Inference for numerical data - exercises solution guide

E. Pastucha

October 2020

```
library(tidyverse)
```

```
## -- Attaching packages ------------ tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts --------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

1. Feeding chickens

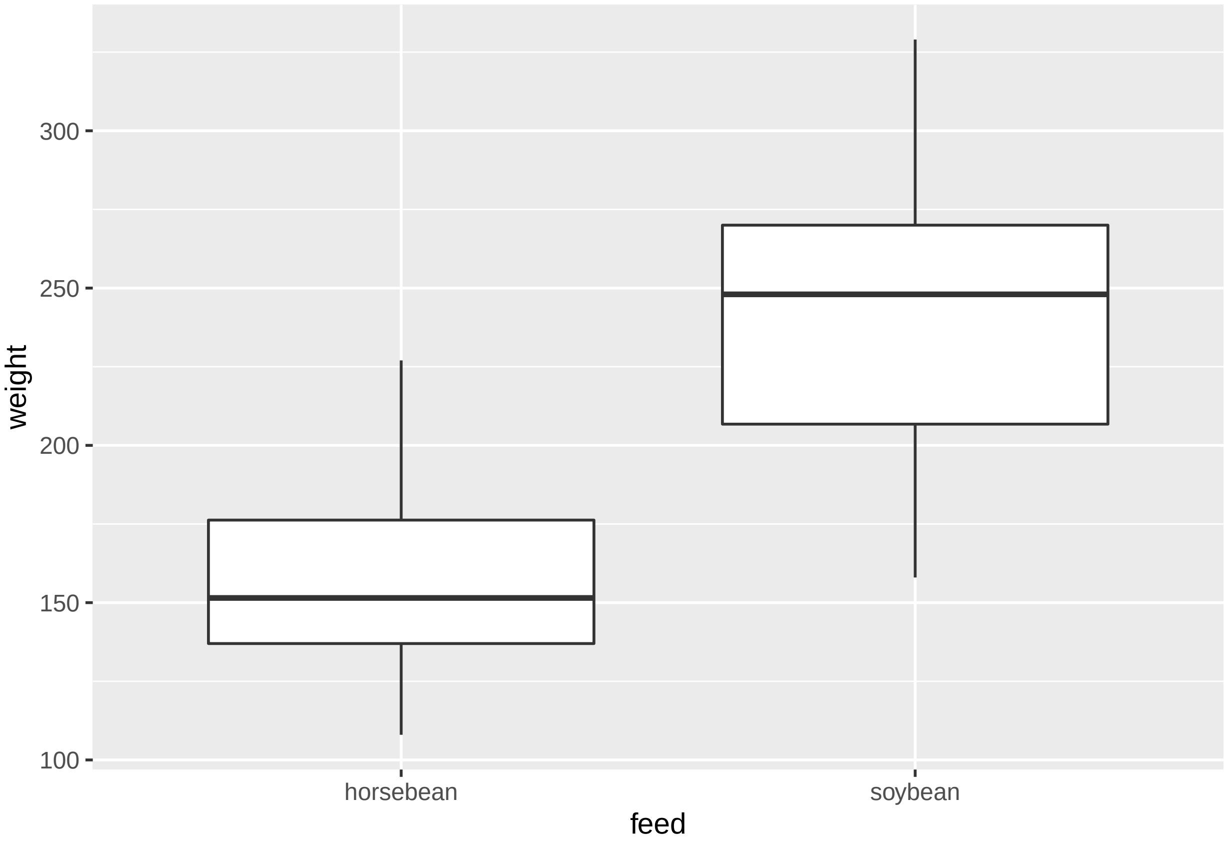
File ‘chickenweights.csv’ contains information about 36 chickens, their weight and primary food. Compare the two weight distributions and examine whether there is a significant difference in their average weight.

AD. 1

```
chicken_weights <- readr::read_csv('chickwts.csv')
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   weight = col_double(),
##   feed = col_character()
## )
```

