Session 14 - *t-test* and ANOVA

Alex Mounsey

18/11/2020

Exercise 1

A new computer software package has been developed to help systems analysts reduce the time required to design, develop, and implement an information system. To evaluate the benefits of the new software package, a random sample of twenty-four system analysts is selected. Each analyst is given specifications for an information system. Twelve of the analysts are instructed to produce the information system by using current technology. The other twelve analysts are trained in the use of a new software package and then instructed to use it to produce the information system. The completion times in hours for the analysts to produce the information system is provided in this table:

Table 1: Completion Times in Hours

Current Technology	New Software
300	274
280	220
344	308
385	336
372	198
360	300
288	315
321	258
376	318
290	310
301	332
283	263

```
ct_data <- read.csv('../data/completion_time.csv') %>%
  rename('Current Technology' = Current.Technology,
         'New Software' = New.Software)  # Remove '.' in column headers
```

Does the new software provide a *statistically significant* different mean project completion time? This means: is there an underlying difference in mean project completion time for analysts using the current technology and analysts using the new software. Answer the question following these steps:

Using the `gather()` function from the `tidyr` package, transform the data into a more convenient format:

```
ct_long <- ct_data %>%
  gather('Software', 'Completion_Time', 1:2)

head(ct_long)
```

```
##             Software Completion_Time
## 1 Current Technology             300
## 2 Current Technology             280
## 3 Current Technology             344
## 4 Current Technology             385
## 5 Current Technology             372
## 6 Current Technology             360
```

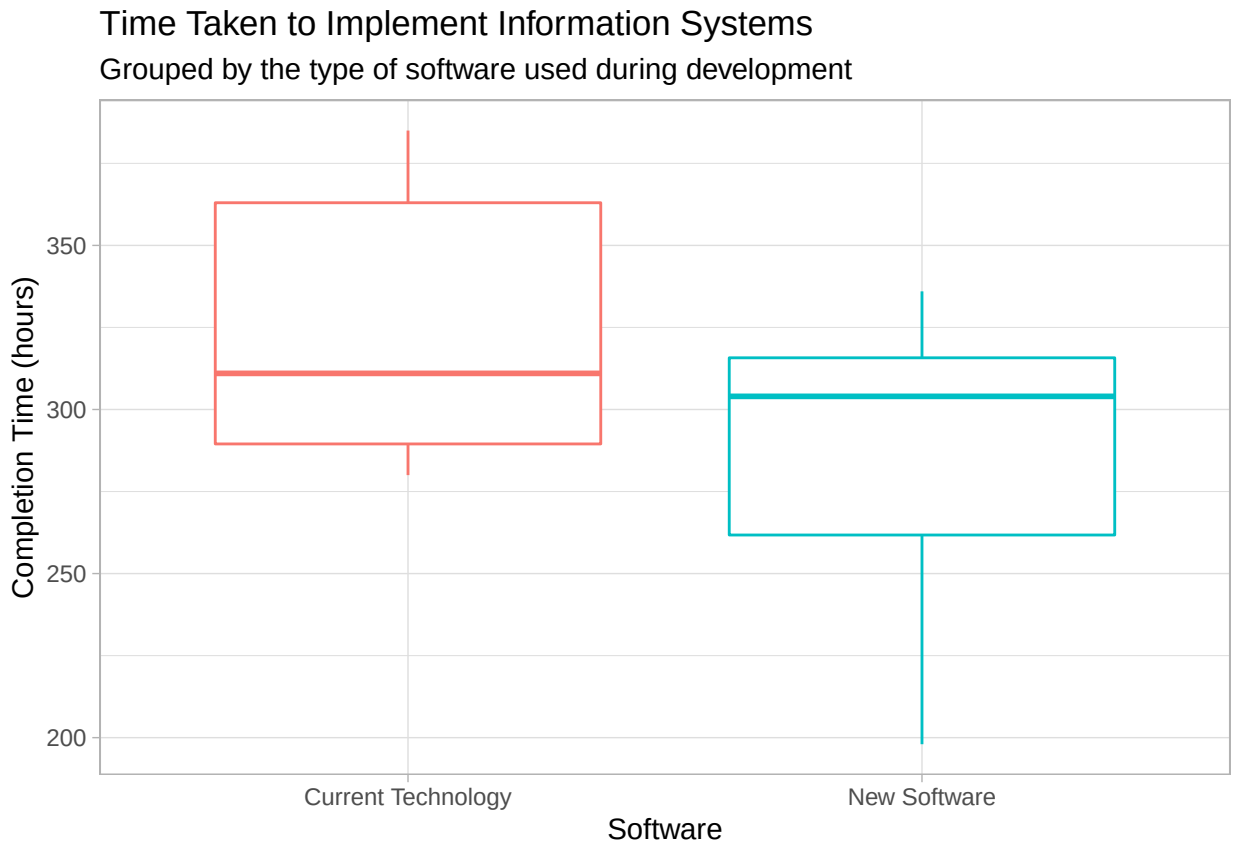
Calculate the sample mean and standard deviation for both types of software:

```
ct_long %>%
  group_by(Software) %>%
  summarise(mean = mean(Completion_Time), standard_deviation = sd(Completion_Time))
```

```
## # A tibble: 2 x 3
##   Software            mean standard_deviation
##   <chr>              <dbl>              <dbl>
## 1 Current Technology   325               40.0
## 2 New Software         286               44.0
```

