A középérték és variancia azonosságának próbái: t-próba, F-próba
Logikai vektorok az R-ben

A normális eloszlás jelentősége

A Fogbarát Cukorgyár ellen panasz érkezik. A márka vásárlói azt állítják, hogy a csomagoláson szereplő 1000 g helyett a csomagban lévő kristálycukor valójában mindig kevesebb. A cukorgyár szerint az értékek átlaga pontosan 1000 g, és valóban előfordul ingadozás, ennek törvényben rögzített szórása 10 g.
A fogyasztóvédelem 50 mintát vesz az ország különböző helyein található üzletekből.
Az eredményt a `cukor` vektorban rögzítik. Letölthető innen:
http://clara.nytud.hu/~mady/courses/statistics/materials/cukor.RData

Igaz-e, hogy a fogyasztók átlagosan 1000 g cukrot visznek haza, és hogy a gyártó nem károsítja meg őket a saját javára?

Nullhipotézis: a minta átlaga beleesik az összes cukorcsomagnál megállapított 10 g-os szórás által megszabott konfidenciaintervallumba.

Konfidenciahatárok megállapítása:
`library(gmodels) ci(cukor)`

Igazat mond a cukorgyár?

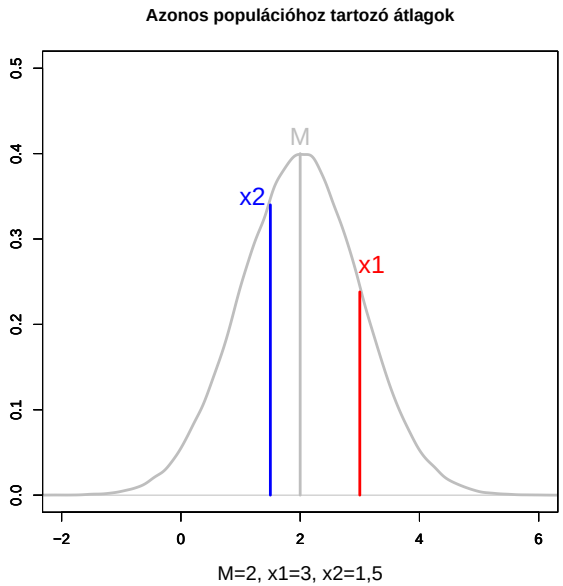
A kritikus fogyasztók maguk kezdenek mérőkampányba az interneten, és a `cukor.ism` objektumban található méréseket folytatják a külön e célból beszerzett patikamérlegen.

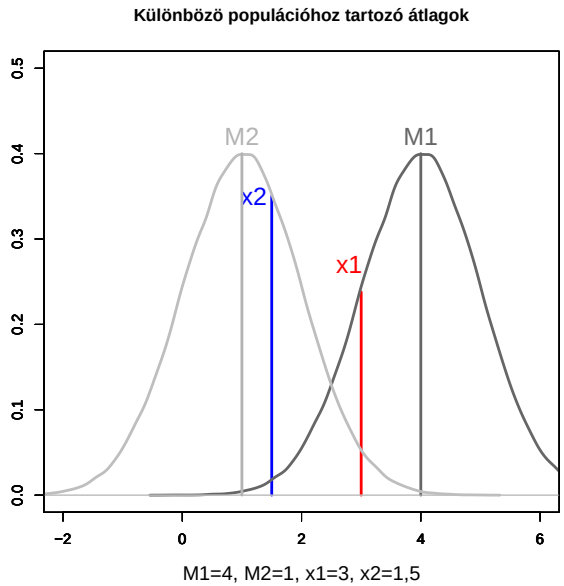
Hasonlít-e az átlag a fogyasztóvédelem által megállapítotthoz? Igaz-e rá, hogy beleesik a törvény által megengedett ingadozásba?

És az igaz, hogy a fogyasztók ugyanolyan arányban visznek haza kicsivel több cukrot, mint kevesebbet?

