Seminar_5

Zoltan Kekecs

March 9, 2021

Ket valtozo kapcsolatanak vizsgalata, statisztikai inferencia

Az ora celja

Az ora celja hogy megismerkedjunk a **statisztikai inferencia alapjaival** ket valtozo kapcsolatanak elemzesen keresztul.

Package-ek betoltese

```
if (!require("tidyverse")) install.packages("tidyverse")
library(tidyverse) # for dplyr and ggplot2
```

Hipotezisteszteles

A statisztikai inferencia, es hipotezis teszteles soran az a celunk, hogy megallapitsuk, letezik-e egy bizonyos hatas vagy kapcsolat. De ezt a **null-hipotezis szignifikancia teszteles (NHST)** soran egy forditott logikaval tesszuk: azt allapitjuk meg, hogy **mekkora a valoszinusege hogy az altalunk megfigyelt adatot/trendet figyeljünk meg (vagy annal meg extremebb trendet), amennyiben a null-hipotezis igaz**.

Egy egyszeru pelda: az a sejtesem, hogy **egy penzerme cinkelt** (vagyis ki van sulyozva hogy az egyik oldalara nagyobb esellyel essen mint a masik oldalara), megpedig ugy hogy nagy valoszinuseggel **fej** legyen az eredmeny amikor feldobjuk. Ebben az esetben a **null-hipotezisem** az, hogy az **erme nem cinkelt**. Vagyis a null-hipotezis szerint ugyan akkora a valoszinusege fejet es irast kapni eredmenykent.

- H1: cinkelt erme (fej fele)
- H0: nem cinkelt erme

Tegyuk fel hogy 10-szer feldobjuk az ermet, es 9-szer fejet dobunk. Mekkora a valoszinusege, hogy az erme cinkelt? Ezt nem tudjuk megmondani. Tobbek kozott azert sem mert nem tudjuk, mennyire lehet cinkelve. Viszont azt meg tudjuk mondani, hogy mekkora a valoszinusege, hogy ezt az eredmenyt kapnank, ha az erme **NINCS** cinkelve.

Annak a valoszinusege, hogy **legalabb 9-szer** (vagy tobbszor) fejet dobok **10 dobasbol** egy nem cinkelt ermevel, p = 0.0107 (**nagyjabol 1%**). (Ezt a kod reszt nem fontos megerteni, a lenyege hogy a pbinom() funkcioval kiszamoltuk a valoszinuseget, hogy 10 feldobasbol legalabb 9 fej lesz).

```
probability_of_heads_if_H0_is_true <- 0.5

heads <- 9
total_flips <- 10
probability_of_result = 1-pbinom(heads-1, total_flips, probability_of_heads_if_H0_is_true)

probability_of_result
```

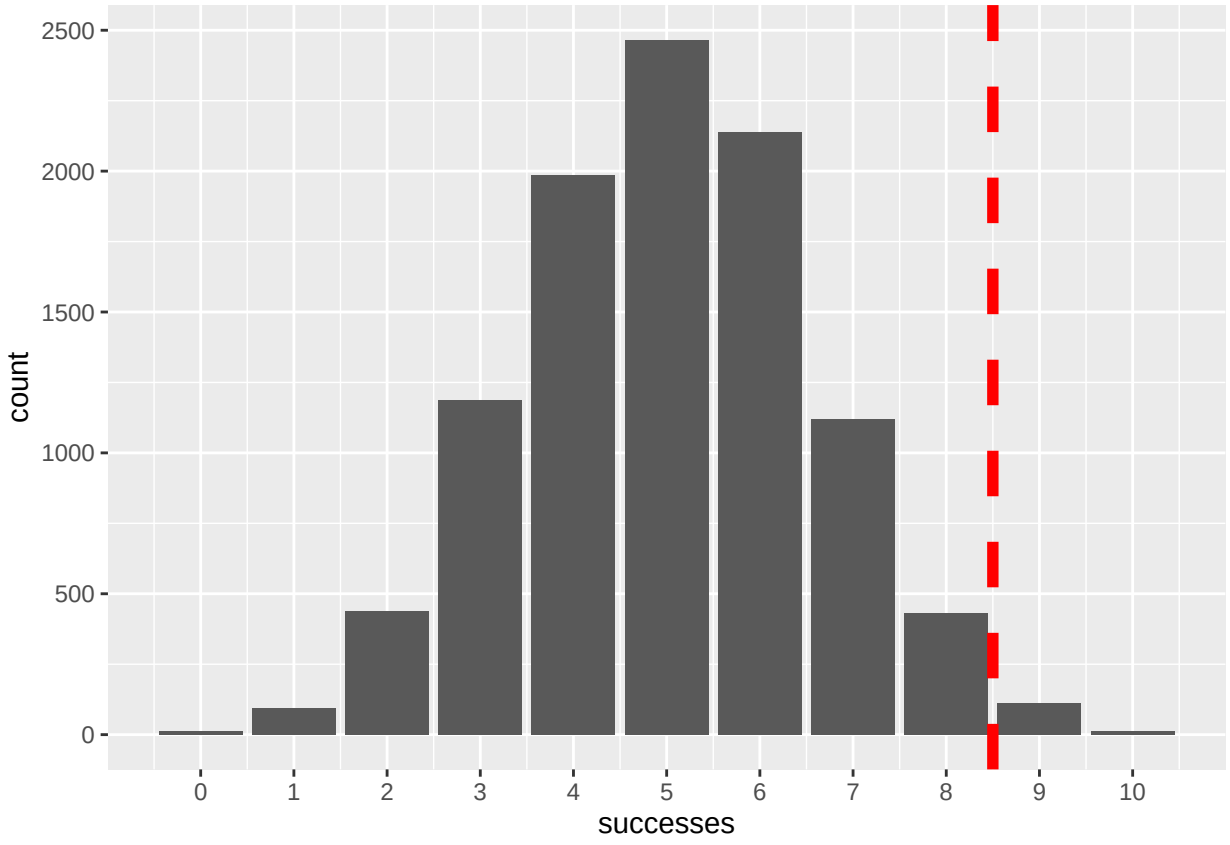
```
## [1] 0.01074219
```

Ez a valoszinuseg **maskepp mondva** azt jelenti, hogy ha ugyan ezt a kiserletet 100 szor megismetelnenk (mindegyikben 10 feldobassal), akkor a 100 kiserletbol csak atlagosan nagyjabol 1-szer varnanak, hogy 9 vagy tobb fejet kapjunk.

Ezt le is ellenorizhetjuk, ha **randomizalunk 10.000 hasonlo kiserletet** az rbinom() funkcioval. Az abran lathato hogy csak a kiserletek igen kis szazalekaban kaptunk 9 vagy tobb "sikert". (Ezt a kodreszt sem fontos megerteni, a lenyege hogy az rbinom() funkcioval 10000-szer azt szimulaltuk, hogy egymas utan 10-szer feldobtunk egy ermet (vagyis hogy veletlenszeruen valasztottunk egy szamot 0 es 1 kozul), ez utan ennek a 10000 kiserletnek az eredmenyet abrazoltuk a ggplot-tal)

```
successes = rbinom(n = 10000, size = 10, prob = 0.5)
random_flips = data.frame(successes)

ggplot(data = random_flips) +
  aes(x = successes) +
  geom_bar()+
  scale_x_continuous(breaks = 0:10)+
  geom_vline(xintercept = 8.5, col = "red", linetype = "dashed", size = 2)
```



Vagyis a 9 fej 10 feldobasbol egy **eleg meglepo** (nagyon ritka) eredmeny, hiszen ez csak az esetek nagyjabol 1%-aban fordul elo ha az erme nem cinkelt. De mit mond ez nekunk arrol hogy **az erme valoban cinkelve van-e** vagy sem? Mekkora ennek az eselye? Ezt sajnos nem tudjuk meg. Amit megtudunk ebből a szamitasbol, az az, hogy **milyen ritka ez az eredmeny amit kaptunk ha azt feltetelezzuk hogy az erme nincs cinkelve**. Ezt a fajta forditott logikat kell megerteni ahhoz, hogy az NHST-t teljesen meg tudjuk erteni.

Tegyuk fel hogy egy **"igen-vagy-nem" dontest** kell hoznunk arrol, hogy cinkelt-e az erme vagy sem.

Mondjuk minket biztak meg hogy ellenorizzuk az ermet egy fontos penzfeldobas elott, es el kell dontenunk, hogy megbizunk-e ennek az ermenek a hitelessegeben, vagy kerjunk egy uj ermet a penzfeldobashoz, mert ezt cinkeltnek iteljuk. Itt jon az NHST **teszt** resze. Ezt a dontest az NHST-ben egy elore megahatarozott valoszinusegi kuszubertek, **dontesi kuszobertek**, figyelembevetelevel hozzuk meg. Ha az altalunk megfigyelt eredmeny **kelloen meglepo**, **kelloen ritka** a null hipotezis helyesseget feltetelezve, akkor elvetjuk azt a feltetelezest, hogy a null-hipotezis helyes. Ilyenkor kizarasos alapon az alternativ hipotezis helyesseget fogadjuk el.

A pszichologia tudomanyaban a dontesi kuszobertek tradicionalisan 5%, vagyis ha annak a valoszinusege hogy az altalunk megfigyelt eredmenyt (vagy annal extremebb eredmenyt) kapjunk a null hipotezis helyessege eseten **kisebb mint 5%** ($p < 0.05$), akkor **elvetjuk a null-hipotezist**.

Fontos azonban hangsulyozni, hogy egy-egy NHTS teszt soran nem tudjuk meg a null hipotezis helyessegenek, vagy az alternativ hipotezis helyessegenek a valodi valoszinuseget. Csak azt tudjuk, hogy mennyire valoszinu vagy valoszinutlen hogy az altalunk megfigyelt eredmenyt latjuk "egy olyan vilagban" ahol a null hipotezis helyes. Se tobbet, se kevesebbet. Es ez alapjan hozzuk meg a dontesunket a null-hipotezis elveteserol, vagy megtartasarol.

