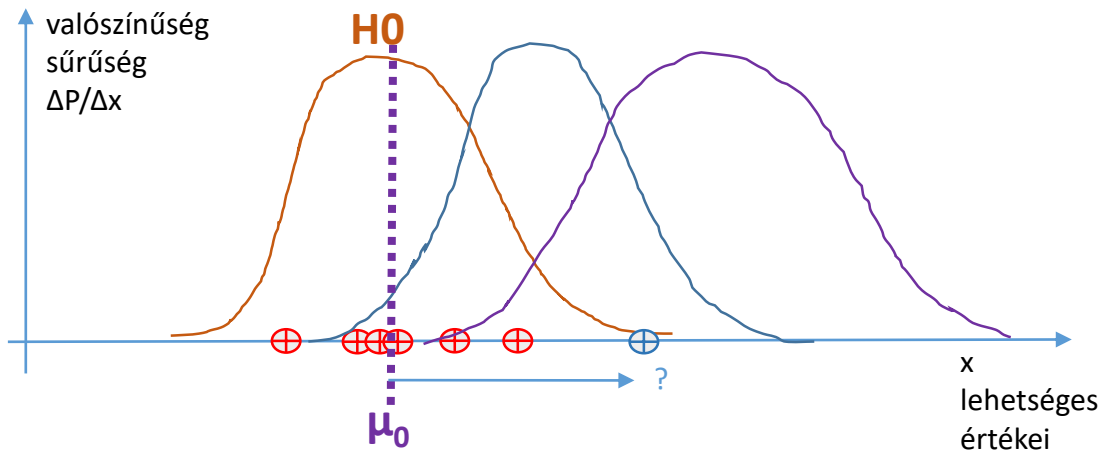
Példa: „Ha néhány mérésből átlagban 6.2-es vércukrot mértem a páciensnek akkor cukorbeteg-e?”

1. I/N kérdés: cukorbeteg?
2. Előfeltevés, null-hipotézis: A páciens a normál vércukor-eloszláshoz tartozik. (normál tartomány **4.0-5.9** mmol/L)



Nem egyértelmű a válasz!

Több eloszláshoz is tartozhat a mért értékünk!

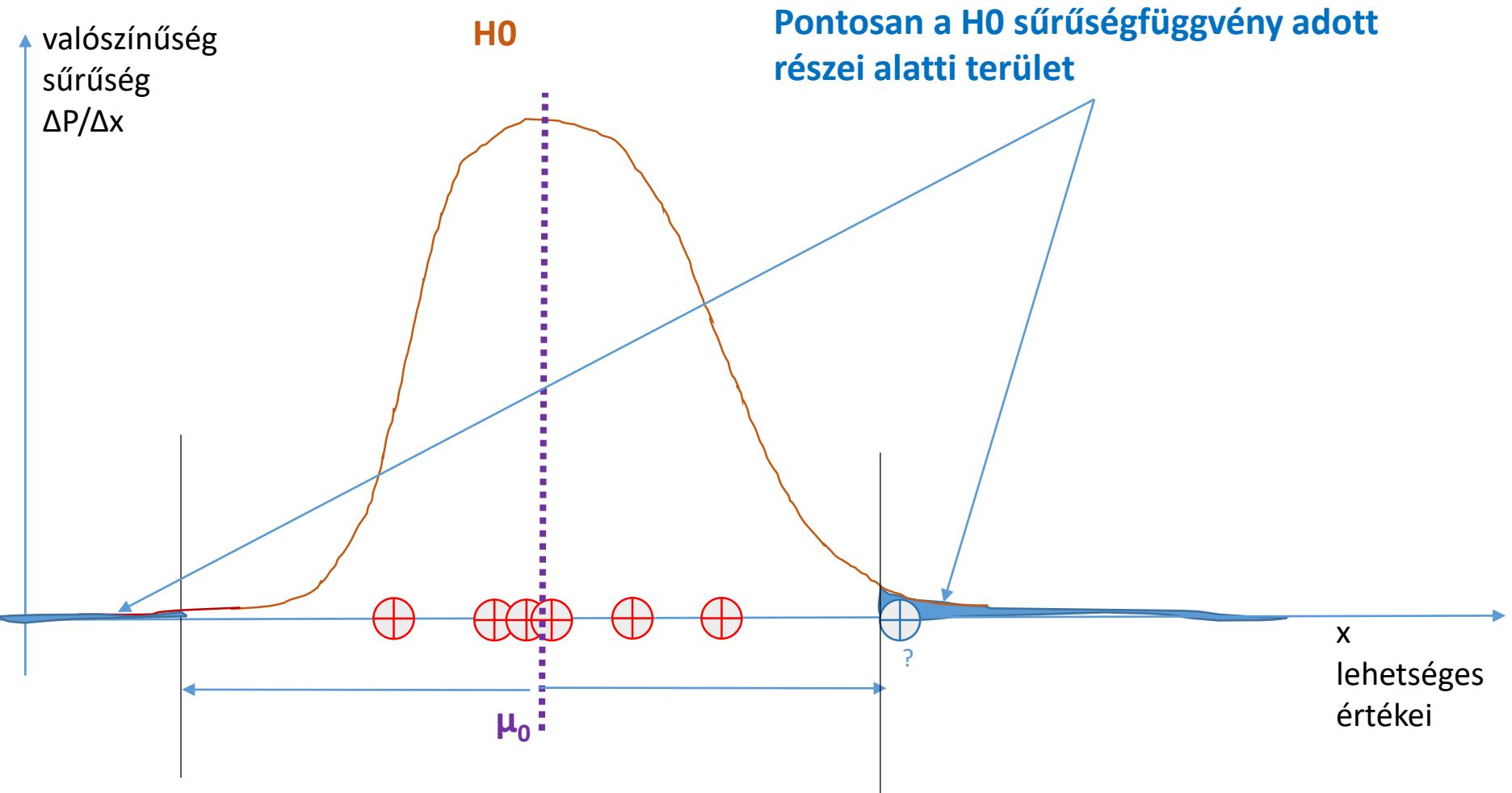


Pontosítjuk a kérdést:

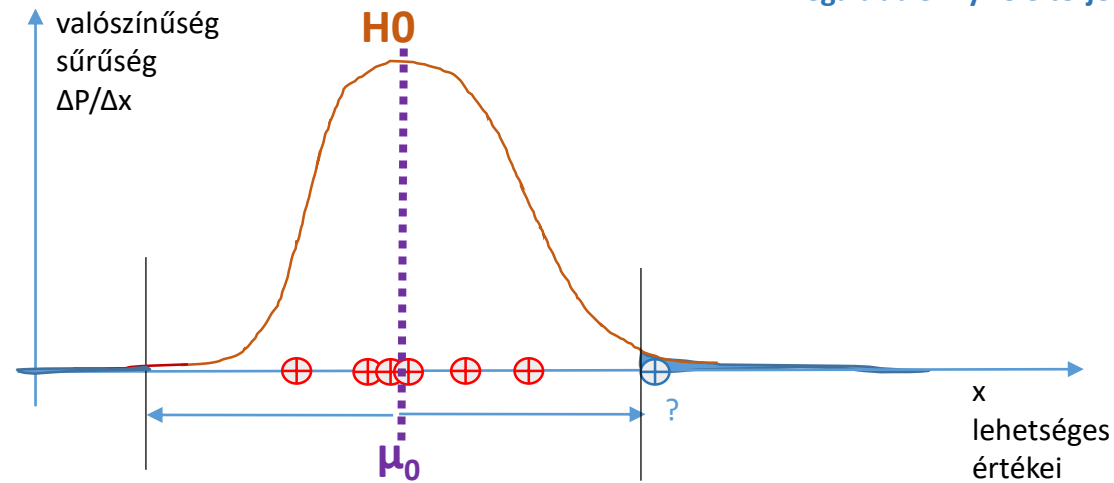
Mekkora a valószínűsége, hogy **HA a H_0 eloszláshoz tartozik** a vizsgált páciens, akkor **legalább ennyire eltérjen** a várható értéktől (μ_0) mint amit látunk?

Ez már egyértelműen kiszámolható!

Mekkora a valószínűsége, hogy **HA a H_0 eloszláshoz tartozik** a vizsgált páciens, akkor **legalább ennyire eltérjen** a várható értéktől (μ_0) mint amit látunk?



Mekkora a valószínűsége, hogy **HA a H_0 eloszláshoz tartozik** a vizsgált páciens, akkor **legalább ennyire eltérjen** a várható értéktől (μ_0) mint amit látunk?



Döntésünk is valószínűségi lesz:

Ha túl kicsi a valószínűség, akkor elvetjük az eredeti feltevést.

Ha pedig nem túl kicsi, akkor megtartjuk az eredeti feltevést.

A dichotóm (I/N) döntési fa

Kérdés



Eldöntendő kérdéssé alakítás ha szükséges

H0 megfogalmazása

Az adataink H0-hoz tartozása számítható kell hogy legyen, ezért H0 többnyire egy ismert, referencia-helyzetet mint előzetes feltevést jelent.



Eldöntjük mennyire ragaszkodunk H0-hoz, azaz mekkora az a valószínűség ami alatt már nem tartjuk meg az előfeltevést. -> ezt hívjuk **szignifikancia szint**nek

α vagy α_{\max}



Kiszámítjuk a valószínűségét annak, hogy véletlenszerűen *legalább ennyire* eltérjen egy eredmény a H0-tól mint amennyire a mi adataink eltérnek.

-> Ez valójában egy feltételes valószínűség: **$P(\text{legalább ekkora eltérés} \mid H0 \text{ igaz})$**



