Undergrad Biostatistics - R Training - Week IV

Ege Ulgen

R tips

- If you can't figure out how to solve an issue, Google is your friend. e.g., "how to calculate mode r"
- If you need help with the usage of a function, type `?function_name`. e.g. `?quantile`
- If you get an error, and cannot fix it. C/P the error into Google. Someone else most likely had a similar problem
- Some resources for learning the basic syntax of R;
  - Codecademy - https://www.codecademy.com/learn/learn-r
  - RStudio Cloud Primers - https://rstudio.cloud/learn/primers
  - Dataquest - https://www.dataquest.io/course/introduction-to-data-analysis-in-r/
  - R for Data Science - https://r4ds.had.co.nz/index.html
- Interesting read: https://www.dataquest.io/blog/learn-r-for-data-science/

Confidence Interval Example

Prepare data

We'll read in the AIDS data as follows:

```
aids_df <- read.delim("../data/aids_dataset.txt", sep = " ")
```

We can take a look at the first 6 rows via `head()`:

```
head(aids_df)
```

```
##   id treatment   age gender week_1 cd4_1 week_2 cd4_2
## 1  1      trt2 36.43   male      0    23   7.57    21
## 2  2      trt4 47.85   male      0    21   8.00    49
## 3  4      trt3 36.60   male      0    61   7.14    61
## 4  5      trt1 35.95   male      0    36   8.00    31
## 5  6      trt2 38.40   male      0    11   7.29    11
## 6  7      trt2 45.08   male      0    11   9.00    41
```

We will use the first 10 patients (first 10 rows):

```
sub_df <- aids_df[1:10, ]
```

We'll define `perc_benefit` and add it to the data frame:

```
sub_df$perc_benefit <- (sub_df$cd4_2 - sub_df$cd4_1) / sub_df$cd4_1 / (sub_df$week_2 - sub_df$week_1) *
sub_df
```

```
##    id treatment   age gender week_1 cd4_1 week_2 cd4_2 perc_benefit
## 1   1      trt2 36.43   male      0    23   7.57    21     -1.14870
## 2   2      trt4 47.85   male      0    21   8.00    49     16.66667
## 3   4      trt3 36.60   male      0    61   7.14    61      0.00000
## 4   5      trt1 35.95   male      0    36   8.00    31     -1.73611
## 5   6      trt2 38.40   male      0    11   7.29    11      0.00000
## 6   7      trt2 45.08   male      0    11   9.00    41     30.30303
## 7   8      trt3 37.20   male      0    16   7.71    11     -4.05318
## 8  11      trt2 42.25   male      0    16   4.14    21      7.54831
## 9  12      trt4 31.46   male      0    46  16.14    51      0.67346
## 10 13      trt4 41.86   male      0     1  17.00     1      0.00000
```

Calculate necessary values

Recall that the general formula for the confidence interval of the mean is:

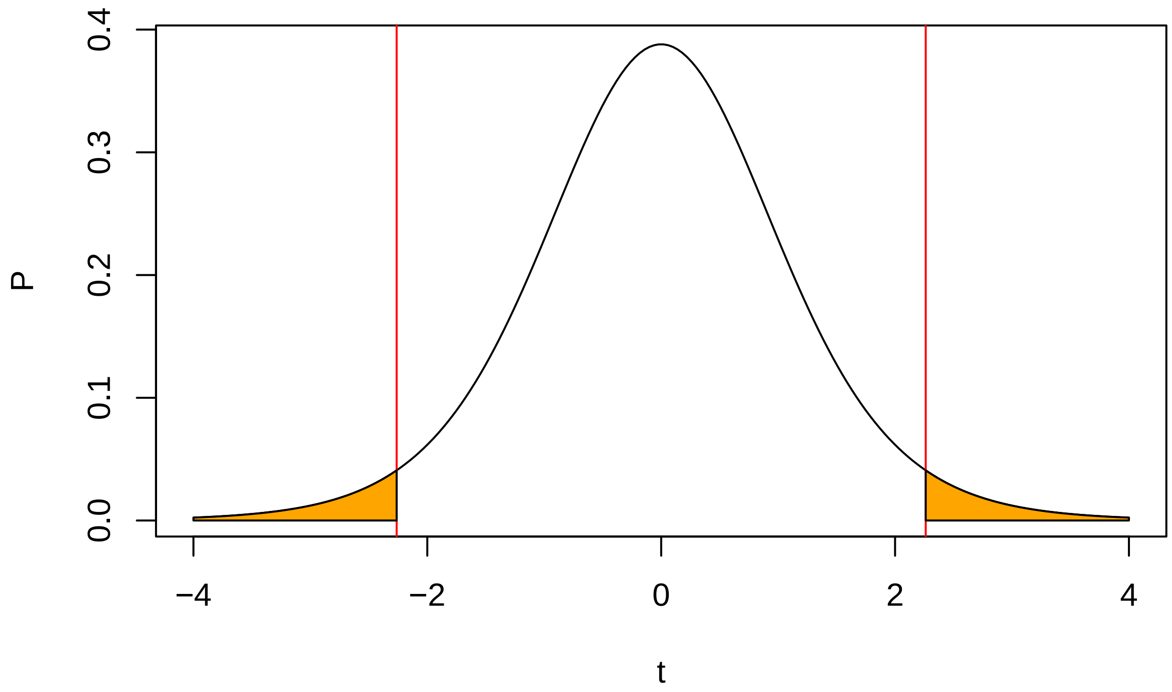
$$(1-\alpha)\%\ CI = [\bar{X} - t^* \frac{s}{\sqrt{n}}, \bar{X} + t^* \frac{s}{\sqrt{n}}]$$

where $\bar{X}$ is the sample mean, s is the sample standard deviation, n is the number of samples, and t^* is the critical value (that is dependent on the confidence level).

Let's first calculate the sample mean (stored in `x_bar`) and the sample standard deviation (stored in `s_x`):

```
x_bar <- mean(sub_df$perc_benefit)
s_x <- sd(sub_df$perc_benefit)
```

Next, we want to find the critical value t^*, such that 95% (our confidence level) of the area below the corresponding t distribution (displayed below) lies between $-t^*$ and $+t^*$. The degrees-of-freedom of the corresponding t distribution (displayed below) is $n - 1 = 10 - 1 = 9$



In the above plot of t_9, the white area is 95%. The left critical value is the 2.5th percentile ($(1 - .95)/2 = 0.05/2 = 0.025$):

```
qt(0.025, df = 10 - 1)
```

```
## [1] -2.2622
```

The 95% Confidence Interval

Recall that the general formula for the confidence interval of the mean is:

$$(1 - \alpha)\% \; CI = [\bar{X} - t^* \frac{s}{\sqrt{n}}, \bar{X} + t^* \frac{s}{\sqrt{n}}]$$

```
x_bar - 2.2622 * s_x / sqrt(10)
```

```
## [1] -2.8698
```

```
x_bar + 2.2622 * s_x / sqrt(10)
```

```
## [1] 12.521
```