Biofizika 2.
fogorvostan hallgatóknak

14. (utolsó) előadás:

Biostatisztika III.

2021. Május 17.

Veres Dániel

VÁLTOZÉKONYSÁG

Megismeréséhez és leírásához:

Más **Gondolkozásmód**

új **Nevezéktan**

csekély **matematika**

A statisztika kulcsszavai

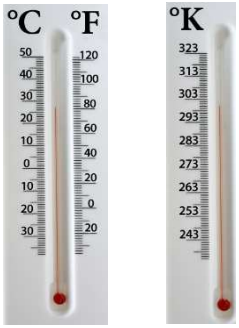
VÁLTOZÉKONYSÁG

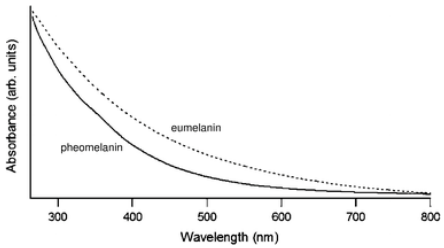
Sztohasztikus

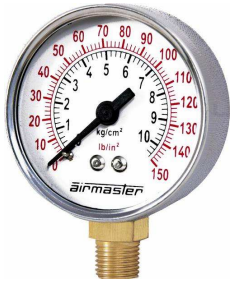
Véletlen

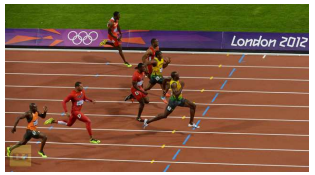
Változók, kimenetelek

Amit meg tudunk mérni vagy meg tudunk figyelni.









Változók típusai, mérési skálák

kvalitatív (kategoriális)

kvantitatív (számszerű)

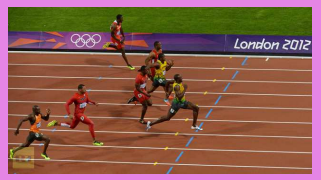
nominális „skála”	ordinális skála	intervallum-skála	arány-skála
= ≠	= ≠ < >	= ≠ < > + −	= ≠ < > + − × ÷

természetes sorrend

ismert távolság

természetes nullérték

(KVÁZI-) DISZKRÉT



(KVÁZI-) FOLYTONOS

Becslések

Alapsokaság és minta

Alapsokaság (populáció)

Az **alapsokaság** rendszerint olyan méretű, hogy az összes eleme nem vizsgálható meg.

„VÉLETLEN”

Minta

Emiatt az alapsokaságnak csak egy részhalmazát vizsgáljuk, ezt nevezzük mintának.

Alapsokaság és minta

Alapsokaság (populáció)

Az **alapsokaság** rendszerint olyan méretű, hogy az összes eleme nem vizsgálható meg.

„VÉLETLEN”

Minta

Emiatt az alapsokaságnak csak egy részhalmazát vizsgáljuk, ezt nevezzük mintának.

A minta elemein méréseket végzünk, majd az így keletkező adathalmazt (amit szintén mintának nevezünk) grafikusan és matematikailag jellemezzük

A minta jellemzői alapján az alapsokaságra vonatkozó következtetést vonhatunk le

BIZONYTALANSÁG! (P és H)

Becslés

Alapsokaság (populáció)	← Becslés	Minta
Valódi érték		Becslés
Valószínűség		Relatív gyakoriság
Várható érték (populációs átlag)		Minta átlaga
Elméleti variancia (szórásnégyzet)		Minta variancia
Két várható érték különbsége		Mintaátlagok különbsége

Hiba

Alapsokaság (populáció)