Döntés

$P < \text{szign.}$

$P \geq \text{szign.}$

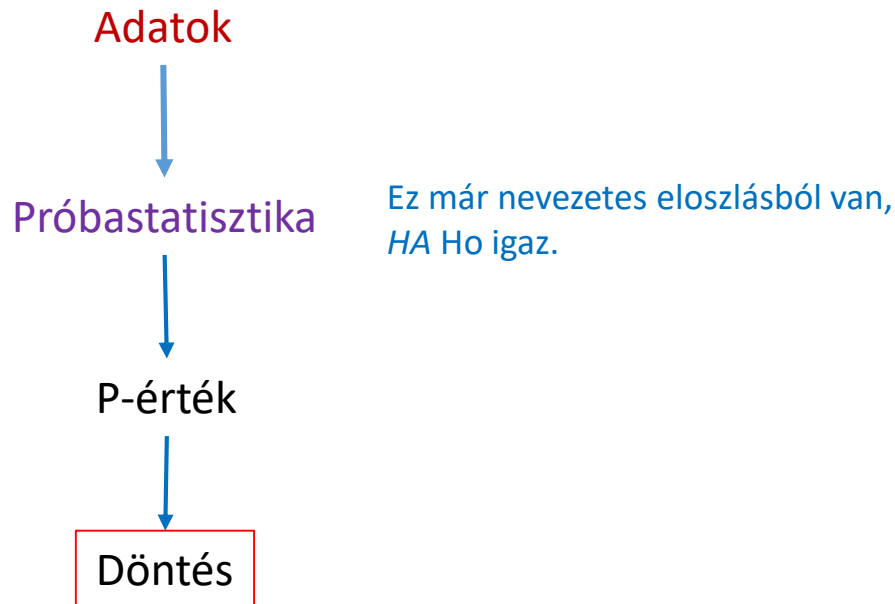
H0-t elvetjük

H0-t megtartjuk

Azaz nem vetjük el!

Gyakran kell az adatainkat konvertálni, átszámítani, hogy valamely ismert eloszláshoz jussunk, és így a görbe adott szakaszai alatti terület kiszámítható legyen.

A kérdésfeltevéstől és az eredeti adatoktól függ, hogy milyen számolással és milyen nevezetes eloszláshoz lehet jutni



Frequency counts (test is for first level):

DIABETES	
diabetes	egészséges
9	91

1-sample proportions test with continuity correction

data: rbind(.Table), null probability 0.105

X-squared = 0.10641, df = 1, **p-value = 0.7443**

alternative hypothesis: true p is not equal to 0.105

95 percent confidence interval:

0.04455927 0.16833464

sample estimates:

p
0.09

nem kell kiszámolni a próbastatisztikát külön, a program rögtön a P-értéket is megadja

Kérdésfelvetés: eltérnek-e egymástól két mérési sorozatban kapott átlagok?

- A) Azonos sokaságon mértük, csak közben valamilyen hatás érte őket. (pl. terápia, idő, rizikó, stb) -> „előtte – utána” jellegű összehasonlítás
- B) Két eltérő sokaságot hasonlítottunk össze.

Feltételek:

- i) Az adatok normális eloszlású populációból vett minták. Nem kell azonos eloszlásnak lennie, csak legyen Gauss.
(sőt! elég ha nagy a minta, ld. később)
- ii) A két csoport szórása azonos (homoszkedasztikus). Ha nem akkor erre korrigálni kell: Welch-teszt.
- iii) A minták elemei függetlenek (tehát az egyes mérések eredményei egymást semmilyen sem befolyásolják)

Ha a fentiek teljesülnek akkor **t-próbát** lehet csinálni.

Egy mintás (párosított, előtte-utána)

Páciens azonosító	Kezelés előtt mért	Kezelés után mért
1	X1	Y1
2	X2	Y2
3	X3	Y3
...

H0: x és y adatok azonos eloszlásból származnak.

$$D1 = Y1 - X1$$

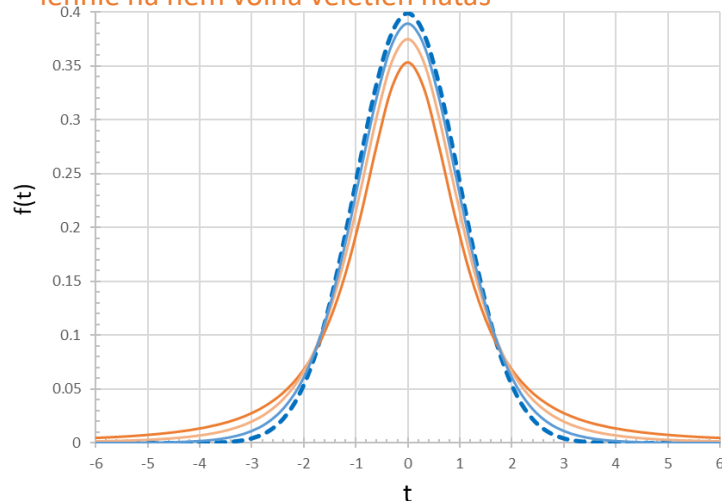
$$D2 = Y2 - X2$$

...

$$t_{n-1} = \frac{\bar{D}}{s_{\bar{D}}}$$

Azaz: az átlag szórásának hányszorosával tér el az átlag a várt 0-tól?

Ez egy n-1 szabadság fokú t-eloszlás függvényből vett véletlen szám lesz, ha H0 igaz. 0-nak kellene lennie ha nem volna véletlen hatás



Adatok



Próbastatisztika

P-érték

Döntés

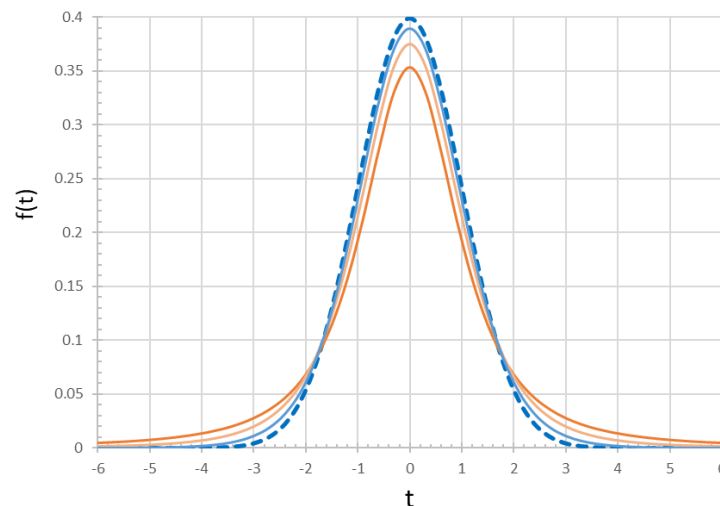
Két csoport összehasonlítása

A csoport	B csoport
X1	Y1
X2	Y2
X3	Y3
...	...

