Kiváltott agyi jelek informatikai feldolgozása 2018

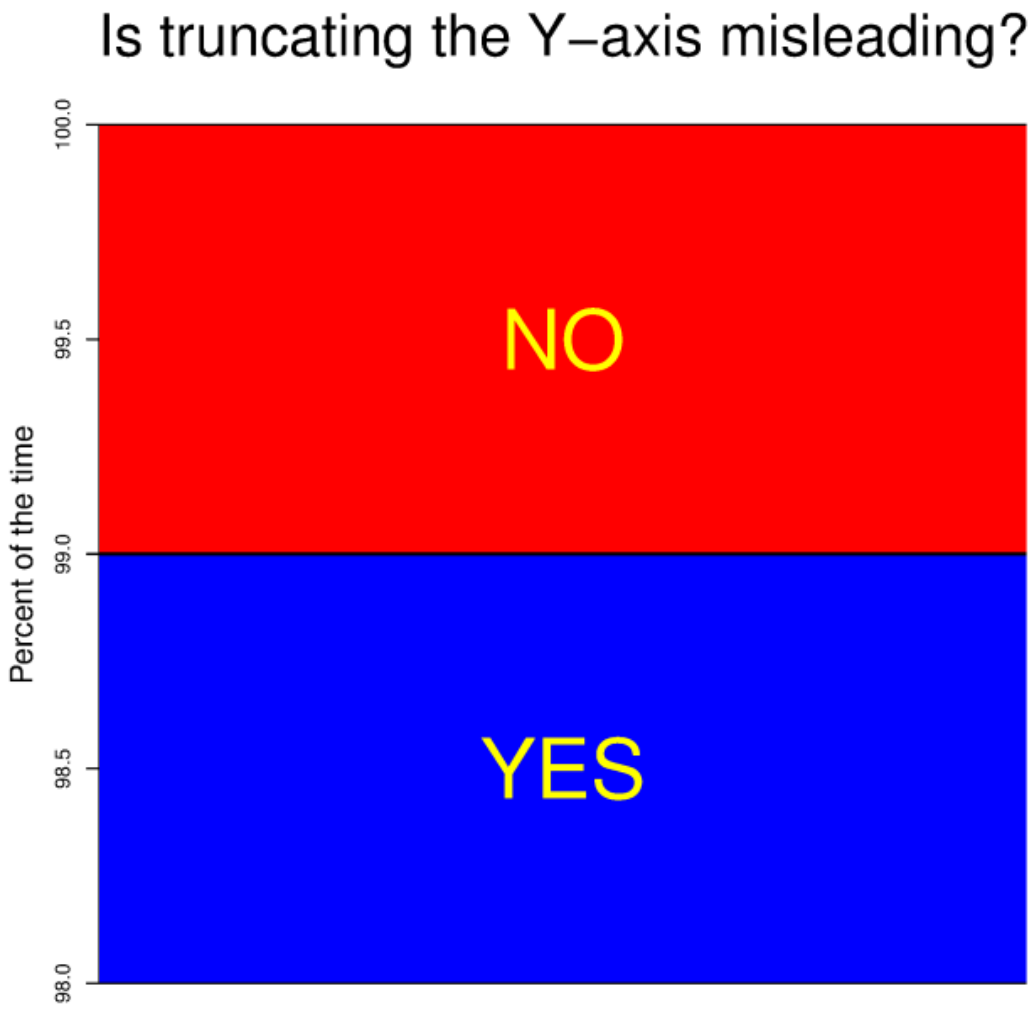
Statisztika

Kiss Gábor
IB.157.
kiss.gabor@tmit.bme.hu

Bevezető

- Miért tanuljuk ezt?
- „A **statisztika** a **valóság számszerű információinak megfigyelésére**, **összegzésére**, **elemzésére** és **modellezésére** irányuló **gyakorlati tevékenység** és **tudomány**.”
- Főbb részterületei:
  - mintavétel (reprezentatív)
  - becsléselmélet
  - hipotézisvizsgálat
  - idősorelemzés
  - korreláció- és regressziószámítás

Statisztikai hibák



Statisztikai hibák

- Amerikában az egyik elnökválasztás előtt (~60-as évek) az akkori statisztikai hivatal telefonos felmérést készített, hogy az egyes jelölteket az emberek hány százaléka támogatja. A felmérés eredménye alapján megállapították, hogy „A” jelöltre 60% fog szavazni „B” jelöltre 40%. A választásokat végül a „B” jelölt nyerte 60-40 arányban. ***Mi lehetett a hiba?***

Statisztikai hibák

- „A” országban a felmérések bizonyítják, hogy nyáron dupla annyi gutaütés éri az embereket mint télen. Ezért megvizsgálják, hogy ennek vajon mi lehet az oka. Észreveszik, hogy nyáron az aszfalt hőmérséklete 20 fokkal magasabb mint télen, így arra a következtetésre jutnak, hogy nyáron hűteni kell az aszfaltot, hogy megelőzzék a gutaütések magas számát. ***Mi lehetett a hiba?***

Statisztikai hibák

- Tipikus hibák:
  - Nem tetsző adatok kihagyása
  - Befolyásoló kérdés
  - Túláltalánosítás
  - Torzított mintavétel
  - A becsült hiba félreértelmezése vagy félreértése
- Emiatt szokták mondani: *„A statisztika az a tudományág, ahol bármit és annak ellenkezőjét is be lehet bizonyítani.”*
- **Ez** természetesen **nem igaz**, megfelelő **matematikai és statisztikai tudással**, **pontos eredmények** állíthatók elő, amikből **helyes következtetéseket** lehet levonni, illetve a **valóságot jól szimuláló modelleket** lehet létrehozni.