Produce boxplots of the completion time for both types of software:

```
ct_long %>%
  group_by(Software) %>%
  ggplot(aes(x = Software, y = Completion_Time, colour=Software)) +
    theme_light() + geom_boxplot() +
    labs(x = "Software", y = "Completion Time (hours)",
         title = "Time Taken to Implement Information Systems",
         subtitle = "Grouped by the type of software used during development") +
    theme(legend.position = 'none')
```



Does the new software provide a statistically significant different mean projcet completion time? Use a *t-test*, assuming that all the test assumptions are met:

```
t.test(Completion_Time ~ Software, data = ct_long, var.equal = T)
```

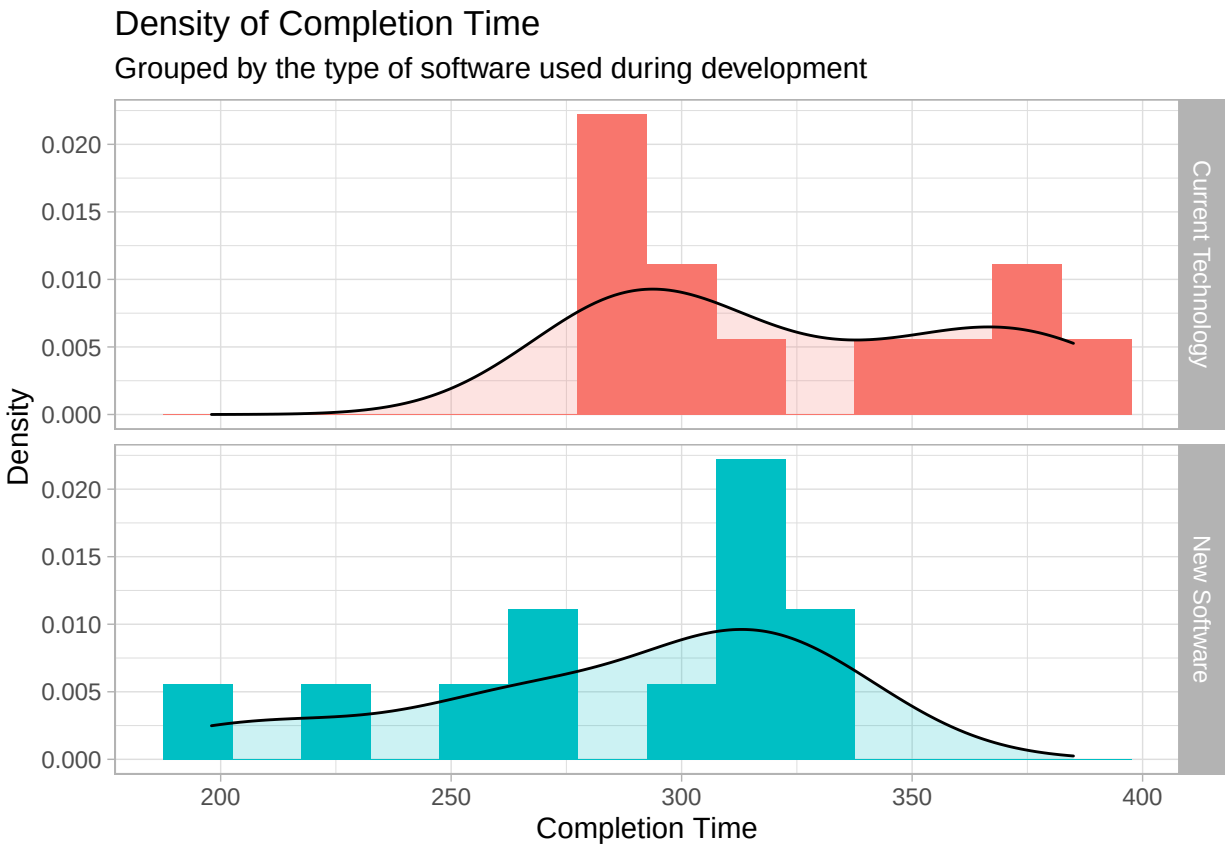
```
##
##  Two Sample t-test
##
## data:  Completion_Time by Software
## t = 2.2721, df = 22, p-value = 0.0332
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.402936 74.597064
## sample estimates:
## mean in group Current Technology       mean in group New Software
##                              325                              286
```