Az NHST modszer fo elonye, hogy ha **konzisztensen hasznaljuk a fent emlitett dontesi kuszobot** a kutatasainkban, akkor **elegge biztosak** lehetunk abban, hogy a statisztikai donteseinknek **csak 5%-aban vetjuk el hibasan a null hipotezist**. Vagyis a statisztikai donteseknek csak 5%-a lesz hibas, ha a null hipotezis valojaban igaz, igy tehet az elsofaju hiba (alpha-error) valoszinusege 5%. (Masszoval csak a teszteknek csak 5%-aban allitjuk hibasan, hogy van hatas, amikor valojaban nincs hatas.)

Ket fontos kitetelt erdemes megfigyelni a fenti allitasban. Egyreszt hogy azt irtam hogy "elegge biztosak" lehetunk. Azert csak "elegge biztosak" lehetunk ebben, es nem teljesen biztosak, mert ahhoz hogy ez az allitas helyes legyen, az altalunk hasznalt statisztikai tesztek **elofelteveseinek teljesulnie kell**, es ebben nem lehetunk teljesen biztosak a populacio szintjen. A masik, hogy **"ha a null hipotezis valojaban igaz"**. Arrol az NHST-ben nem kapunk garanciat, hogy a statisztikai donteseinknek hany szazaleka hibas ha az alternativ hipotezis az igaz. Azt is fontos megerteni, hogy ez nem jelenti azt, hogy az osszes publikalt null hipotezis-tesztelesben csak 5%-nyi lenne az elsofaju hiba, mert nem minden statisztikai dontest publikalnak.

Statisztikai tesztek

Nem kell jonak lennunk valoszinusegszamitasbol hogy jo statisztikai donteseket tudjunk hozni. A megfigyeles valoszinuseget a null-hipotezis helyesseget feltetelezve altalaban egy **statisztikai teszt** mondja meg nekunk. Ezen az oran 5 statisztikai tesztet fogunk megismerni.

- binomialis teszt
- khi-negyzet teszt
- t-teszt
- egyszempontos ANOVA
- korrelacios teszt

binomialis teszt

A hipotezist, hogy az erme cinkelt, tesztelhetjuk a **binomialis teszttel**, aminek R-ben binom.test() a funkcioja. Az x helyere a megfigyelt "celmegfigyelesek" vagy "sikerek" szamat (a mi esetunkben a fejek szamat, $x = 9$), az n helyere az osszes megfigyeles szamat ($n = 10$), a p helyere pedig a **null-hipotezes** helyesseget feltetelezve a "celmegfigyelesek" eleresenek valoszinuseget kell beirni (mivel a hipotezisunk az hogy az erme cinkelt, az null hipotezisunk az, hogy az erme "nem cinkelt"). Ezt valoszinusegkent kell megadni, amit egy 0 es 1 kozotti szammal jellemezhetunk (ahol a 0 azt jelenti hogy a megfigyelesek 0%-a lesz "siker", az 1 pedig azt hogy a megfigyelesek 100%-a lesz "siker", vagyis a 0.6 jelentese hogy a megfigyelesek 60%-a lesz "siker"). A mi esetunkben a null hipotezis helyessege eseten a fej valoszinusege 50% ($p = 0.5$).

```
binom.test(x = 9, n = 10, p = 0.5, alternative = "greater")
```

```
##
```

```
##  Exact binomial test
##
## data:  9 and 10
## number of successes = 9, number of trials = 10, p-value = 0.01074
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.6058367 1.0000000
## sample estimates:
## probability of success
##                    0.9
```

Ennek a tesztnek az eredmenye a kovetkezot mutatja:

- p-value: p-ertek, annak a valoszinusege, hogy az altalunk megfigyelt, vagy extremebb eredmenyt kapunk, feltetelezve hogy a null-hipotezis helyes. Altalaban ha ez az ertek 0.05 alatti, akkor elvetjuk a null-hipotezist.
- alternative hypothesis: Itt irja le, hogy mi volt a H1, ami a mi esetunkben az volt, hogy a fej valoszinusege nagyobb mint 0.5 (50%). Ez egyben azt is jelenti, hogy a null-hipotezisunk az volt, hogy a fej valoszinusege 0.5.
- 95 percent confidence interval (vagy roviden 95% CI): a 95%-os konfidencia intervallum. Ez azt jelenti, hogy ha a kiserletet sokszor megismeteljuk es ugyan igy kiszamoljuk a konfidencia intervallumot minden kiserletnel, az igy kapott konfidencia intervallumok 95%-a tartalmazni fogja a valos hatasmeretet (ami a mi esetunkben a "siker"/fej valoszinusege). Fontos, hogy nem tudjuk, hogy a mi konkrét kiserletunkban a konfidencia intervallum tartalmazza-e a valos hatasmeretet.
- sample estimates: A "siker" ("celmegfigyeles", a mi esetunkben a fej) valoszinusegenek becsult merteke a populacioban a megfigyelt valoszinuseg alapjan. Ez egy pontbecsles, ami mindig megegyezik a megfigyelt valoszinuseggel.

Az eredmenyt igy irhatjuk le:

"A kutatasunkban 9 fejet figyeltunk meg 10 penzfeldobasbol (90%). Ez alapjan ugy iteljuk, hogy annak a valoszinusege, hogy fejet dobunk az ermevel szignifikansan tobb mint 50%. A fej dobas valoszinusege 0.9 volt a mintaban (95% CI = 0.61, 1)."

Adatgeneralas az orahoz

Az alabbi kod **adatokat general** a szamunkra. Az adatgeneralashoz hasznlalt kod megertese ezen a szinten meg nem szukseges.

```
n_per_group = 40

base_height_mean = 164
base_height_sd = 10
base_anxiety_mean = 18
base_anxiety_sd = 2
resilience_mean = 7
resilience_sd = 2

treatment_effect = - 3
resilience_effect = - 0.8

gender_bias = 0.7
gender_effect = - 1
gender_effect_on_height = 12


treatment <- rep(c(1, 0), each = n_per_group)
```

```r
set.seed(1)

gender_num <- rbinom(n = n_per_group * 2, size = 1, prob = 0.7)
gender <- NA
gender[gender_num == 0] = "female"
gender[gender_num == 1] = "male"

set.seed(2)
home_ownership <- sample(c("own", "rent", "friend"), n_per_group * 2, replace = T)

set.seed(3)
resilience <- rnorm(mean = resilience_mean, sd = resilience_sd, n = n_per_group*2)

set.seed(6)
anxiety_base <- rnorm(mean = base_anxiety_mean, sd = base_anxiety_sd, n = n_per_group*2)
anxiety_baseline <- anxiety_base + resilience * resilience_effect + gender_num * gender_effect + rnorm(n
anxiety_post <- anxiety_base + treatment * treatment_effect + resilience * resilience_effect + gender_nu
participant_ID <- paste0("ID_", 1:(n_per_group*2))

set.seed(5)
height_base <- rnorm(mean = base_height_mean, sd = base_height_sd, n = n_per_group*2)
height <- height_base + gender_num * gender_effect_on_height


group <- rep(NA, n_per_group*2)
group[treatment == 0] = "control"
group[treatment == 1] = "treatment"

health_status <- rep(NA, n_per_group*2)
health_status[anxiety_post < 11] = "cured"
health_status[anxiety_post >= 11] = "anxious"

data <- data.frame(participant_ID)
data = cbind(data, gender, group, resilience, anxiety_baseline, anxiety_post, health_status, home_owners
data = as_tibble(data)

data = data %>%
  mutate(gender = factor(gender))

data = data %>%
  mutate(group = factor(group))

data = data %>%
  mutate(health_status = factor(health_status))

data = data %>%
  mutate(home_ownership = factor(home_ownership),
         anxiety_baseline = round(anxiety_baseline, 2),
         anxiety_post = round(anxiety_post, 2),
         resilience = round(resilience, 2),
         height = round(height, 2))
```

Az adatok egy (kepzeletbeli) randomizalt kontrollalt klinikai kutatas eredmenyeibol szarmaznak, ahol a **pszichoterapia hatekonysagat** teszteltek. Olyan szemelyeket vontak be a kutatasba, akik egy **hurrikan**

aldozatai voltak, es **szorongassal** kuszkodtek. A szemelyeknel felmertek a reziliancia (psziches ellenallokepesseg) szintjet, majd veletlenszeruen osztottak a szemelyeket egy kezelesi vagy egy kontrol csoportba. Ezt kovetoen a kezelesi csoport **pszichoterapiat kapott 6 heten keresztul** heti egyszer, mig a kontrol csoport nem kapott kezelest. A vizsgalat vegen megmertek a szemelyek **szorongasszintjet**, es a klinikai kriteriumok alapjan meghataroztak, hogy a szemely **gyogyultnak, vagy szorongonak** szamit-e.

Lathatjuk, hogy 8 valtozo van az adattablaban.