```
ggplot(chicken_weights) +
  geom_boxplot(aes(feed, weight))
```



```
chicken_weights %>%
  group_by(feed) %>%
  summarise(mean_weight = mean(weight),
            sd_weight = sd(weight))
```

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

```
## # A tibble: 2 x 3
##   feed      mean_weight sd_weight
##   <chr>           <dbl>     <dbl>
## 1 horsebean        160.      38.6
## 2 soybean          246.      54.1
```

AD.2

A difference of means t-test is used to test whether the difference in average weights from the two distributions are significant.

AD.3

We assume that observations are independent - but if we were to conduct the experiment ourself we would have to pay attention to living conditions, lighting conditions, etc. so that no other thing influences chicken weight.

AD.4

H0: There is no difference in between average chicken weights depending on the food.

HA: There is significant difference in between average chicken weights depending on the food.

AD.5

```
testresult <- t.test(weight ~ feed, data = chicken_weights)
testresult
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by feed
## t = -4.5543, df = 21.995, p-value = 0.0001559
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -125.49476  -46.96238
## sample estimates:
## mean in group horsebean   mean in group soybean
##                160.2000                246.4286
```

AD.6

A p value of 0.0001559 is found. This value is lower than typical used significance levels (eg 0.05), so the difference is significant.

2. Fireworks

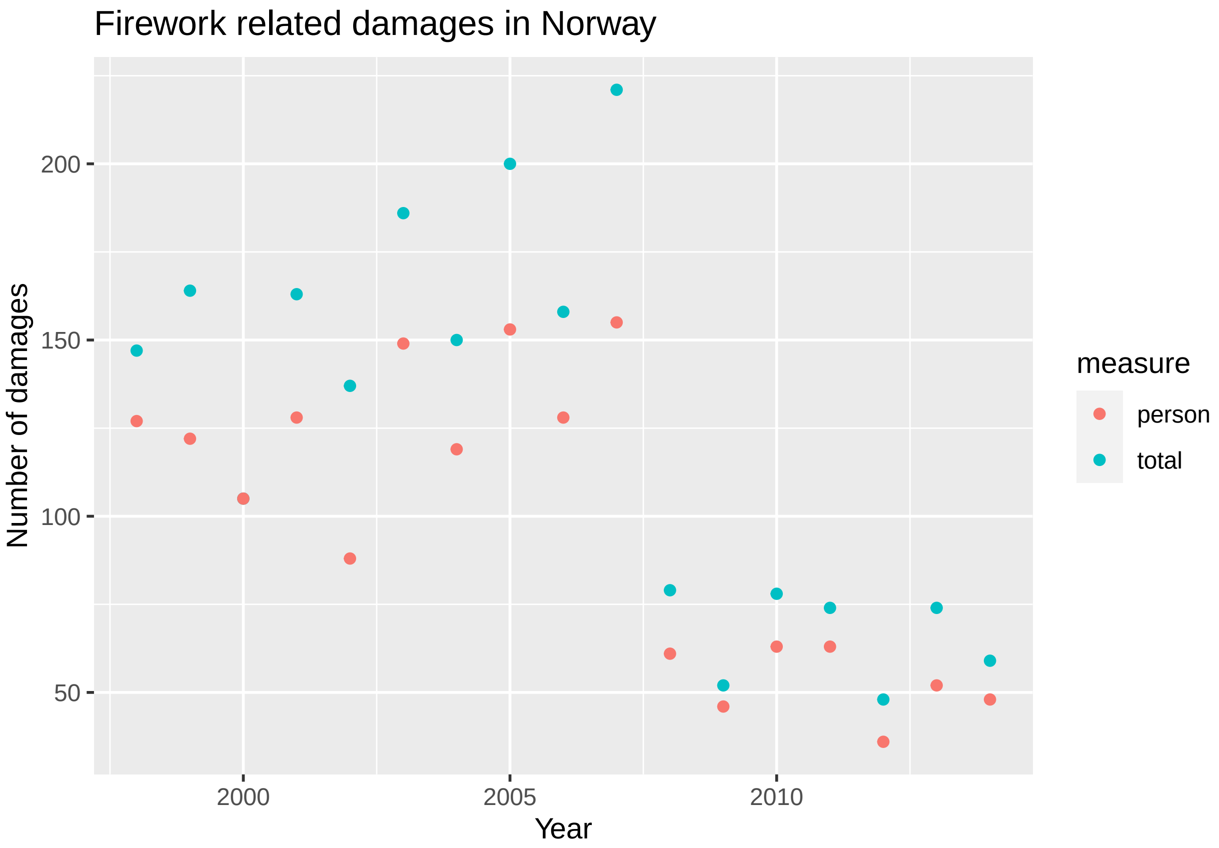
The file *fireworks.csv* contains data about fireworks damages in Norway. Use a t.test to compare the total number of damages per year for the two groups: group 1: years prior to and including 2007, group 2: years after 2007.

AD. 1