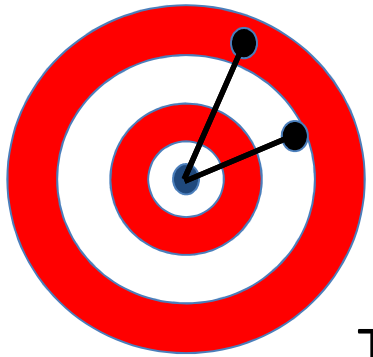
- Valódi érték

Becslés

Minta

- Becslések

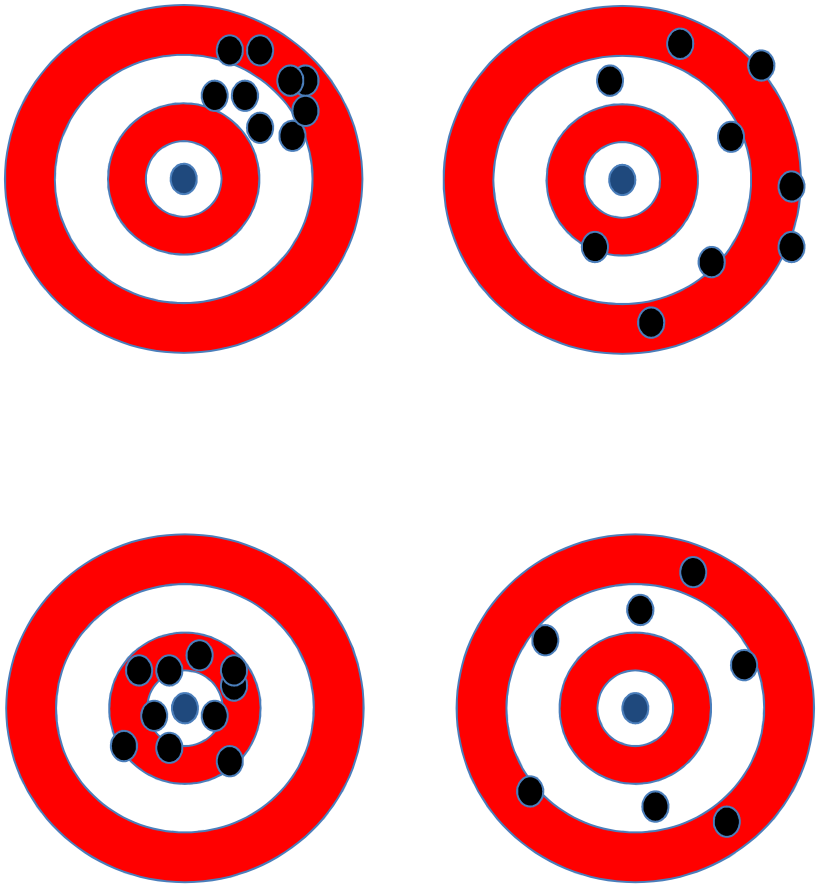
Hiba



Több véletlen mintavételt képzelünk el –becslések

Hiba – 2 dimenziója

„Átlagos eltérés” (torzítás)
Szisztematikus hiba



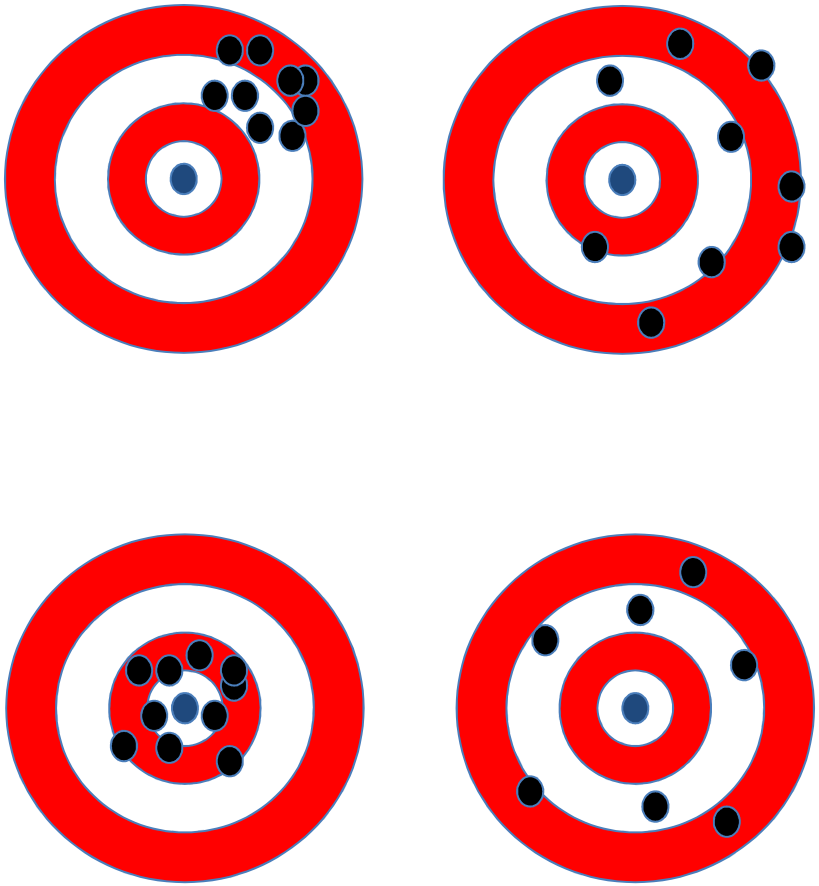
„Változékonyság” (szóródás);
Véletlen hiba (mintavételi hiba)

1. A becslések változékonysága

2. A becslések „közepének” eltérése a valódi értéktől

Hiba – 2 dimenziója

„Átlagos eltérés” (torzítás)
Szisztematikus hiba



„Változékonyság” (szóródás);
Véletlen hiba

Jó becslés, ha:

Torzítatlan:

A becslések „közepe” (várható értéke) a valódi érték

Hatásos:

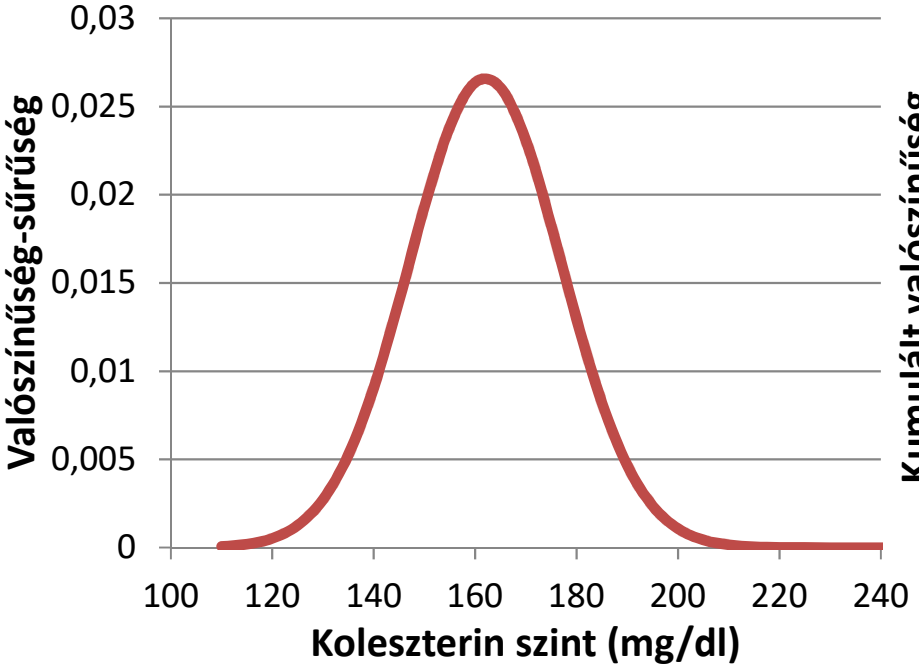
A becslések változékonysága a lehető legkisebb

