20. Bootstrapping

- **Definition:** Bootstrapping is a robust statistical technique that enhances our understanding of data by repeatedly sampling from the original dataset.
  - Introduced by Bradley Efron in 1979.
- Estimate the sampling distribution of a statistic.
  - **Nonparametric Bootstrappin**g: Directly samples from the original data, with replacement.
  - **Parametric Bootstrapping**: Assumes a distribution (e.g., normal, binomial) and generates new datasets based on estimated parameters from the original sample.

- How does it work:
  - Generate multiple samples from the original dataset by randomly selecting observations (with replacement).
  - For each bootstrap sample, compute a statistic (e.g., mean, proportion, variance).
- Collect the computed statistics to form **an empirical distribution of the statistic.**
  - This empirical distribution, can be used to estimate confidence intervals and standard errors for instance.

- Consider the test scores of five randomly selected students from a Mathematical Statistics I class:
  `c(74.4,76.0,92.0,98.4,66.4)`
  - Produce a 95% confidence interval for the exam average in the class.
- Let's formulate the following hypothesis: "The average for the Mathematical Statistics I class Test 2 is less than 85%."
- The null hypothesis (H_0) represents the status quo or **default assumption**—there is no effect, no difference, or no change. Here, "The class average score is 85 (or more)."
- The alternative hypothesis (H_A or H_1) challenges the status quo and represents what we try to show, such as "The class average score is less than 85."

- You can use confidence intervals to test hypotheses:
  - Construct a confidence interval for the parameter of interest (e.g., the mean).
  - Check if the null hypothesis boundary value (e.g., $\mu_0 = 85$) lies within the interval:
    - If it does, fail to reject H_0 because the null value is plausible. Not enough evidence.
    - If it does not, reject H_0, suggesting evidence for the alternative hypothesis.
  - Do we have enough evidence in this data to conclude our statement?
  - What if we have one more observation, another exam result that is 58.4%. Do we have enough evidence now?

- **Major Applications of Bootstrapping**
  - **Hypothesis Testing:** Test of statistical hypotheses by generating a distribution of test statistics from resampled data.
  - **Confidence Interval Estimation** for population parameters without stringent assumptions.
  - **Regression Analysis:** Evaluates the variability and stability of regression coefficients, particularly in small samples or complex models.
  - **Model Validation**: Assesses the predictive performance across bootstrap samples.
  - **Time Series Analysis**: Employs block bootstrapping to preserve temporal dependencies in the data.

Pros:

- **Simplicity**: Intuitive and straightforward to implement.
- **Few Assumptions**: Does not need assumptions for the data.
- **Versatility**: Applicable across a wide range of statistical analyses, including confidence intervals and hypothesis testing.
- **Small Sample Efficacy**: Performs well even with limited sample sizes.
- Provides **reliable estimates** of standard errors and confidence intervals even for small or non-standard datasets.

Cons:

- **Computationally Intensive**: Can be computationally-demanding with large datasets. I would say this is not a major concern anymore.
- **Not Ideal for Very Small Samples**: May yield unreliable estimates with extremely small samples.
  - **Specifically, if the population variance is high or the population is very-skewed, etc.**
- **Random Variability**: Results can exhibit variability across different bootstrap iterations. Resolve by running
- **Theoretical Limitations**: Some applications lack a robust theoretical foundation.

Very powerful statistical tool to derive results from data.

21. R Markdown Brief

- Although you can copy and paste from the console and save your results, sometimes you might want to set up an interactive environment to work and share your code for a project with others.
  - Computational Notebooks are ideal when you want to combine text with formatting, graphics, executable computations, etc.

  In the Machine Learning community **Project Jupyter** computational notebooks are prevailing. Actually, JuPyteR stands for Julia, Python, and R. Yes, you can do R Jupyter notebooks.

- **R Markdown Notebook** (Posit project) has been getting some traction. It is typically required for STAT518.

  https://rmarkdown.rstudio.com/
  - Combines R Markdown language with computational notebook capabilities.
  - R Markdown and ggplot2 can be used together to create dynamic reports with embedded plots.
  - "(T)he primary difference between *R Markdown Notebook* and *R Markdown* is that when executing code in an R Markdown document, all the code is sent to the console at once, but in a notebook, only one line at a time is sent."
  - I will just briefly introduce *R Markdown Notebook* here.

- ***Markdown*** is a formatting syntax for authoring primary HTML, PDF, and MS Word documents.
  - The ultimate reference is http://rmarkdown.rstudio.com.
- R Markdown supports dozens of static and *dynamic output formats* including HTML , PDF , MS Word, Beamer, HTML5 slides, Tufte-style handouts, books, dashboards, shiny applications, scientific articles, websites, and more.
- Besides including R code and output, you can also write mathematical equations with embedded LaTeX code, format text, embed pictures, etc.

- Let's do a default document with R Markdown with target export to MS Word.
  - In RStudio, go to File->New File-> R Markdown
    - Select Word
  - You can change the title, edit the text and the code…
  - You can save the file (we use an `.Rmd` extension for R markdown files).
    - In R Markdown, instead of explicit code cells you indicate embedded *code chunks* in the RStudio text editor.
    - Add a new code chunk by clicking the <Insert Chunk> button on the toolbar or by pressing <Ctrl+Alt+I>.