```
fireworks <- readr::read_delim('fireworks.csv', delim='\t')
```

```
## Parsed with column specification:
## cols(
##   measure = col_character(),
##   year = col_double(),
##   count = col_double()
## )
```

```r
ggplot(fireworks) +
  geom_point(aes(year, count, color=measure)) +
  labs(x = 'Year',
       y = 'Number of damages',
       title = 'Firework related damages in Norway')
```



AD.2

A difference of means t-test is used to test whether the difference in average total damages per year from the two distributions are significant.

AD.4

H0: There is no difference in between average total damages per year before and after 2007.

HA: There is significant difference in between average total damages per year before and after 2007

AD.3

We might raise a question whether experiments are independent, cause ther might be a lot of different things influencing number of damages. Also, we would expect most damages to occure at the end of December and beggining of January, so maybe different date generalization would be usefull.

AD.5

```
fireworks %>% filter(measure == 'total') %>%
  t.test(count ~ year > 2007, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  count by year > 2007
## t = 8.4139, df = 12.512, p-value = 1.675e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   71.85702 121.77155
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           163.10000            66.28571
```

AD.6

A p value of 1.675e-06 is found. This value is lower than typical used significance levels (eg 0.05), so the difference is significant.

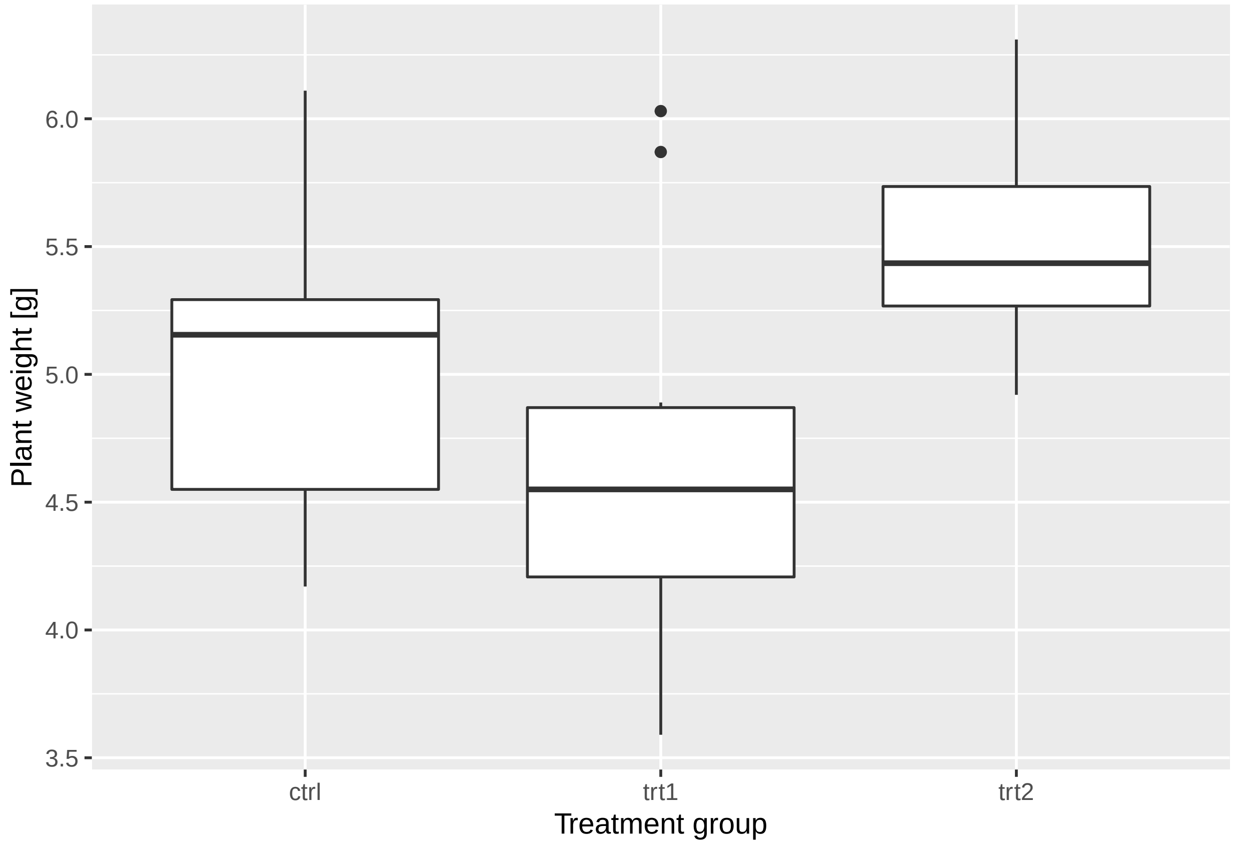
3. Plant growth

The file: *plantgrowth.csv* contains results of an experiment where plants were treated with different chemical compunds. All you're provided is category: treatment or control group, and weight of a plant. Choose a suitable statistical test and use that to determine whether there is a significant difference in the mean plant weight for the control group *ctrl* and the first treatment group *trt1*. ## AD. 1

```
plant_growth <- readr::read_csv('plantgrowth.csv')
```

```
## Parsed with column specification:
## cols(
##   weight = col_double(),
##   group = col_character()
## )
```

```
plant_growth %>%
  ggplot() +
  geom_boxplot(aes(group, weight)) +
  labs(x = "Treatment group",
       y = "Plant weight [g]")
```



## AD.2

A difference of means t-test is used to test whether the difference in average weight of plants within ctrl and tr1 groups.

AD.4

H0: There is no difference in between average weight of plants within ctrl and tr1 groups.

HA: There is significant difference in between weight of plants within ctrl and tr1 groups.

AD.3

We assume that observations are independent - but if we were to conduct the experiment ourself we would have to pay attention to all conditions, so that no other thing influences plant growth.

AD.5