Comments:

The *p-value* (0.0332044) is less than `0.05`. Therefore we can reject the null hypothesis *(i.e. that there is no difference)* and conclude that there is a relationship between the type of software used and project completion times.

Now, check the *t-test* assumptions:

```
ct_long %>%
  group_by(Software) %>%
  ggplot(aes(x = Completion_Time, fill = Software)) +
    theme_light() + facet_grid(Software ~ .) +
    geom_histogram(aes(y=..density..), binwidth = 15) +
    geom_density(alpha = 0.2) +
    labs(x = "Completion Time", y = "Density",
         title = "Density of Completion Time",
         subtitle = "Grouped by the type of software used during development") +
    theme(legend.position = 'none')
```




```
curr_tech <- ct_long %>%
  filter(Software == 'Current Technology')

shapiro.test(curr_tech$Completion_Time)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  curr_tech$Completion_Time
## W = 0.8725, p-value = 0.07031
```

```
new_soft <- ct_long %>%
  filter(Software == 'New Software')
```

```
shapiro.test(new_soft$Completion_Time)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_soft$Completion_Time
## W = 0.9034, p-value = 0.1755
```

Comments:

The *p-value* for both the new software (0.1754761) and current technology (0.0703102) are *above* the threshold of `0.05`. Thus we can accept the null hypothesis.

Testing for a shorter mean project completion time, rather than a *different* one

It's also possible to answer a more relevant question: **"Does the new software provide a statistically significant *shorter* mean project completion time?"** In order to answer this, we must provide an additional argument to the `t.test()` function: `alternative = 'greater'`.

```
t.test(Completion_Time ~ Software, data = ct_long, var.equal = TRUE,
       alternative = 'greater')
```

```
##
##  Two Sample t-test
##
## data:  Completion_Time by Software
## t = 2.2721, df = 22, p-value = 0.0166
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  9.526018      Inf
## sample estimates:
## mean in group Current Technology       mean in group New Software
##                              325                              286
```

What do you conclude from the new *p-value* of 0.0166022?

Comments:

The *p-value* (0.0166022) is less than the threshold of `0.05`, which suggests that we should reject the null hypothesis.

You may be wondering why, when we are testing for a significantly *shorter* mean project completion time, we specify `alternative = 'greater'`. This is because the groups are ordered alphabetically as "Current Technology" and "New Software".

We want to test whether the new software provides a statistically significant *shorter* mean project completion time than the current technology. Using the specified group order, our question would be: **"is the underlying mean project completion time associated with the *new software*?"** Of course, if we were to specify the levels of the factor `Software` in the opposite order, we would replace `alternative = 'greater'` with `alternative = 'less'`, giving the same *p-value*:

```r
ct_long_2 <- ct_long %>%
  mutate(Software_f = factor(Software, levels = c("New Software", "Current Technology")))

t.test(Completion_Time ~ Software_f, data = ct_long_2, var.equal = TRUE,
       alternative = 'less')
```

```
##
##  Two Sample t-test
##
## data:  Completion_Time by Software_f
## t = -2.2721, df = 22, p-value = 0.0166
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -9.526018
## sample estimates:
##       mean in group New Software mean in group Current Technology
##                              286                               325
```

Exercise 2

Stress has become a large problem in today's workplace. In a study designed to measure stress, 15 property agents, 15 architects, and 15 stockbrokers were selected at random, and their stress levels were measured using an established continuous scale which takes into account issues such as ambiguity and role conflict. Higher values indicate a higher degree of stress.

