

Seminar_5

Zoltan Kekecs

October 6, 2020

Ket változó kapcsolatának vizsgálata, statisztikai inferencia

Az óra célja

Az óra célja hogy megismerkedjünk a **statisztikai inferencia alapjaival** két változó kapcsolatának elemzésén keresztül.

Package-ek betöltése

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

Hipotezistesztes

A statisztikai inferencia, és hipotézis tesztelés során az a célunk, hogy megállapítsuk, letezik-e egy bizonyos hatás vagy kapcsolat. De ezt a **null-hipotézis szignifikancia tesztelés (NHST)** során egy fordított logikával tesszük: azt állapítjuk meg, hogy **mekkora a valószínűsége hogy az általunk megfigyelt adatot/trendet figyeljük meg (vagy annál meg extremer trendet), amennyiben a null-hipotézis igaz.**

Egy egyszerű példa: az a sejtésem, hogy **egy pénzérme cinkelt**, megpedig úgy hogy nagy valószínűséggel **fej** legyen az eredmény amikor feldobjuk. Ebben az esetben a **null-hipotézisem** az, hogy az **érme nem cinkelt**. Vagyis a null-hipotézis szerint ugyanakkor a valószínűsége fejnek és irásnak kapni egyenlő.

- H1: cinkelt érme (fej fele)
- H0: nem cinkelt érme, vagy irás fele cinkelt

Tegyük fel hogy 10-szer feldobjuk az érmét, és 9-szer fejet dobunk. Mekkora a valószínűsége, hogy az érme cinkelt? Ezt nem tudjuk megmondani. Többek között azért sem mert nem tudjuk, mennyire lehet cinkelve. Viszont azt meg tudjuk mondani, hogy mekkora a valószínűsége, hogy ezt az eredményt kapnánk, ha az érme **NINCS** cinkelve.

Annak a valószínűsége, hogy **legalább 9-szer** (vagy többször) fejet dobok **10 dobásból** egy nem cinkelt érmevel, $p = 0.0107$ (**nagyjából 1%**). (Ezt a kódot részben nem fontos megérteni, a lényege hogy a `pbinom()` függővel kiszámoltuk a valószínűséget, hogy 10 dobásból legalább 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

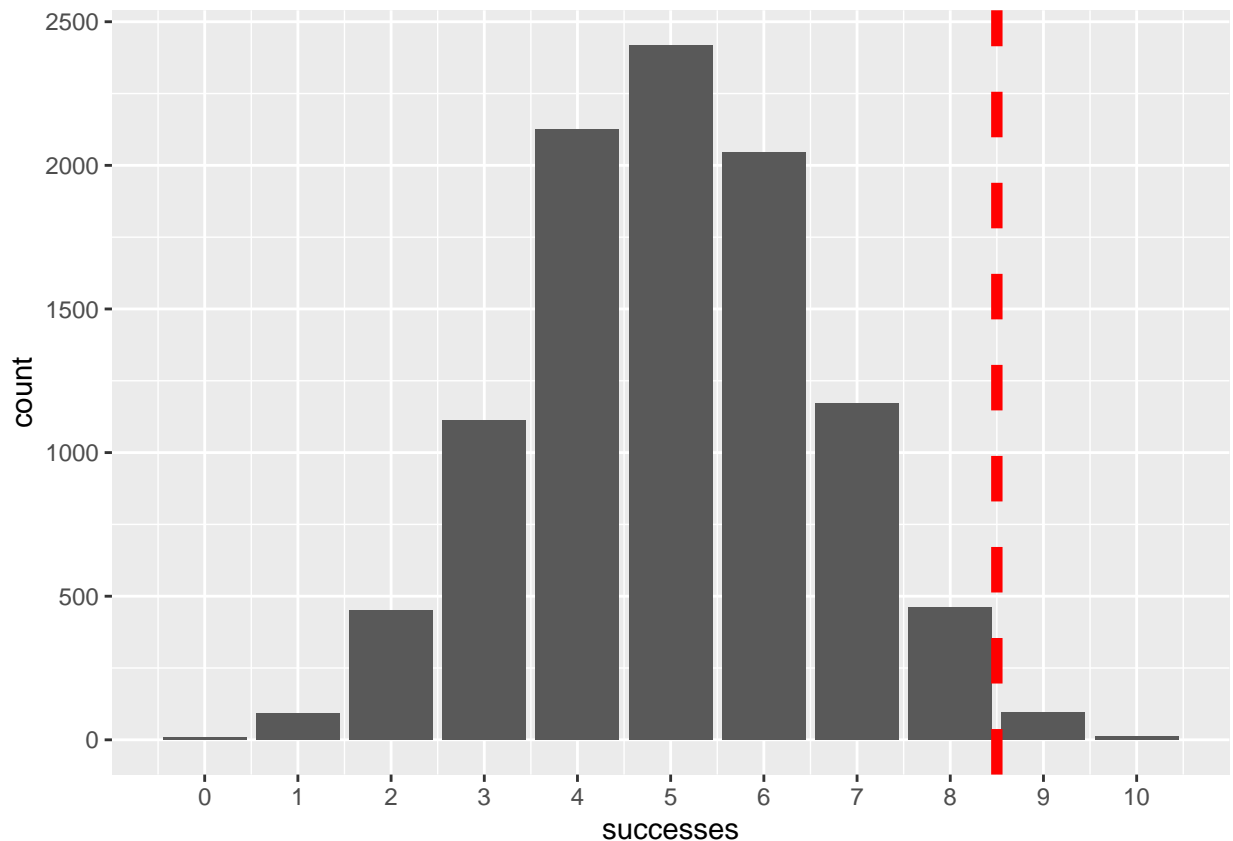
## [1] 0.01074219
```

Ez a valószínűség **maskepp mondva** azt jelenti, hogy ha ugyan ezt a kísérletet 100 szor megismételnenk (mindegyikben 10 feldobással), akkor a 100 kísérletből csak átlagosan 1-szer varnának, hogy 9 vagy több fejet kapjunk.

Ezt le is ellenőrizhetjük, ha **randomizálunk mondjuk 10.000 hasonló kísérletet** az `rbinom()` funkcióval. Az ábrán látható hogy csak a kísérletek igen kis százalékában kaptunk 9 vagy több “sikert”. (Ezt a kódrész sem fontos megérteni, a lényege hogy az `rbinom()` funkcióval 10000-szer azt szimuláltuk, hogy egymás után 10-szer feldobtunk egy érmet (vagyis hogy véletlenszerűen választottunk egy számot 0 és 1 közül), ez után ezt ábrázoltuk 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 ez egy elég meglepo (bar nem lehetetlen) eredmény, ha az érme nem lenne cinkelve. Amikor NHST-t csinálunk, ez alapján hozzuk meg a **dontesunket** arról, hogy **elvetjük-e** a null-hipotézist, vagy, kello cáfoló bizonyíték híján **megtartjuk** azt.

A pszichológiában általában $p < 0.05$ határértéket használunk a döntéshozásban, vagyis ha egy olyan meglepo eredményt figyelünk meg, **aminek a valószínűsége kisebb mint 5% ha a null-hipotézis igaz, akkor elvetjük a null-hipotézist**. Vagyis a fenti eredmény esetén elvethetnek a hull-hipotézist, mert a megfigyelt eredmény (vagy annál extremebb eredmény) valószínűsége 1% ($p = 0.01$) ha a H_0 igaz, ami kisebb mint 5% ($p < 0.05$).

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 ot statisztikai tesztet fogunk megismerni.

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

binomialis teszt

A hipotézist, hogy az érme cinkelt, tesztelhetjük a **binomialis teszt**tel, aminek R-ben `binom.test()` a funkciója. Az `x` helyére a megfigyelt “celmegfigyelesek” vagy “sikerek” számát (a mi esetünkben a fejek számát), az `n` helyére az összes megfigyeles számát, a `p` helyére pedig a **null-hipotézis** helyesseget feltételezve a “celmegfigyelesek” elérésének valószínűségét kell beírni (mivel a hipotézisünk az hogy az érme cinkelt, az null hipotézisünk az, hogy az érme “nem cinkelt”). Ezt valószínűségként kell megadni, ami 0 és 1 közötti szám (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 jelentése hogy a megfigyelesek 60%-a lesz “siker”).

