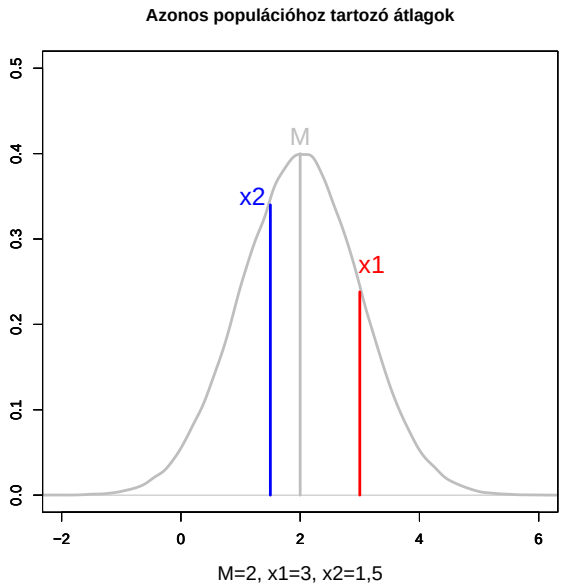
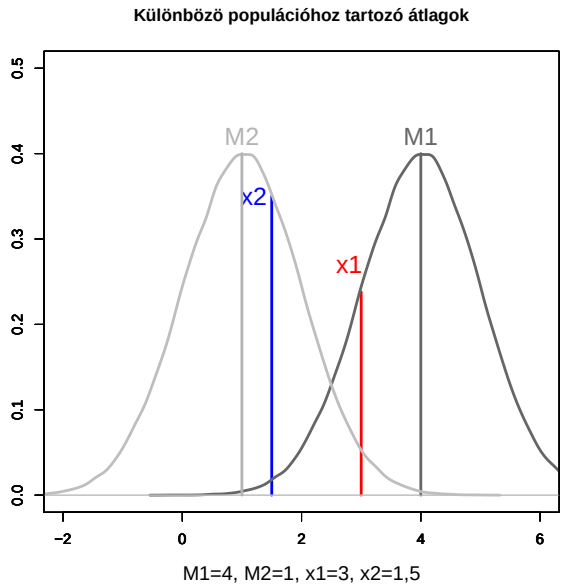
A középérték és variancia azonosságának próbái: t-próba, F-próba
Logikai vektorok az R-ben

Hipotézisállítás

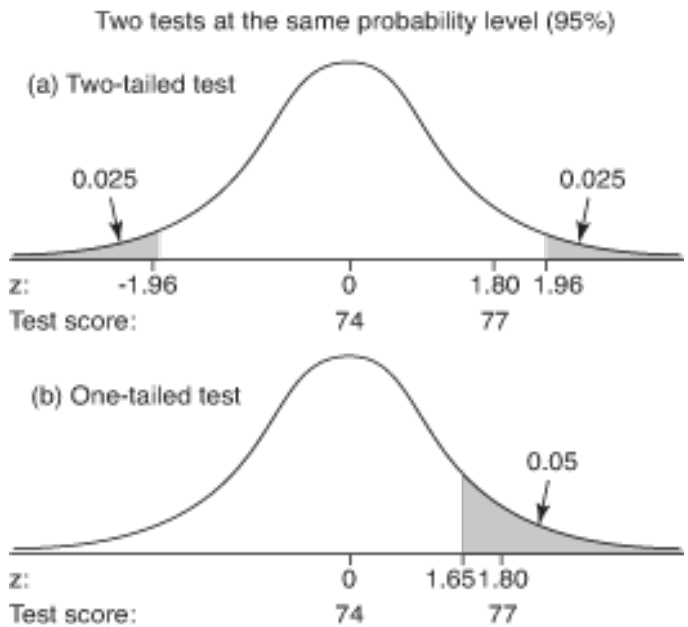
- Feltételezés: *a* minta egy adott szempont alapján más populációhoz tartozik, mint *b* minta.
- Nullhipotézis (H_0): *a* minta és *b* minta egyazon populációhoz tartozik, azaz az átlaguk ugyanazon μ populációátlag körül szór.
- Alternatív hipotézis (H_1): *p* valószínűséggel állítható, hogy *b* minta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