Hipotézisállítás

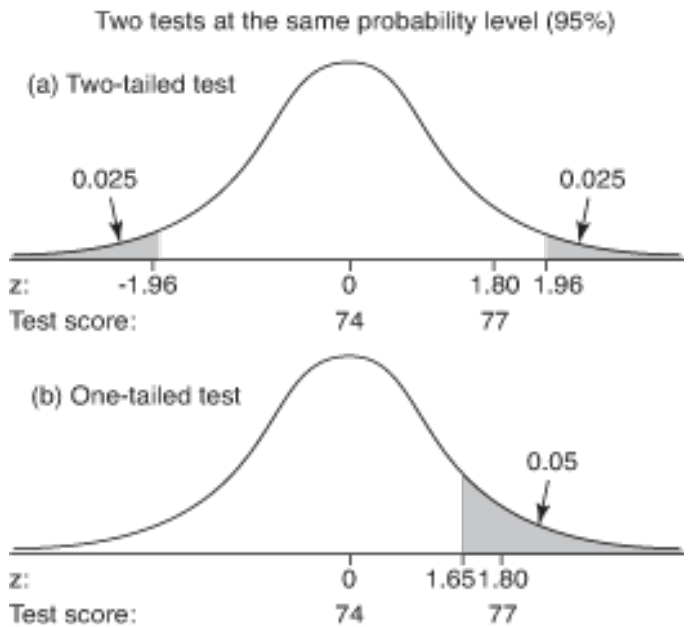
- Feltételezés: *a* minta egy adott szempont alapján más populációhoz tartozik, mint *b* minta.
- Nullhipotézis (H_0): *a* minta és *b* minta egyazon populációhoz tartozik, azaz az átlaguk ugyanazon μ populációátlag körül szór.
- Ellenhipotézis (H_1): *p* valószínűséggel állítható, hogy *b* minta átlaga nem ugyanahhoz a populációhoz tartozik, mint az *a* minta.





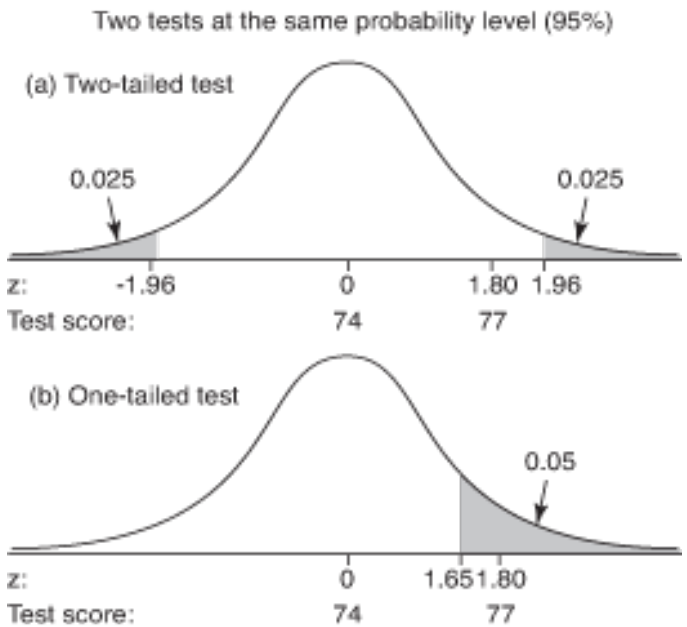
Hipotézis tesztelése $p = 95\%$-os megbízhatósággal

1. H_1: *a* nagy valószínűséggel **eltér** *b*-től.
   H_0: *a* és *b* ugyanazon populáció része. Elutasítás: ha $\bar{x}$ a sűrűségfüggvény két szélén $\alpha/2$-be esik $\Rightarrow$ kétoldali teszt (felső ábra).



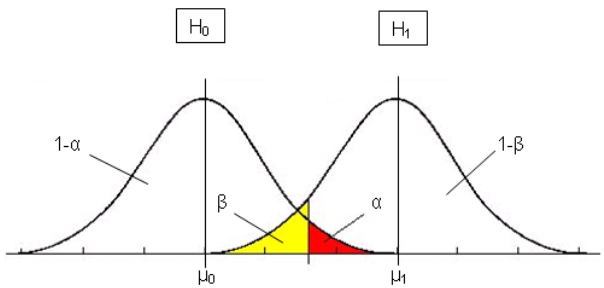
Hipotézis tesztelése $p = 95\%$-os megbízhatósággal

1. H_1: *a* nagy valószínűséggel **eltér** *b*-től.
   H_0: *a* és *b* ugyanazon populáció része. Elutasítás: ha $\bar{x}$ a sűrűségfüggvény két szélén $\alpha/2$-be esik $\Rightarrow$ kétoldali teszt (felső ábra).
2. H_1: *a* nagy valószínűséggel **nagyobb**, mint *b*.
   H_0: *b* nem kisebb, mint *a*. Elutasítás: ha $\bar{x}$ a sűrűségfüggvény jobb szélén α-ba esik $\Rightarrow$ egyoldali teszt (alsó ábra).