```
binom.test(x = heads, n = total_flips, p = probability_of_heads_if_H0_is_true, alternative = "greater")
```

```
##
## Exact binomial test
##
## data: heads and total_flips
## 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 eredménye a következőt mutatja:

- p-value: p-érték, annak a valószínűsége, hogy az általunk megfigyelt, vagy extremer eredményt kapunk, feltételezve hogy a null-hipotézis helyes. Általában ha ez az érték 0.05 alatti, akkor elvetjük a null-hipotézist.
- alternative hypothesis: Itt írja le, hogy mi volt a H_1 , ami a mi esetünkben az volt, hogy a fej valószínűsége nagyobb mint 0.5 (50%). Ez egyben azt is jelenti, hogy a null-hipotézisünk az volt, hogy a fej valószínűsége 0.5.
- 95 percent confidence interval (vagy röviden 95% CI): a 95%-os konfidencia intervallum. Ez azt jelenti, hogy ha a kísérletet sokszor megismételjük és ugyan így kiszámoljuk a konfidencia intervallumot minden kísérletnél, az így kapott konfidencia intervallumok 95%-a tartalmazni fogja a való hatásmeretet (ami a mi esetünkben a “siker”/fej valószínűsége). Fontos, hogy nem tudjuk, hogy a mi konkrét kísérletünkben a konfidencia intervallum tartalmazza-e a való hatásmeretet.
- sample estimates: A “siker” (“celmegfigyeles”, a mi esetünkben a fej) valószínűségének becslt merteke a populációban a megfigyelt valószínűség alapján. Ez egy pontbecslés, ami mindig megegyezik a megfigyelt valószínűséggel.

Az eredményt így írhatjuk le:

“A kutatásunkban 9 fejet figyeltünk meg 10 pénzfeldobásból (90%). Ez alapján úgy ítéltük, hogy annak a valószínűsége, hogy fejet dobunk az érmevel szignifikánsan több mint 50%. A fej dobás valószínűsége 0.9 volt a mintában (95% CI = 0.61, 1).”

Adatgeneralas az orahoz

Az alábbi kód **adatokat general** a számunkra. Az adatgeneralashoz használt kód megértése ezen a szinten még nem szükséges.

```
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)
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 <- anxiety_base + treatment * treatment_effect + resilience * resilience_effect + gender_num * gender_effect
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 < 11] = "cured"
health_status[anxiety >= 11] = "anxious"

data <- data.frame(participant_ID)
```

```

data = cbind(data, gender, group, resilience, anxiety, health_status, home_ownership, height)
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))

```

Az adatok egy (kepzeletbeli) randomizalt kontrollalt klinikai kutatas eredményeibol szarmaznak, ahol a **pszichoterapia hatekonysagat** teszteltek. Olyan személyeket vontak be a kutatásba, akik egy **hurrikan aldozatai** voltak, es **szorongással** kusz kodtek. A személyeknel felmerte a reziliencia (psziches ellenal-lokepesseg) szintjet, majd veletlenszeruen osztottak a személyeket egy kezelési vagy egy kontrol csoportba. Ezt kovetoen a kezelési csoport **pszichoterapiat kapott 6 heten keresztul** heti egyszer, mig a kontrol csoport nem kapott kezelést. A vizsgalat vegen megmerte a személyek **szorongasszintjet**, es a klinikai kriteriumok alapjan meghataroztak, hogy a személy **gyogyultnak, vagy szorongonak** számít-e.

Lathatjuk, hogy 8 változó van az adattáblában.

- participant_ID - részvevo azonositoja
- gender - nem
- group - csoporttagság, ez egy faktor változó aminek két szintje van: “treatment” (kezelt csoport), es “control” (kontrol csoport). A “treatment” csoport kapott kezelést, mig a “control” csoport nem kapott kezelést.
- resilience - reziliencia: a nehezsegekkel valo megküzdés kepessege, ez egy személyes kepesseg, olyasmi mint a személyiségvonások
- anxiety - szorongás szint
- health_status - a klinikai kriteriumok alapjan szorongonak vagy gyogyultnak tekintheto a személy
- home_ownership - lakhatási helyzet: három szintje van az alapján hogy a személy hol lakik: “friend” - barát nál vagy családnál lakik, “own” - saját tulajdonu lakásban lakik, “rent” - bérlet lakásban lakik,
- height - magasság

Adatellenorzes

Mint mindig, elemzés előtt **ellenorizzuk**, hogy az adattal minden rendben van-e!

```

data

## # A tibble: 80 x 8
##   participant_ID gender group resilience anxiety health_status home_ownership
##   <fct>          <fct> <fct>      <dbl>   <dbl> <fct>          <fct>
## 1 ID_1          male  trea~      5.08    10.5  cured          own
## 2 ID_2          male  trea~      6.41     7.61  cured          friend
## 3 ID_3          male  trea~      7.52     9.72  cured          rent
## 4 ID_4          female trea~      4.70    14.7  anxious        rent
## 5 ID_5          male  trea~      7.39     8.14  cured          own
## 6 ID_6          female trea~      7.06    10.1  cured          own
## 7 ID_7          female trea~      7.17     6.64  cured          own
## 8 ID_8          male  trea~      9.23     8.09  cured          own
## 9 ID_9          male  trea~      4.56    10.4  cured          own

```

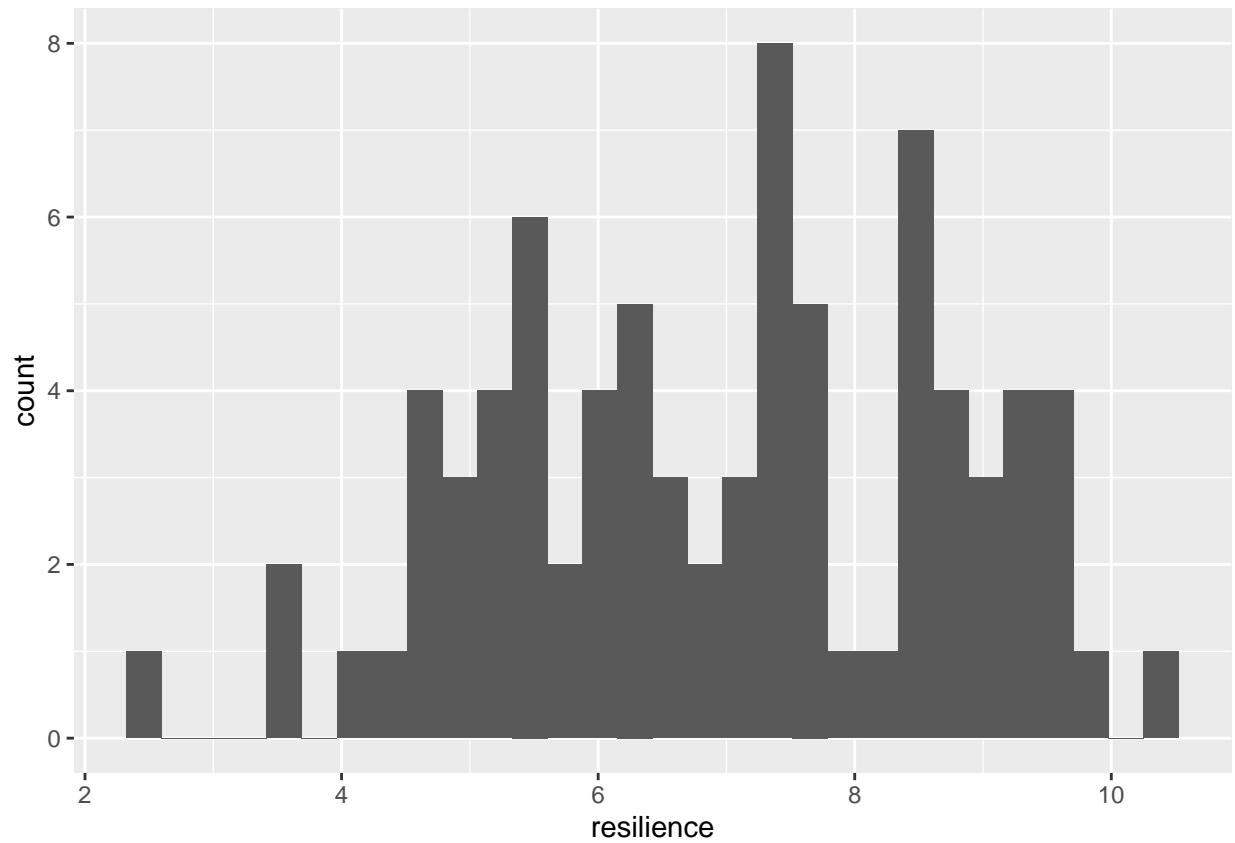
```
## 10 ID_10          male   trea~      9.53   4.28 cured          rent
## # ... with 70 more rows, and 1 more variable: height <dbl>
```