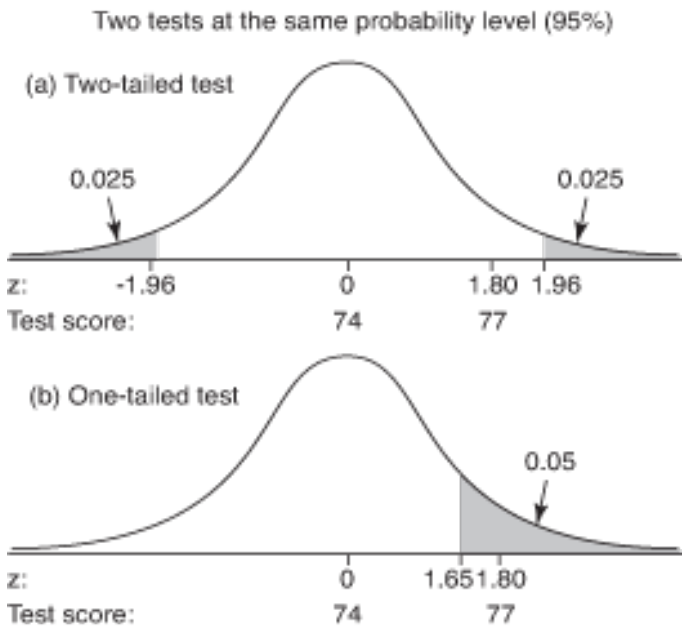
Hipotézis tesztelése $p = 95\%$-os megbízhatósággal

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).



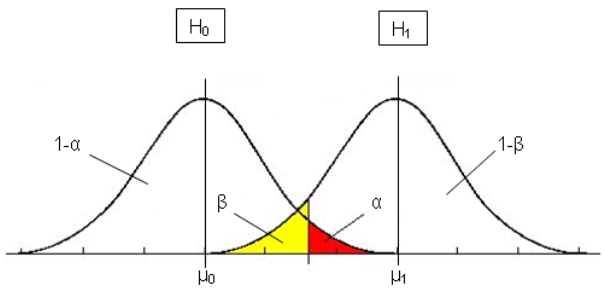
Hipotézis tesztelése $p = 95\%$-os megbízhatósággal

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).



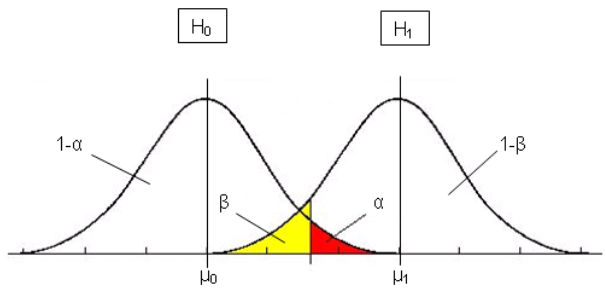
Hibatípusok

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.

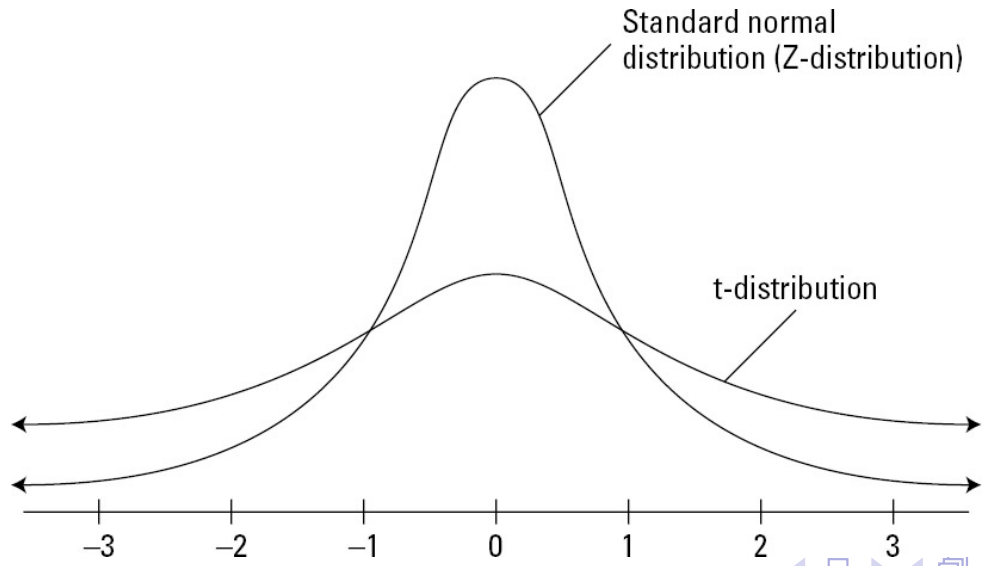
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

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

t-eloszlás

Jellemzők:

- Szimmetrikus, átlaga 0, aszimptotikus, de nem standard 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, mert a görbe egyre meredekebb a nagy elemszám miatt z-értékeket lehet használni.