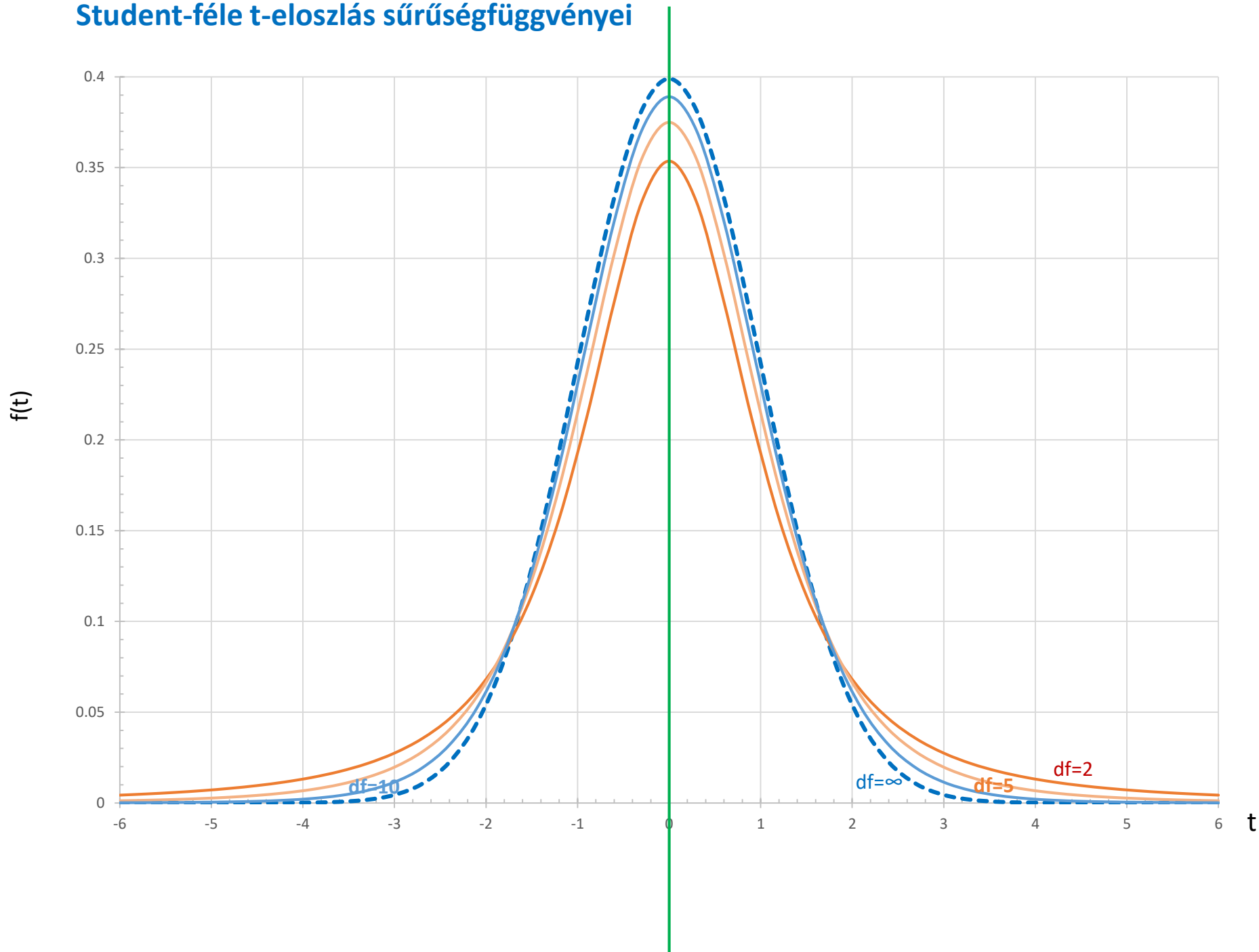
H0: x és y adatok azonos eloszlásból származnak.