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

```
## participant_ID    gender      group      resilience      anxiety
## ID_1   : 1      female:25    control :40    Min.   : 2.469    Min.   : 3.914
## ID_10  : 1      male  :55    treatment:40   1st Qu.: 5.516    1st Qu.: 8.220
## ID_11  : 1                                Median : 7.124    Median :10.108
## ID_12  : 1                                Mean    : 6.981    Mean    :10.212
## ID_13  : 1                                3rd Qu.: 8.479    3rd Qu.:12.255
## ID_14  : 1                                Max.    :10.398    Max.    :16.706
## (Other):74
## health_status home_ownership      height
## anxious:32    friend:22          Min.   :142.2
## cured  :48    own   :31          1st Qu.:163.4
##                                Median :173.0
##                                Mean    :172.3
##                                3rd Qu.:179.7
##                                Max.    :198.2
##
```

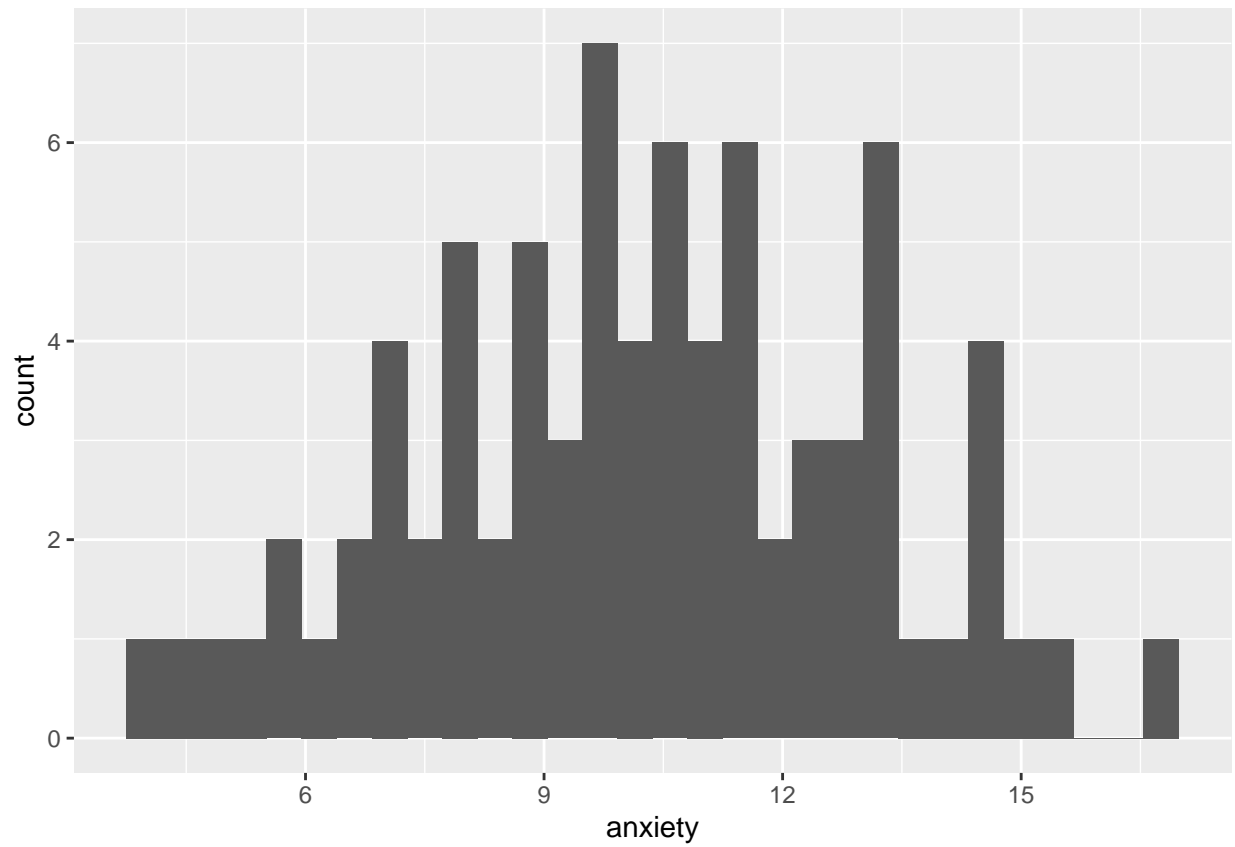
```
data %>%
  ggplot() +
    aes(x = resilience) +
    geom_histogram()
```

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

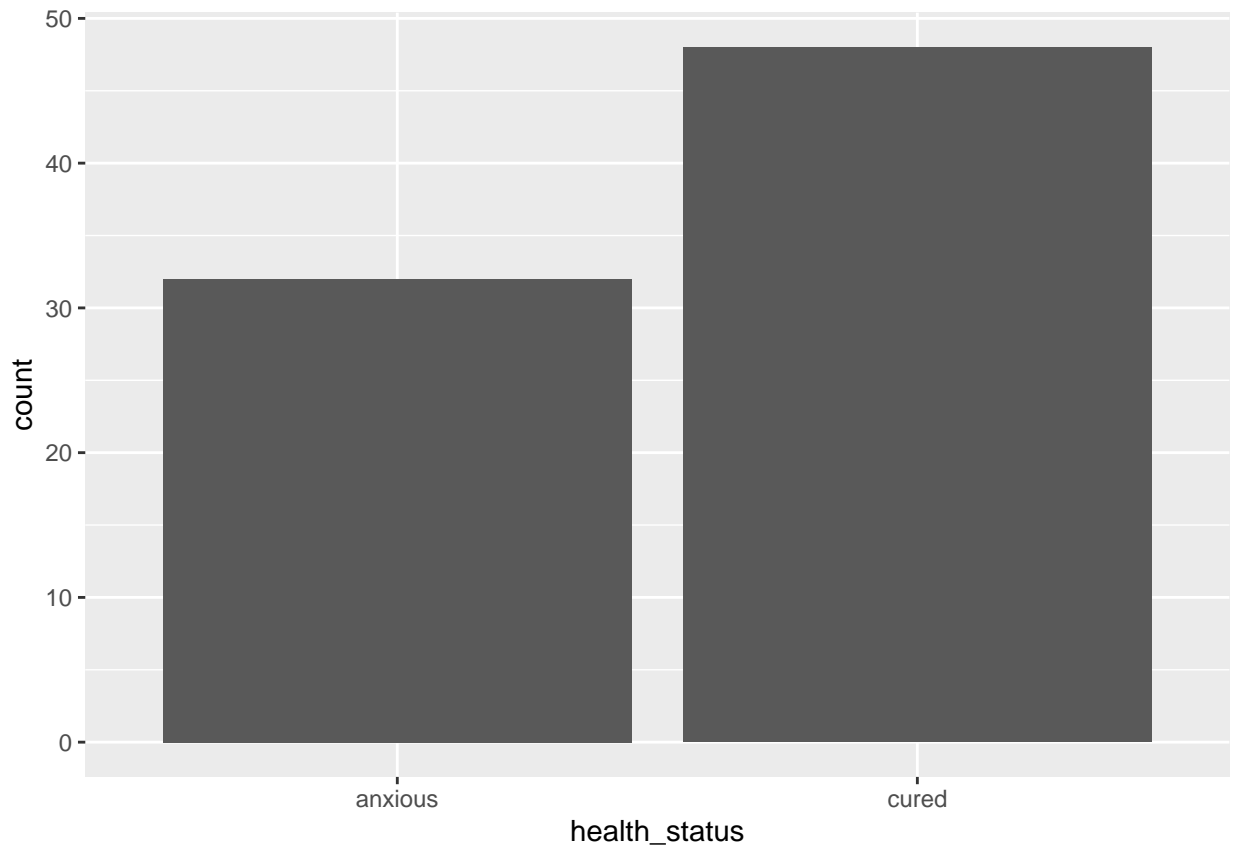


```
data %>%  
  ggplot() +  
    aes(x = anxiety) +  
    geom_histogram()
```