```
plant_growth %>%
  filter(group %in% c("ctrl", "trt1")) %>%
  t.test(weight ~ group, data = .)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  weight by group
## t = 1.1913, df = 16.524, p-value = 0.2504
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2875162  1.0295162
## sample estimates:
## mean in group ctrl mean in group trt1
##              5.032              4.661
```

AD.6

A p value of 0.2504 is found. This value is higher than typical used significance levels (eg 0.05), so the difference is insignificant.

4. Major League Baseball players

The file *major_league_baseball_players.csv* contains information about major league baseball players.

a) For all player types (positions), calculate the average and standard deviation of the weight. Sort the list according to the average weight.

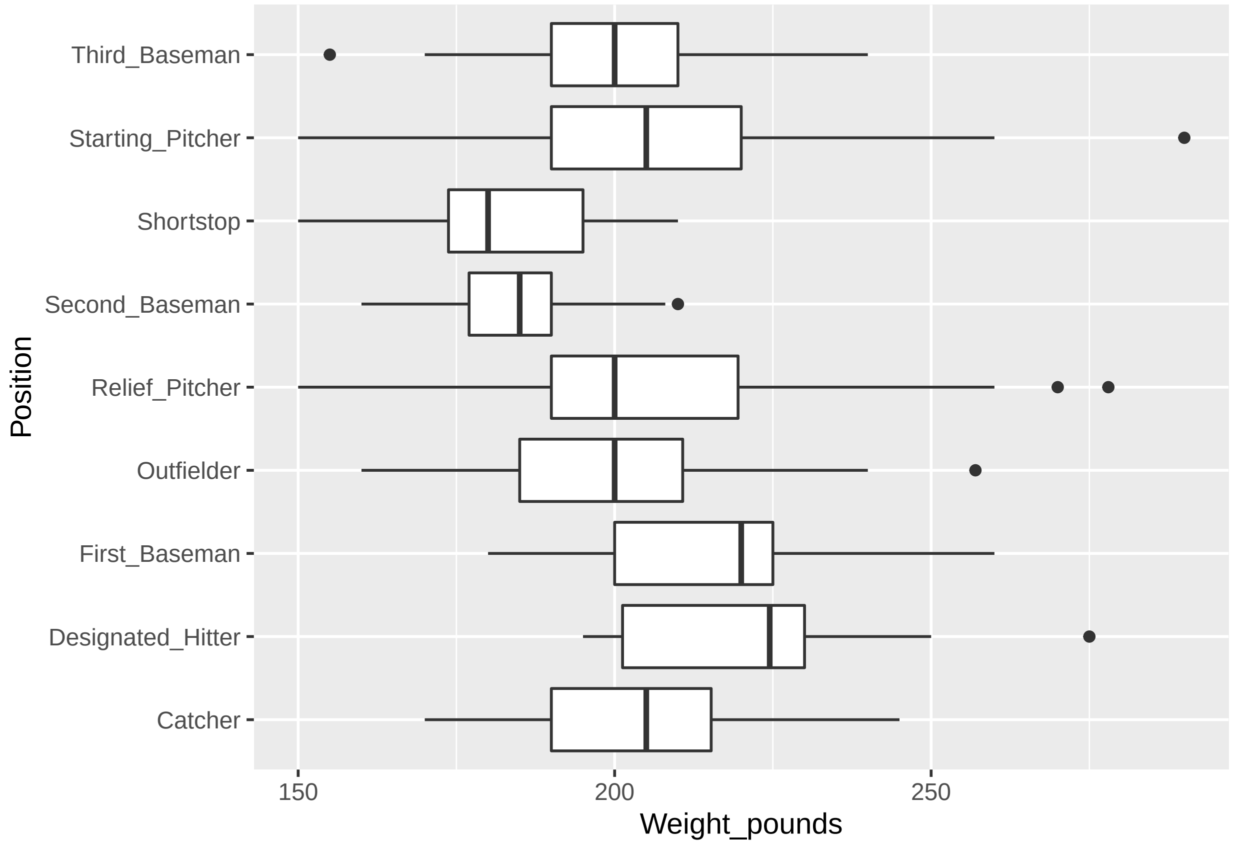
b) Use t-test to compare the average weight of a *Designated_Hitter* and a *Third_Baseman*

AD.1

```
mlb_players <- readr::read_delim('major_league_baseball_players.csv', delim = '\t')
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   Team = col_character(),
##   Position = col_character(),
##   Height_inches = col_double(),
##   Weight_pounds = col_double(),
##   Age = col_double()
## )
```

```
mlb_players %>%
  ggplot() +
  geom_boxplot(aes(Position, Weight_pounds)) +
  coord_flip()
```