$$t \sim \frac{\bar{x} - \bar{y}}{s^*}$$

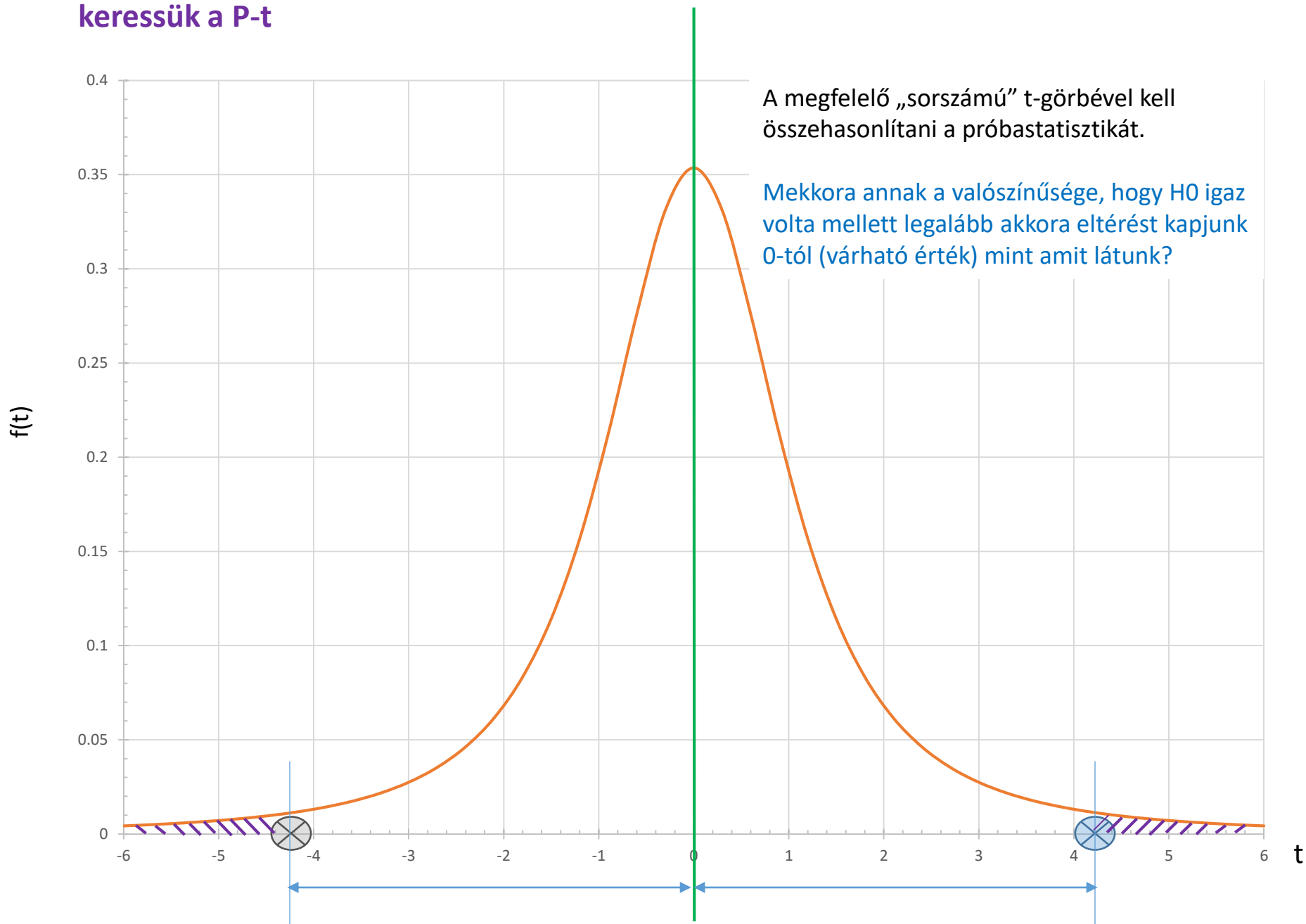
Az átlagok különbsége a korrigált közös szórás hányszorosával tér el a várt nullától.



Student-féle t-eloszlás sűrűségfüggvényei



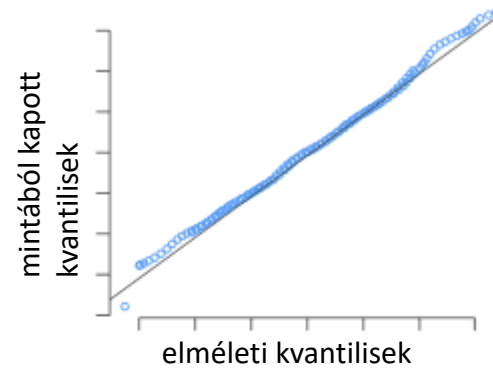
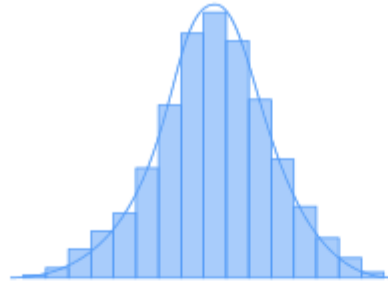
keressük a P-t



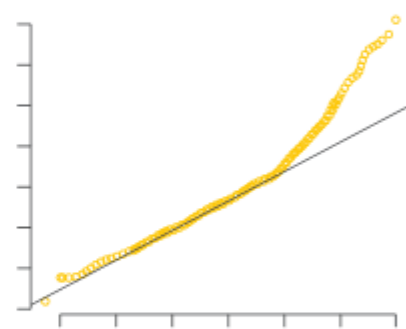
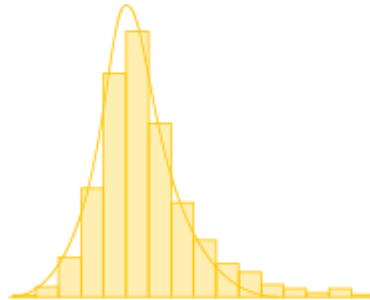
Ha P kicsi, akkor H_0 -t elvetjük, ellenkező esetben megtartjuk