Szabadsági fokok

Szabadsági fok, *degree of freedom, df*: a szabadon változtatható elemek száma, ami mellett a minta átlaga változatlan marad. Vagyis ha más mintát vennénk, hány olyan eleme van, ami eltérhet a mostani mintától?

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 átlaga változatlan marad. Vagyis ha más mintát vennénk, hány olyan eleme van, ami eltérhet a mostani mintától?

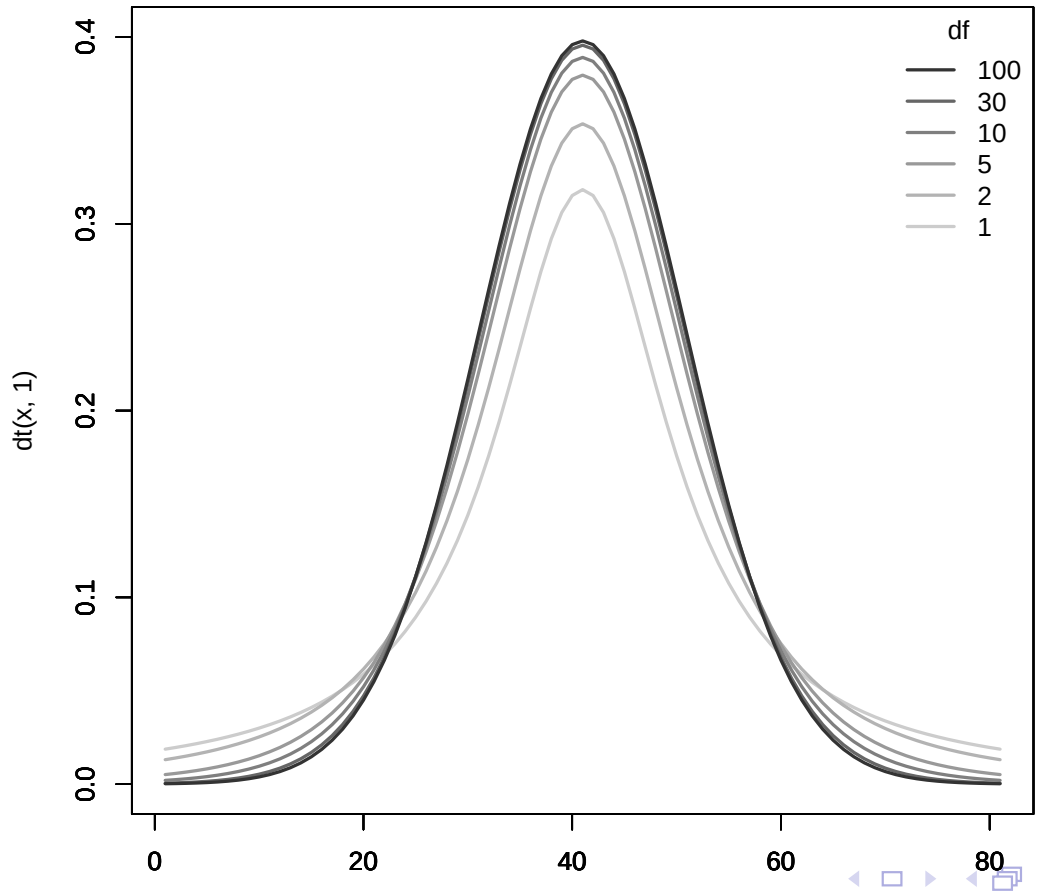
Pl. egy $n = 5$ elemű minta átlaga $\bar{x} = 10$. Hány elem változtatható szabadon a mintaátlag változatlansága mellett?

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. Vagyis annál nagyobb a nemkívánatos nullhipotézis elutasításának lehetősége.



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 15 (az IQ-teszt rögzített értékei). Okosabbak-e az oda járó gyerekek az átlagnál?

Feladat

Átlag: 108, populáció átlaga: 100, szórás: 15, elemek száma 60.

$$t_{minta} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{108-100}{15/\sqrt{60}} = \frac{8}{1{,}29} = 4,13$$

Kritikus értékhez tartozó t meghatározása ($p = 1 - \alpha = 0,95$): adott kvantilishez (0,95) tartozó t-érték 59-es szabadsági fok mellett kétoldali minta esetén, ahol a 0,05-ös α-érték felét vesszük figyelembe, mert megfelezzük a 0,05-ös valószínűséget:

`qt(p,df)`, itt: `qt(0.975,59)` $\rightarrow$ 2,000995 a t-eloszlás kritikus értéke pozitív és negatív irányban.