Valószínűség számítás alapjai

- **Eseménytér**: Egy véletlen esemény kimeneteli lehetőségeinek a halmaza: S = {s1, s2, ..., sn}, ahol S a biztos esemény, Ø a lehetetlen esemény. Ha a véletlen esemény kimeneteli lehetőségeinek a száma „n” véges vagy megszámlálhatóan végtelen, akkor diszkrét eseménytérről beszélünk, ellenkező esetben folytonosról.
- **Esemény**: az S eseménytér részhalmaza.

Valószínűség számítás alapjai

- **Valószínűségi mérték (Pr)**: Az események egy valós függvénye, amelynek az értékkészlete a [0,1] tartományba esik, és amely teljesíti az alábbi feltételeket:
  - Pr(S) = 1
  - Pr(A U B) = Pr(A) + Pr(B), ha A ∩ B = Ø
- **Elemi esemény**: si elemi esemény, amire igaz, hogy: pi = Pr(si), 1,2… n és $(\sum_{i=1}^{n} pi)$=1

Valószínűség számítás alapjai

- **Valószínűségi változó**: az eseménytér leképzése egy halmazra, azaz az eseménytér egy függvénye:
  - diszkrét valószínűségi változó
  - folytonos valószínűségi változó

Valószínűség számítás alapjai példa

- Vegyünk egy szabályos pénzérmét, ami véletlenszerűen feldobunk.
- Eseménytér eseményei:
  - s1: fej
  - s2: írás
  - S:= {s1, s2}
- Pr(s1)=1/2 és Pr(s2)=1/2
- Legyen a következő valószínűségi változó F(si):
  - F(s1) = 0, F(s2) = 1

Valószínűség számítás alapjai példa

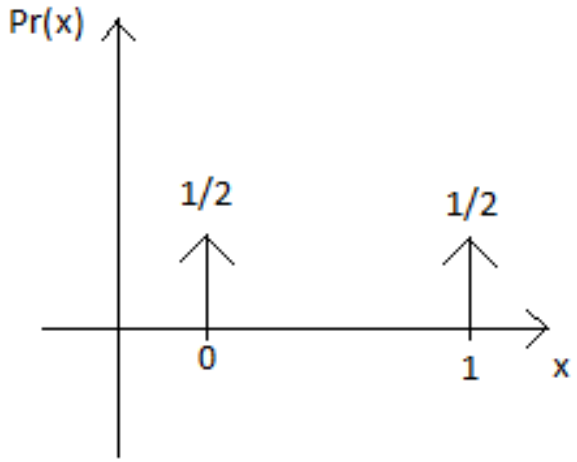
- Vegyünk egy szabályos dobókockát, amit véletlenszerűen feldobunk.
- Eseménytér eseményei:
  - s1: első oldal, s2: második oldal, s3: harmadik oldal, s4: negyedik oldal, s5: ötödik oldal, s6: hatodik oldal
  - S:= {s1, s2, s3, s4, s5, s6}
- Pr(s1)=Pr(s2)=Pr(s3)=Pr(s4)=Pr(s5)=Pr(s6)=1/6
- Legyen a következő valószínűségi változó F(si):
  - F(s1) = 1, F(s2) = 2, F(s3) = 3, F(s4) = 4, F(s5) = 5, F(s6) = 6

Valószínűség számítás alapjai példa

- Vegyünk egy szabályos véletlen szám generátort, ami 0 és 1 között add vissza egy valós számot
- Eseménytér eseményei (megszámlálhatatlanul végtelen sok):
  - x valós számok halmaza [0-1] intervallumban
  - S:= x>0 és x<1
- Pr(x)=0
- Legyen a következő valószínűségi változó F(x):
  - F(x) = x

Eloszlás

- Diszkrét valószínűségi változó esetén egy jó leírása a valószínűségi változónak az eloszlása pl: pénz feldobás

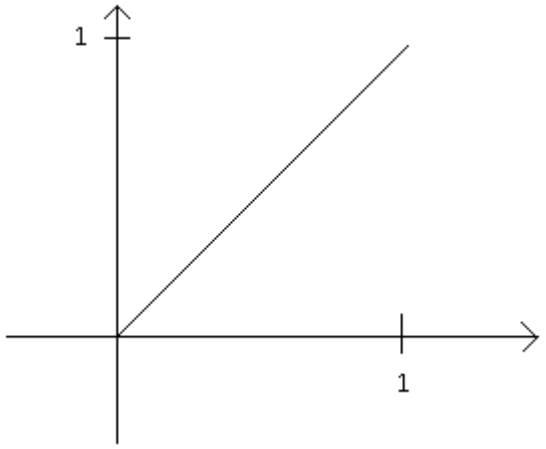


Eloszlásfüggvény

- Folytonos valószínűségi változónak jó leírása az eloszlás függvénye:
  - Legyen adott X valószínűségi változó, képezzük ebből az az F(X) függvényt, hogy F(x)=$\int_{-\infty}^{x} \Pr(x)$, az F(X) függvényt nevezzük az X valószínűségi változó eloszlásfüggvényének

Eloszlásfüggvény

- Példa: Vegyünk egy szabályos véletlen szám generátort, ami 0 és 1 között add vissza egy valós számot. Adjuk meg az eloszlásfüggvényét!
- F(X):

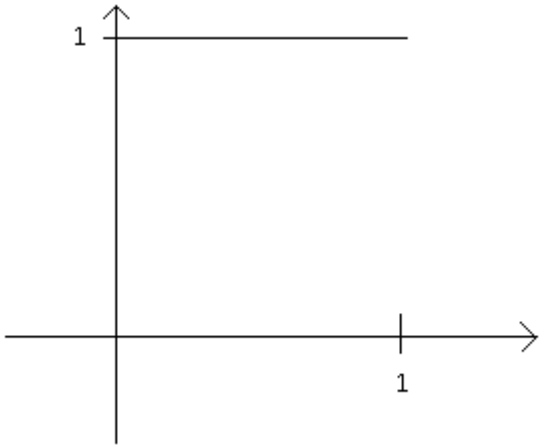


Sűrűségfüggvény

- Legyen adott F(X) eloszlásfüggvény, vegyük a függvény x szerinti deriváltját, így kapjuk meg az f(X) sűrűségfüggvényt.
- Az eloszlásfüggvény definíciójából adódik, hogy a sűrűség függvény alatti terület az 1.
- Megjegyezés: nem minden eloszlásfüggvénynek létezik sűrűségfüggvénye.