- participant_ID - reszvevo azonositoja
- gender - nem
- group - csoporttagsag, ez egy faktor valtozo aminek ket szintje van: "treatment" (kezelt csoport), es "control" (kontrol csoport). A "treatment" csoport kapott kezelest, mig a "control" csoport nem kapott kezelest.
- resilience - reziliancia: a nehezsegekkel valo megkuzdes kepessege, ez egy szemlyes kepesseg, olyasmi mint a szemelyisegvonasok
- anxiety_baseline - szorongas szint a terapia elott
- anxiety_post - szorongas szint a terapia utan
- health_status - a klinikai kriteriumok alapjan szorongonak vagy gyogyultnak tekintheto a szemely
- home_ownership - lakhatasi helyzet: harom szintje van az alapjan hogy a szemely hol lakik: "friend" - baratnal vagy csaladnal lakik, "own" - sajat tulajdonu lakasban lakik, "rent" - berelt lakasban lakik,
- height - magassag

Adatellenorzes

Mint mindig, elemzes elott **ellenorizzuk**, hogy az adattal minden rendben van-e!

```
data
```

```
## # A tibble: 80 x 9
##    participant_ID gender group resilience anxiety_baseline anxiety_post
##    <chr>          <fct>  <fct>      <dbl>            <dbl>        <dbl>
##  1 ID_1           male   trea~       5.08            18.7         10.5
##  2 ID_2           male   trea~       6.41            10.5          7.61
##  3 ID_3           male   trea~       7.52            17.2          9.72
##  4 ID_4           female trea~       4.7             17.7         14.7
##  5 ID_5           male   trea~       7.39             8.04         8.14
##  6 ID_6           female trea~       7.06            10.3         10.1
##  7 ID_7           female trea~       7.17            12.6          6.64
##  8 ID_8           male   trea~       9.23            10.6          8.09
##  9 ID_9           male   trea~       4.56            12.8         10.4
## 10 ID_10          male   trea~       9.53             5.82         4.28
## # ... with 70 more rows, and 3 more variables: health_status <fct>,
## #   home_ownership <fct>, height <dbl>
```

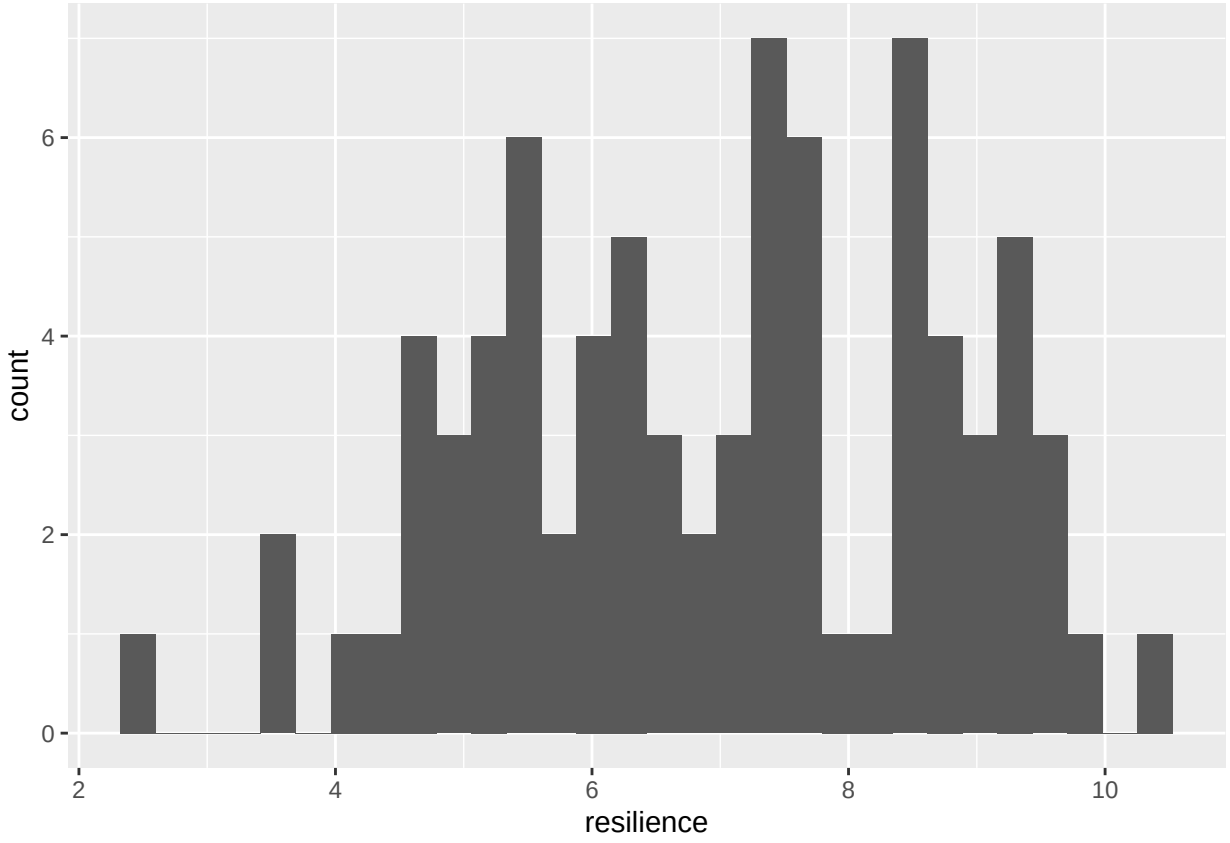
```
data %>%
  summary()
```

```
##  participant_ID        gender         group      resilience
##  Length:80          female:25   control  :40   Min.   : 2.470
##  Class :character   male  :55   treatment:40   1st Qu.: 5.518
##  Mode  :character                                Median : 7.125
##                                                  Mean   : 6.981
##                                                  3rd Qu.: 8.477
##                                                  Max.   :10.400
##  anxiety_baseline  anxiety_post    health_status home_ownership     height
##  Min.   : 4.650   Min.   : 3.910   anxious:32    friend:22      Min.   :142.2
##  1st Qu.: 9.668   1st Qu.: 8.223   cured  :48    own   :31      1st Qu.:163.4
```

```
##  Median :11.155   Median :10.110                rent  :27      Median :173.0
##  Mean   :11.393   Mean   :10.212                              Mean   :172.3
##  3rd Qu.:12.730   3rd Qu.:12.255                              3rd Qu.:179.7
##  Max.   :19.320   Max.   :16.710                              Max.   :198.2
```

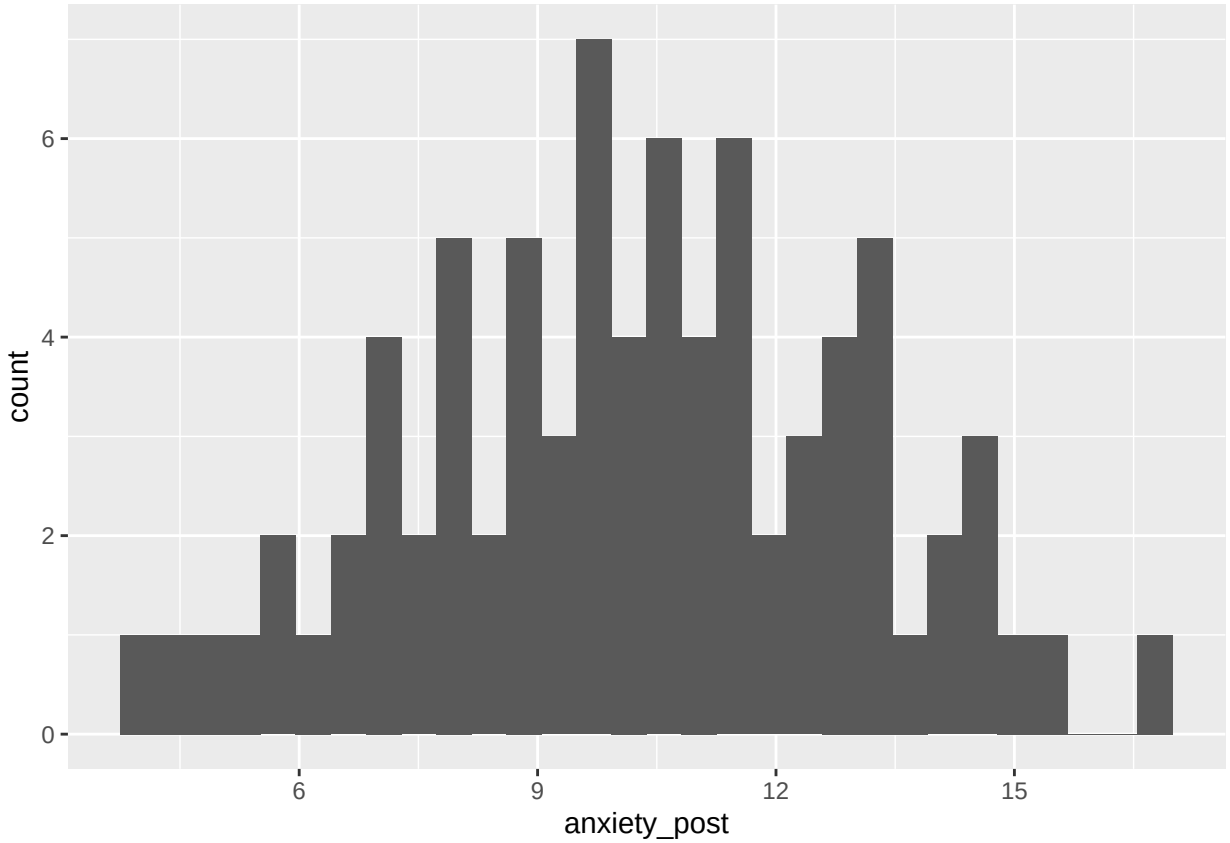
```
data %>%
  ggplot() +
    aes(x = resilience) +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

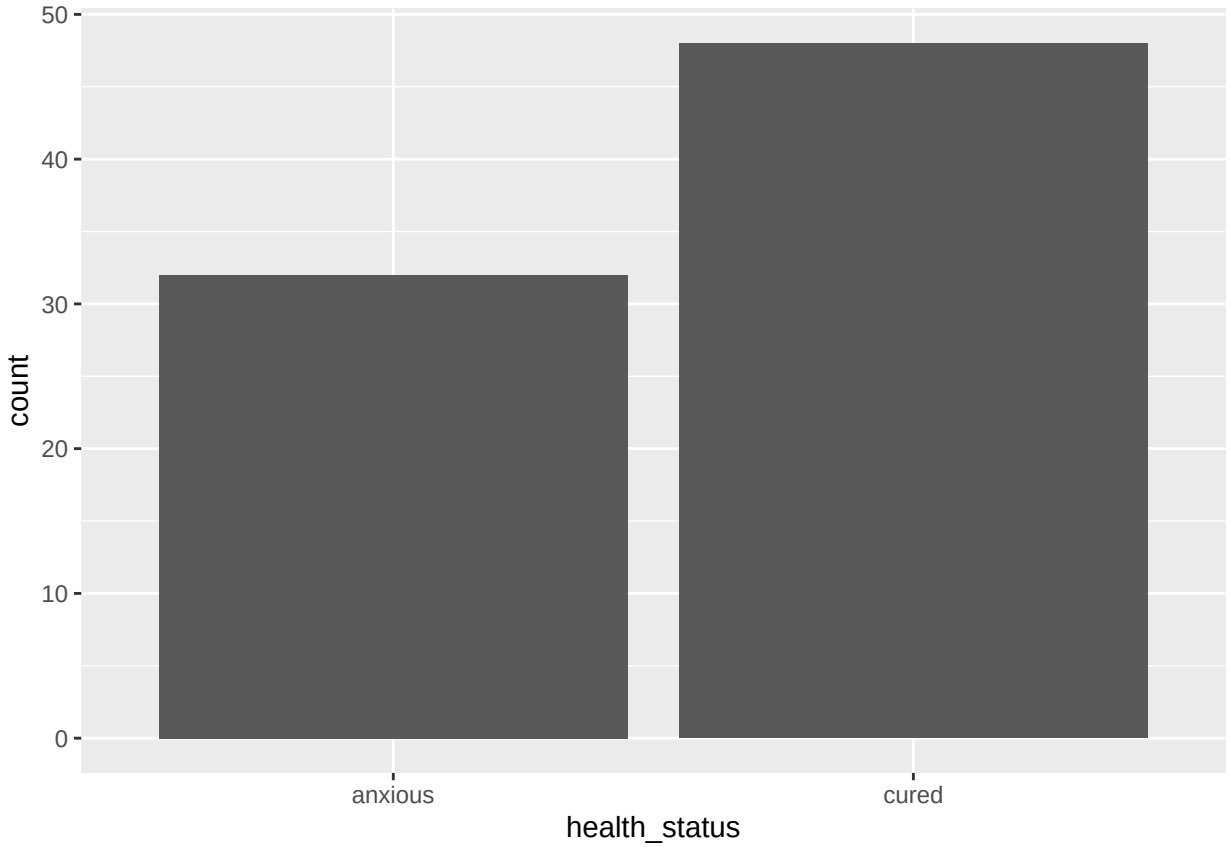