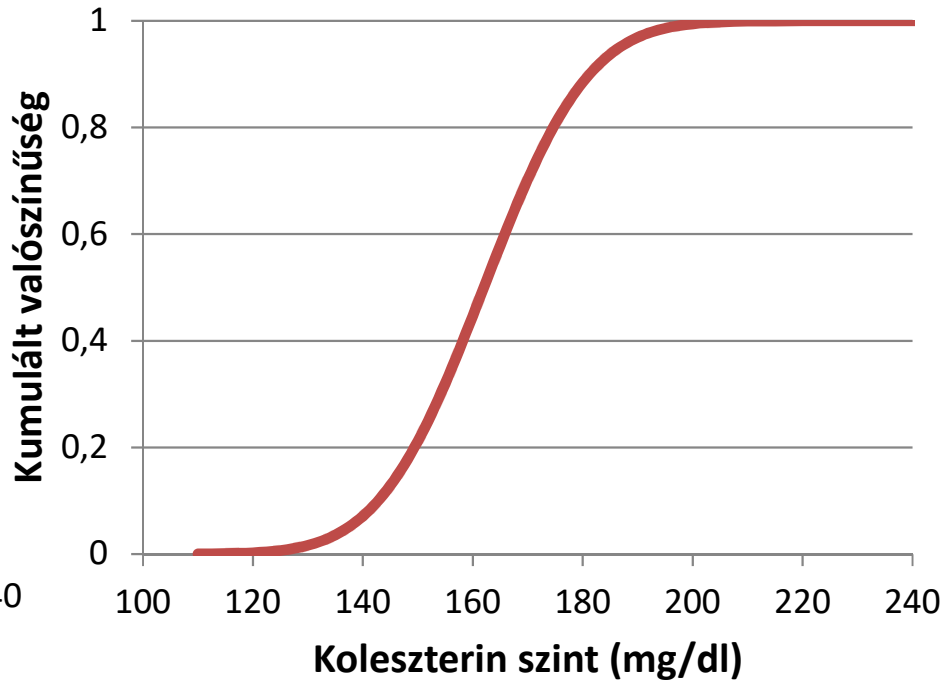
Konzisztens:

az elemszám növelésével csökken a becslés változékonysága

()

Normál (Gauss) eloszlás I.

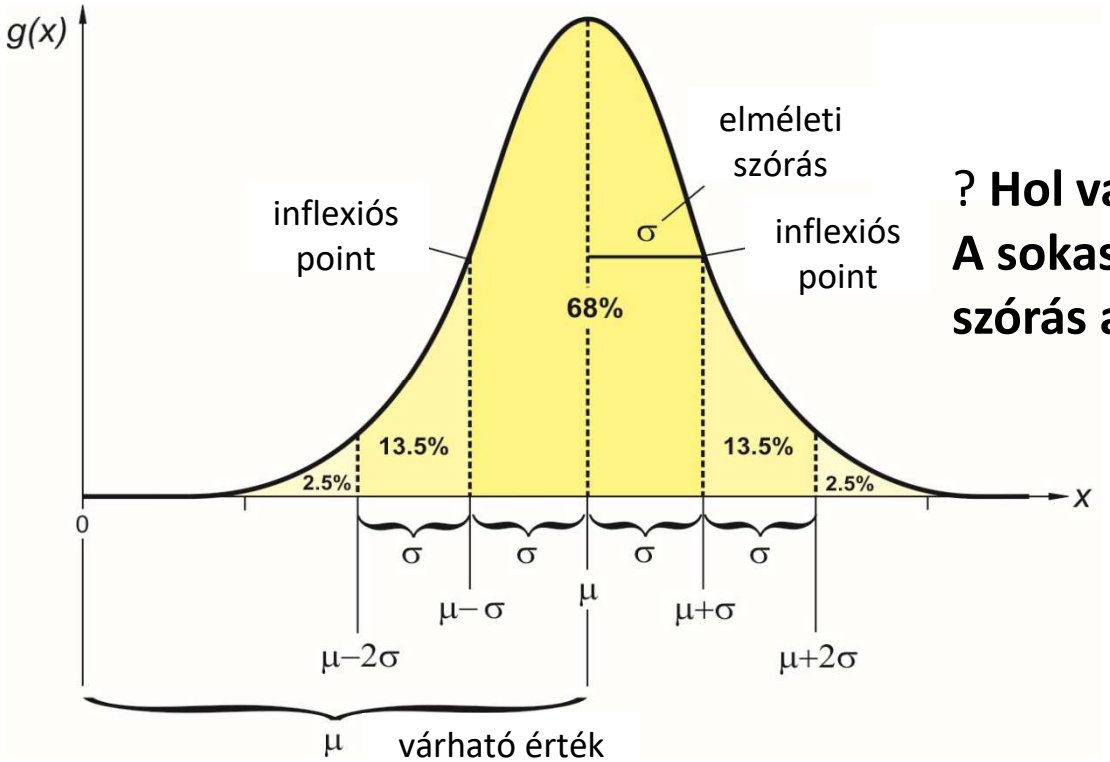




Koleszterinszin, vércukorszint, állkapocsízületi szög....

Centrális határeloszlás tétele változókra: ha sok független valószínűségi változót összegzünk, akkor elég általános feltételek teljesülése esetén az összeg normális eloszlású valószínűségi változó lesz.

Becslési (predikciós) tartományok



? Hol van a sokaságban az adatok X %-a?
A sokaságot ismerve: várható érték a szórás alapján megadható!

$$pl : 95\ \% : \mu \pm \sim 2 * \sigma$$

A minta alapján?

Sokaság	**<----**	**minta**
várható érték (μ)	<----	átlag ($\bar{x}$)
elméleti szórás (σ)	<----	tapasztalati szórás (s)
normál eloszlás	<----	t-eloszlás (pontosabban...)

$$pl : 95\ \% : \bar{x} \pm \sim 2 * s$$

$$95\ \% : \bar{x} \pm t * s$$

Becslési (predikciós) tartományok

Nevezéktan: „saját név” csak a 95%-os becslési tartománynak

95% -os becslési tartomány

= referencia tartomány = normál tartomány

$$pl\ :95\ \%\ :\overline{x}\pm\ \sim\ 2\ *\ s$$

Átlag becslése

Mintából az átlag alapján a várható érték becslése is bizonytalan! Hogyan becsülhető ez a bizonytalanság?