Table 2: Stress Levels

Property Agent	Architect	Stockbroker
86	43	65
53	63	48
73	60	57
74	52	91
59	54	70
67	77	67
81	68	83
61	57	75
66	61	53
70	80	71
69	50	54
74	37	72
88	73	65
90	84	58
80	58	58

```r
stress_data <- read.csv('../data/stress.csv') %>%
  rename('Property Agent' = Property.Agent)
```

Transform this data into a more convenient format:

```r
sd_long <- stress_data %>%
  gather('Profession', 'Stress', 1:3)

head(sd_long)
```

```
##       Profession Stress
## 1 Property Agent     86
## 2 Property Agent     53
## 3 Property Agent     73
## 4 Property Agent     74
## 5 Property Agent     59
## 6 Property Agent     67
```

Calculate the overall sample mean and standard deviation:

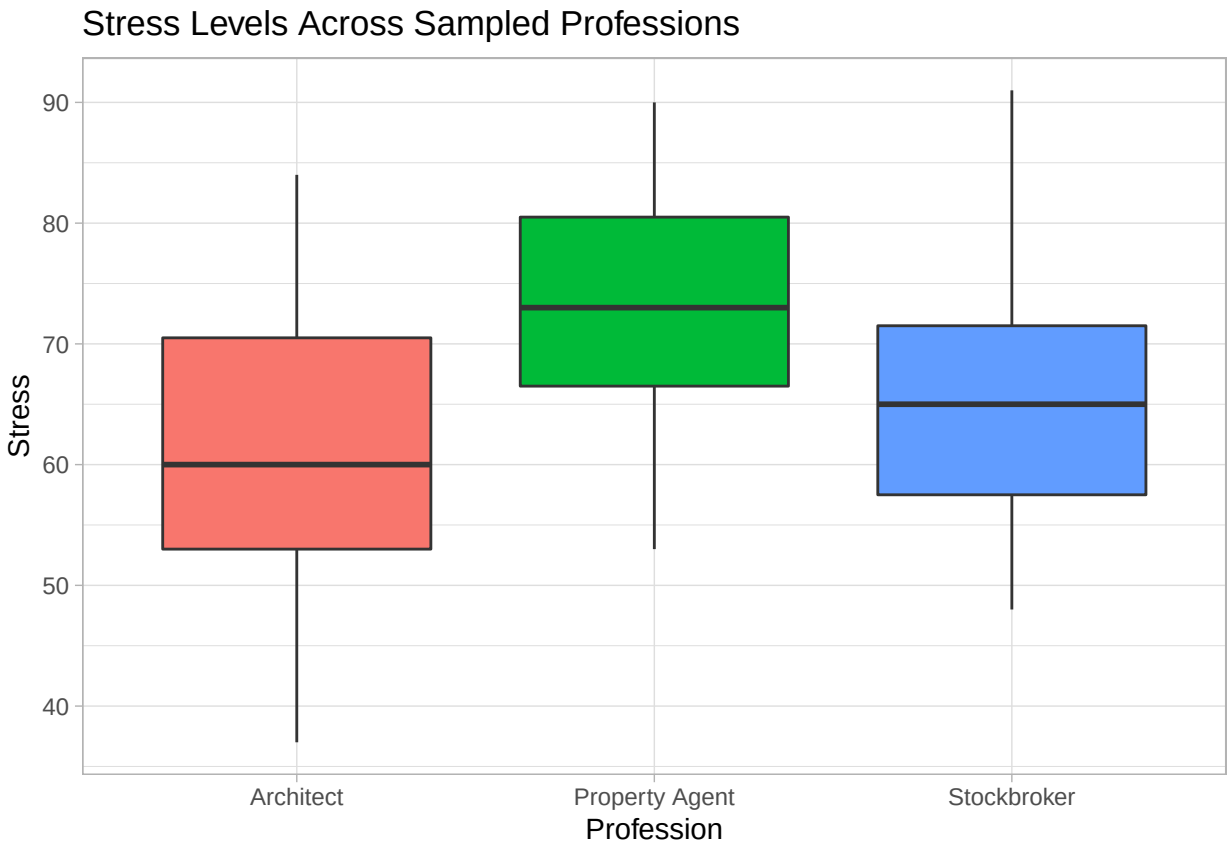
```r
sd_long %>%
  group_by(Profession) %>%
  summarise(mean = mean(Stress), standard_deviation = sd(Stress))
```

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

```
## # A tibble: 3 x 3
##   Profession      mean standard_deviation
##   <chr>          <dbl>              <dbl>
## 1 Architect       61.1               13.4
## 2 Property Agent  72.7               10.8
## 3 Stockbroker     65.8               11.7
```

Produce boxplots of the stress levels for each profession

```r
sd_long %>%
  group_by(Profession) %>%
  ggplot(aes(x = Profession, y = Stress, fill = Profession)) +
    theme_light() + geom_boxplot() +
    labs(x = "Profession", y = "Stress",
         title = "Stress Levels Across Sampled Professions") +
    theme(legend.position = 'none')
```


Stress Levels Across Sampled Professions
90
80
70
60
50
40
Stress
Architect
Property Agent
Stockbroker
Profession