Sűrűségfüggvény

- Példa: Vegyünk egy szabályos véletlen szám generátort, ami 0 és 1 között add vissza egy valós számot. Adjuk mega sűrűségfüggvényét!
- f(X):



Empirikus eloszlásfüggvény

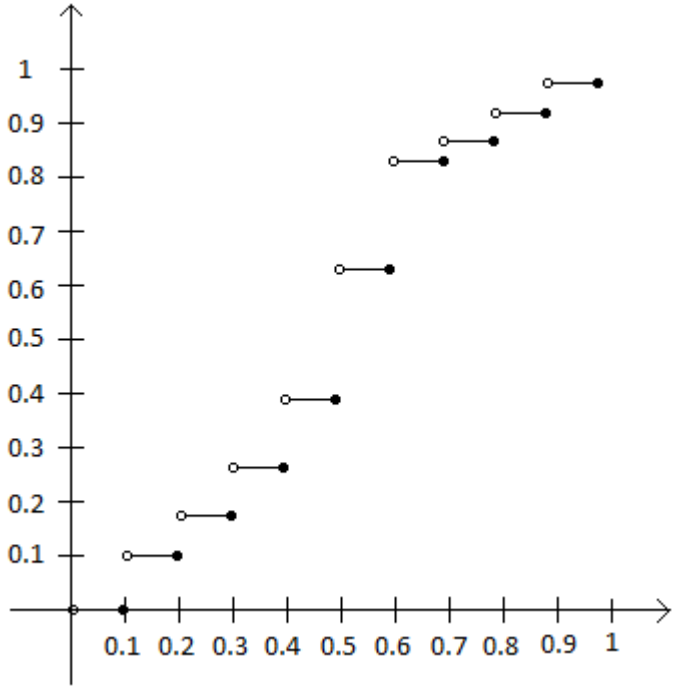
- Adott egy nem szabályos véletlen szám generátor ami 0 és 1 között add egy véletlen számot, de nem egyenletesen, a következő 30 számot generálta: {0.1, 0.5, 0.2, 0.3, 0.5, 0.6, 0.4, 0.5, 0.6, 0.3, 0.2, 0.1, 0.4, 0.5, 0.6, 0.3, 0.1, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.4, 0.5, 0.6, 0.9, 0.5, 0.6, 0.8}, mi lehet véletlen szám generátor eredeti eloszlásfüggvénye?

Empirikus eloszlásfüggvény

- Legyen az empirikus eloszlásfüggvény:
  - Rendezzük a mintát: X*1<=X*2<=…<=X*n
  - Legyen az empirikus eloszlásfüggvény F*(x):=
    - 0, ha x <= X*1
    - k/n, ha X*k<x<=X*k+1, k=1, …, n-1
    - 1, ha x>X*n

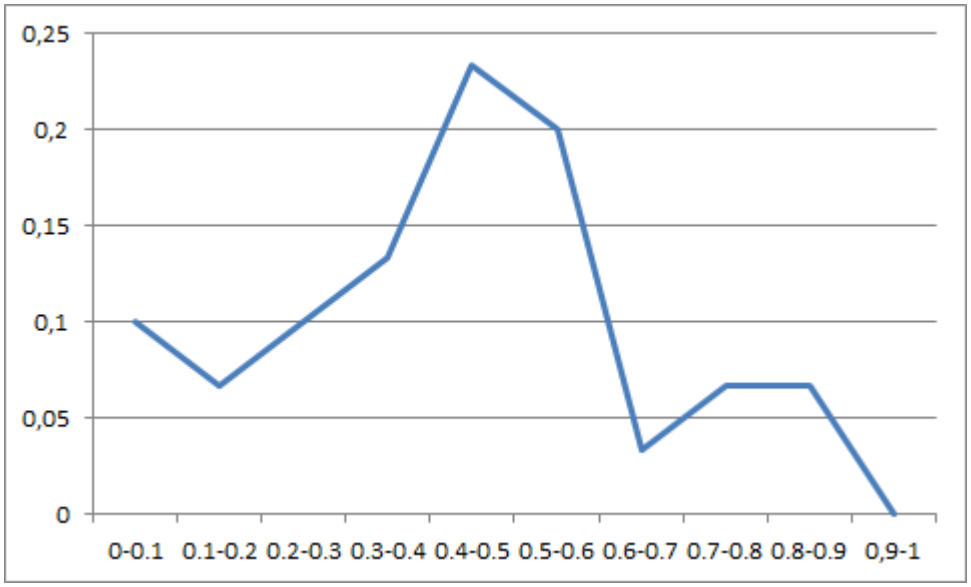
Empirikus eloszlásfüggvény

- 0.1: 3 => 0.1
- 0.2: 2 => 0.17
- 0.3: 3 => 0.27
- 0.4: 4 => 0.4
- 0.5: 7 => 0.63
- 0.6: 6 => 0.83
- 0.7: 1 => 0.87
- 0.8: 2 => 0.93
- 0.9: 2 => 1



Hisztogram (empirikus sűrűségfüggvény)

- 0-0.1: 3 => 0.1
- 0.1-0.2: 2 => 0.067
- 0.2-0.3: 3 => 0.1
- 0.3-0.4: 4 => 0.133
- 0.4-0.5: 7 => 0.233
- 0.5-0.6: 6 => 0.2
- 0.6-0.7: 1 => 0.033
- 0.7-0.8: 2 => 0.067
- 0.8-0.9: 2 => 0.067
- 0.9-1: 0 => 0