Normalitás ellenőrzése: QQ-plot

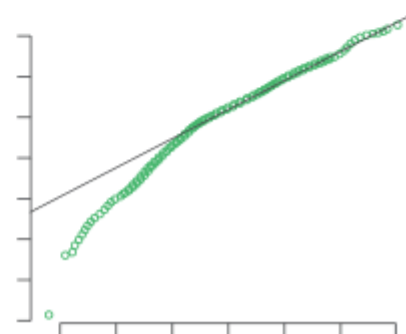
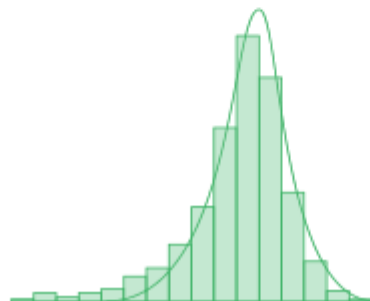
normál eloszlás



Jobbra ferde eloszlás



Balra ferde eloszlás



Megjegyzés:

Ha elég nagy elemű a minta, akkor a centrális határeloszlás tétel miatt az átlag eloszlása már jó közelítéssel Gauss, azaz normál eloszlású lesz. Mivel a képletekben csak az átlagok eltérését nézzük, így nagy elemű (pl. több mint 30...50) minták esetén nem kell az alapsokaságnak normál eloszlásúnak lennie.

Ez sokszor segítség!

Ha viszont a minta normál eloszlású, akkor nem kell az elemszámmal foglalkozni.

Függetlenség:

Meglevő mintavételből nehéz kimutatni ha nem teljesül az egyes elemek függetlensége, legjobb ha a kísérlet tervezésekor gondolunk rá.

Ha kérdéses, akkor ez egy külön kérdés, amit saját hipotézis-vizsgálatával meg kell nézni.

Szórás homogenitás:

Erre is létezik külön próba, pl. F-teszt, Cochrane-teszt, stb.
Ezek többnyire nagyon érzékenyek a normalitásra is.

Legjobb a homogenitás hiányára felkészülni, és pl. módosított próbát (Welch-teszt) végezni.
(heteroszkedasztikus korrekció)

Mit tegyünk ha több csoportunk van?

Naív ötlet: szépen összehasonlítgatjuk mindet. DE! Ilyenkor több döntést hozunk, és minden döntésnek van hibája. Ezek a hibalehetőségek össze fognak adódni! A végén meglepően nagy valószínűséggel hozunk hibás döntést (sokkal valószínűbben mint a szignifikancia-szintből gondolnánk!)

Megoldás:

- A) Korrigálunk a többszörös döntés okozta hibalehetőségre (effektíven lejebb visszük a döntések számától függően az egyedi szignifikancia-szinteket).
- B) „egyben” eldöntjük, hogy minden csoportunk egyforma-e vagy sem! Ha mind egyformák akkor készen vagyunk, ha nem akkor lehet keresni eltérést *post-hoc* tesztekkel.

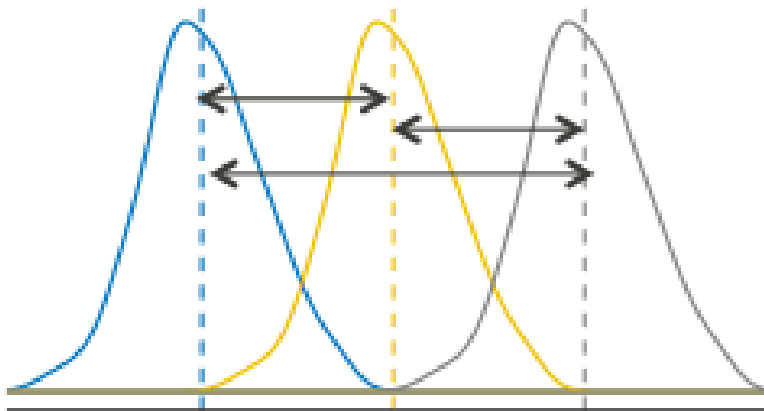
Több csoport vizsgálata: van-e „kakukktojás” ?

H_0 : nincsen, minden csoport egyforma.

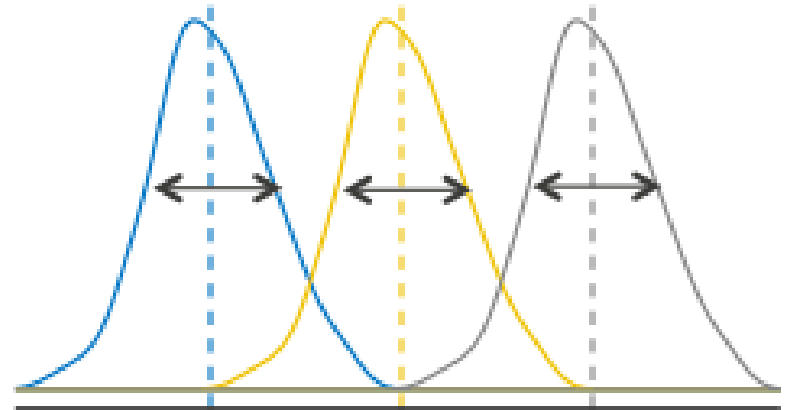
ANalysis Of VAriance (ANOVA)

Alapötlet: vizsgáljuk meg a **szórásokat**!