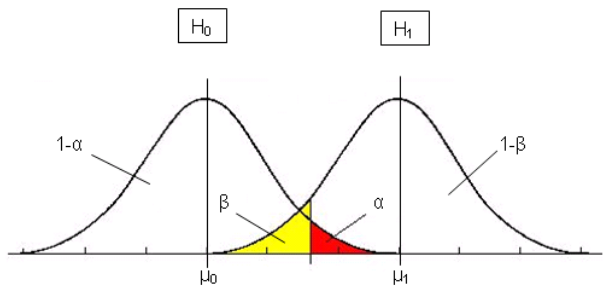
Hibatípusok

1. α**-hiba (első fajta, elsőfajú hiba):** elutasítjuk H_0-t, mert az átlag a megadott konfidenciaintervallumon kívül esik $\rightarrow \alpha$ része (piros tartomány).
2. β**-hiba (második fajta, másodfajú hiba):** megtartjuk H_0-t, holott az átlag más populációhoz tartozik (sárga tartomány).



Hibatípusok

1. α-**hiba (első fajta, elsőfajú hiba):** elutasítjuk H_0-t, mert az átlag a megadott konfidenciaintervallumon kívül esik $\rightarrow \alpha$ része (piros tartomány).
2. β-**hiba (második fajta, másodfajú hiba):** megtartjuk H_0-t, holott az átlag más populációhoz tartozik (sárga tartomány).



	H_0-t megtartjuk	H_0-t elvetjük
H_0 igaz	helyes döntés	α-hiba (álpozitív)
H_1 igaz	β-hiba (álnegatív)	helyes döntés

Összehasonlítás alapjai

- Átlagok,
- szórások,
- minta populációval $\leftrightarrow$ minta mintával,
- azonos varianciák $\leftrightarrow$ eltérő varianciák,
- független $\leftrightarrow$ párosított minták,
- parametrikus $\leftrightarrow$ ordinális vagy nem normális eloszlású minták.

Ha a populáció σ szórása ismert: átlagok z-eloszlás szerint szórnak μ körül.

Gyakorlatban: a populáció szórása nem ismert, ezért a mintaátlag szórását a Student-féle t eloszlással jellemezzük.

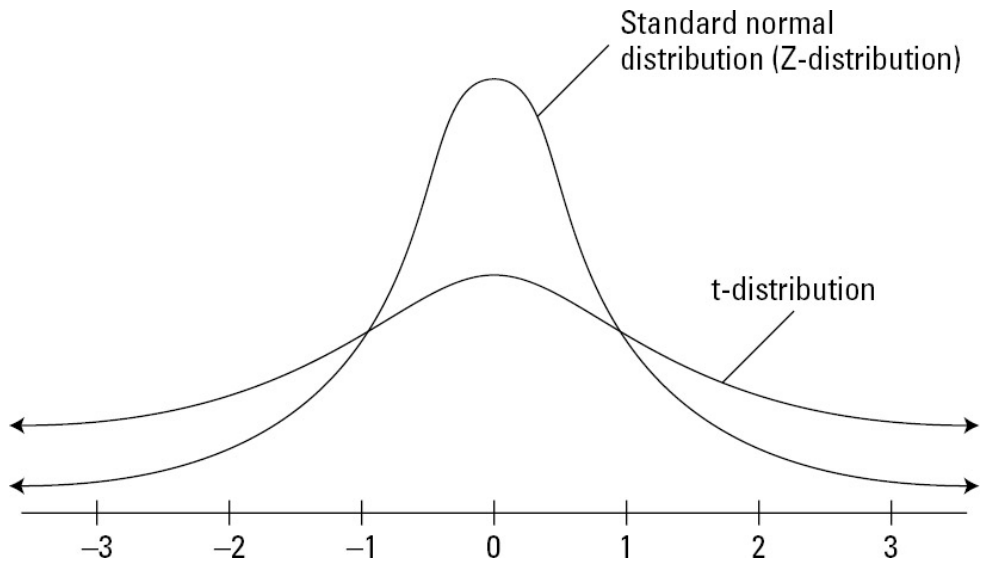
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$\sim$ normalizálás a z értékre, de σ helyett s.

t-eloszlás

Jellemzők:

- Szimmetrikus, átlaga 0, aszimptotikus, de nem normális eloszlású.
- Függ a minta méretétől, n-től.
- A t-eloszlás laposabb, mint $z \Rightarrow$ adott szignifikanciaszint határértékei messzebb esnek az átlagtól.
- $n = \infty$ esetén t eloszlás azonos z eloszlással.
- $n \geq 100$ esetén a különbség elhanyagolható, z-értékeket lehet használni.