```
data %>%
  ggplot() +
    aes(x = anxiety_post) +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
data %>%
  ggplot() +
    aes(x = health_status) +
    geom_bar()
```




```
set.seed(Sys.time())
```

Hipotezisek

Vizsgaljuk meg kutatasban szereplo valtozok osszefuggeset a hipotezisek menten.

A kutatas hipotezise a kovetkezok voltak:

1. Tobb a ferfi mint a no ebben a klinikai mintaban (**gender** vs. 50%).
2. A pszichoterapiat kapo csoportban a terapia utan kevesebb lesz a klinikai kriteriumok alapjan szorongonak szamito szemely (**health_status** vs. **group**)
3. A terapias csoportban alacsonyabb lesz a szorongas atlaga a kutatas vegere mint a kontrol csoportban (**anxiety_post** vs. **group**)
4. A reziliancia es a kutatas vegen mert szorongasszint negativ osszefuggest fog mutatni (vagyis aki reziliensebb, annal alacsonyabb szorongasszintet fognak merni a kutatas vegen) (**anxiety_post** vs. **resilience**)

Gyakorlas

Teszteld a hipotezist, hogy "Tobb a ferfi mint a no ebben a klinikai mintaban" (**gender** valtozo)

- Ezt ugyan ugy teheted meg, mint a fenti peldaban, hiszen a null-hipotezis az, hogy a ferfiak ("male") elvart valoszinusege 50% vagy kevesebb (p = 0.5). Szoval a ferfiak ekvivalensek a "fejekkel" a penzfeldobasos peldaban.
- Meg kell hataroznod a ferfiak szamat a mintaban, es a teljes mintaelemszamot, hogy ki tudd tolteni a binom.test() fuggveny parametereit.
- Ez utan vegezd el a tesztet

- Es ird le a fentiek szerint az eredmenyeket.

Ket kategorikus valtozo kapcsolata: Khi-negyzet proba (Chi-squared test)

Ket kategorikus valtozo kapcsolatanak vizsgalatara tobbfajta teszt is alkalmazhato. Olyan esetben, ahol mind a ket kategorikus valtozonak ket-ket szintje van csak (vagyis a kulonbozo kategoria-szint kombinaciok egy 2x2-es tablazatban abrazolhatoak) a Fisher tesztet erdemes alkalmazni vagy a likelihood ratio tesztet, de olyan tablazatokra ahol a kategoriak kombinacioja tobb mint 2x2-es tablazatot alkot, vagyis ahol az egyik csoportosito valtozonak tobb mint 2 szintje van, a **Khi-negyzet proba** javasolt.

Peldaul megvizsgalhatjuk, hogy van-e kapcsolat abban, hogy a szemelyek lakhatasi helyzete (**home_ownership**) es a kozott, hogy a kutatas vegen az egyes szemelyek meggyogyultak-e (**health_status**).

A Khi-negyzet proba elofeltetelei:

- Minden megfigyeles fuggetlen a tobbi megfigyelestol (pl. egy megfigyeles szemelyenkent)
- A kategoria-kombinaciok abrazolasaval kapott tablazatban nem tobb mint a cellak 20%-aban kisebb a varhato ertek 5-nel, es minden cellaban magasabb a varhato ertek mint 1.

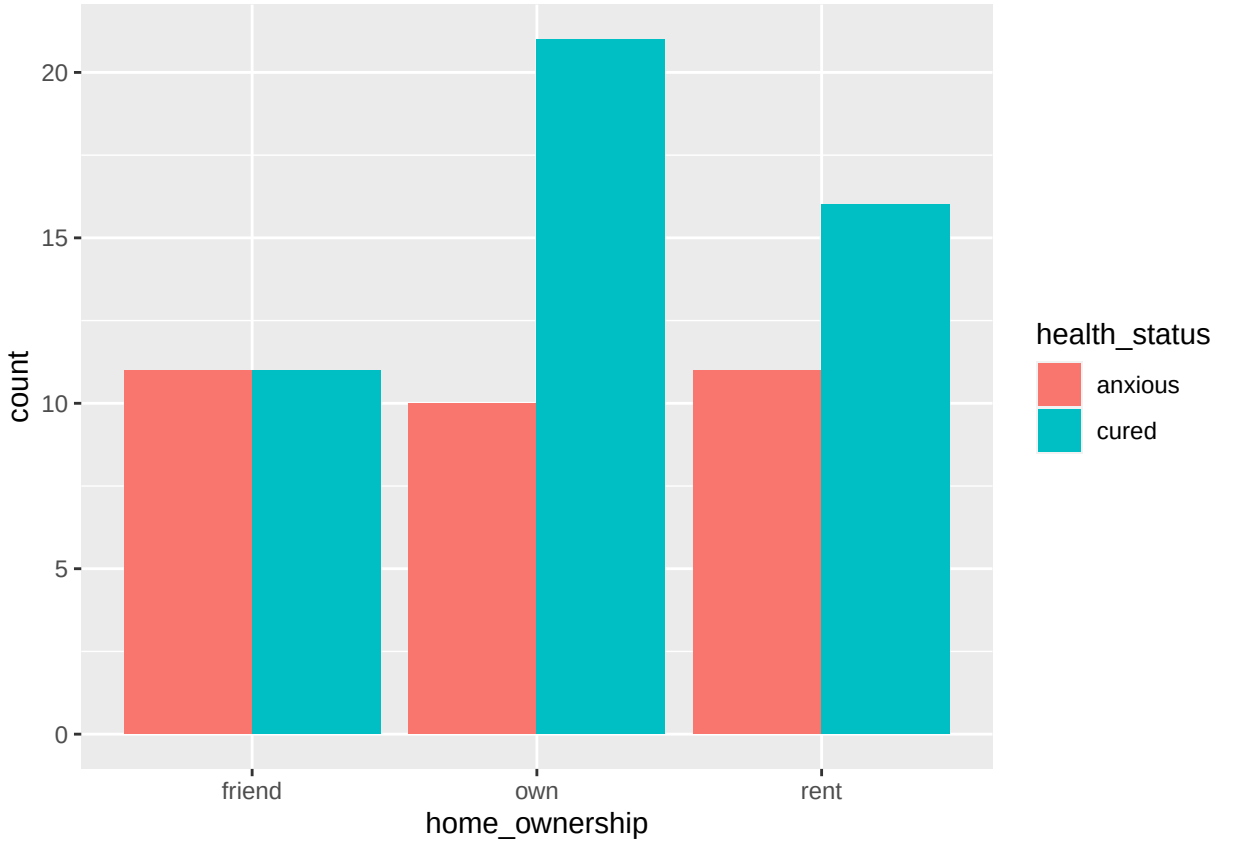
Eloszor feltaro elemzest vegzunk:

- tablazatot rajzolunk a ket valtozo kapcsolatarol
- abrat keszitunk (pl. geom_bar)

```
table(data$home_ownership, data$health_status)
```

```
##
##          anxious cured
##   friend      11    11
##   own         10    21
##   rent        11    16
```

```
data %>%
  ggplot() +
    aes(x = home_ownership, fill = health_status) +
    geom_bar(position = "dodge")
```



Ez utan elvegezzuk a Khi-negyzet probat. Ehhez eloszor keszitenunk kell egy **tablazatot a ket valtozo kapcsolatarol**, amit egy uj objektumban elmentunk.

A Khi-negyzet proba azt a **null-hipotezist** teszteli, hogy **a csoportokban ugyan olyan a masik kategorikus valtozo eloszlasa** (vagyis a mi esetunkben a null hipotezis hogy ugyan olyan aranyban gyogyulnak meg akik baratnal laknak, akiknak sajat lakasuk van, es akik berlik a lakast).

