ISE 3293/5013 Laboratory 11: Classical confidence intervals

In this lab we will learn how to use R to create confidence intervals. This will cover the remainder of chapter 7. Bootstrap intervals are easy to make once an R script has been created, because it relies on simulation each run of the algorithm will result in slightly different intervals. Classical analytical methods give constant results and everyone will calculate the same values for the endpoints of the confidence intervals (provided there are no arithmetic mistakes).

There are two interpretations that you will need to grapple with:

- Theoretical – a long run probabilistic understanding. This relates to many future samples.
- Practical – a statement of "confidence". This relates to the sample or samples in hand.

You will need practice at using these interpretations.
For this course it is NOT required that you memorize formulae for confidence interval calculation. These will always be supplied. You will need to select the correct formula, make the calculation and interpret the interval made. The most important part is the interpretation. The selection of the correct formula proceeds from a good understanding of the sampling situation and the intent of the experiment.

There are some easy ways to calculate intervals in R, however they rely on the use of a function that is made to primarily carry out a test (Ch 8). The confidence interval is a by-product and it is this that we want to retrieve.

Tasks
All output made please copy and paste into **this word file**. Save and place in the dropbox when completed. Anything you are asked to make should be recorded under the question in this document. There will be two files you need to upload:

- a pdf of this document (pdf) or the word file (docx)
- a text file of all the code you used to create answers (txt)

Note: All plots you are asked to make should be recorded in this document.

You are expected to adjust the functions as needed to answer the questions within the tasks below.

- Task 1
    - Make a folder LAB11
    - Download the file "lab11.r"
    - Place this file with the others in LAB11.

- Start Rstudio
- Open “lab11.r” from within Rstudio.
- Go to the “session” menu within Rstudio and “set working directory” to where the source files are located.
- Issue the function `getwd()` and copy the output here.
- Create your own R file and record the R code you used to complete the lab.

```
F:/Google Drive - Saied/Courses/02 OU/11 Fundamentals of Engin
eering Statistical Analysis/02 Labs/11 Lab 11
```

- Task 2
  - This relates to one sample from a population where we want to estimate μ.
  - Suppose we wish to build a confidence interval for the mean.
  - $ci\ for\ \mu = \bar{y} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$
  - A ram press makes washers for agricultural machinery. The mean diameter (cm) of the washers is of interest. Suppose the following are the diameters of washers from a random sample taken from the press:

d=c(5.0581, 4.9707, 5.0893, 4.9334, 4.9777, 5.0285, 4.8555, 4.9565,
4.9769, 4.9722, 4.999, 4.9925, 4.9686, 5.0662, 4.9239, 4.9781,
5.0485, 5.0014, 4.9957, 5.0195, 5.0118, 4.9928, 5.0361, 5.0185,
4.9879)

  - Create the following intervals for the mean μ using R as a calculator
    - 95% ci

      ```
      4.974137 5.014607
      ```
    - 90% ci

      ```
      4.974137 5.014607
      ```
    - 80% ci

      ```
      4.981452 5.007292
      ```
    - 50% ci

      ```
      4.987658 5.001086
      ```
  - Create an 80% ci for the mean using `t.test()`using the option `conf.level=0.80`, you will need to make an object and then extract the confidence interval using `$`

    ```
    [1] 4.981452 5.007292
    attr(,"conf.level")
    [1] 0.8
    ```
  - Now we will concentrate on the population variance σ^2.
  - $ci\ for\ \sigma^2 = \left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\right]$
  - Create the following intervals for the population variance σ^2 using R as a calculator
    - 95% ci