Centrális határeloszlás tétele (mintavételi átlagokra): ha egy adathalmazból *n* elemű mintákat veszünk, akkor elég általános feltételek teljesülése esetén a minták átlagai normál eloszlásúak lesznek, és az eloszlás varianciája az eredeti eloszlás varianciájának *n*-ed része lesz.

Konfidencia tartományok

Tehát ha mintákat veszünk, akkor ha a minták átlagait nézzük, az is normál eloszlást követ!

Ebben az esetben az átlagok eloszlásának szórása a standard hiba (más néven az átlag szórása)!

Az átlagokra vonatkozó predikciós intervallumot hívjuk konfidencia intervallumnak.

?Mekkora (hol van) a sokaság átlaga egy X%-os biztonsággal?

Az átlag 95% szintű konfidencia intervallumának határai tehát hasonlóan becsülhetők:

$$pl : 95\%CI : \bar{x} \pm \sim 2 * SEM$$

Más becslésre is megadható adott szintű konfidencia intervallum!
Mutatja a becslés értékét és annak hibáját egyszerre.

Hipotézisvizsgálatok

Hipotézisvizsgálat - menete

Mi a kérdés: A hatos dobás valószínűsége eltér 1/6-tól, méghozzá nagyobb?

Nullhipotézis: H_0: A 6-os dobás valószínűsége 1/6.

Szignifikancia szint:10%

A minta: 24 dobásból 6 darab 6-os.

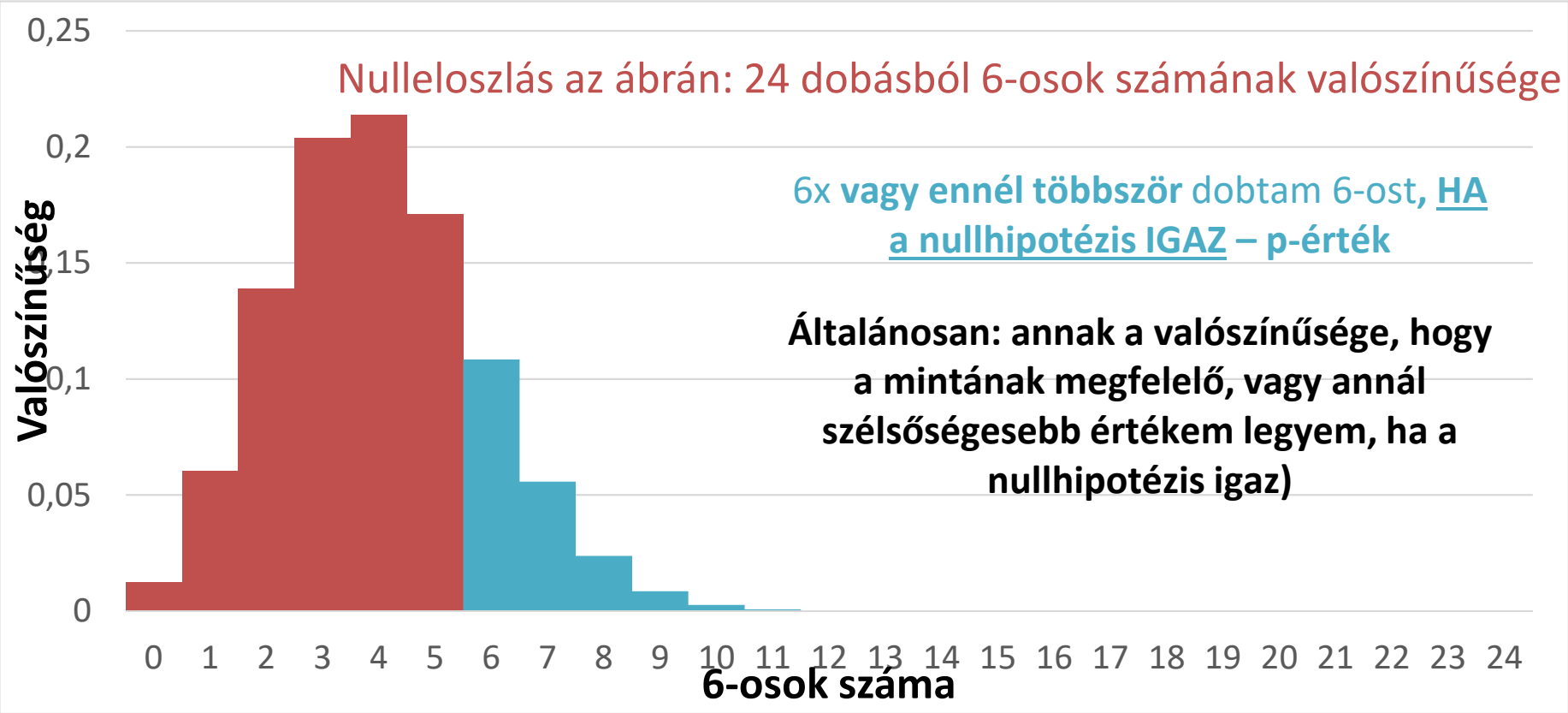
Lényeges a különbség egyáltalán – releváns: minta alapján: 1/4 ez **1,5**-szeres az 1/6-nak

Mennyi a bizonyíték – p-érték: 0,1995

Döntés: nincs elég bizonyíték az elvetésre – megtartjuk a H_0-t

		A populációban (a valóságban) a null hipotézis:	
		Igaz	Hamis
A döntés: a null hipotézist:	Megtartom (Nem vetem el)	Helyes döntés	Hiba (**másod fajú**) (β) (álnegatív eredmény)
	Elvetem	Hiba (**első fajú**) (α) (álpozitív eredmény)	Helyes döntés (**erő**) (1-β)