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



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

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

Hipotezisek

Vizsgáljuk meg kutatásban szereplő változók összefüggését a hipotezisek mentén.

A kutatás hipotézise a következők voltak:

1. Több a férfi mint a nő ebben a klinikai mintában (**gender** vs. 50%).
2. A pszichoterápiát kapó csoportban a terápia után kevesebb lesz a klinikai kritériumok alapján szorongónak számító személy (**health_status** vs. **group**)
3. A terápiás csoportban alacsonyabb lesz a szorongás átlaga a kutatás végére mint a kontrol csoportban (**anxiety** vs. **group**)
4. A reziliencia és a kutatás végen mért szorongásszint negatív összefüggést fog mutatni (vagyis aki reziliensebb, annál alacsonyabb szorongásszintet fognak mérni a kutatás végen) (**anxiety** vs. **resilience**)

Gyakorlás

Teszteld a hipotézist, hogy “Több a férfi mint a nő ebben a klinikai mintában” (**gender** változó)

- Ezt ugyan úgy teheted meg, mint a fenti példában, hiszen a null-hipotézis az, hogy a férfiak (“male”) elvárt valószínűsége 50% vagy kevesebb ($p = 0.5$). Szóval a férfiak ekvivalensek a “fejekkel” a pénzfeldobásos példában.
- Meg kell határozni a férfiak számát a mintában, és a teljes mintaelemszámot, hogy ki tudj tölteni a `binom.test()` függvény paramtereit.
- Ez után vegezd el a tesztet
- És írd le a fentiek szerint az eredményeket.

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

A Khi-negyzet proba ket kategorikus valtozo kapcsolatat hivatott megvizsgalni.

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

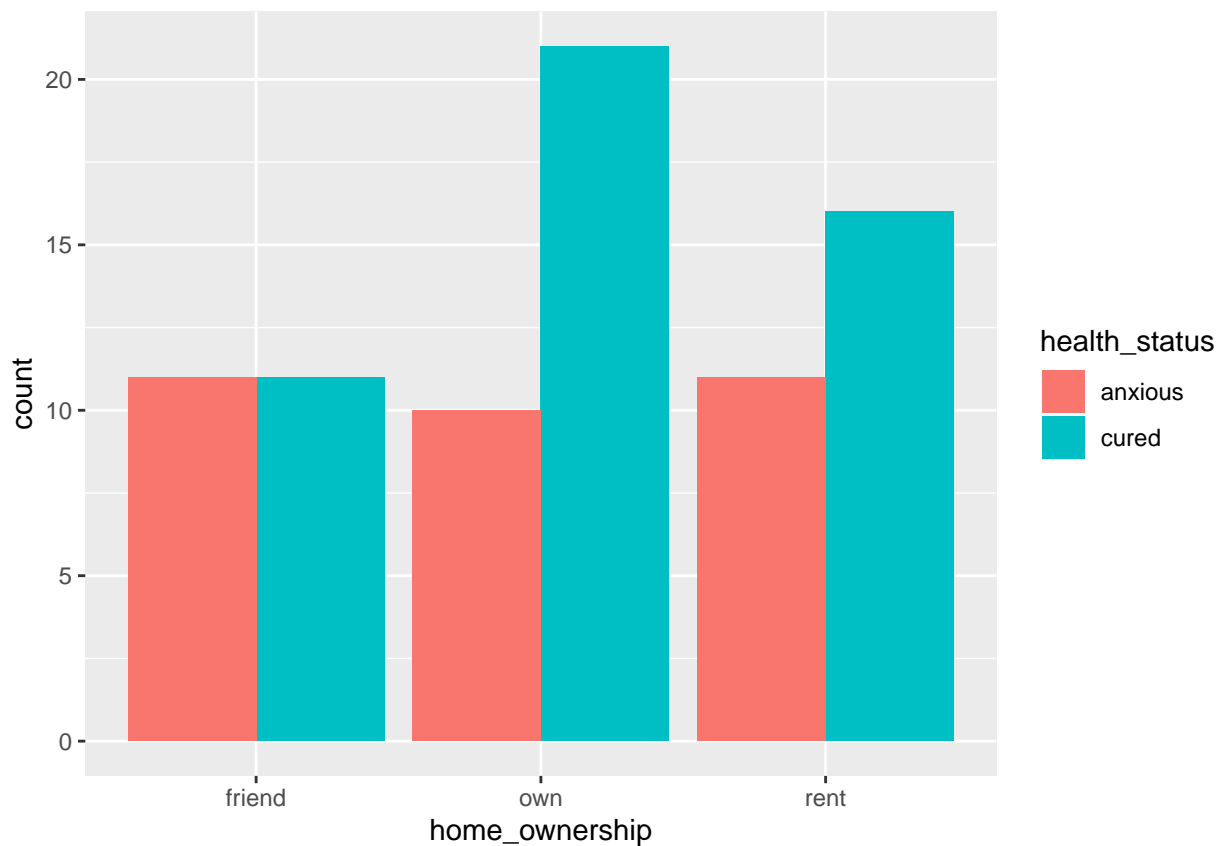
Eloszor feltaro elemzest vegzunk:

- 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 másik kategorikus változó eloszlása (vagyis a mi esetünkben a null hipotézis hogy ugyan olyan arányban gyógyulnak meg akik barátnál laknak, akiknek saját lakásuk van, és akik berlik a lakást).

```
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 eredményt így írhatjuk le:

“Nem volt szignifikáns eltérés abban, hogy a különböző lakhatási csoportokban (barátnál, saját lakásban, vagy berlemben lakók) milyen arányban voltak azok akik meggyógyultak a kutatás végére ($X^2 = 1.7$, $df = 2$, $p = 0.428$).”

Gyakorlás

Teszteld a 2. hipotézist, hogy “A pszichoterápiát kapó csoportban a terápia után kevesebb lesz a klinikai kritériumok alapján szorongónak számító személy” (**health_status** vs. **group**)

- Ezt ugyan úgy teheted meg, mint a fenti példában, hiszen a null-hipotézis az, hogy nincs különbség a csoporttagság szerint (treatment vs. control) abban hogy milyen arányban gyógyultak meg a kutatás végére.
- Eloszor vegezzünk egy feltárási elemzést egy táblázattal a két változó kapcsolatáról a `table()` funkcióval és egy ábrával (mondjuk `geom_bar()` használatával)
- A táblázatot mentsd el egy új objektumba
- Ez után vegezd el a tesztet, `chisq.test()`
- És írd le a fentiek szerint az eredményeket.

Egy numerikus változó átlagának különbsége csoportok között: anova és t-teszt

Tesztelhetjük például, hogy van-e különbség a nemek között (**gender**) a kutatás végén mért szorongás szintjében (**anxiety**).

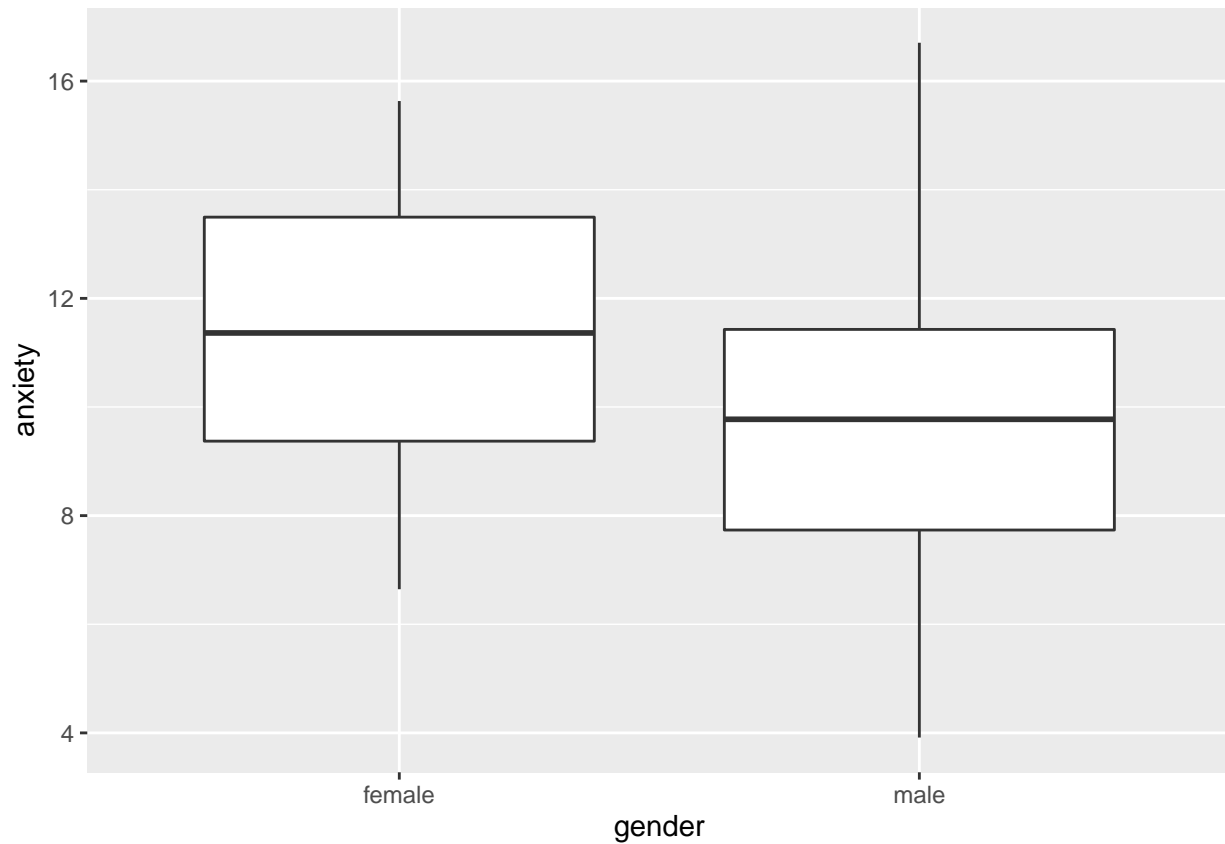
Eloszor szokás szerint feltárási elemzést végünk átlagok csoportonkénti összehasonlításával és ábrával. Erre pl. remek a `geom_boxplot()` és a `geom_density()`

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

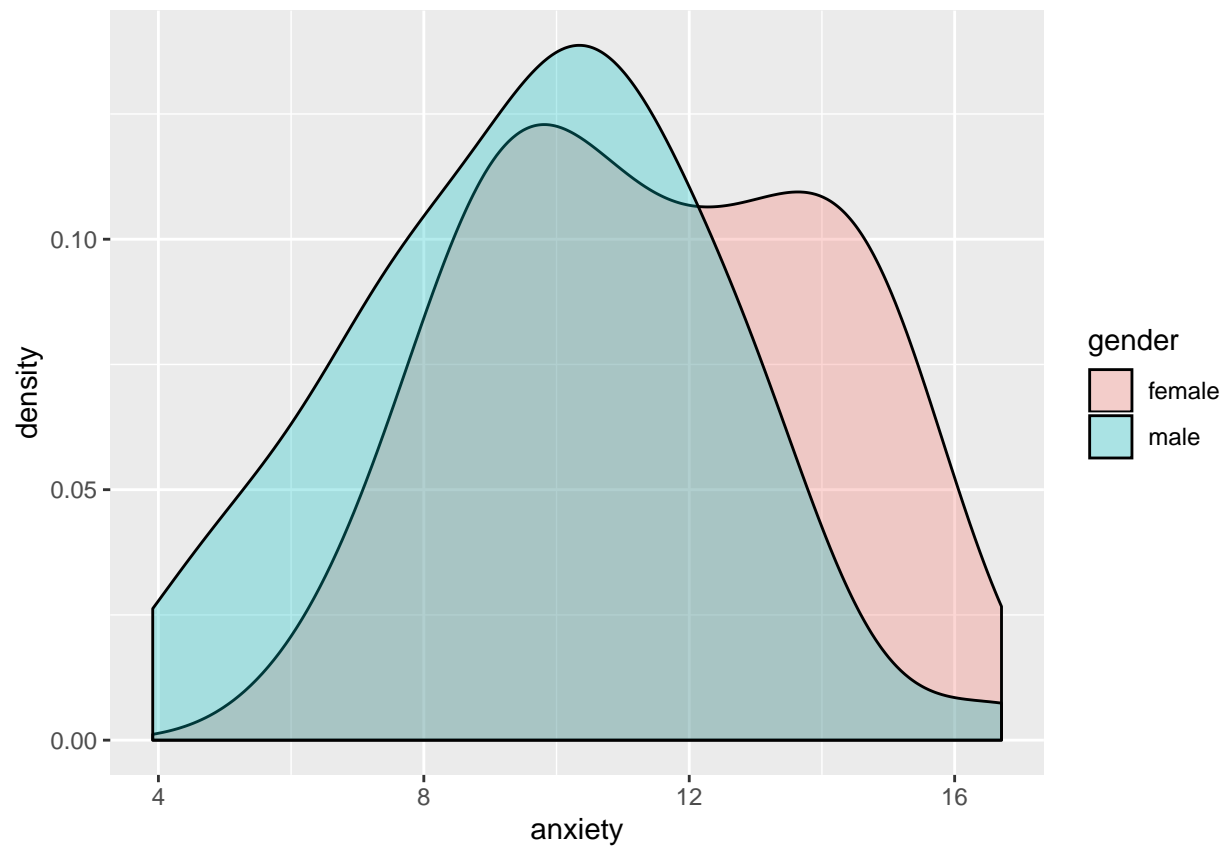
```
## # A tibble: 2 x 3
##   gender mean    sd
```