Az óvodásokra számolt t-érték ennél nagyobb, tehát kívül esik a populáció átlagához tartozó konfidenciaintervallumon. Vagyis 0,05-nél kisebb az esélye, hogy ehhez a populációhoz tartozik a mintánk. Ezért a nullhipotézist a meghatározott szignifikanciaszintet figyelembe véve elutasíthatjuk.

Feladat

Átlag: 108, populáció átlaga: 100, szórás: 15, elemek száma 60.

$$t_{minta} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{108-100}{15/\sqrt{60}} = \frac{8}{1{,}29} = 4,13$$

Kritikus értékhez tartozó t meghatározása ($p = 1 - \alpha = 0,95$): adott kvantilishez (0,95) tartozó t-érték 59-es szabadsági fok mellett kétoldali minta esetén, ahol a 0,05-ös α-érték felét vesszük figyelembe, mert megfelezzük a 0,05-ös valószínűséget:

`qt(p,df)`, itt: `qt(0.975,59)` $\rightarrow$ 2,000995 a t-eloszlás kritikus értéke pozitív és negatív irányban.

Az óvodásokra számolt t-érték ennél nagyobb, tehát kívül esik a populáció átlagához tartozó konfidenciaintervallumon. Vagyis 0,05-nél kisebb az esélye, hogy ehhez a populációhoz tartozik a mintánk. Ezért a nullhipotézist a meghatározott szignifikanciaszintet figyelembe véve elutasíthatjuk.

A Kincskereső óvodában tesztelt gyerekek tehát kortársaik átlagához képest szignifikánsan intelligensebbek.

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!
A varianciák azonosságát a `var.test()` függvénnyel ellenőrizhetjük

Welch-próba

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

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

Példa

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)
```

1. lépés: normális eloszlásúak-e a minták külön-külön?

```
shapiro.test(bika), shapiro.test(uszo)
```

Ha p 0,05-nél nagyobb, elfogadjuk a normális eloszlás feltételezését.

2. lépés: azonosak-e a varianciák?

```
var.test(bika,uszo)
```

Ha p 0,05-nél nagyobb, elfogadjuk a varianciák azonosságának feltételezését.

Két minta összehasonlítása t-próbával:

`t.test(bika,uszo)`

alapbeállítás: kétoldali (alternative=two.sided), varianciák nem egyenlőek (var.equal=FALSE).

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

A mintáinkban a varianciák azonosak, ezért használhatjuk a pontosabb, többnyire alacsonyabb p-értékhez vezető függvényt:
`t.test(bika,uszo,var.equal=TRUE)`

$p > 0,05 \Rightarrow$ a különbség nem szignifikáns egyik teszt esetében sem.

A p értéke azt mutatja, hogy mekkora valószínűséggel tévedünk, ha megtartjuk a nullhipotézist, vagyis azt, hogy a minták egyazon populációhoz tartoznak.

Kétmintás páros t-próba

A minta egyazon elem vagy összetartozó elemek kétszeri megfigyeléséből áll, például pulzus reggel és este egyazon személynél.

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 az összes mérés száma, mert az kétszer annyi).

Itt voltaképpen azt teszteljük, hogy igaz-e, hogy az egyazon elemen végzett két mérés különbségének átlaga páronként 0, vagyis a mérések nem különböznek szisztematikusan egyik irányban sem.

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.

Feladat

Töltsük le a trans.RData fájlt innen közvetlen eléréssel:
`load(url(phon.nytud.hu/mady/courses/statistics/materials/trans.RData))`

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$Class == "plant", ha karaktervváltozó
z = testmagassag$height < 170, ha numerikus változó

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) – ez segít a tengelyek szélső értékeinek meghatározásában.

Házi feladat

Végezzük el a megfelelő próbákat az `evszakok` adatmátrixra. Normális az eloszlása a tavaszra és az őszre adott pontszámoknak? Azonosak a varianciák? Ha az eloszlások normálisak, melyik t-próbát lehet elvégezni rajtuk?

Készítsünk pontdiagramot, amiben az évszakok külön-külön színnel vannak ábrázolva. Az x tengelyen a válaszadók kora, az y tengelyen az általuk adott pontszám szerepeljen. Van különbség a két évszakra adott pontszámokban a kor megoszlása szerint?