```
ownership_health_status_table = table(data$home_ownership, data$health_status)
ownership_health_status_table
```

```
##
##          anxious cured
##   friend      11    11
##   own         10    21
##   rent        11    16
```

```
chisq.test(ownership_health_status_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ownership_health_status_table
## X-squared = 1.697, df = 2, p-value = 0.428
```

Az eredmenyt igy irhatjuk le:

“Nem volt szignifikans elteres abban, hogy a kulonbozo lakhatasi csoportokban (baratnal, sajat lakasban, vagy berlemenyben lakok) milyen aranyban voltak azok akik meggyogyultak a kutatas vegere (X^2 = 1.7, df = 2, p = 0.428).”

Gyakorlas

Teszteld a 2. hipotezist, hogy "A pszichoterapiat kapo csoportban a terapia utan kevesebb lesz a klinikai kriteriumok alapjan szorongonak szamito szemely" (**health_status** vs. **group**)

- Ezt ugyan ugy teheted meg, mint a fenti peldaban, hiszen a null-hipotezis az, hogy nincs kulonbseg a csoporttagsag szerint (treatment vs. control) abban hogy milyen aranyban gyogyultak meg a kutatas vegere.
- Eloszor vegezzunk egy feltaro elemzest egy tablazattal a ket valtozo kapcsolatarol a table() funkcioval es egy abraval (mondjuk geom_bar() hasznalataval)
- A tablazatot mentsd el egy uj objektumba
- Ez utan vegezd el a tesztet, chisq.test()
- Es ird le a fentiek szerint az eredmenyeket.

Egy numerikus valtozo atlaganak kolunbsege csoportok kozott: anova es t-teszt

Tesztelhetjuk peldaul, hogy van-e kulonbseg a nemek kozott (**gender**) a kutatas vegen mert szorongas szintjeben (**anxiety_post**).

Eloszor szokas szerint feltaro elemzest vegzunk atlagok csoportonkenti osszehasonlitasaval es abraval. Erre pl. remek a geom_boxplot() es a geom_density()

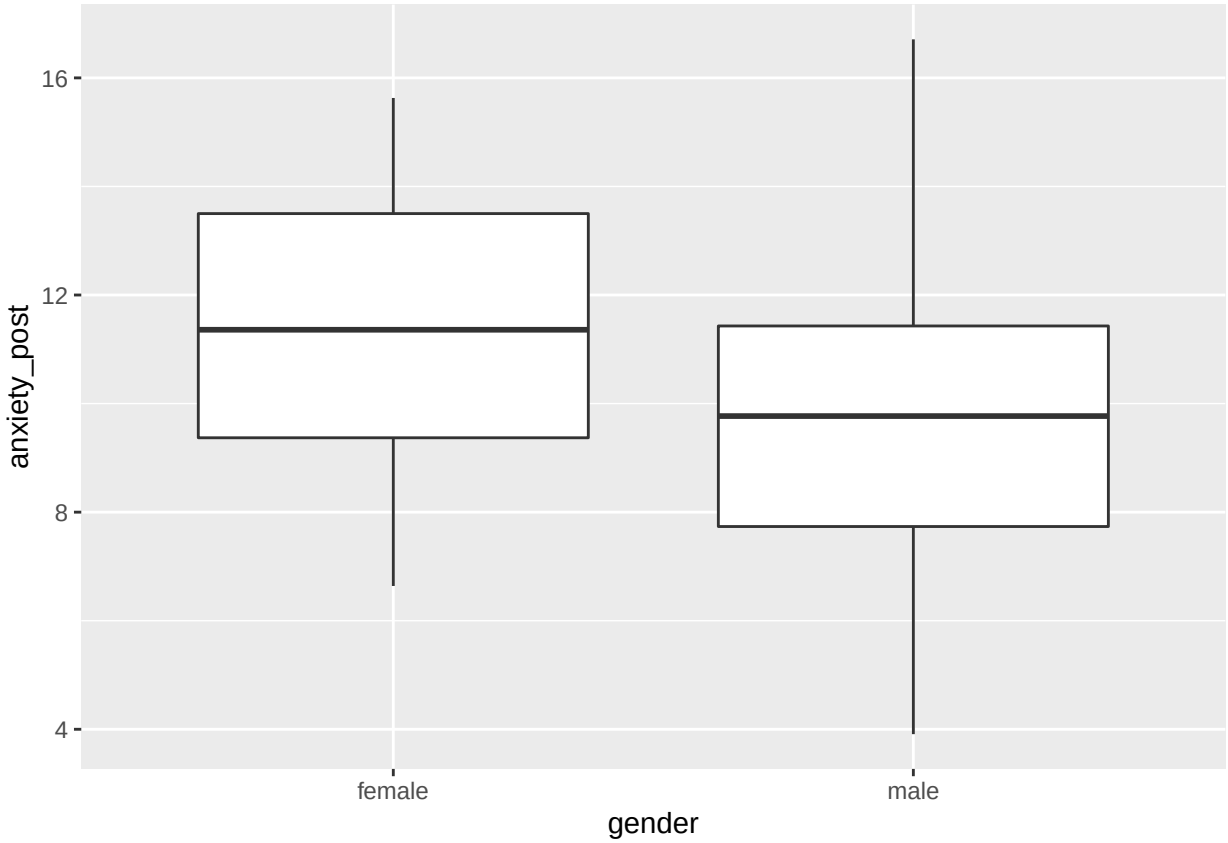
```
summary = data %>%
  group_by(gender) %>%
    summarize(mean = mean(anxiety_post), sd = sd(anxiety_post))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

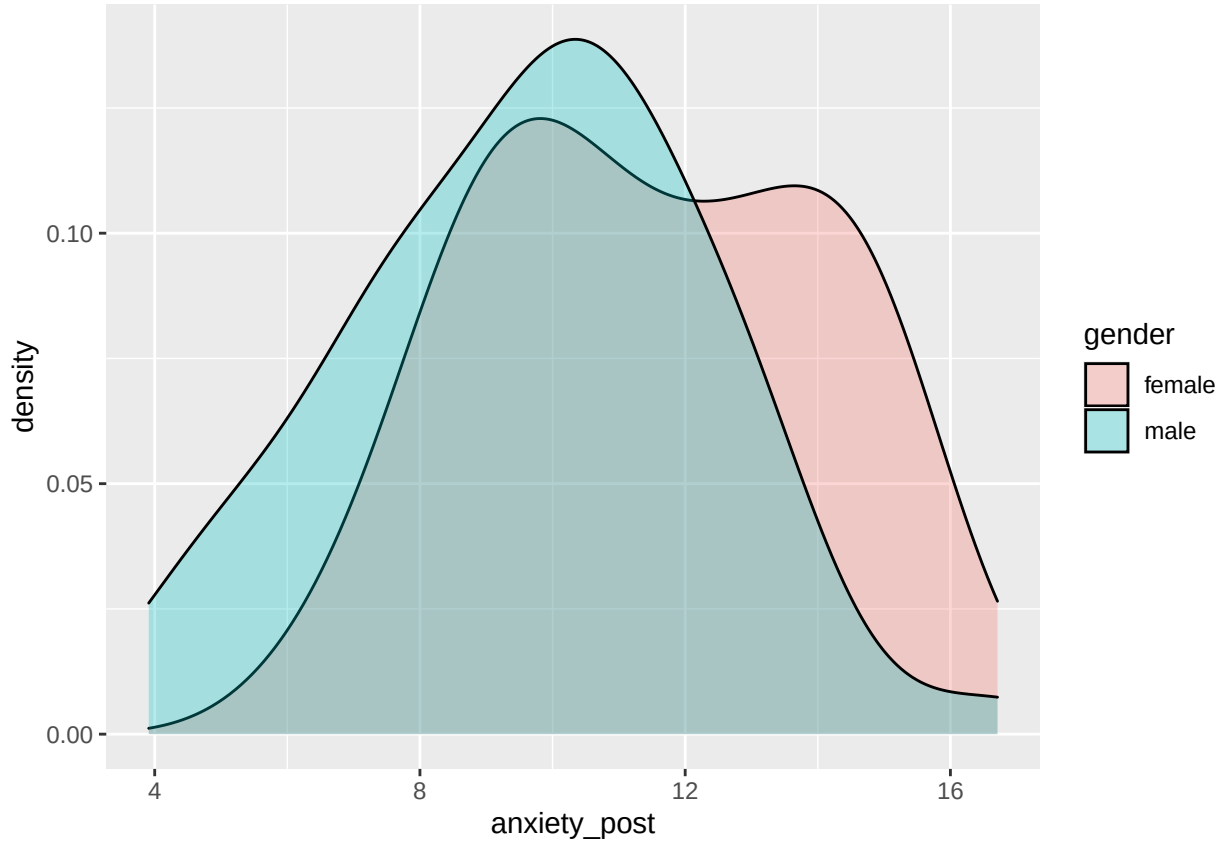
```
summary
```

```
## # A tibble: 2 x 3
##   gender  mean    sd
##   <fct>  <dbl> <dbl>
## 1 female 11.5   2.58
## 2 male    9.64  2.70
```