Hipotézisvizsgálat – 0 eloszlás



Egymintás Student-féle t-próba

Mire vagyok kíváncsi

A minta várható értéke megegyezik-e egy ismert populációátlaggal

Változó típusa

1 számszerű és folytonos

Feltételek

Egymástól független megfigyelések

átlagok eloszlása legyen normál azaz:

normál eloszlású változó vagy nagy minta (CLT miatt)

Megjegyzés: Számolás:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

Párosított t-próba

Mire vagyok kíváncsi

Két várható érték megegyezik-e - párosított csoportokban

Változó típusa

1 számszerű és folytonos, valamint 1 bináris („csoportok”)

Feltételek

Egymástól független megfigyelések a csoportokban, párosított csoportok

átlagok különbségének eloszlása legyen normál azaz:

normál eloszlású különbség vagy nagy minta

Megjegyzés:

párosított próbák ereje általában nagyobb, mint nem párosítotté

Student-féle 2 mintás t-próba

Mire vagyok kíváncsi

Két várható érték megegyezik-e

Változó típusa

1 számszerű és folytonos, valamint 1 bináris („csoportok”)

Feltételek

Csoportonként és egymástól is független megfigyelések

átlagok eloszlása legyen normál mindkét csoportban azaz:

normál eloszlás a csoportokban vagy nagy minta

szórások azonosak legyenek a csoportokban

Megjegyzés:

szórásokat általában nem ismerjük, így inkább Welch próbát használjunk!

Welch-féle t-próba

Mire vagyok kíváncsi

Két várható megegyezik-e

Változó típusa

1 számszerű és folytonos, valamint 1 bináris („csoportok”)

Feltételek

Csoportonként és egymástól is független megfigyelések

átlagok eloszlása legyen normál mindkét csoportban azaz:

normál eloszlás a csoportokban vagy nagy minta

Megjegyzés:

nem érzékeny az eltérő varianciákra (robosztus a varianciák eltérésére)

Multiplicitás – vizsgára NEM kell tudni!

…Chocolate Helps Weight Loss.
(…A csoki segít a lefogyásban.)

„A Bild címlapján (Európa egyik lenagyobb napilapja) a következő jelent meg: német kutatók egy csoportja azt találta, hogy az alacsony szénhidráttartalmú diétát tartók 10%-kal gyorsabban veszítenek testsúlyukból, ha mellé csokoládét is fogyasztanak.”

„*a csokoládéfogyasztás*
statisztikailag szignifikánsan fogyasztó hatása tényleg kimutatható a *a mért adatok alapján*”
Nade hogyan??

…Chocolate Helps Weight Loss.

„… véletlenszerűen osztották be az alanyokat három étrendcsoport egyikébe. Az egyik csoport alacsony szénhidráttartalmú diétát követett. Egy másik ugyanazt az alacsony szénhidráttartalmú étrendet követte, plusz napi 1,5 egység sötét csokoládét is evett. A többieket – a kontrollcsoportot – arra utasították, hogy ne változtassanak jelenlegi étrendjükben.”

„A tanulmányban 18 különböző mutatót vizsgáltak 15 emberen: testtömeget, koleszterinszintet, nátriumszintet, plazmafehérjék szintjét, stb.”

…Chocolate Helps Weight Loss.

Az általában használt szignifikancia-szint: 5%, tehát a 5% a valószínűsége, hogy a mintánk (illetve a p) a véletlen mintavétel miatt eltérjen az elfogadható H0-tól, ha H0 igaz. H0 igazsága esetén:

Hibázás valószínűsége 1 próba esetében:	p
Nincs hiba 1 próba esetében:	$1-p$
Nincs hiba k egymástól független próbából:	$(1-p)^k$
Van hiba k egymástól független próbából:	$1 - (1-p)^k$

Ha k = 18, akkor **60%** a valószínűsége legalább 1 hibának, azaz hogy legalább egy próba eredménye szignifikáns, pedig H0 mindegyikre igaz!

Ezt a többszörös összehasonlítási hibát hívják multiplicitási problémának

I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How.

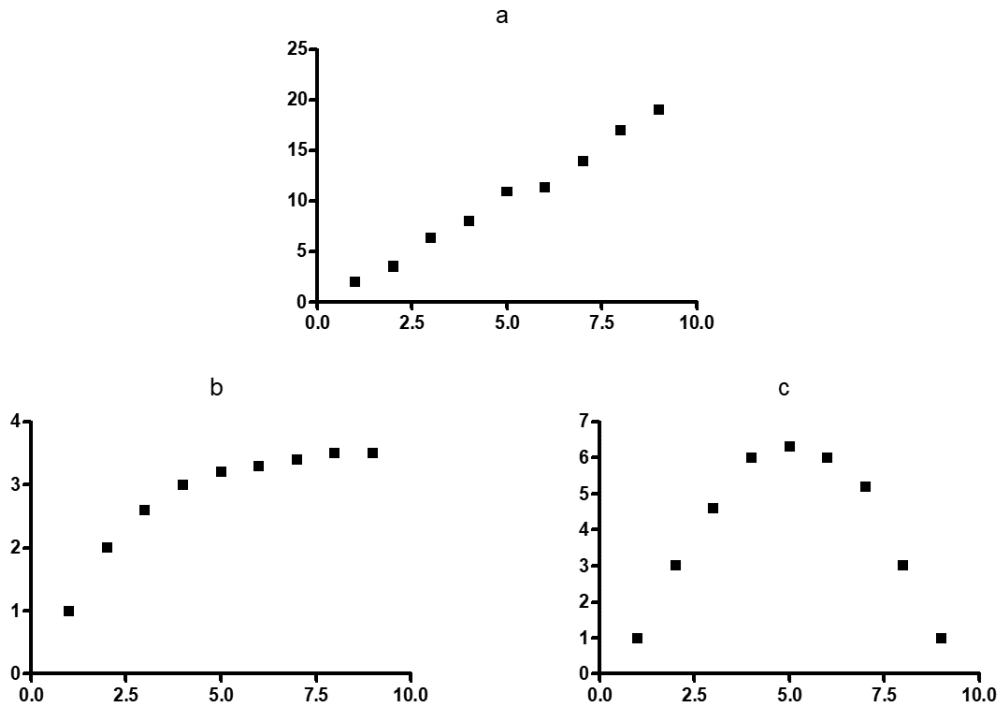
- Multiplicity
  - ☺*eg: Chocolate Helps Weight Loss*
  - *https://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800*