```
##   <fct> <dbl> <dbl>
## 1 female 11.5   2.58
## 2 male   9.64  2.70
```

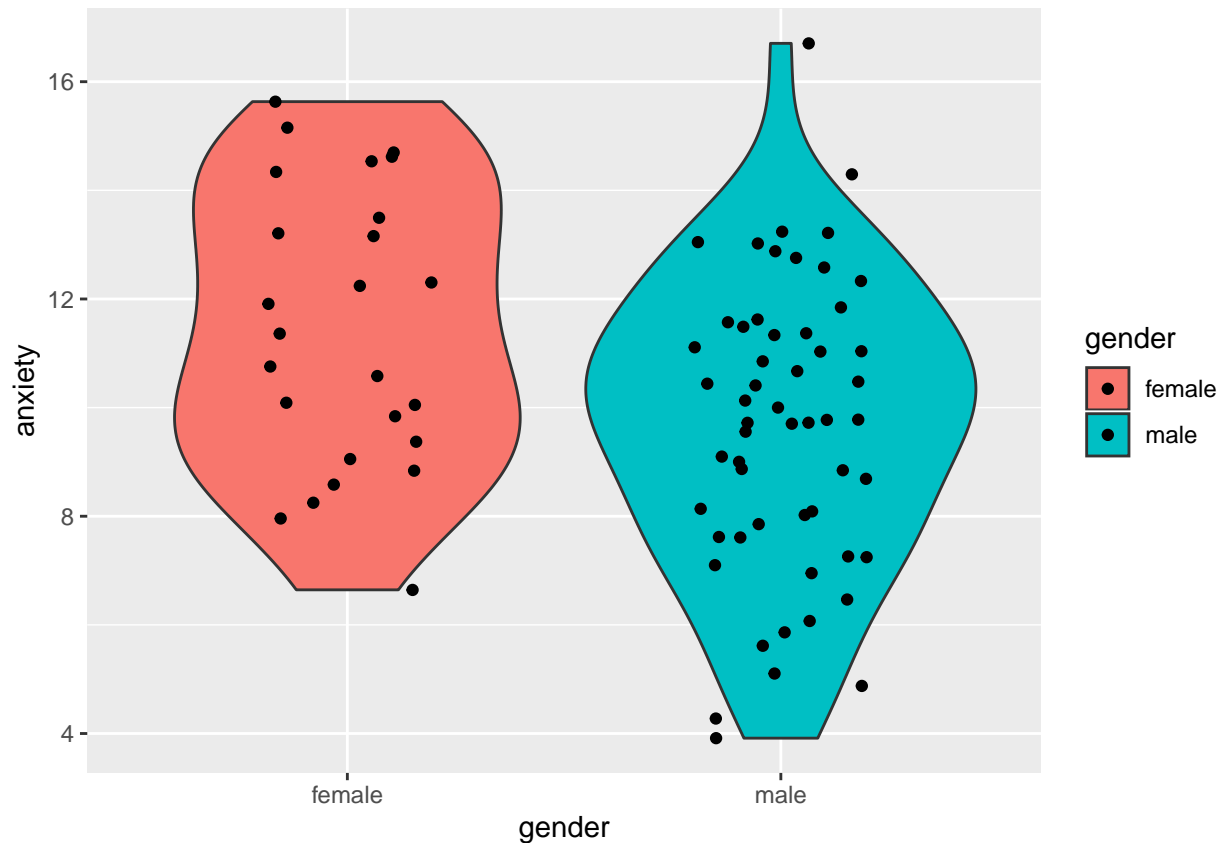
```
data %>%
  ggplot() +
    aes(x = gender, y = anxiety) +
    geom_boxplot()
```



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



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



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

Arra, hogy meghatározzuk van-e különbség két csoport között valamilyen numerikus változó átlagában, használhatjuk a t-tesztet, `t.test()`.

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

##
##  Welch Two Sample t-test
##
## data:  anxiety by gender
## t = 2.8896, df = 48.521, p-value = 0.005751
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5554417 3.0942134
## sample estimates:
## mean in group female    mean in group male
##           11.466392           9.641564

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

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

```
## 1      1.82
```

Az eredményt így írhatjuk le:

“A férfiak és nők szignifikánsan különböztek a szorongás szintjükben ($t = 2.89$, $df = 48.52$, $p = 0.006$). A csoportok szorongás szintjének átlaga és szórása a következő volt: nők: 11.47(2.58), férfiak: 9.64(2.7). A nők átlagosan 1.82 ponttal voltak szorongóbbak (95% CI = 0.56, 3.09).”

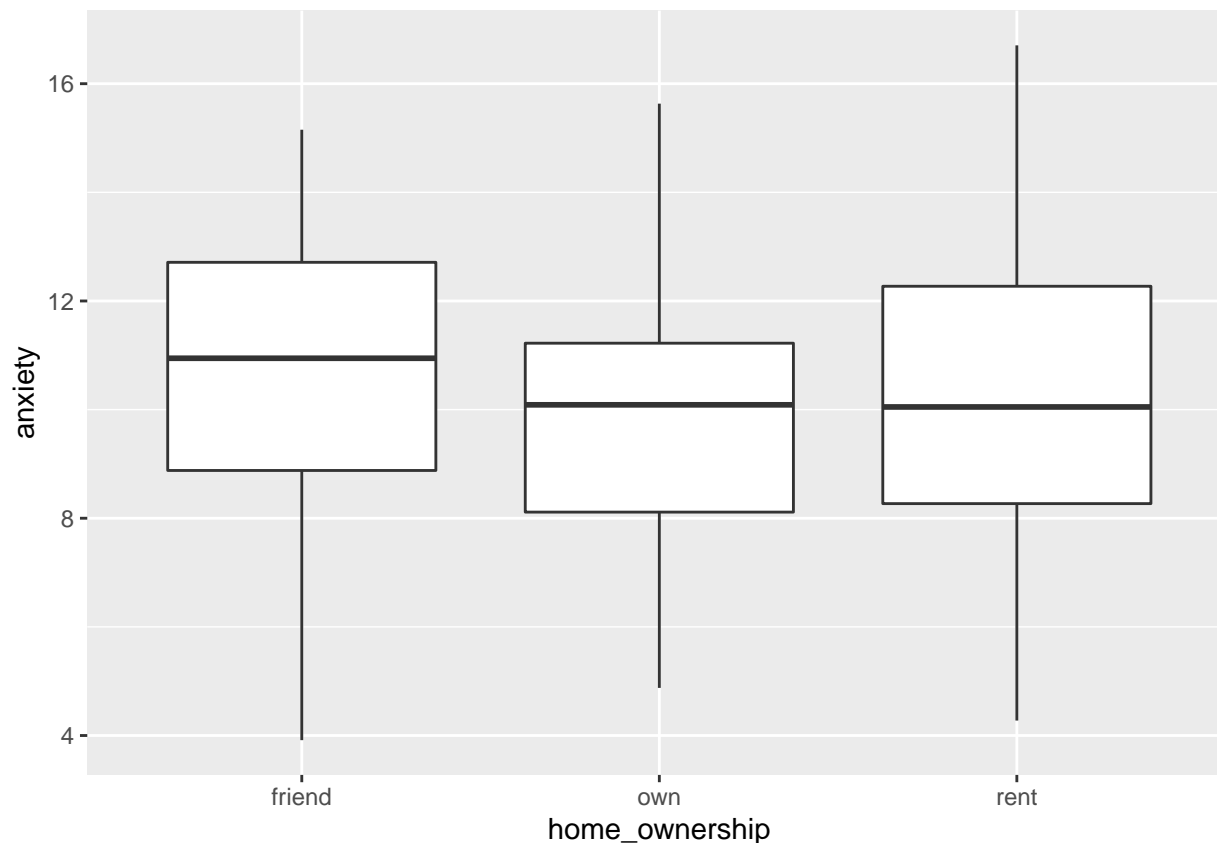
Ha egy kategorikus változón belül **három vagy több csoportunk** is van, a `t.test` nem használható. Helyette használhatjuk az **egyszempontos ANOVA**-t (one-way ANOVA) az `aov()` függővel. A formula ugyan úgy néz ki, mint a `t`-teszt esetén.

Mondjuk alább azt teszteljük, hogy van-e különbség a lakhatási helyzet csoportjai között a szorongásszintben. Itt valójában azt teszteljük, hogy páronként akarmelyik csoport között van-e szignifikáns különbség.

```
summary_home_ownership_vs_anxiety = data %>%
  group_by(home_ownership) %>%
  summarize(mean = mean(anxiety), sd = sd(anxiety))
summary_home_ownership_vs_anxiety
```

```
## # 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) +
  geom_boxplot()
```



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

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## home_ownership  2    7.5   3.768   0.481   0.62
## Residuals      77  603.3   7.835
```

Az eredményt így írhatjuk le:

“A lakhatási csoportonk szerint nem volt szignifikáns különbség a szorongás átlagos szintjében ($F(2, 77) = 0.48$, $p = 0.62$). A szorongás átlagát és szórását az egyes csoportok szerinti bontásban lásd az 1. táblázatban”

Alább látható, hogyan produkálnánk a megfelelő táblázatot a szorongás átlagával home_ownership csoportok szerint.