Nevezetes eloszlások

- Egyenletes eloszlás (folytonos/diszkrét)
- Binomiális eloszlás (diszkrét)
- Poisson eloszlás (diszkrét)
- Normális eloszlás (Gauss eloszlás) (folytonos)
- Exponenciális eloszlás (folytonos)
- …

Valószínűségi változót jellemző értékek

- Várható érték
- Szórás
- Kvantilisek
- Momentumok
- Ferdeség
- Lapultság
- Medián (középső érték)
- Módusz (leggyakoribb érték)

Kvantilis

- P kvantilis az X val. változó azon értéke amelynél a kisebb mintaelemek hányada p.
  - 0.1 kvantilis = decilis
  - 0.25 kvantilis = első kvartilis
  - 0.5 kvantilis = második kvartilis = medián
  - 0.75 kvantilis = harmadik kvartilis
  - 0.9 kvantilis

Becsléselmélet

- Vegyünk egy szabályos hatoldalú dobókockát. Dobjunk vele tízszer és adjuk össze a dobott értékeket. Mi a legvalószínűbb összeg amit kaphatunk, illetve becsüljük meg a kapott összeg nagyságát!
  - Létezik:
    - Pontbecslés
    - Intervallumbecslés

Várhatóérték

- Vegyük a $\sum_{i=1}^{n} (\Pr(x) * F(x))$összeget (folytonos esetben szumma helyett integrál -$\infty + \infty$ között), ahol a Pr(x) az x esemény valószínűségi mértéke míg az F(x) a valószínűségi változó x helyen vett értéke, ezt hívjuk a valószínűségi változó várhatóértékének, jelölése E(X).
- Véges minta esetén nem tudunk várhatóértéket számítani, ekkor az átlag egy jó becslése a várhatóértéknek jelölése: $\bar{x}$ vagy m(x)-el.

Becsléselmélet - Pontbecslés

- Vegyünk egy szabályos hatoldalú dobókockát. Dobjunk vele tízszer és adjuk össze a dobott értékeket. Mi a legvalószínűbb összeg amit kaphatunk?
- Számítsuk ki 1 kocka várhatóértékét:
  - E(kocka)=1/6*1+1/6*2+1/6*3+1/6*4+1/6*5+1/6*6=

    1/6*(6*7/2)=3,5 (természetesen 3,5-t nem lehet dobni a kockával)
  - Ha tízszer dobunk a kockával akkor igaz, hogy a legvalószínűbb összeg amit kaphatunk az 10*3,5=35.

Becsléselmélet - Pontbecslés

- Egy véletlen szám generátor az alábbi számokat generálta nekünk: {0.1, 0.4, 0.2, 0.7, 0.5, 0.6, 0.2, 0.3, 0.5, 0.7}, ha tudjuk hogy a véletlen szám generátor nullánál nagyobb számokat generál egyenletes eloszlás szerint, akkor vajon mennyi lehet a maximuma?
- Számítsuk ki a generált számok átlagát! Szorozzuk meg kettővel és ezt az értéket használjuk, mint becsült érték:
  - m(X) = 0,42 => feltehetőleg 0,84 ig generál számokat
  - Lenne jobb becslés is…

Játék a várhatóértékkel

- Következő játékot játszhatunk a bankkal:
  - 1 játék 1024 forintba kerül
  - Addig dobhatunk fel egy szabályos pénzérmét amíg írást nem dobunk
  - A bank összeadja a dobott fejek számát, és 2^x forintot fizet nekünk, ahol az x a dobott fejek száma
- Megéri részt venni a játékban? Ha állítható, hogy egy játék mennyibe kerül, hány forinttól éri meg az nekünk?

Játék a várhatóértékkel

- Számítsuk ki a várható nyereményünket egy játék esetén:
  - $E(X) = 1/2*2^0+1/4*2^1+1/8*2^2+\ldots+1/2^n*2^{n-1}+\ldots$

  $E(X) = 1/2+1/2+1/2+\ldots+1/2+\ldots = \infty$
- Számítsuk ki a várható profitját egy játéknak:
  - $E(X) = \infty - 1024 = \infty$
- Tehát bármekkora összegért megéri részt venni a játékban.

Szórás

- Legyen a $\sqrt{E((X-E(X))^{2)}}$=D(X) az X valószínűségi változó szórása.
- Egyszerűbben (véges mintára):
  - D(X)=$\sqrt{\frac{\sum_{i=1}^{n}(E(X)-xi)^2}{n}}$, vagyis $\sqrt{\frac{\sum_{i=1}^{n}(\bar{x}-xi)^2}{n}}$
- Az angol irodalomban a szórás alatt általában a szórás négyzetet értik.
- A szórás jól leírja, hogy egy adott eloszlás a várhatóérték körül, mekkora ingadozást vesz fel, feltéve ha ismerjük az eloszlást.
- Véges minta szórását s(x)-el szokás jelölni.
- Hipotézisvizsgálat során sokszor kell egy minta korrigált szórása, ami annyiban különbözik a szórás eredeti képletétől, hogy n helyett (n-1)-el kell osztani. (valójában ez a helyes becslése az eredeti valószínűségi változó szórásának)

Szórás

- Legyen adott egy hatoldalú szabályos dobókocka. Számítsuk ki a kocka, mint valószínűségi változó szórását!
  - E(X) = 3,5
  - $\sqrt{E((X-E(X))^{2)}}=\sqrt{E((X-3{,}5))^{2)}}=$
  - $=\sqrt{\frac{(3{,}5-1)^2+(3{,}5-2)^2+(3{,}5-3)^2+(3{,}5-4)^2+(3{,}5-5)^2+(3{,}5-6)^2}{6}}=$
  - $=\sqrt{\frac{17{,}5}{6}} \approx \sqrt{2{,}9}$

A játék szórása

- Vajon mekkora az előzőekben felvázolt játék nyereményének, mint valószínűségi változónak a szórása?