      0.001465121 0.004650629
    - 90% ci

      0.001583773 0.004164600
    - 80% ci

      0.001737340 0.003683142
    - 50% ci

      0.002042167 0.003029489

- Task 3
  - In this task we will examine confidence intervals for $\mu_1 - \mu_2$. Suppose two large ($n_1\ or\ n_2 > 30$) independent random samples are taken from two populations (not necessarily normal).
  - $ci\ for\ \mu_1 - \mu_2 \approx \bar{y}_1 - \bar{y}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \bar{y}_1 - \bar{y}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  - For small sample sizes $n_1 or\ n_2 < 30$ and $\sigma_1 = \sigma_2$ the following is the appropriate formula:
  - $ci\ for\ \mu_1 - \mu_2 \approx \bar{y}_1 - \bar{y}_2 \pm t_{\frac{\alpha}{2}}\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}$ where $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
  - Suppose that two random samples of fish were caught, 20 blue cod fish and 15 snapper. The lengths of the fish were measured in inches. Assume pop. Variances are equal.
    blue=c(21.65, 17.48, 20.1, 21.57, 14.82, 19.17, 21.08, 18.23, 22.93,
    15.66, 20.89, 21.66, 18.5, 20.59, 18.63, 18.91, 19.53, 17.7, 16.5, 19.03)
  - snapper=c(31.65, 27.48, 30.1, 31.57, 24.82, 29.17, 31.08, 28.23, 32.93,
  - 25.66, 30.89, 31.66, 28.5, 30.59, 28.63)
  - Choosing the correct ci formula make a 95% confidence interval for $\mu_{blue} - \mu_{snapper}$ by using R as a calculator.

```
> x1 = snapper
> x2 = blue
> n1 = length(x1)
> n2 = length(x2)
>
> if (n1 < n2) n = n1 else n = n2
>
> alpha = 1 - 0.95
>
> if (n > 30){
+   za = qnorm(1 - alpha/2)
+   ci[1] = mean(x1) - mean(x2) - za * sqrt(var(x1)/n1 + var(x2)/n2)
+   ci[2] = mean(x1) - mean(x2) + za * sqrt(var(x1)/n1 + var(x2)/n2)
+ } else {
+   ta = qt(1 - alpha/2, n-1)
+   varp = ((n1 - 1) * var(x1) + (n2 - 1) * var(x2)) / (n1 + n2 - 2)
+   ci[1] = mean(x1) - mean(x2) - ta * sqrt(varp * (1/n1 + 1/n2))
+   ci[2] = mean(x1) - mean(x2) + ta * sqrt(varp * (1/n1 + 1/n2))
+ }
> ci
[1]  8.674031 11.924303
```

  - What is the 95% ci for $\mu_{snapper} - \mu_{blue}$?

    [8.795967, 11.802366]

  - Give a practical interpretation of the above interval.

    With 95% confidence, we can say that the average size of blue cod fish is smaller than snappers between 8.80 and 11.80 inches.

  - N.B small sample ci for the difference in means with unequal variances are given on page 291 MS. We will use `t.test()` and let the software manage the correct formula.
  - Use `t.test(,var.equal=TRUE)$conf.int` to calculate a
    - 95% ci for $\mu_{snapper} - \mu_{blue}$

```
[1]  8.757585 11.840749
attr(,"conf.level")
[1] 0.95
```

        - 85% ci for $\mu_{snapper} - \mu_{blue}$

          [1] 9.182414 11.415919
          attr(,"conf.level")
          [1] 0.85

        - 75% ci for $\mu_{snapper} - \mu_{blue}$

          [1] 9.411911 11.186422
          attr(,"conf.level")
          [1] 0.75

        - 25% ci for $\mu_{snapper} - \mu_{blue}$

          [1] 10.05570 10.54263
          attr(,"conf.level")
          [1] 0.25

    - What happens to the interval as the confidence level decreases?

      The interval become smaller and the higher and lower values converge to the me an value.

- Task 4
    - Paired samples (dependent) occur when two measurements are made on the same experimental units.
    - $(1-\alpha)100\%\ confidence\ interval\ for\ \mu_d = (\mu_1 - \mu_2)\ is\ \bar{d} \pm z_{\frac{\alpha}{2}} \frac{\sigma_d}{\sqrt{n}}, n \geq 30$
    - $(1-\alpha)100\%\ confidence\ interval\ for\ \mu_d = (\mu_1 - \mu_2)\ is\ \bar{d} \pm t_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}, n < 30$
    - If the population variance is unknown and $n \geq 30$ use s_d
    - In 2012 MATH 4753 had two exams, the following were the results:

      Exam1=c(40.98, 59.36, 46.69, 41.8, 61.63, 65.31, 62.96, 60.21, 56.89,
      78.41, 53.44, 75.2, 60.54, 52.43, 41.41, 70.79, 73.55, 55.65,
      61.43, 63.84, 58.07, 53.79, 54.45, 67.18, 44.46)

      Exam2=c(50.22, 66.19, 58.75, 51.88, 66.61, 70.86, 74.25, 70.23, 69.55,
      87.18, 63.62, 81.7, 70.5, 66.02, 51.35, 80.92, 85.65, 65.44,
      74.37, 75.28, 67.86, 59.92, 64.42, 73.57, 57.15)
    - Make a 95% ci for $\mu_d = \mu_1 - \mu_2$

      -10.680441 -8.761159
    - Interpret the interval practically

      The average for second exam is higher than first exam. With confidence level of 95%, the difference between averages are between 8.76 and 10.68.
    - Make the following intervals using the appropriate options in `t.test()`:
        - 90% ci for μ_d

          [1] -10.558485 -8.883115
          attr(,"conf.level")
          [1] 0.9

        - 80% ci for μ_d

          [1] -10.366041 -9.075559
          attr(,"conf.level")
          [1] 0.8

        - 70% ci for μ_d

          [1] -10.239465 -9.202135
          attr(,"conf.level")
          [1] 0.7

        - 60% ci for μ_d

          [1] -10.140335 -9.301265

```
attr(,"conf.level")
[1] 0.6
```

- 10% ci for μ_d

```
[1] -9.782981 -9.658619
attr(,"conf.level")
[1] 0.1
```

- Task 5
  - $n \geq 30\ (1-\alpha)100\%\ ci\ for\ a\ population\ proportion\ p$

    $$\hat{p} \pm z_{\frac{\alpha}{2}}\sigma_{\hat{p}} \approx \hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}, n\hat{p} \geq 4, n\hat{q} \geq 4$$
  - $n_1 \geq 30\ and\ n_2 \geq 30\ (1-\alpha)100\%$ ci for a population $p_1 - p_2$
  - $(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}}\sigma_{\hat{p}_1-\hat{p}_2} \approx (\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
  - $n_i p_i \geq 4\ and\ n_i q_i \geq 4$
  - The following table is derived from the data set NZBIRDS.
  - Verify the table by reading in the data set and using a suitable function. HINT: `with()`, `table()`.

Bird population	Number of species sampled	Number of flightless species
Extinct	38	21
Non extinct	78	7

  - Use R as a calculator and find a 95% ci for the difference in proportion of flightless birds for extinct and non-extinct species.

```
0.1381492 0.4090094
```

- Task 6
  - A $(1-\alpha)100\%$ ci for the ratio of two population variances, $\frac{\sigma_1^2}{\sigma_2^2}$.
  - Two samples taken independently from Normal populations.
  - $\frac{s_1^2}{s_2^2}\frac{1}{F_{\frac{\alpha}{2}(\nu_1,\nu_2)}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2}F_{\frac{\alpha}{2}(\nu_2,\nu_1)}$
  - You will need `qf(…,df1,df2)` in R
  - Use the following simulated samples to obtain a 95% ci for the ratio (1 to 2) of the population variances.

```
0.2257732 1.0309437
```

  - `set.seed(35);sam1=rnorm(25,mean=10,sd=5);`
    `set.seed(45);sam2=rnorm(34,mean=40,sd=8)`
  - Use `var.test()` to create ci's for the same population variances with the following confidences
    - 80%

```
[1] 0.2920457 0.7831464
attr(,"conf.level")
[1] 0.8
```

    - 70%

```
[1] 0.3204384 0.7105597
attr(,"conf.level")
[1] 0.7
```

    - 60%

```
[1] 0.3449688 0.6581343
attr(,"conf.level")
[1] 0.6
```

- 50%

```
[1] 0.3675269 0.6165110
attr(,"conf.level")
[1] 0.5
```

################## LAB FINISHES HERE ##############################

- Task 7: Extra for experts!
    - Make a bootstrap() function for 2 sample intervals