```r
mlb_players %>%
  group_by(Position) %>%
  summarise(mean = mean(Weight_pounds),
            sd = sd(Weight_pounds),
            n = n()) %>%
  arrange(mean) %>%
  knitr::kable()
```

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

Position	mean	sd	n
Shortstop	182.9231	14.49793	52
Second_Baseman	184.3448	11.01124	58
Outfielder	199.1134	18.43761	194
Third_Baseman	200.9556	18.13955	45
Relief_Pitcher	203.5175	21.73719	315
Catcher	204.3289	15.74326	76
Starting_Pitcher	205.1636	22.11263	220
First_Baseman	213.1091	19.02695	55
Designated_Hitter	220.8889	22.11970	18

AD.2

A difference of means t-test is used to compare the average weight of *Designated_Hitter* and a *Third_Baseman*

AD.4

H0: There is no difference in between average weight of *Designated_Hitter* and a *Third_Baseman*

HA: There is significant difference in between weight of *Designated_Hitter* and a *Third_Baseman*

AD.3

We could possibly assume that the observations are independent. But there is a possibility that a whole team was put on some kind of food regime :)

AD.5

```r
mlb_players %>%
  filter(Position %in% c('Designated_Hitter', 'Third_Baseman')) %>%
  t.test(Weight_pounds ~ Position, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  Weight_pounds by Position
## t = 3.394, df = 26.632, p-value = 0.00217
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.874739 31.991927
## sample estimates:
## mean in group Designated_Hitter     mean in group Third_Baseman
##                        220.8889                        200.9556
```

AD.6

A p value of 0.00217 is found. This value is lower than typical used significance levels (eg 0.05), so the difference is significant.

5. Nobel prizes

Data: *nobel_laureates.csv*

Choose a suitable statistical test and use that to determine whether there is a significant difference in the mean number of people sharing the award for Chemistry and the mean number of people sharing the award for Peace (for all available years).

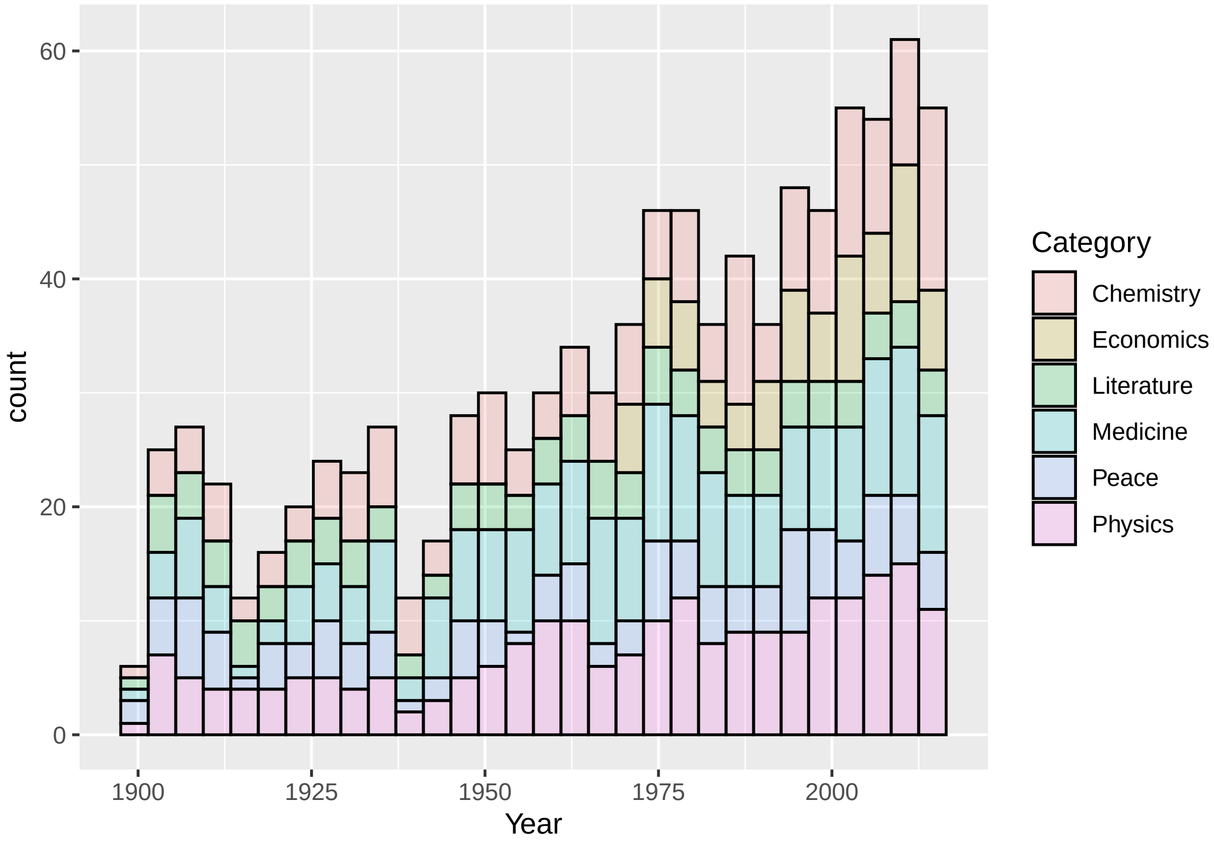
AD. 1