Ha homogén a társaság akkor mindegy csoporton belül, vagy „keresztben” nézzük a szórást.



csoportok közötti szórások



csoportokon belüli szórások

Függetlenség vizsgálata *kategóriális változó esetén*

-> Folytonos esetre regressziós modellt érdemes inkább használni!

χ^2 - próbák

Alaphelyzet: vannak gyakoriság adataink, bizonyos változók szerint csoportosítottuk a megfigyeléseinket. Meg szeretnénk tudni, hogy függetlenek-e a változóink egymástól.

pl. van-e hatása a mobiltelefon használatnak a reakcióidőre? A telefon használati szokások kategóriális változó.
van-e összefüggés a dohányzás és a tüdőrák előfordulása között? Itt mindkét változó kategóriális.

	B1 pl. dohányzik	B2 pl. nem dohányzik
A1 pl. rákos	n1	n2
A2 pl. nem rákos	n3	n4

A és B a két változónk, akár több lehetséges értékeik is lehetnek, nem csak igen/nem.

n-ek a megfigyelt gyakoriságok.

	B1 pl. dohányzik	B2 pl. nem dohányzik	
A1 pl. rákos	n1	n2	$P_{A1}=(n1+n2)/N$
A2 pl. nem rákos	n3	n4	$P_{A2}=(n3+n4)/N$
	$P_{B1}=(n1+n3)/N$	$P_{B2}=(n2+n4)/N$	

A P valószínűségekre a függetlenség tételét használhatjuk (ezért jó H_0 az hogy függetlenek)

Tehát pl. $P(A1 \text{ és } B1) = P_{A1} * P_{B1}$, stb.

Továbbá ha az együttes várt valószínűségek ismertek, akkor az **abszolút gyakoriság várható értéke** a $P * N$ lesz minden esetben!

	B1 pl. dohányzik	B2 pl. nem dohányzik	
A1 pl. rákos	n1	n2	$P_{A1}=(n1+n2)/N$
A2 pl. nem rákos	n3	n4	$P_{A2}=(n3+n4)/N$
	$P_{B1}=(n1+n3)/N$	$P_{B2}=(n2+n4)/N$	$N=n1+n2+n3+n4$

megfigyelt gyakoriságok
(kontingencia táblázat)



	B1 pl. dohányzik	B2 pl. nem dohányzik	
A1 pl. rákos	$P(A1*B1)*N$	$P(A1*B2)*N$	
A2 pl. nem rákos	$P(A2*B1)*N$	$P(A2*B2)*N$	
			$N=n1+n2+n3+n4$

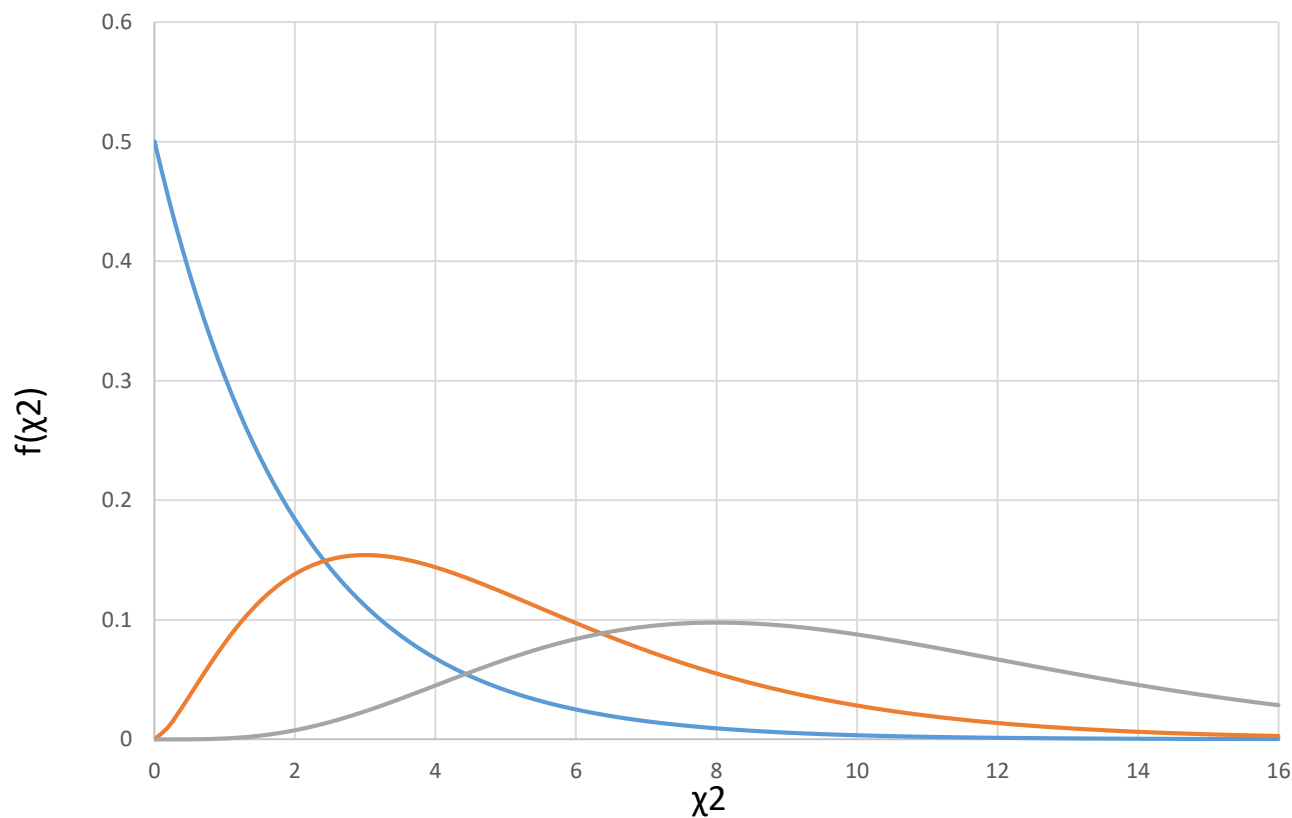
H0 igaz volta esetén várt
gyakoriságok
(segéd-kontingencia táblázat)