Becsléselmélet – Intervallum becslés

- Tudjuk azt, hogyha tízszer dobunk egy szabályos hatoldalú kockával, akkor a dobások összegének a várhatóértéke az 35. De vajon mekkorát tévedünk általában.
- Ehhez a négyzetes átlagos tévesztésnek a gyökét szokás megadni (Root Mean Square Error, RMSE), ami megegyezik a szórással.
- Mekkora lesz a szórás?

Becsléselmélet – Intervallum becslés

- Azt tudjuk, hogy egy kockának a szórás négyzete az körülbelül 2.9. Ekkor a tíz kocka összegének a szórás négyzete az 10*2.9=29 lesz körülbelül.
- Vagyis a D(X)=$\sqrt{29} \approx 5{,}4$
- Tehát mondhatjuk azt, hogy ha a dobások összegét 35-tel becsüljük, akkor az RMSE $\approx 5{,}4$
- De az is igaz, hogy a dobások összege az nagy valószínűséggel $35 \pm 5{,}4$ intervallumba esik.
- Általánosságban egy valószínűségi változót az: E(X) $\pm$ D(X) intervallummal szokás becsülni.

Hipotézisvizsgálat - Alapjai

- Van három pénzérménk, amit feldobunk 20-szor. És leírjuk a dobások eredményét egy papírra (fej:=0, írás:=1). Kérdés szabályosak-e a pénzérmék?
  - 11001100011110010011
  - 10101010101010101010
  - 00000100100100010010
- Állítsunk fel két hipotézist:
  - H0: az adott pénzérme szabályos (null hipotézis)
  - H1: az adott pénzérme nem szabályos (alternatív hipotézis)
- Készítsünk egy próba statisztikát (függvényt), aminek a minta realizációk a bemenetei, és a kimenete pedig 0 ha a H0-t tartja igaznak, 1 ha a H1-et.

Hipotézisvizsgálat - Alapjai

	H0-t tartjuk igaznak	H1-t tartjuk igaznak
Eredetileg a H0 igaz	Jó	Elsőfajú hiba
Eredetileg a H1 igaz	Másodfajú hiba	Jó

- Elsőfajú hiba:
  - Elsőfajú hibát követünk el akkor, ha a null hipotézist elvetjük, bár az volt igaz.
  - Az elsőfajú hiba nagysága mindig felülről becsülhető.
- Másodfajú hiba:
  - Másodfajú hibát követünk el, ha elfogadjuk a null hipotézist, de az alternatív hipotézis volt az igaz.
  - Másodfajú hiba nagysága általában nem becsülhető felülről
- Emiatt a gyakorlatban a „pesszimista” null hipotéziseket szoktak felállítani. Illetve csak szigorú feltételek teljesülése esetén fogadják el.

Hipotézisvizsgálat - Alapjai

- Készítsünk rögtön két ilyen függvényt is:
  - f1: számoljuk össze az 1-esek számát, ha az 10 ± 4 értéket vesz fel akkor elfogadjuk a H0-t egyébként a H1-t fogadjuk el
  - f2: nézzük meg, hogy mennyi a leghosszabb 0 vagy 1 sorozat. Ha ez nagyobb vagy egyenlő mint 4 akkor a H0-t fogadjuk el, ha kisebb akkor a H1-t.

Hipotézisvizsgálat - Alapjai

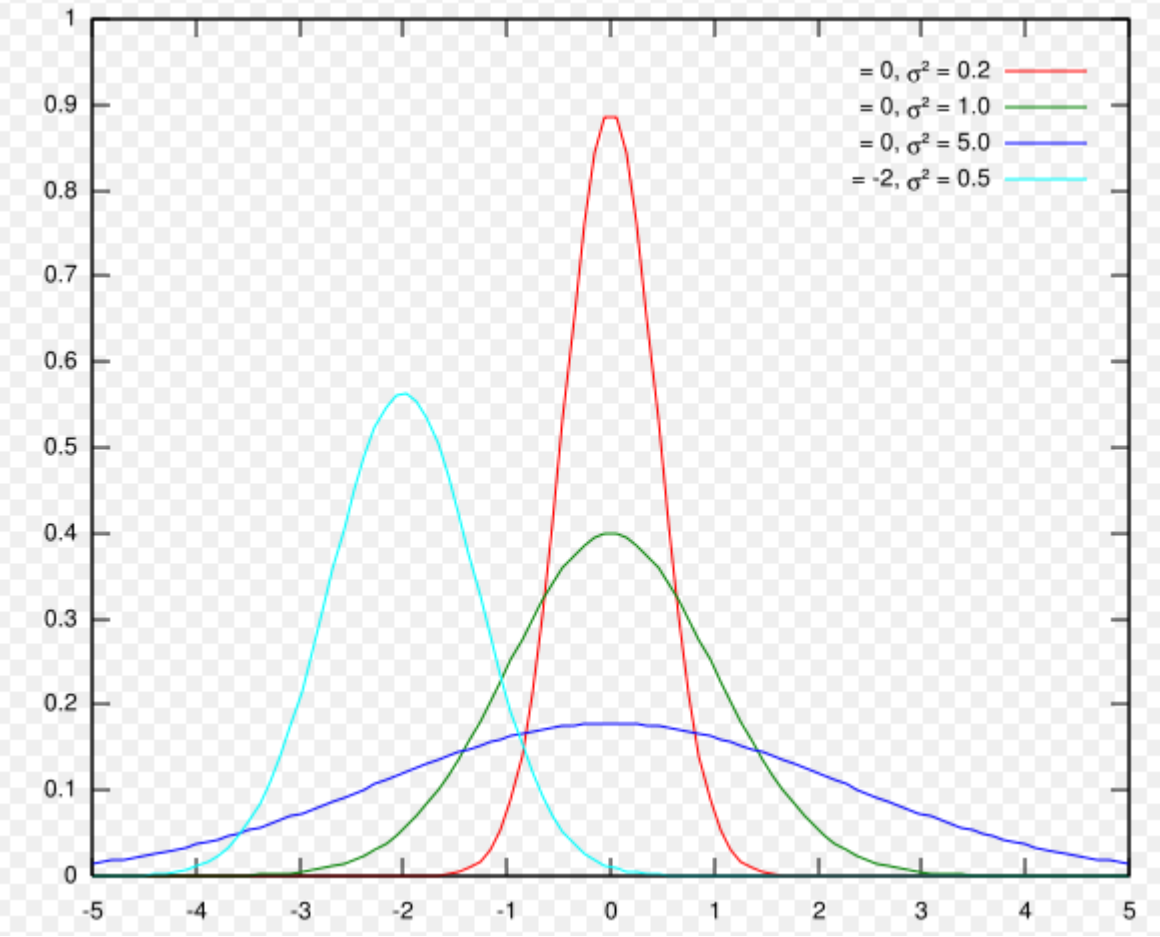
- Számítsuk ki, hogy szabályos pénzérme esetén, mekkora az esélye, hogy f1 illetve f2 elfogadja a sorozatot:
  - f1: 2^{20} lehetséges sorozat van, az hogy 10 ± 4 darab 1-es van benne annak a lehetséges darabszáma: $\binom{20}{6} + \binom{20}{7} + \binom{20}{8} + \binom{20}{9} + \binom{20}{10} + \binom{20}{11} + \binom{20}{12} + \binom{20}{13} + \binom{20}{14}$ =>