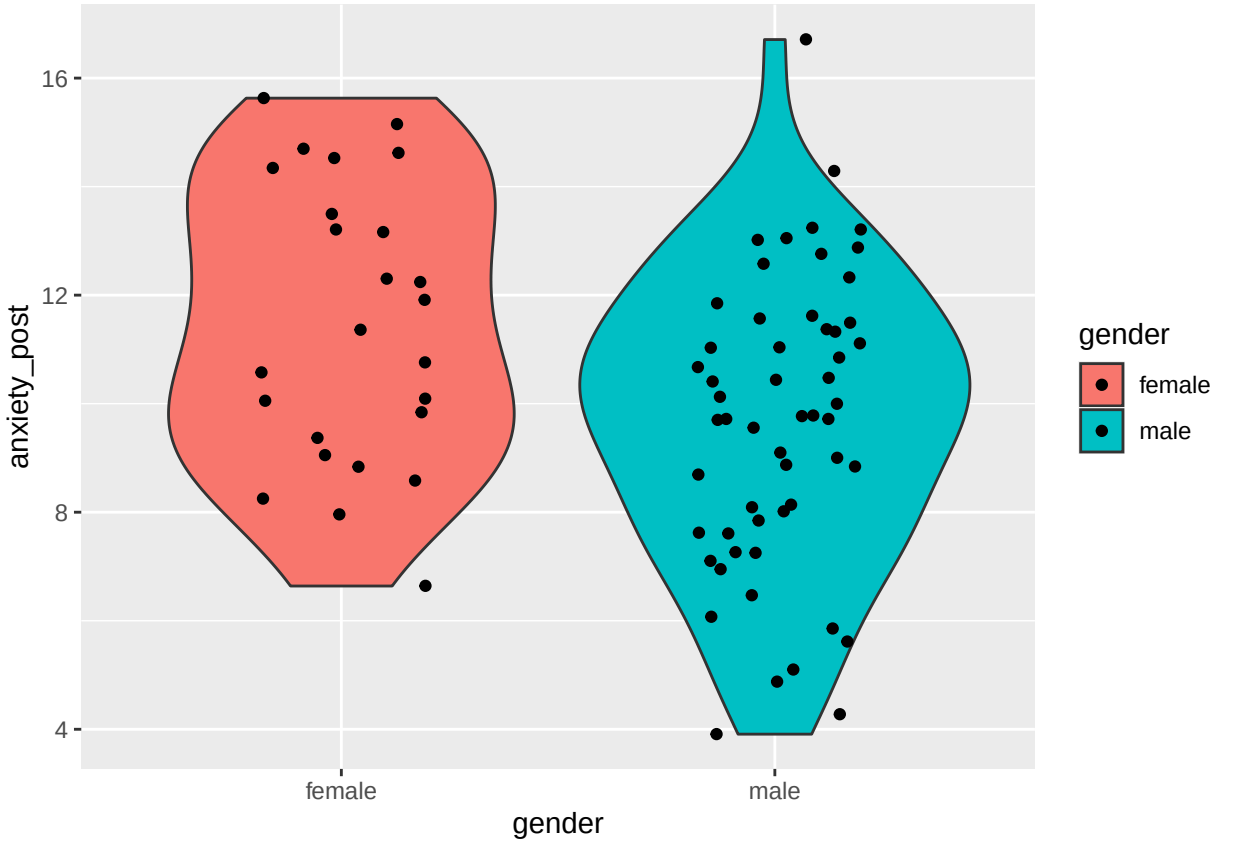
```
data %>%
  ggplot() +
    aes(x = gender, y = anxiety_post) +
    geom_boxplot()
```




```
data %>%
  ggplot() +
    aes(x = anxiety_post, fill = gender) +
    geom_density(alpha = 0.3)
```




```
data %>%
  ggplot() +
    aes(x = gender, y = anxiety_post, fill = gender) +
    geom_violin() +
    geom_jitter(width = 0.2)
```



Lathatjuk a feltaro elemzes alapjan, hogy a nok szorongasszintje nagyobb valamivel mint a ferfiake atlagosan. Most nezzuk meg, ez a kulonbseg statisztikailag szignifikans-e.

Fuggetlen mintas t-teszt

Arra, hogy meghatarozzuk van-e kulonbseg ket csoport kozott valamilyen numerikus valtozo atlagaban, hasznalhatjuk a fuggetlen mintas **t-tesztet**, t.test().

A t-teszt elofeltetelei:

- A foggo valtozo intervallum vagy aranyskalan mozog
- A fuggetlen valtozo ket egymastol fuggetlen kategorikus csoportot reprezental
- A megfigyelesek fuggetlenek egymastol. Minden megfigyeles csak az egyik csoportba sorolhato, es a csoportok kozott nincs osszefugges az egyes megfigyelesek között.
- Nincsenek jelentos kiugro esetek.
- Csoportonkent normalis eloszlast mutat a fuggo valtozo eloszlasa.
- Variancia-homogenitas: a fuggo valtozo varianciaja azonos a ket csoportban. (A welch t-teszt-et lehet alkalmazni, ha ez a feltetel serul).

```
t_test_results = t.test(anxiety_post ~ gender, data = data)
t_test_results
```

```
##
##  Welch Two Sample t-test
##
## data:  anxiety_post by gender
## t = 2.8895, df = 48.518, p-value = 0.005754
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  0.5553479 3.0941793
## sample estimates:
## mean in group female   mean in group male
##            11.466400             9.641636
```

```
mean_dif = summary %>%
    summarize(mean_dif = mean[1] - mean[2])
mean_dif
```

```
## # A tibble: 1 x 1
##   mean_dif
##      <dbl>
## 1     1.82
```

Az eredmenyt igy irhatjuk le:

"A ferfiak es nok szignifikansan kulonboztek a egymastol a szorongas szintjukben (t = 2.89, df = 48.52, p = 0.006. A csoportok szorongas szintjenek atlaga es szorasa a kovetkezo volt:"nok: 11.47(2.58), ferfiak: 9.64(2.7). A nok atlagosan 1.82 ponttal voltak szorongobbak (95% CI = 0.56, 3.09)."

Egyszempontos ANOVA

Ha egy kategorikus valtozon belul **harom vagy tobb csoportunk** is van, a t.test nem hasznalhato. Helyette hasznalhatjuk az **egyszempontos ANOVA**-t (one-way ANOVA) az aov() funkcioval. A formula ugyan ugy nez ki, mint a t-teszt eseten.

Az egyszempontos ANOVA elofeltetelei majdnem ugyan azok, mint a fuggetlen mintas t-tesztei:

- A foggo valtozo intervallum vagy aranyskalan mozog
- A fuggetlen valtozo ket vagy tbb egymastol fuggetlen kategorikus csoportot reprezental
- A megfigyelesek fuggetlenek egymastol. Minden megfigyeles csak az egyik csoportba sorolhato, es a csoportok kozott nincs osszefugges az egyes megfigyelesek között.
- Nincsenek jelentos kiugro esetek.
- Csoportonkent normalis eloszlast mutat a fuggo valtozo eloszlasa.
- Variancia-homogenitas: a fuggo valtozo varianciaja azonos a csoportokban.

Igy teszteljuk hogy van-e kulonbseg a **lakhatasi helyzet csoportjai** kozott a **szorongasszintben**.

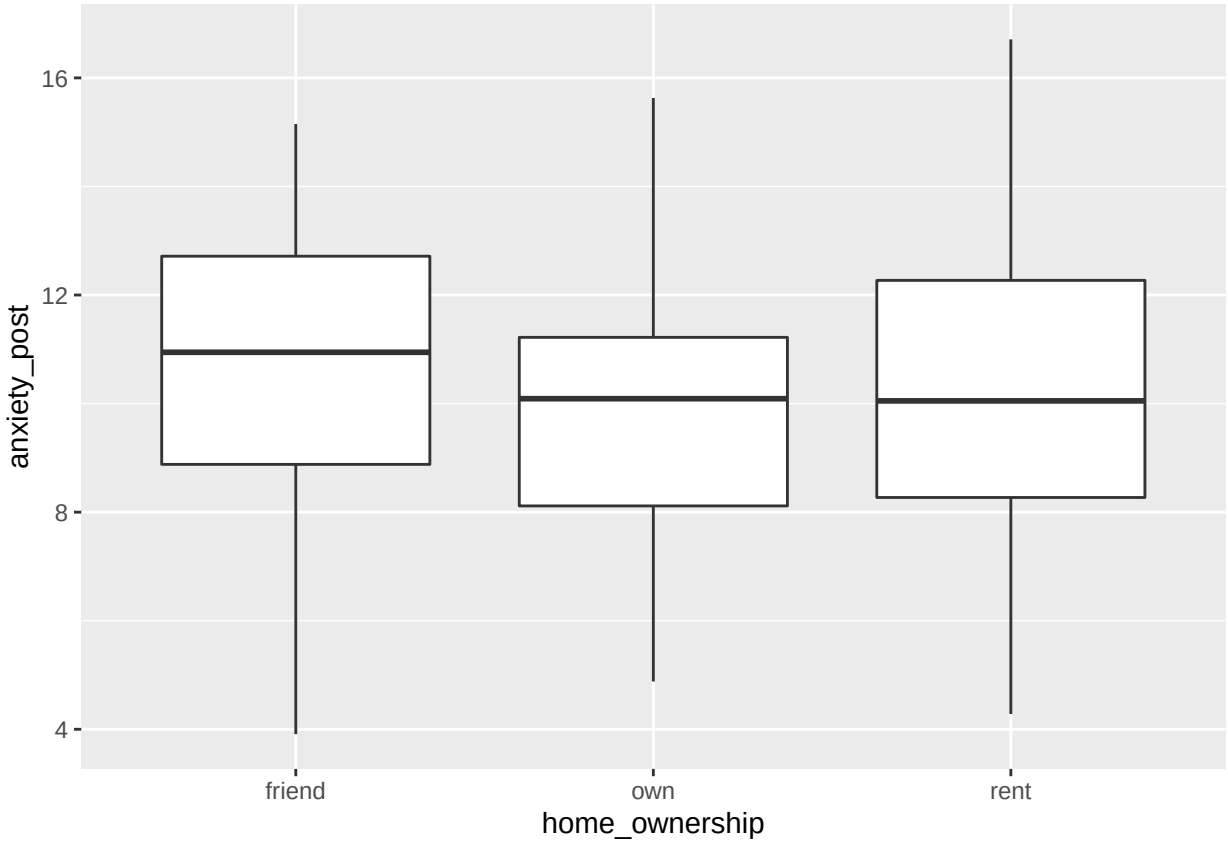
```
summary_home_ownership_vs_anxiety_post = data %>%
  group_by(home_ownership) %>%
    summarize(mean = mean(anxiety_post), sd = sd(anxiety_post))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
summary_home_ownership_vs_anxiety_post
```

```
## # A tibble: 3 x 3
##   home_ownership  mean    sd
##   <fct>          <dbl> <dbl>
## 1 friend         10.6   2.93
## 2 own             9.86  2.57
## 3 rent           10.3   2.94
```

```
data %>%
  ggplot() +
    aes(x = home_ownership, y = anxiety_post) +
    geom_boxplot()
```




```
ANOVA_result = aov(anxiety_post ~ home_ownership, data = data)
summary(ANOVA_result)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## home_ownership  2    7.5   3.760    0.48  0.621
## Residuals      77  603.3   7.835
```

Az eredmenyt igy irhatjuk le:

"A lakhatasi csoportonk szerint nem volt szignifikans kulonsbeg a szorongas atlagos szintjeben (F (2 ,77) = 0.48, p = 0.621). A szorongas atlagat es szorasat az egyes csoportok szerinti bontasban lasd az 1. tablazatban"

Alabb lathato, hogyan produkalnank a megfelelo tablazatot a szorongas atlagaval home_ownership csoportok szerint.