Szabadsági fokok

Szabadsági fok, *degree of freedom, df*: a szabadon változtatható elemek száma, ami mellett a minta egy adott tulajdonsága változatlan marad.

Pl. egy $n = 5$ elemű minta átlaga $\bar{x} = 10$. Hány elem változtatható szabadon a mintaátlag változatlansága mellett?

Szabadsági fokok

Szabadsági fok, *degree of freedom, df*: a szabadon változtatható elemek száma, ami mellett a minta egy adott tulajdonsága változatlan marad.

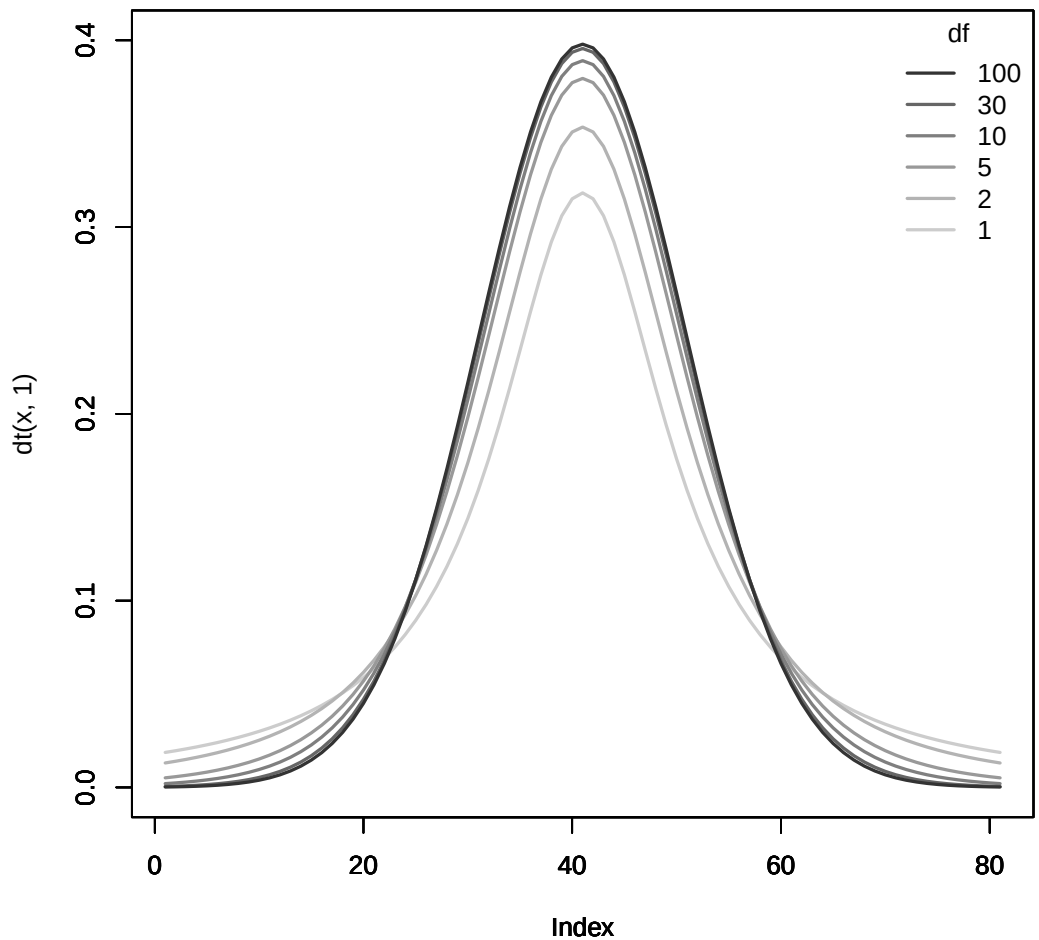
Pl. egy $n = 5$ elemű minta átlaga $\bar{x} = 10$. Hány elem változtatható szabadon a mintaátlag változatlansága mellett?