```
nobel <- readr::read_csv('nobel_laureates.csv')
```

```
## Parsed with column specification:
## cols(
##   Year = col_double(),
##   Category = col_character(),
##   Prize = col_character(),
##   `Laureate ID` = col_double(),
##   `Laureate Type` = col_character(),
##   `Full Name` = col_character(),
##   `Birth Date` = col_date(format = ""),
##   `Birth City` = col_character(),
##   `Birth Country` = col_character(),
##   Sex = col_character(),
##   `Organization Name` = col_character(),
##   `Organization City` = col_character(),
##   `Organization Country` = col_character(),
##   `Death Date` = col_date(format = ""),
##   `Death City` = col_character(),
##   `Death Country` = col_character()
## )
```

```
nobel %>%
  ggplot()+
  geom_histogram(mapping = aes(x = Year, fill = Category),
                 colour = 'black',
                 alpha = 0.2)
```

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



AD.2

A difference of means t-test is used to test whether there is a difference in average number of people sharing the award for Chemistry and the mean number of people sharing the award for Peace.

AD.4

H0: There is no difference in between average number of people sharing the award for Chemistry and the mean number of people sharing the award for Peace.

HA: There is significant difference in between number of people sharing the award for Chemistry and the mean number of people sharing the award for Peace.

AD.3

We could assume that observations are independent - the comission was just each year (?).

AD.5

```
nobel %>%
  filter(Category %in% c('Chemistry', 'Peace')) %>%
  group_by(Year, Category) %>%
  tally() %>% head(5)
```

```
## # A tibble: 5 x 3
## # Groups:   Year [3]
##    Year Category      n
##   <dbl> <chr>     <int>
## 1  1901 Chemistry     1
## 2  1901 Peace         2
## 3  1902 Chemistry     1
## 4  1902 Peace         2
## 5  1903 Chemistry     1
```

```
nobel %>%
  filter(Category %in% c('Chemistry', 'Peace')) %>%
  group_by(Year, Category) %>%
  tally() %>%
  t.test(n ~ Category, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  n by Category
## t = 4.0099, df = 159.96, p-value = 9.299e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2314606 0.6807196
## sample estimates:
## mean in group Chemistry     mean in group Peace
##                1.796296                1.340206
```

AD.6

A p value of 9.299e-05 is found. This value is lower than typical used significance levels (eg 0.05), so the difference is significant.