    =>vizsgált esetek/összes ≈ 0.96 => tehát szabályos pénzérme esetén az esetek 96%-ban teljesül ez a feltétel.
  - f2: ≈ 0.94 (házi feladat kiszámítani), tehát szabályos pénzérme esetén az esetek 94%-ban teljesül ez a feltétel.
  - Vagyis annak az esélye, hogy az f1-en megbukik egy „jó” sorozat az körülbelül 4% míg, hogy az f2-ön az körülbelül 6%, ez a elsőfajú hiba felső becslése egyben.
  - 11001100011110010011 => f1:H0, f2: H0
  - 10101010101010101010 => f1: H0, f2: H1
  - 00000100100100010010 => f1: H1, f2: H0
  - Miért nem becsülhető felülről a másodfajú hiba?

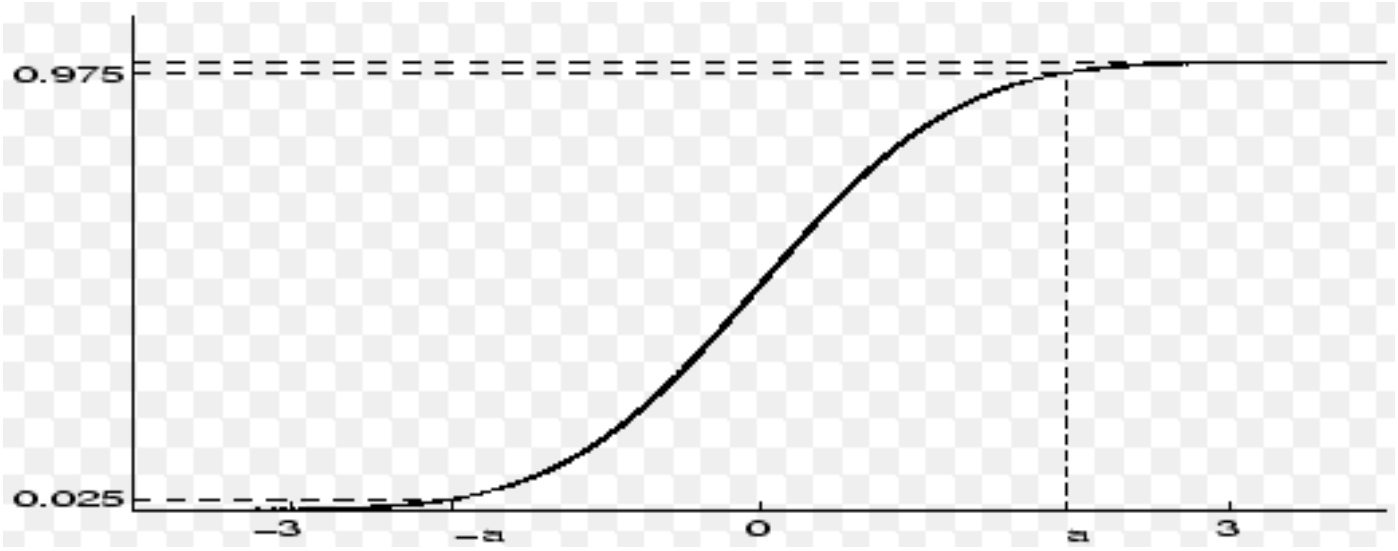
Normális (Gauss) eloszlás

- X valószínűségi változó normális eloszlású, ha a sűrűség függvénye pontosan:
  - f(x)=$\frac{1}{\sigma*\sqrt{2\pi}} * e^{-\frac{(x-m)^2}{2\sigma^2}}$, ahol a két paraméter σ és m valós számok és $\sigma >$ 0. Jelölése: X∼N (m, σ^2).
- Max(X)=E(X)=m
- D(X)= σ
- Szimmetrikus függvény
- Az X∼N(0,1)akkor X standard normális eloszlású
- Tetszőleges normális eloszlás standardizálható:
  - Ha X∼N(m, σ^2) akkor a (X-m)/σ ∼N(0,1)
- Független normális eloszlású val. változók összege is normális, illetve ha független val. változók összege normális, akkor egyenként is azok.

