

Egységtesztelés a statisztikában

Dr. Baranyai László

Összefoglaló. Az egységtesztelés módszere programozó körökben ismert elsősorban. Célja, hogy automatikusan ellenőrizzük az algoritmusok működését. Röviden: olyan számításokat végzünk, amelyek eredményét előre ismerjük. Ezt a technikát kitűnően használhatjuk tanuláshoz, a statisztikai mérőszámok és próbák mélyebb megértésére. Egy példán keresztül követheti az egységtesztelés módszerét [R környezetben](#). A normális eloszlásból vett véletlen mintákat elemezve figyelheti meg az elemszám és a szórás hatását a számított átlag megbízhatóságára.

Átlag számítása

Első lépésként létrehozuk a teszt adatsort az `rnorm()` függvény segítségével. A függvénynek három paramétere van: a minta elemszáma, az eloszlás középértéke és végül a szórás. Ezek alapján az egységtesztet a következő:

```
> x <- rnorm(999, mean=8, sd=1)
> mean(x)
[1] 8.038424
> sd(x)
[1] 0.9767589
```

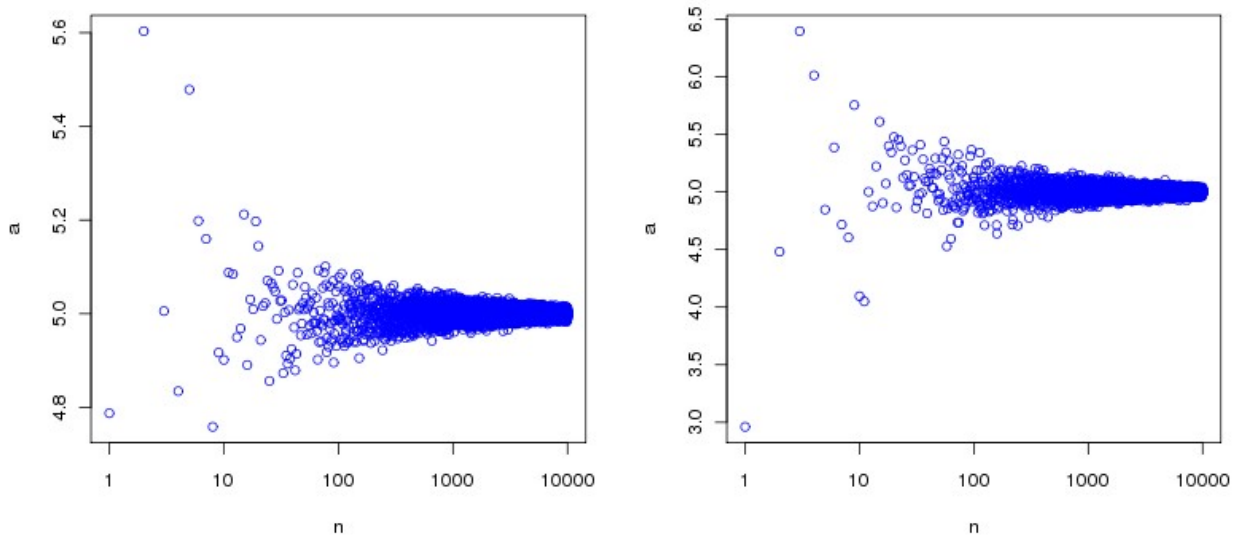
Az eredmények a véletlen számok miatt próbánként eltérőek lehetnek. Itt érdemes megvizsgálni az elemszám és a szórás hatását az átlag megbízhatóságára. A következő `UTest.mean()` függvény a maximális elemszámig növelve a véletlen minta méretét újraszámolja az átlagot és a végén grafikonon ábrázolja az adatsort.

```
UTest.mean <- function(Items, Mean, SD)
{
  n <- seq(1, Items, by=1)
  a <- rep(0, Items)
  for (i in 1:Items) a[i] <- mean( rnorm(i, mean=Mean, sd=SD) )
  plot(n, a, log="x", col="blue")
}
```

Próbáljuk ki 10% és 30% variációs tényezővel, $N=10^4$ elemszámig. A kód futása a számítógép teljesítményétől függően több másodperc is lehet.

```
> UTest.mean(1e4, 5, 0.5)
> UTest.mean(1e4, 5, 1.5)
```

A grafikonokat az 1. ábrán hasonlíthatjuk össze. Jól látható a növekvő elemszám pozitív hatása a becslés pontosságára. A variációs tényezőtől függetlenül $N=100$ elemszám alatt jelentős eltérések lehetnek az eredményben.