Egyoldalu vs. ketoldalu tesztek

Fontos, hogy ha van elozetes elkepzelesunk a hipotezisalkotaskor arrol, hogy **milyen iranyu** lesz a hatas, akkor **egy-oldalu (one-sided) tesztet** kell hasznalnunk az alapertelmezett ket-oldalu teszt helyett.

Peldaul tegyuk fel hogy amikor a hipotezisunket meghataroztuk (idealis esetben ez meg az adatgyujtes elott megtortenik), ugy gondoltuk, hogy a noknek magasabb lesz a szorongasszintjuk, mint a ferfiaknak. Ezt az alternative = "greater" parameterrel hatarozhatjuk meg.

Ha osszehasonlitjuk ezt az eredmenyt a korabbi t-teszt eredmenyevel, eszrevehetjuk hogy minden szam valtozatlan maradt, kiveve **a p-erteket**, ami pontosan felere csokkent, es a 95%-os **konfidencia intervallumot**, aminek a felso hatara most egy vegtelen nagy szam (inf).

A p-ertek azert felezodott meg, mert azzal, hogy meghataroztuk, melyik iranyban fog a ket csoport kulonbozni egymastol fele akkora lett az eselye hogy a most megfigyelt, vagy annal nagyobb kulonbseget kapunk a

null-hipotezis helyesseget feltetelezve. Vagyis amikor tudjuk, milyen iranyu hatast varunk el, mindig erdemes egy-oldalu tesztet alkalmazni, mert ezzel no a statisztikai eronk a hatas kimutatasara.

Az egyoldalu tesztek eseten amikor az a hipotezisunk, hogy **a referencia-csoport** atlaga magasabb lesz, (alternative = "greater"), akkor a konfidencia intervallumnak csak az also hatarat szamoljuk ki. Ezert irja a teszt eredmenye hogy a 95% CI 1.11, Inf, vagyis felfele a vegtelensegig tart a konfidencia intervallum.

Fontos megjegyezni, hogy amikor azt irjuk a tesztben hogy **alternative = "greater"**, ez alatt azt ertjuk hogy az alternativ hipotezisunk az, hogy **a referencia-csoport** atlaga magasabb lesz. Ha az alternativ hipotezisunk az lenne hogy a referencia-csoport atlaga alacsonyabb lesz, akkor azt kellene irnunk: **alternative = "less"**.

Ahogy korabban mar volt rola szo, a referencia-csoport (vagy referencia-szint egy faktor valtozoban) alapertelmezett modon a faktorszintek nevenek ABC sorrendje alapjan dol el, az ABC sorrendben az elso faktorszint lesz a referencia-szint. A peldankban a gender valtozoban a "female" a referencia-szint, mert az ABC sorrendben a "male" elott van. Azt, hogy mi legyen a referencia-szint a korabban tanultak szerint a **factor() funkcioban a levels =** parameter beallitasaval lehet befolyasolni. Nagyon fontos, hogy amikor kategorikus/csoportosito valtozokkal dolgozunk, mindig tudjuk, mi a referencia-szint.

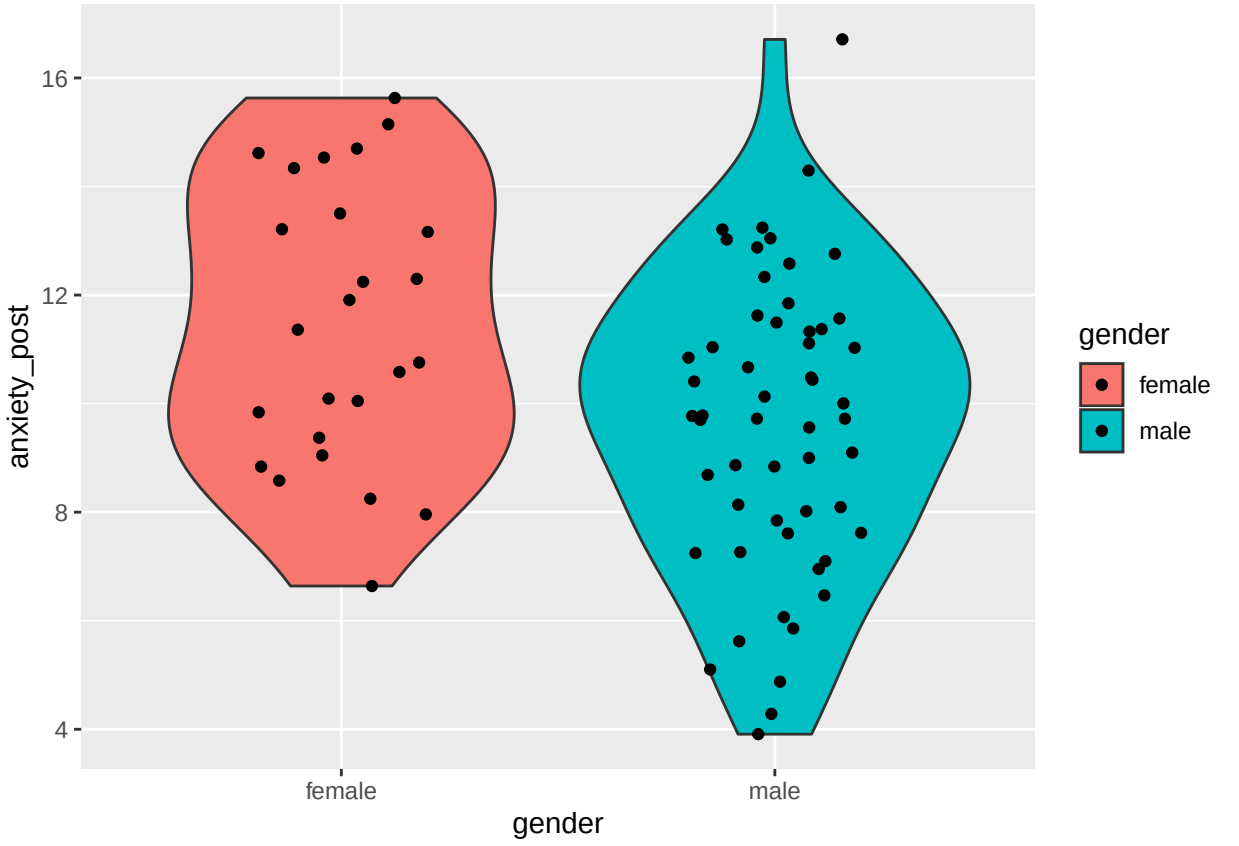
```
summary = data %>%
  group_by(gender) %>%
    summarize(mean = mean(anxiety_post), sd = sd(anxiety_post))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
summary
```

```
## # A tibble: 2 x 3
##   gender  mean    sd
##   <fct>  <dbl> <dbl>
## 1 female 11.5   2.58
## 2 male    9.64  2.70
```

```
data %>%
  ggplot() +
    aes(x = gender, y = anxiety_post, fill = gender) +
    geom_violin() +
    geom_jitter(width = 0.2)
```




```
t_test_results_one_sided = t.test(anxiety_post ~ gender, data = data, alternative = "greater")
t_test_results_one_sided
```

```
##
##  Welch Two Sample t-test
##
## data:  anxiety_post by gender
## t = 2.8895, df = 48.518, p-value = 0.002877
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7657774       Inf
## sample estimates:
## mean in group female   mean in group male
##            11.466400             9.641636
```

Az eredmenyt igy irhatjuk le:

“A ferfiak es nok szignifikansan kulonboztek a egymastol a szorongas szintjukben (t = 2.89, df = 48.52, p = 0.003. A csoportok szorongas szintjenek atlaga es szorasa a kovetkezo volt:”nok: 11.47(2.58), ferfiak: 9.64(2.7). A nok atlagosan 1.82 ponttal voltak szorongobbak (95% CI = 0.77, inf).”

Nezzuk meg, mi tortenne, ha azt tippeltuk volna a hipotezisalkotaskor, hogy a noknek alacsonyabb lesz a szorongasszintjuk. Ezt ugy hatarozhatjuk meg, hogy a t.test() funkcioban alternative = “less” parametert allitunk be.

A p-ertek itt majdnem eleri az 1-et, vagyis nagyon nagy a valoszinusege, hogy a null-hipotezis helyesseget feltetelezve ilyen, vagy ennel extremebb kulonbseget figyelunk meg. Nem is csoda, hiszen a null hipotezisunk itt az hogy a nok szorongasanak atlaga nem fog kulonbozni, vagy nagyobb lesz mint a ferfiake, es azt

tapasztaltuk, hogy valoban nagyobb volt, vagyis a megfigyeles egyaltalan nem segit abban, hogy elutasitsuk a null-hipotezist.

```
t_test_results_one_sided = t.test(anxiety_post ~ gender, data = data, alternative = "less")
t_test_results_one_sided
```

```
##
##  Welch Two Sample t-test
##
## data:  anxiety_post by gender
## t = 2.8895, df = 48.518, p-value = 0.9971
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf 2.88375
## sample estimates:
## mean in group female   mean in group male
##            11.466400             9.641636
```

Azt is erdemes megjegyezni, hogy a "greater" es a "less" minding a kategorikus valtozo **referencia-szintjere** vonatkozik. Ha ezt nem allitottuk be maskepp, pl. a **factor (levels =)** funkcioval, akkor a referencia-szint az ABC sorrendben elorebb levo szint lesz. A fenti esetben a ket szint a "female" es a "male", amik kozul a "female" jon elobb ABC sorrendben. Ha azt tippeltuk volna, hogy az lenne a hipotezisunk, hogy a ferfiak ("male") szorongasszintje lesz magasabb, akkor alternative = "less"-t kellene beallitanunk, mert ezzel egyben azt tippeljuk, hogy a referenciaszint ("felame") atalaga lesz az alacsonyabb. Vagy at kellene allitani a referenciaszintet.