Egyoldalu vs. kétoldalu tesztek

Fontos, hogy ha van előzetes elképzelésünk a hipotézisalkotásról arról, hogy **milyen irányu** lesz a hatás, akkor **egy-oldalu (one-sided) tesztet** kell használnunk az alapértelmezett két-oldalu teszt helyett.

Például tegyük fel, hogy amikor a hipotézisünket meghatároztuk (ideális esetben ez még az adatgyűjtés előtt megtörténik), úgy gondoltuk, hogy a nőknek magasabb lesz a szorongásszintjük, mint a férfiaknak. Ezt az $\text{alternative} = \text{“greater”}$ paraméterrel határozhatjuk meg.

Ha összehasonlítjuk ezt az eredményt a korábbi t-teszt eredményével, észrevehetjük, hogy minden szám változatlan maradt, kivéve a **p-értéket, ami pontosan felére csökkent**, és a 95%-os konfidencia intervallumot, aminek a felső határa most egy végtelen nagy szám (∞).

A p-érték azért feleződött meg, mert azzal, hogy meghatároztuk, melyik irányban fog a két csoport különbözni egymástól, fele akkora lett az esélye, hogy a most megfigyelt, vagy annál nagyobb különbséget kapunk a

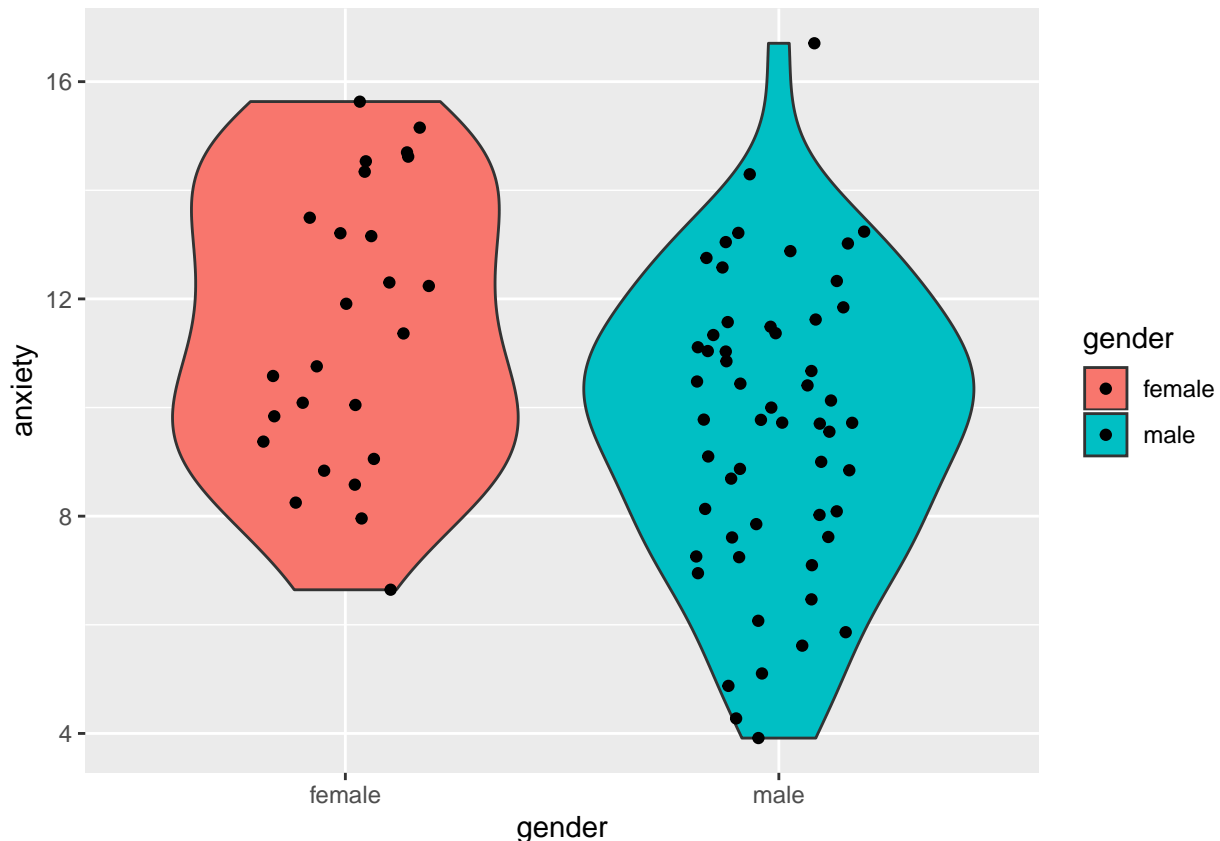
null-hipotezis helyesseget feltetelezve. Vagyis amikor tudjuk, milyen irányu hatást várunk el, mindig érdemes egy-oldali tesztet alkalmazni, mert ezzel nő a statisztikai erőnk.

Az egyoldali tesztek esetén amikor az a hipotézisünk, hogy **a referencia-csoport** átlaga magasabb lesz, (alternative = "greater"), akkor a konfidencia intervallumnak csak az alsó határt számoljuk ki. Ezért írja a teszt eredménye hogy a 95% CI 1.11, Inf, vagyis felfelé a végtelenségig tart a konfidencia intervallum.

```
summary = data %>%  
  group_by(gender) %>%  
  summarize(mean = mean(anxiety), sd = sd(anxiety))  
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, fill = gender) +  
    geom_violin() +  
    geom_jitter(width = 0.2)
```



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

```
##
```

```
## Welch Two Sample t-test
##
## data: anxiety by gender
## t = 2.8896, df = 48.521, p-value = 0.002876
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.7658658      Inf
## sample estimates:
## mean in group female    mean in group male
##      11.466392          9.641564
```

Az eredményt így írhatjuk le:

“A férfiak és nők szignifikánsan különböztek a szorongás szintjükben ($t = 2.89$, $df = 48.52$, $p = 0.003$). A csoportok szorongás szintjének átlaga és szórása a következő volt:” nők: 11.47(2.58), férfiak: 9.64(2.7). A nők átlagosan 1.82 ponttal voltak szorongóbbak (95% CI = 0.77, inf).”

Nezzük meg, mi történe, ha azt tippeltük volna a hipotézisalkotáskor, hogy a nőknek alacsonyabb lesz a szorongásszintjük. Ezt úgy határozhatjuk meg, hogy a `t.test()` funkcióban `alternative = "less"` paramétert allítunk be.

A p-érték itt majdnem eléri az 1-et, vagyis nagyon nagy a valószínűsége, hogy a null-hipotézis helyessegét feltételezve ilyen, vagy ennél extremerbb különbséget figyelünk meg. Nem is csoda, hiszen a null hipotézisünk itt az, hogy a nők szorongásának átlaga nem fog különbözni, vagy nagyobb lesz mint a férfiaké, és azt tapasztaltuk, hogy valóban nagyobb volt, vagyis a megfigyeles egyáltalán nem segít abban, hogy elutasítsuk a null-hipotézist.

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

```
##
## Welch Two Sample t-test
##
## data: anxiety by gender
## t = 2.8896, df = 48.521, p-value = 0.9971
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 2.883789
## sample estimates:
## mean in group female    mean in group male
##      11.466392          9.641564
```

Azt is érdemes megjegyezni, hogy a “greater” és a “less” mind a kategorikus változó **referencia-szintjére** vonatkozik. Ha ezt nem allítottuk be maskepp, pl. a `factor(levels =)` funkcióval, akkor a referencia-szint az ABC sorrendben előbb levő szint lesz. A fenti esetben a két szint a “female” és a “male”, amik közül a “female” jön előbb ABC sorrendben. Ha azt tippeltük volna, hogy az lenne a hipotézisünk, hogy a férfiak (“male”) szorongásszintje lesz magasabb, akkor `alternative = "less"`-t kellene beallítanunk, mert ezzel egyben azt tippeljük, hogy a referenciaszint (“female”) átlaga lesz az alacsonyabb. Vagy át kellene allítani a referenciaszintet.