Négy, hiszen az ötödik elemet úgy kell kiválasztani, hogy a minta átlaga 10 maradjon, tehát csak négy elem változtatható szabadon.

Tehát $df = n - 1$.

t-eloszlás és szabadsági fokok

A t-eloszlás laposssága függ a szabadsági fokoktól. Minél nagyobb a szabadsági fok, annál közelebb esik a kritikus érték (= szignifikanciahatár, konfidenciaintervallum szélső értéke) az átlaghoz.



Egymintás Student-féle t-próba

- Feltétel: normális eloszlású változó, ismeretlen szórással.
- Alkalmazás: populáció vagy nagyszámú referenciaminta átlaga ismert, pl. IQ $= 100$.
- Eljárás: ha $t_{minta} > t_{1-\alpha(n-1)} \Rightarrow H_0$ elvetése.

A Kincskereső óvodába 60 okos és ügyes gyerek jár. Átlagos IQ-juk 108, a szórás 10. Okosabbak-e az oda járó gyerekek az átlagnál?

Feladat

Átlag: 108, populáció átlaga: 100, szórás: 10, elemek száma 60.

$$t_{minta} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{108-100}{10/\sqrt{60}} = \frac{8}{1{,}29} = 6,2$$

Kritikus értékhez tartozó t meghatározása ($p = 1 - \alpha = 0,95$):
adott kvantilishez (0,95) tartozó t-érték 59-es szabadsági fok mellett:
`qt(p,df)`, itt: `qt(0.975,59)` $\rightarrow$ 2,000995

Feladat

Átlag: 108, populáció átlaga: 100, szórás: 10, elemek száma 60.

$$t_{minta} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{108-100}{10/\sqrt{60}} = \frac{8}{1{,}29} = 6,2$$

Kritikus értékhez tartozó t meghatározása ($p = 1 - \alpha = 0,95$): adott kvantilishez (0,95) tartozó t-érték 59-es szabadsági fok mellett:
`qt(p,df)`, itt: `qt(0.975,59)` $\rightarrow$ 2,000995

Mivel $t_{minta} > t_{0.95(59)} \Rightarrow H_0$-t elutasítjuk.

A Kincskereső óvodába tehát az átlagnál szignifikánsan intelligensebb gyerekek járnak.

Kétmintás független t-próba

- Két minta alapján két ismeretlen μ értéket hasonlítunk össze.
- Minták kiválasztása egymástól független (pl. spanyol óvodások és cseh óvodások).
- Feltétel: normális eloszlás, azonos varianciák

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

ahol s mindkét mintában azonos: a közös variancia becslése a mintánkénti szórásokból.

DE: a szórás egyenlőségét ritkán állíthatjuk biztosan!

Welch-próba

Mint a kétmintás független t-próba, de nem feltételezzük a varianciák egyenlőségét.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Eltér-e az alábbi mintában a nőstény és hím borjak születéskor mért testtömege?

bika (kg)	46	37	39	37	33	48	35		
üsző (kg)	27	37	35	41	35	34	43	38	40

```
bika = c(46,37,39,37,33,48,35)
uszo = c(27,37,35,41,35,34,43,38,40)
```

Feladat

Normális eloszlásúak-e a minták?
`shapiro.test(bika), shapiro.test(uszo)`
Ha p az adott szignifikanciaszintnél nagyobb, elfogadjuk a normális eloszlás feltételezését.

Két minta összehasonlítása t-próbával:
`t.test(bika,uszo)`
alapbeállítás: kétoldali (alternative=two.sided), varianciák nem egyenlőek (var.equal=FALSE).

Feladat

Normális eloszlásúak-e a minták?
`shapiro.test(bika), shapiro.test(uszo)`
Ha p az adott szignifikanciaszintnél nagyobb, elfogadjuk a normális eloszlás feltételezését.

Két minta összehasonlítása t-próbával:
`t.test(bika,uszo)`
alapbeállítás: kétoldali (alternative=two.sided), varianciák nem egyenlőek (var.equal=FALSE).

$p > 0,05 \Rightarrow$ különbség nem szignifikáns.

Kétmintás páros t-próba

A minta egyazon elem vagy összetartozó elemek kétszeri megfigyeléséből áll.

Feltétel: egy elem két értékének különbsége normális eloszlású, $n \geq 30$ esetén feltételezni szokás a normális eloszlást.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

ahol $\bar{d}$ a különbségek átlaga, s_d a különbségek becsült szórása, n a párok száma (tehát az elemszám, nem a mérések száma).