Normális (Gauss) eloszlás



Normális (Gauss) eloszlás



Normális eloszlás – Pont becslés

- Adottak az elmúlt hat év júliusi középhőmérsékletei „A” városban: 24, 28, 25, 26, 22, 25 fok. Feltéve hogy tudjuk, hogy a középhőmérséklet egy adott év adott napján normális eloszlást követ, mi lehet a normális eloszlás várhatóértéke?
- Számítsuk ki a minta átlagát:
  - 1/6(24+28+25+26+22+25) = 25 fok.
  - Belátható, hogy ez egy jó becslése az eredeti eloszlás várhatóértékének.
  - De vajon mekkorát tévedünk?

Normális eloszlás – Intervallum becslés – Konfidencia Intervallum

- Becsüljük meg, hogy a mért értékek alapján, egy adott valószínűség mellett, az eredeti eloszlás várhatóértéke mekkora intervallumból származhat
- Ehhez először is meg kell adnunk a megbízhatósági szintet (szignfikancia szint) vagyis, hogy az esetek hány százalékára legyen igaz az intervallum.
  - Pl: 95%-os megbízhatósági szint szerint keressük az adott intervallumot, akkor annak az esélye, hogy az eredeti eloszlás várhatóértéke mégsem ebből az intervallumból származik maximum 5%.
- Ezután megkeressük azokat az x értékeket amire igaz, hogy ha x az adott normális eloszlás várhatóértéke, akkor annak az esélye, hogy a mért adatok ilyen várhatóértékkel és szórással rendelkeznek legalább 95%.
- Ezt például tudja az excel is:
  - $25 \pm 2{,}67$ (95%) vagy 3,5 (99%)
  - Ha a megbízhatósági szintet 100%-ra vennénk akkor az intervallum végtelen lenne.

Normális eloszlás – Egymintás u-próba

- Adott egy sörcsapoló gép. Tudjuk, hogy a gép $X \sim N(5, 1)$ normális eloszlás szerint csapolja a sört (dl). Kikérünk 5 sört, és a következő értékeket kapjuk: 5, 4.4, 4, 5.3, 5.7 . Az a gyanúnk támad, hogy a gép várható értékét elállították. Hogyan tudjuk ez bizonyítani?
- $\bar{x}$=4,88

Normális eloszlás – Egymintás u-próba

- Legyen:
  - H0:= a gép 5 egységet csapol
  - H1:= a gép nem 5 egységet csapol
- Legyen a próbastatisztikánk a következő:
  - u:=$\frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}}$, ahol az $\bar{x}$ a minta átlaga, m a viszgált val. változó ismert várható értéke, σ a vizsgált val. változó ismert szórása, n a mintadarabszáma

Normális eloszlás – Egymintás u-próba

- Válasszunk szignifikancia szintet legyen 0.05 (95%-os)
- u=$\frac{\bar{x}-m}{\frac{\sigma}{\sqrt{n}}} \approx \frac{4,88-5}{\frac{1}{2,236}} \approx$ -2,68.
- Valójában ezzel standardizáltuk az eloszlást, ezek után az a kérdés, a X∼N(0 , 1) ből ha elhagyjuk a 0.05/2 kvantilist és a (1-0.05/2) kvantilist akkor milyen intervallumot kapunk:
  - -1,96($-u_{p/2}$)– +1,96($u_{p/2}$)
- Mivel az |u| > $u_{p/2}$ ezért elvethetjük a nullhipotézist. Tehát kijelenthetjük, hogy a gép várható értéke szignifikáns eltérést mutat az 5-höz képest 0.05 (95%)-os szignifikancia szint mellett u próba alapján. Elvetjük a H0-t, a tévedés esélye maximum 5%.
- Válasszuk a szignifikancia szintet 0.01-re. Ekkor a próbastatisztikánk értéke ugyanaz marad. Viszont a standard normális eloszlás 99%-os tartományát kell ki választanunk, így az $u_{p/2}$-re 2,81-et kapunk. Tehát kijelenthetjük, hogy a gép várható értékében nem tudtunk kimutatni szignifikáns eltérést 0.01 (99%) –os szignifikancia szint mellett az u próbával. Megtartjuk a H0-t.
- Mekkora az esélye az egyes esetekben, hogy hamisan vádoljuk meg a csapolót.
- Vegyük észre, hogy ha a szignifikancia szintet 100%-ra emeljük, akkor $u_{p/2}$-re ∞ kapunk vagyis mindig megtartjuk a H0 hipotézist.

Normális eloszlás – kétmintás t próba

- Van két pék „X” és „Y”, akik kenyeret sütnek és ugyanannyiért árulják a kenyeret. Mindkettőtől veszünk 10-10 kenyeret, amiknek a következő a tömege:
  - „X”: 92, 95, 97, 102, 103, 100, 96, 97, 101, 104
  - „Y”: 93, 94, 101, 98, 104, 101, 97, 102, 96, 100
- Van e szignifikáns eltérés a két pék kenyereinek a tömege között, feltéve ha tudjuk, hogy normális eloszlást követ a kenyereiknek a tömege.