Korreláció, regresszió

Változók közötti reláció (kapcsolat)

A kapcsolat típusa:

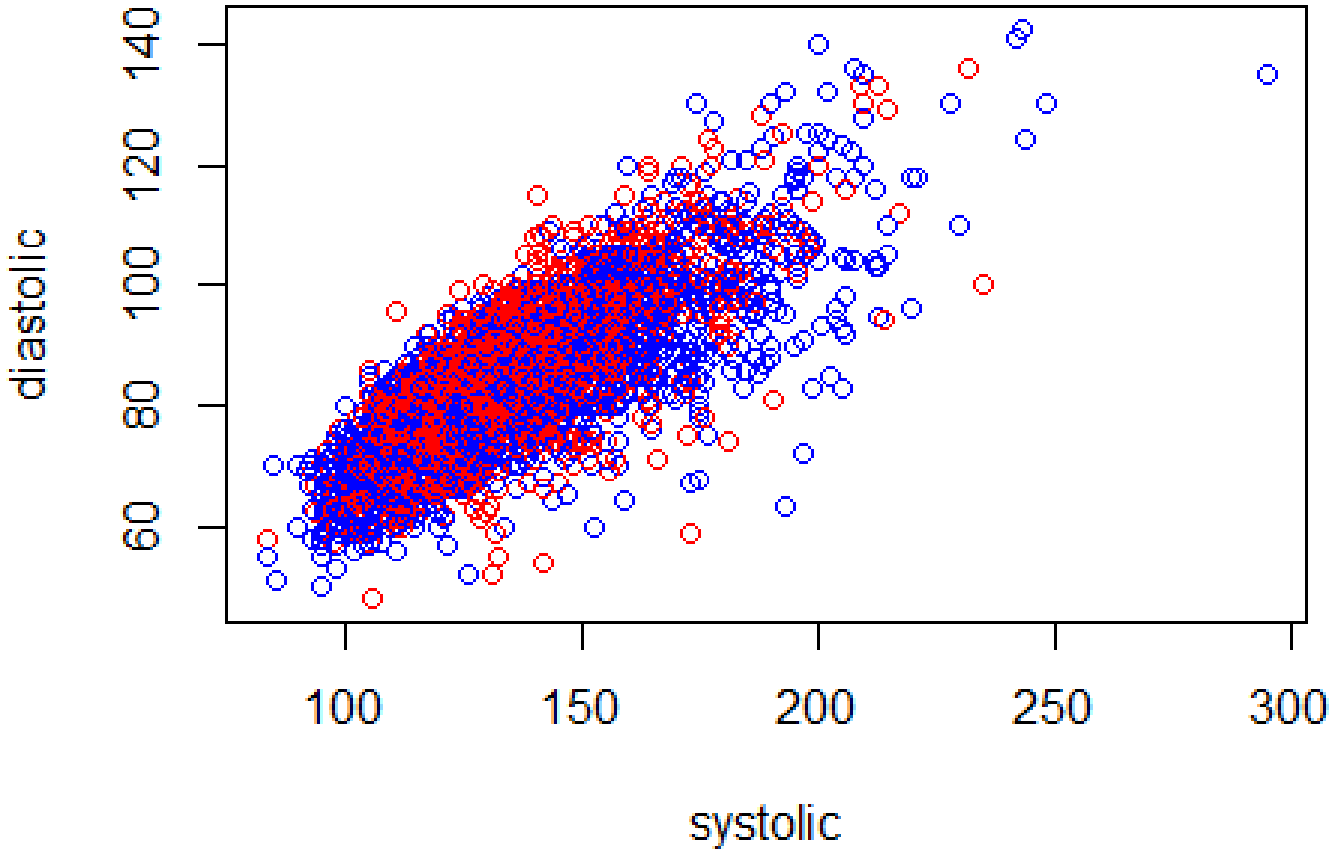
- monoton
  - pozitív
  - negatív
  - Pozitív lineáris
  - …
- nem monoton
  - parabolikus
  - …
- Nincs kapcsolat



Korreláció

Monoton,

szimmetrikus (nem megmondható, hogy melyik függ melyiktől) kapcsolat **2 véletlen** (véletlen hibájú, „nem beállított”) változó között.



Korreláció

Korreláció **erősségének** kifejezésére:

Korrelációs együtthatók (r):

ha **lineáris korrelációt** feltételezünk: **Pearson** r

ha **monoton** (nem feltétlenül lineáris): **Spearman rang** r,

A korrelációs együtthatók **értéke**:

-1 től +1

negatív: negatív korreláció

pozitív: pozitív korreláció

minél közelebb esik|1|-hez annál erősebb a korreláció

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

„Középtől vett távolság” – mint y és mind x irányban

„Korrelációs” t-próba (Pearson r -re)

Mire vagyok kíváncsi

2 változó (lineárisan) korrelált-e (r értéke különbözik 0-tól)

Változó típusa

2 számszerű változó (X és Y)

Feltételek

Egymástól független megfigyelések (x és y párok)