Gyakorlas

Teszteld a 3. hipotezist, hogy "A terapias csoportban alacsonyabb lesz a szorongas atlaga a kutatas vegere mint a kontrol csoportban" (**anxiety_post** vs. **group**)

- Eloszor vegezzunk egy feltaro elemzest egy tablazattal a ket valtozo kapcsolatarol a summarize(mean(), sd()) funkciokkal, es keszitsunk abrat, mondjuk geom_boxplot() segitsegevel.
- egy- vagy ketoldalu tesztet kell alkalmaznunk? (gondolj arra, hogy a hipotezisunkben megjosoljuk-e a hatas vagy kulonbseg iranyat vagy sem)
- Mi a null-hipotezis ebben az esetben?
- Melyik tesztet erdemes hasznalni, az egyvaltozos ANOVA-t, vagy a t-tesztet? (gondolj arra, hogy hany csoport (szint) van a kategorikus valtozon belul)
- Ez utan vegezd el a tesztet
- Es ird le a fentiek szerint az eredmenyeket.

Ket numerikus valtozo kozotti kapcsolat, korrelacio, cor.test()

Vizsgaljuk meg, van-e egyuttjaras a reziliencia (**resilience**) es a magassag (**height**) kozott.

Eloszor vegezzunk feltaro elemzest a korrelacios egyutthato kiszamitasaval, es egy pontdiagrammal. Hasznaljunk geom_point() es geom_smooth() geomokat egyszerre, es hasznaljuk az "lm" modszert a trendvonal megrajzolasara.

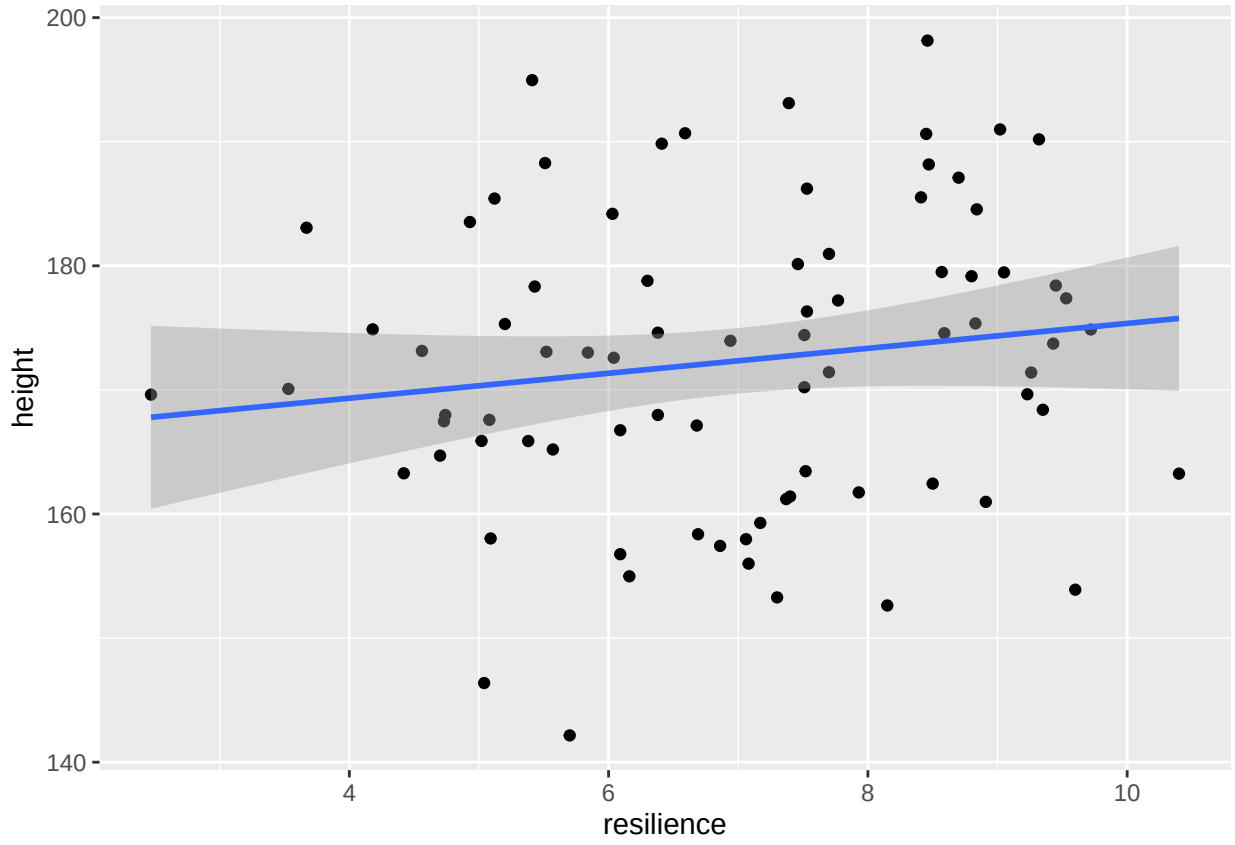
```
data %>%
  select(resilience, height) %>%
    cor()
```

```
##            resilience   height
## resilience   1.000000 0.146929
```

```
## height       0.146929 1.000000
```

```
data %>%
  ggplot() +
    aes(x = resilience, y = height) +
      geom_point() +
      geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



A ket valtozo fuggetlennek tunik egymastol a feltaro elemzes alapjan, de elkepzelheto, hogy a hatas, barmilyen kicsit is, megis statisztikailag szignifikans, szoval vegezzuk el a statisztikai tesztet is.

Ezt a **pearson korrelacios teszt** segitsegevel tehetjuk meg.

The assumptions of pearson's correlation are:

- Ket folytonos skalaju valtozo. Ha barmelyik valtozo ordinalis skalaju, akkor a spreaman korrelaciot lehet hasznalni.
- Minden megfigyelesi egyseghez ket ertek tartalmazzon.
- Nincsenek jelentos kiugro ertekek
- Linearitas. A ket valtozo kapcsolata egy egyenes vonallal jellemezheto.
- Normalitas: mindket vatozo normalis eloszlast mutat. Nem normalis eloszlas eseten a Spearman korrelacio hasznalhato.

A tesztet a cor.test() funkcioval vegezhetjuk el a kovetkezo keppen:

```
correlation_result = cor.test(data$resilience, data$height)
correlation_result
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$resilience and data$height
## t = 1.3119, df = 78, p-value = 0.1934
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.07521612  0.35517967
## sample estimates:
##      cor
## 0.146929
```

Az eredmenyt igy irhatjuk le:

"A reziliencia es a magassag kozott nem talaltunk szignifikans egyuttjarast (r = 0.15, 95% CI = -0.08, 0.36, df = 78, p = 0.193)"

Hasonloan a t-teszthez, a korrelacios teszt eseteben is erdemes egyoldalu tesztet hasznalni amikor a hipotezisunk megmondja a kapcsolat iranyat is, nem csak azt, hogy van kapcsolat a ket valtozo kozott.

Peldaul feltetelezzuk, hogy a ket valtozo kozotti **kapcsolat pozitiv iranyu** lesz. Vagyis egy ember minel magasabb, annal magasabb a rezilienciaja. Ezt ugy adhatjuk meg a statisztikai teszt specifikaciojakor, hogy a formulahoz hozzatesszuk az **alternative = "greater"** parametert. Ha az eredmenyt osszehasonlitjuk az elozo korrelacios teszt eredmenyevel, lathatjuk, hogy a p-ertek is megvalozott. A konfidencia intervallumnak itt is csak az also hatara erdekes, a felso hatara a leeheto legmagasabb erteket veszi fel ilyenkor, ami a korrelacional 1.

```
correlation_result_greater = cor.test(data$resilience, data$height, alternative = "greater")
correlation_result_greater
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$resilience and data$height
## t = 1.3119, df = 78, p-value = 0.0967
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  -0.03942784  1.00000000
## sample estimates:
##      cor
## 0.146929
```

Gyakorlas

Teszteld a 4. hipotezist, hogy "A reziliancia es a kutatas vegen mert szorongasszint negativ osszefuggest fog mutatni (vagyis aki reziliensebb, annal alacsonyabb szorongasszintet fognak merni a kutatas vegen)" (**anxiety_post** vs. **resilience**)

- Eloszor vegezzunk egy feltaro elemzest a korrelacios egyutthato meghatarozasaval es egy pontdiagrammal a ket valtozo kapcsolatarol.
- egy- vagy ketoldalu tesztet kell alkalmaznunk? (gondolj arra, hogy a hipotezisunkben megjosoljuk-e a hatas vagy kulonbseg iranyat vagy sem)
- Mi a null-hipotezis ebben az esetben?
- Ez utan vegezd el a tesztet
- Es ird le a fentiek szerint az eredmenyeket.

A statisztikai tesztek eredmenyenek kozleserol altalaban

A statisztikai tesztek eredmenyenek kozlese soran a kovetkezo informaciokat szoktuk megadni altalanossagban. Ez tesztrol tesztre valtozhat, de az alabbiak kozul minel tobb informaciot megadnunk, annal jobb.

- az eredmeny szoveges leirasa
- teszt-statisztika
- szabadsagfok (ez egyszeru teszteknel altalaban az elemszammal is megadhato)
- p-ertek
- hatas merteke (parameterbecsles)
- hatasmertek 95%-os konfidencia intervalluma