Celláról cellára nézve az eltéréseknek ha H0 igaz nagyon kicsinek (valójában nullának) kellene lennie

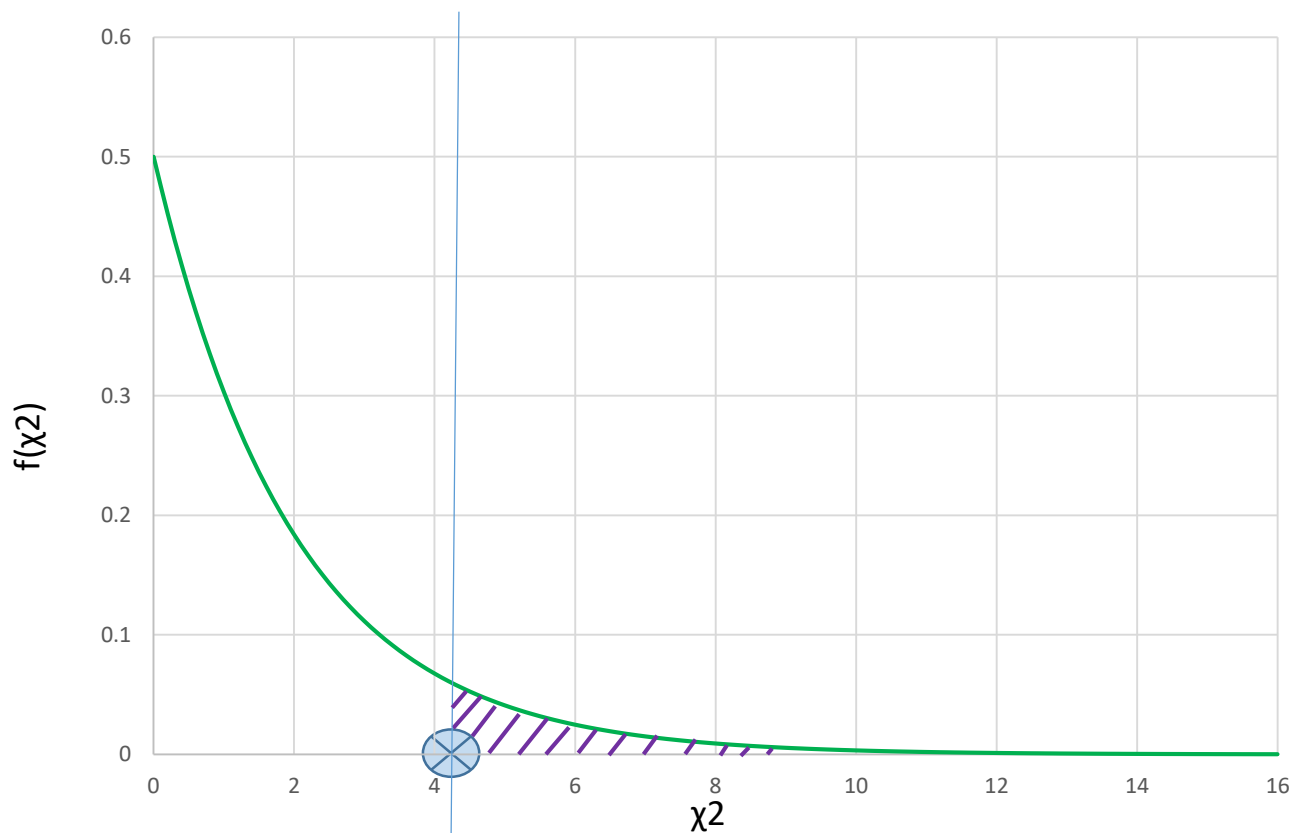
$$\chi^2 = \sum_{A,B} \frac{(Megfigyelt - Várt \text{ érték})^2}{Várt \text{ érték}}$$

χ^2 értékét össze tudjuk vetni a megfelelő sűrűség függvényekkel
A „szabadságfok” itt a táblázat sorai és oszlopai számának (1-el csökkentve!) szorzata.

pl. itt $(2-1)*(2-1) = 1-1 = 1$ lesz a szabadság-fok, tehát az 1-es „sorszámú” görbe kell a görbe-seregéből.



Megkeressük annak a valószínűségét hogy legalább ekkora próbastatisztika paramétert kapunk H_0 igaz volta esetén



Ha ez kicsi, akkor a H_0 -t elvetjük, egyébként megtartjuk

-> A χ^2 próba bármilyen ismert eloszlásra való illeszkedést is tud keresni.