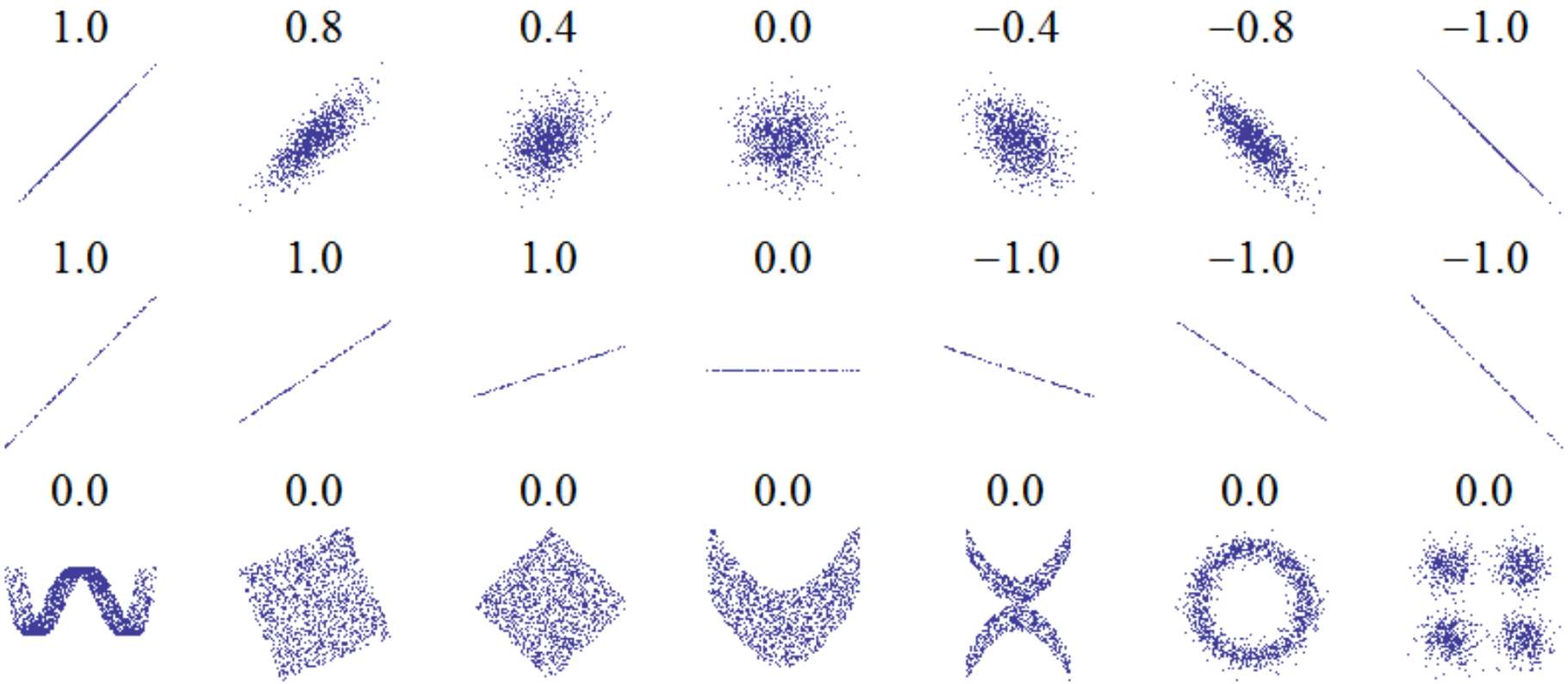
Szimmetrikus, lineáris kapcsolatot feltételezünk

x és y véletlen változók

Megjegyzés:

Megjegyzések

- *Ábrázoljunk mindig!!!!;*
- *Korreláció nem jelent okozati összefüggést*
  ☺ *eg: http://www.fastcodesign.com/3030529/infographic-of-the-day/hilarious-graphs-prove-that-correlation-isnt-causation*



Regresszió

Függvény kapcsolat (NEM szimmetrikus) egy függő (cél, eredmény, Y) változó és egy független (magyarázó, prediktáló, X) változó**(k)** között. [Y véletlen változó, X nem feltétlenül]

Y függ X-től – a függőségi viszony iránya **klinikailag** feltételezett, statisztikailag nem vizsgálható.

Kapcsolódó kérdések:

- Van (adott típusú) függvényszerű kapcsolat? (statisztikai kapcsolat, nem ok-okozati)
- Mekkora Y értéke, ha X:…? (becslés)
- Mekkora X értéke, ha Y:…? (becslés)
- Milyen függvény írja le legjobban Y X-től való függését?

Lineáris regresszió

Lineáris függvénykapcsolat feltételezett.

2 változó esetében a regresszió – korreláció kérdései, számolásai legtöbbször egymással „ekvivalensé” tehetőek.

Lineáris regresszió

Az egyenes becslésére: OLS (Ordinary Least Square method – azaz legkisebb négyzetek módszere)