Normális eloszlás – kétmintás t próba

- Legyen:
  - H0:= E(X)=E(Y)
  - H1:= E(X)≠E(Y)
- Legyen a próbastatisztikánk a következő:
  - $\frac{\bar{x}-\bar{y}}{\sqrt{(n-1)(s*_x)^2+(m-1)(s*_y)^2}}\sqrt{\frac{nm(n+m-2)}{n+m}}$, ahol
    - $\bar{x}$ és $\bar{y}$: az egyik illetve a másik minta átlaga
    - $s*_x$ és $s*_y$: az egyik illetve a másik minta korrigált szórása (n helyett (n-1)-el kell osztani a szórás képletében, mivel ez a pontos becslése az eloszlások szórásának)
    - n és m: az egyik illetve a másik minta darabszáma
- A két mintás t próba feltételei
  - A val. változók normális eloszlást követnek
  - A val. változók szórása megegyezik
  - (függetlenek)

Normális eloszlás – kétmintás t próba

- Az adott példában tegyük, fel hogy a szórásaik megegyeznek. (ezt majd F próbával tudjuk bizonyítani)
- Számítsuk ki a t próbastatisztika értékét:
  - $\bar{x}$ = 98.7; $\bar{y} = 97.6$
  - $(s *_x)^2 \approx 15.1$; $(s *_y)^2 \approx 10.9$
  - m=n=10
  - t=0,68
- Válasszunk szignfikancia szintet 0.05 (95%)-os
- Vessük össze a Student táblázattal f= n + m -2 szabadsági fokkal. Ha |t| < tp komperátor érték ami a Student táblázatból kiolvasható, akkor megtartjuk a nullhipotézist.
- Student(95%) f=18-hoz 2,1 tp érték tartozik, vagyis meg tartjuk a H0-t tehát, nem tudtunk szignifikáns eltérést kimutatni a két minta átlagai között.
- Megjegyzés: 50%-os szignifikancia szint mellett is elfogadnánk a H0-t vagyis feltehetőleg tényleg azonos a várhatóértékük.

Normális eloszlás – F próba

- Van két pék „X” és „Y”, akik kenyeret sütnek és ugyanannyiért árulják a kenyeret.
  Mindkettőtől veszünk 10-10 kenyeret, amiknek a következő a tömege:
  - „X”: 92, 95, 97, 102, 103, 100, 96, 97, 101, 104
  - „Y”: 93, 94, 101, 98, 104, 101, 97, 102, 96, 100
- Azt szeretjük, hogyha várhatóan minimális az adott pék kenyereinek az eltérése az átlagtól.
- Melyiktől vásároljunk?

Normális eloszlás – F próba

- Legyen:
  - H0:= M(s_1^2-s_2^2)=0
  - H1:= M(s_1^2-s_2^2)≠ 0
- Legyen a próbastatisztikánk a következő:
  - $F_s = \frac{s_1^2}{s_2^2}$, ahol s_1^2>=s_2^2, és a szabdsági fokok: n1-1;n2-1
  - Kritikus értéket az F eloszlás táblázatból vesszük
- F próba feltétele a normalitás (erősen függ tőle)

Normális eloszlás – F próba

- $F_s = \frac{s_1^2}{s_2^2} = 1{,}169$
- Legyen szignifikancia szint 95% => α=0,05
- $F_{krit}(\alpha/2\ )$= 4,03
- Tehát a két pék kenyereinek tömegének a szórása között nem tudtunk szignifikáns különbséget kimutatni.
- Megjegyzés α vagy $\alpha/2$?

Párosított T próba

- A két adatsorunk nem független egymástól.
- Ebben az esetben használjuk a következő próba statisztikát:
  - $t = \frac{\bar{x} - m}{\frac{s_*}{\sqrt{n}}}$, ahol az x változó az páronkénti különbségekből képzett változó.
  - (Ez az egymintás t próba.)
  - t-eloszlás táblázatból kiolvasandó a megfelelő kritikus érték
- Megjegyzés létezik kétmintás u próba is. Azt akkor használjuk, ha ismert a két adatsor szórása.

Normális eloszlás – Illeszkedési vizsgálat

- Kérdés, normális eloszlásúak voltak-e az adott példákban szereplő val. változók?
- Illeszkedés vizsgálathoz legalább 30 adat célszerű.
- Kolmogorov-Szmirnov próba:
  - Adott val. változó illeszkedését vizsgálja egy előre meghatározott eloszláshoz
  - Szerkesszük meg az eloszlás függvényét a meghatározott eloszlásnak (pl: normális) és a leíró paramétereit határozzuk meg a mért adatokból
  - Szerkesszük meg az empirikus eloszlás függvényét a mért adatoknak.
  - Keressük meg a legnagyobb eltérést, ez lesz a próba statisztika eredménye.
  - Ezt hasonlítsuk össze egy kritikus értékkel és ez alapján döntsünk.
- Ez sem bonyolult, de papíron már macerásabb számolni.
- Vegyük észre, hogy a legnagyobb eltérés az biztosan az empirikus eloszlás függvény „törés” pontjaiban lehetséges.

Összefoglalás

- Statisztika:
  - Megfigyelt jelenség(ek)ből való **tudás szerzés**, aminek a célja:
    - „becslés” (részből következtetés az egész populációra)
    - „jóslás” (időjárás előrejelzés, tőzsde változásának előrejelzés)
    - „jelenség megértése” (összefüggések kimutatása)
    - „osztályozás és regresszió”
    - „korreláció vizsgálat”
  - Mivel valamiért **nem tudjuk a jelenséget egészében** megfigyelni (nem áll rendelkezésünkre az összes adat, nem determinisztikus, stb…)

Összefoglalás

- Ajánlott lépések:
  - **Hipotézisek megfogalmazása**
  - **Mért változó(k) leíró jellemzőinek kiszámítása**
    - átlag, szórás, stb…
  - **Mért változó(k), empirikus függvényének/hisztogramjának a grafikus megjelenítése**
    - A vizsgált jelenség(ek) eloszlásának megbecsülése, hipotézisek felállítása
    - Illeszkedés vizsgálat
  - **Alaphipotézis(ek) ellenőrzése, összefüggések kimutatása**
    - Próbák elvégzése (hiba becslése)
    - Correláció vizsgálat
    - stb…
  - **Újabb hipotézisek megfogalmazása…**