Gyakorlás

Teszteld a 3. hipotézist, hogy “A terápiás csoportban alacsonyabb lesz a szorongás átlaga a kutatás végére mint a kontrol csoportban” (**anxiety vs. group**)

- Eloszor vegezzünk egy feltároló elemzést egy táblázattal a két változó kapcsolatáról a `summarize(mean(), sd())` funkciókkal, és készítsünk ábrát, mondjuk `geom_boxplot()` segítségével.

- egy- vagy ketoldalú tesztet kell alkalmaznunk? (gondolj arra, hogy a hipotézisünkben megjelöljük-e a hatas vagy különbseg irányát vagy sem)
- Mi a null-hipotézis ebben az esetben?
- Melyik tesztet érdemes használni, az egyváltozós ANOVA-t, vagy a t-tesztet? (gondolj arra, hogy hány csoport (szint) van a kategorikus változon belül)
- Ez után végezd el a tesztet
- Es írd le a fentiek szerint az eredményeket.

Két numerikus változó közötti kapcsolat, korreláció, `cor.test()`

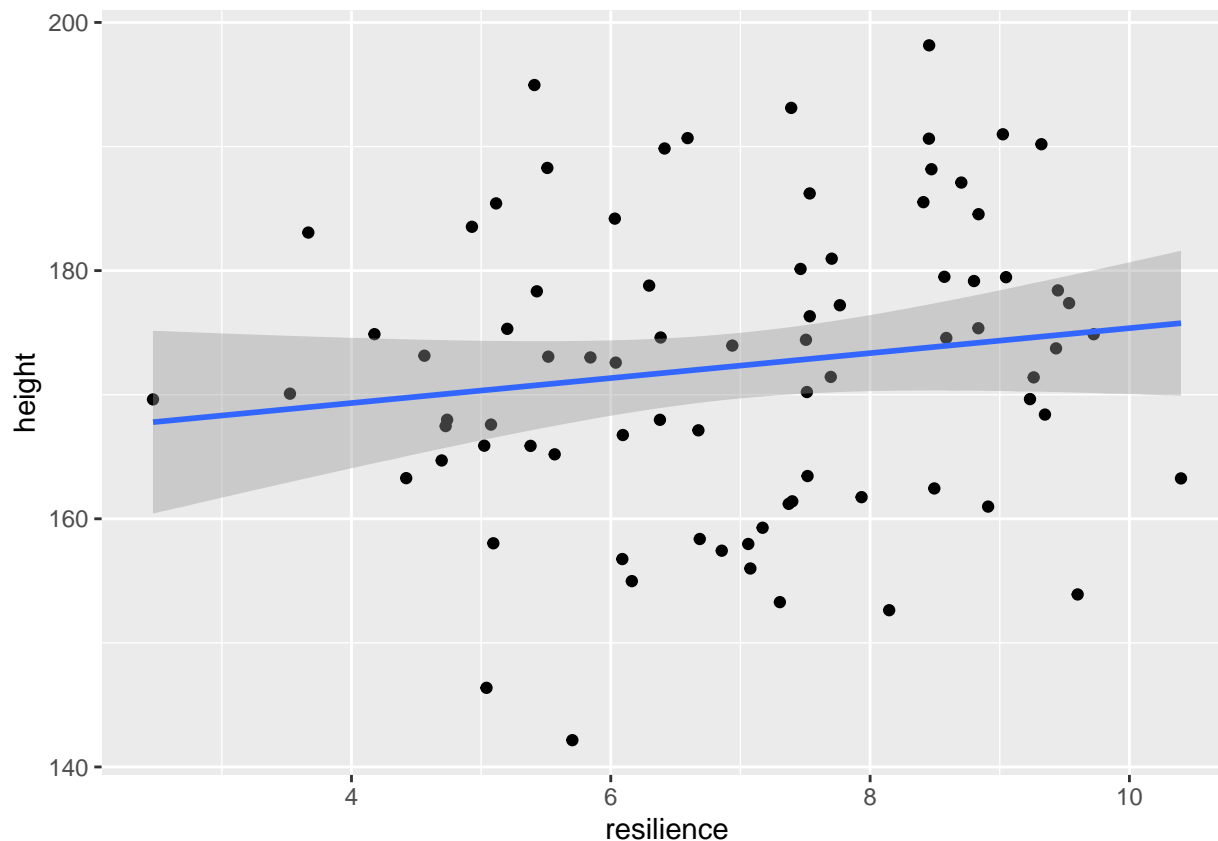
Vizsgáljuk meg, van-e együttjárás a reziliencia (**resilience**) és a magasság (**height**) között.

Eloszor végezzünk feltárási elemzést a korrelációs együttható kiszámításával, és egy pontdiagrammal. Használjunk `geom_point()` és `geom_smooth()` geomokat egyszerre, és használjuk az “lm” módszert a trendvonal megrajzolására.

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

##           resilience    height
## resilience 1.0000000 0.1471229
## height     0.1471229 1.0000000

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



A két változó függetlennek tűnik egymástól a feltárolt elemzés alapján, de elképzelhető, hogy a hatás, bármilyen kicsit is, mégis statisztikailag szignifikáns, szóval vegezzük el a statisztikai tesztet is.

Ezt a Pearson-korrelációs teszt segítségével tehetjük meg, `cor.test()` a következő képpen:

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

```
##
## Pearson's product-moment correlation
##
## data: data$resilience and data$height
## t = 1.3136, df = 78, p-value = 0.1928
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.07501901 0.35535288
## sample estimates:
## cor
## 0.1471229
```

Az eredményt így írhatjuk le:

“A reziliencia és a magasság között nem találtunk szignifikáns együttjárást ($r = 0.15$, 95% CI = -0.08, 0.36, $df = 78$, $p = 0.193$)”

Hasonlóan a t-teszthez, a korrelációs teszt esetében is érdemes egyoldalt tesztet használni amikor a hipotézisünk megmondja a kapcsolat irányát is, nem csak azt, hogy van kapcsolat a két változó között.

Például feltételezzük, hogy a két változó közötti kapcsolat pozitív lesz. Vagyis egy ember minél magasabb,

annal magasabb a rezilienciaja. Ezt úgy adhatjuk meg a statisztikai teszt specifikációjakor, hogy a formulához hozzátesszük az `alternative = "greater"` paramétert. Ha az eredményt összehasonlítjuk az előző korrelációs teszt eredményével, láthatjuk, hogy a p-érték is megváltozott. A konfidencia intervallumnak itt is csak az alsó határa érdekes, a felső határa a lehető legmagasabb értéket veszi fel ilyenkor, ami a korrelációs 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 = 1.3136, df = 78, p-value = 0.09641
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
## -0.03922992 1.00000000
## sample estimates:
## cor
## 0.1471229
```

Gyakorlás

Teszteld a 4. hipotézist, hogy “A reziliencia és a kutatás végen mért szorongásszint negatív összefüggést fog mutatni (vagyis aki reziliensebb, annál alacsonyabb szorongásszintet fognak mérni a kutatás végen)” (**anxiety vs. resilience**)

- Először vegezzünk egy feltárási elemzést a korrelációs együttható meghatározásával és egy pontdiagrammal a két változó kapcsolatáról.
- egy- vagy kétoldali tesztet kell alkalmaznunk? (gondolj arra, hogy a hipotézisünkben megjelöljük-e a hatás vagy különbség irányát vagy sem)
- Mi a null-hipotézis ebben az esetben?
- Ez után vedd el a tesztet
- És írd le a fentiek szerint az eredményeket.

A statisztikai tesztek eredményének közléséről általában

A statisztikai tesztek eredményének közlése során a következő információkat szoktuk megadni általánosságban. Ez tesztől tesztre változhat, de az alábbiak közül minél több információt megadnunk, annál jobb.

- az eredmény szöveges leírása
- teszt-statisztika
- szabadságfok (ez egyszerű teszteknel általában az elemszámmal is megadható)
- p-érték
- hatás mértéke (parameterbecslés)
- hatásmérték 95%-os konfidencia intervalluma