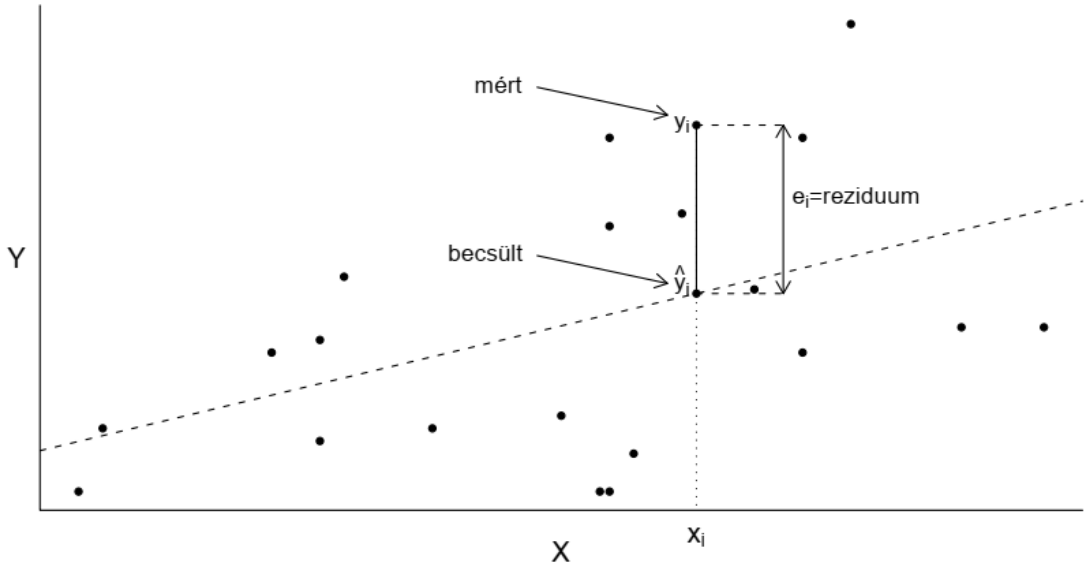
tengelymetszet

Meredekség

Lineáris függvény: $Y = \beta_0 + \beta_1 * X + \epsilon$

Hibatag; reziduum: pont-egyenes **függőleges** távolsága (a becsült és mért értékek különbsége)

Az OLS szerinti legjobb egyenes az, ahol a **legkisebb** a pont-egyenes függőleges távolságok **négyzet**összege.



„Korrelációs” t-próba (meredekségre)

Mire vagyok kíváncsi

Y változó X-tól lineárisan függ-e.

Változó típusa

2 számszerű változó (X és Y)

Feltételek

Egymástól független megfigyelések (x és y párok)

Lineáris kapcsolatot feltételezünk

x értékei „hiba nélkül” mérhetőek

a reziduumok eloszlása minden x-re normál

és varianciájuk minden x-re megegyezik

Megjegyzés:

A próbában a **meredekség** nem 0-t teszteljük.

Meredekség és R^2

Meredekség

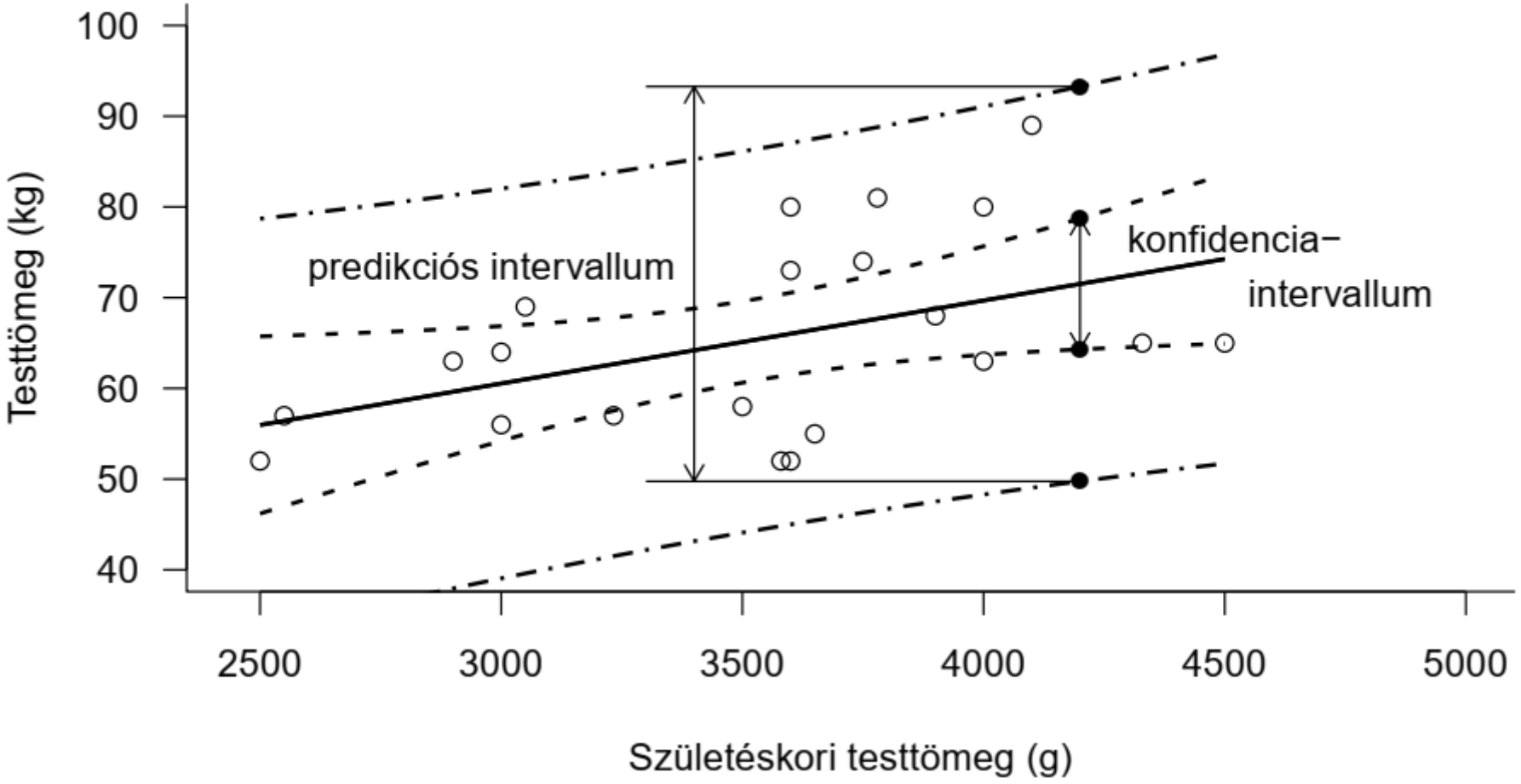
Az Y-ban bekövetkező átlagos változás X egységnyi változtatására

R^2 – meghatározottsági (determinációs) együttható

az r négyzete

Az Y változó varianciájának (változatosságának) hány százaléka magyarázható az X varianciájával (változásával)

Konfidencia és predikciós intervallumok



OMHV
(Oktatók Hallgatói Véleményezése)