Feladat

`ratings` adatmátrix.
Növények és állatok nevének gyakorisága és ismertsége (Frequency, meanFamiliarity). Különböznek-e a méretre és súlyra adott becslések páronként?

Normális eloszlás tesztelése:

```
shapiro.test(ratings$Frequency)
shapiro.test(ratings$meanFamiliarity)
```

páros t-próba:

```
t.test(ratings$Frequency,ratings$meanFamiliarity,
paired=T)
```

$p \ll 0,001$, tehát a megkérdezettek az állatok és növények méretét szignifikánsan nagyobbra becslik egy adott skálán, mint a súlyukat.

Varianciára vonatkozó próbák

Tesztek és feltételeik (legalább) két minta esetén:

- F-**próba:** mindkét mintában normális eloszlás, független minták. R: `var.test()`.
- **Levene-próba:** közelítő próba, de normális eloszlás hiányában is használható, több mintára is. R: `levene.test` a `car` könyvtárban.
- **Bartlett-próba:** normális eloszlás, páros mintákra is használható. R: `bartlett.test()`.

Csomag telepítése: `install.packages("car")`, `library(car)`

Feladat

Töltsük le a trans.RData fájlt innen:
clara.nytud.hu/∼mady/courses/statistics/materials/trans.RData
Letöltés `load("konyvtar/trans.RData")` függvénnyel (NEM `read.table()`).

A mátrixban angol, ill. portugál, kb. 1500 szavas szövegek hossza van megadva, majd a másik nyelre való lefordítás utáni hosszuk.

Ellenőrizzük, azonos-e az angol és portugál szövegek varianciája, majd teszteljük, szignifikánsan különböznek-e.

`var.test(fuggovaltozo∼fuggetlenvaltozo)`, azaz
`var.test(trans$length∼trans$language)`

`t.test(fuggovaltozo∼fuggetlenvaltozo)`, azaz
`t.test(trans$length∼trans$language)`

Logikai vektorok

Szűrés: próbák az adatmátrix adott feltételnek megfelelő elemeire.
Eljárás: az adatmátrixban egy adott változón belüli csoportok definiálása.

Operátorok:

==	azonos
!=	nem azonos
%in%	tartalmazza a vektor valamely elemét
<	kisebb, mint
>	nagyobb, mint
<>	nem egyenlő
|	vagy
&	és

Logikai vektorok definíciója

z = ratings$Class == "plant"
z = testmagassag$height < 170

feltételt teljesítő sorok listázása:
`ratings[z,]`

feltételt NEM teljesítő sorok listázása:
`ratings[!z,]` – főleg akkor praktikus, ha csak két faktorszintünk, azaz kategóriánk van

összes elem feltételt teljesítő elemei vektorként:
`ratings$Class[z]`

Melyik elemekre igaz:
`which(z)`

Összes előfordulás:
`sum(z)`

Feladat

Növények és állatok ismertségi foka (meanFamiliarity):
pontdiagramm készítése az állatokra és a növényekre eltérő színnel.
A tengelyhosszok legyenek azonosak.
logikai vektor: csak növények:

```
z = ratings$Class == "plant"
```

Növények [z] ábrázolása piros színnel:

```
plot(ratings$Frequency[z],ratings$meanFamiliarity[z],
col="red",xlim=range(ratings$Frequency),
ylim=range(ratings$meanFamiliarity))
par(new=T)
```

állatok [!z] ábrázolása kék színnel

```
plot(ratings$Frequency[!z],ratings$meanFamiliarity[!z],
col="blue",xlim=range(ratings$Frequency),
ylim=range(ratings$meanFamiliarity))
```

`range()`: egy adott vektor terjedelme (min. . . max)

Feladatok

Hasonlítsuk össze a bikák és üszők születési súlyának varianciáit a megfelelő tesztekkel. Milyen különbségeket látunk?

`ratings` adatmátrix:
Boxplotok készítése adott csoportra: pl. növények között az egyszerű és összetett szavak gyakorisága.

Igaz-e, hogy a ratings mátrixban szereplő állatnevek gyakorisága alacsonyabb mértékű, mint a növényeké? És az ismertségük?

Milyen értéket vesz fel p, ha a növények és állatok súly- és méretbecslését külön-külön vizsgáljuk?

Teszteljük minden esetben, hogy az adatok normális eloszlásúak-e, és hogy a varianciák homogének (egyenlőek)-e.