1. ábra: Egységtesztelés eredménye 10% (bal) és 30% (jobb) variációs tényezővel

A kívánt pontosság eléréséhez szükséges elemszám ilyen úton is kikövetkeztethető, de ismert az összefüggés a pontosság, szórás és az elemszám között:

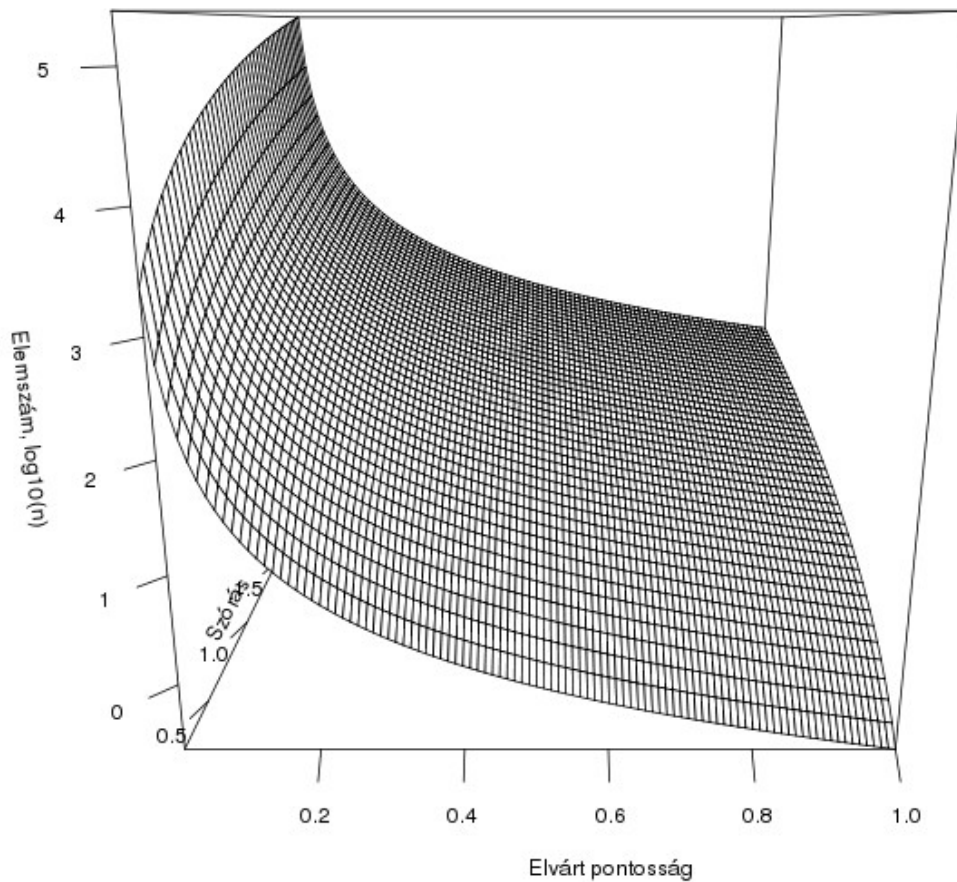
$$n = \left(\frac{t \cdot S}{\Delta} \right)^2 \quad (1)$$

ahol n az elemszám, S az adatok becsült szórása, Δ az elvárt pontosság és t a student eloszlás kritikus értéke ($t=1,96$ ha $p<0,05$ és $df=\infty$). Ha a fenti minták átlagát egy tizedes értékre helyesen kívánjuk meghatározni, a tévedés legfeljebb $\pm 0,05$ lehet a kerekítési hiba miatt. Az 1. egyenletbe behelyettesítve a szükséges legkisebb elemszám (kerekítve) 385 és 3458 db a 10% és 30% variációs tényezővel rendelkező mintákra, ebben a sorrendben. Az 1. ábra pontfelhői ezt megerősítik.

Az elvárt pontosság és a szórás hatásának megfigyeléséhez készítsünk egy ábrát. Az adatok variációs koefficiensét 5% és 50%, a pontosságot pedig 0,01 és 1 között változtatva számítsuk ki a szükséges mintaszámot! A jelentős eltérések miatt az elemszám logaritmusát érdemes ábrázolni, továbbá a könnyebb áttekinthetőség kedvéért transzponáljuk. A 2. ábra a `persp()` utasítás felhasználásával készült.

```
> x <- seq(0.25, 2.5, by=0.05)
> y <- seq(0.01, 1, by=0.01)
> n <- matrix( rep(0, length(x)*length(y)), ncol=length(y) )
> dim(n)
[1] 46 100
> for (i in 1:46) { for (j in 1:100) n[i,j] <- (1.96*x[i]/y[j])^2 }
> persp(y, x, log10(t(n)), xlab="Elvárt pontosság", ylab="Szórás",
+ zlab="Elemszám, log10(n)", ticktype="detailed")
```

Az egységtesztelés módszerével megvizsgálva az átlag függvény viselkedését megerősíthetjük a statisztikai elemzések alap felvetését: kevés adatból nem kapunk megbízható eredményeket. A 95% megbízhatóság eléréséhez legalább 20 elemű mintát javasolnak, figyelembe véve az 1 db hibás adattal bekövetkező 5% veszteséget. A bemutatott számításokkal már ennél pontosabb követelmény is támasztható.



2. ábra: Az elvárt pontosság és a szórás hatása a szükséges elemszámra

Gyakorló feladat

Egészítse ki az `UTest.mean()` függvényt úgy, hogy

- folytonos piros vonallal jelölje az elvárt átlagértéket,
- tetszőleges további paramétereket adhatunk, amelyeket a `plot()` függvény használ fel (természetesen a szín is ilyen paraméter kell legyen)