    - Let's make sure 2+2=4
    - You can execute a chunk of code by clicking the small *play* button on the right or selecting from the run menu in the RStudio text editor window.
  - You can "knit" the final word document.
    - Hit the run button at the RStudio text editor window
    - Press the Knit button and select Knit to Word, select save location.

- To add beautiful formulas with LaTeX formatting:
  - Use `$$` *before* and *after* a LaTeX equation.

- For equations, you need to learn LaTeX first…

  or

- Alternatively, use AI:
  - Upload your math handwriting.
  - Politely ask it to convert to LaTex.
  - Open the .tex file in a text editor, simply copy and next paste the equation with $$ around it in your R Markdown doc.

22. Tests and Models Introduction

Traditional Statistics:

- What specific question are we trying to answer?
- Can this question be measured quantitatively?
- How can we collect the data appropriately?
- What statistical methods should we use to analyze it?
- How can we use the analysis to make decisions, draw conclusions, or make predictions?

Data Science:

- What data sources can we access or extract?
- What machine learning models or algorithms are suitable for this data? What patterns, trends, or insights can we extract from the data?
- How can we use these models to improve decision-making or to automate processes? Do these insights apply to future datasets or scenarios?
- How do we ensure the reliability and ethical use of our data-driven insights?

Statistical Analysis Paradigm:

1. Begins with the formulation of a question of interest and a "significance level".
2. Continues with the collection of relevant data.
3. Follows with the analysis of the data.
4. Concludes with a formal statistical test and interpretation of the results.

Warning: formulation of the problem and the "significance level" **in advance** is a paramount in statistics.

Hypothesis Truth	Reject H_0	Fail to Reject H_0
H_0 is True	**Type I Error (α)**: Occurs when you incorrectly reject a true null hypothesis. This means you conclude that there is an effect or difference when, in reality, there is none.	**Correct Acceptance (1 - α)**: Occurs when you correctly fail to reject a true null hypothesis. This means you correctly conclude there is no effect or difference.
H_0 is False	**Correct Rejection (1 - β)**: Occurs when you correctly reject a false null hypothesis. This means you correctly conclude there is an effect or difference when it actually exists.	**Type II Error (β)**: Occurs when you fail to reject a false null hypothesis. This means you incorrectly conclude there is no effect or difference, when in reality, there is one.

The significance level is the maximum tolerance for the risk of committing Type I error, that is falsely rejecting the null hypothesis (or the status quo) when it is true.

- Type I Error (α): A **false positive**, where you incorrectly reject the null hypothesis when it is actually true.
- Type II Error (β): A **false negative**, where you fail to reject the null hypothesis when it is actually false.
- Correct Rejection of the Alternative (1 - β): Correctly rejecting a false null hypothesis, meaning you detect a true effect.
- Correct Acceptance of the Alternative (1 - α): Correctly failing to reject a true null hypothesis, meaning you do not falsely claim there is an effect when there is none.

- Some functions for statistical tests:
  - t-test: `t.test()`
  - Binomial test: `binom.test()`
  - Chi-squared test: `chisq.test()`

- Some statistical and other functions for models:
  - Fit and save a linear model: `fit<-lm()`
  - Fit generalized linear model: `glm()`
  - ANOVA table: `anova()`
  - Parameter(s) estimate: `coef(fit)`, `summary(fit)`
  - Confidence interval for a parameter: `confint(par)`

  - Residuals: `resid(fit)`
  - Diagnostic plots: `plot(fit)`
  - Predict from fit: `predict(fit,…)`

- Good starting point that I can recommend:
  - Quick-R is a great resource to quickly see how to start common types of analyses.

https://www.statmethods.net/about/sitemap.html

    - You could just modify the R examples provided there.

***

- Let's consider the `iris` dataset yet again.
  - For the Iris data set, what were the variables?
  - What were the 3 different species present?

- Let's construct boxplots to compare all 4 continuous variables stratified by the 3 different species present in the dataset.

```
boxplot(Sepal.Length ~ Species, xlab = "", ylab =
"Sepal Length", col = c("green"), data=iris)
boxplot(Sepal.Width ~ Species, xlab = "", ylab =
"Sepal Width", col = c("green"), data=iris)
boxplot(Petal.Length ~ Species, xlab = "", ylab =
"Petal Length", col = c("green"), data=iris)
boxplot(Petal.Width ~ Species, xlab = "", ylab =
"Petal Width", col = c("green"), data=iris)
```


- If we want to combine all plots in a single one, we can use

```
op<-par(mfrow = c(2,2))
# re-run the same plot code now.
# ...
# At end, reset to previous settings:
par(op)

# let's save the picture
```


You can arguably do better with ggplot2

```
require("ggplot2")
# scatter plot
ggplot(iris, aes(Sepal.Length, Sepal.Width)) +
geom_point(aes(color = Species)) +
theme(legend.position = "top")

# boxplot
ggplot(iris, aes(Species, Sepal.Width)) +
geom_boxplot(aes(fill = Species)) +
theme(legend.position = "top")
```

- Let's compare if *Iris Setosa* and *Iris Virginica* sepal widths.
  - We will perform a formal Student's t-test to check if the two population means are significantly different at alpha level 0.05.

WARNING:

- Again, note that we explored the data first and this is wrong as it brings
  - *Multiple testing issue...*
  - For us, the main goal in this course is to provide an example of the basic commands so we will continue.

- Let

  μ_S be the true population *mean for Iris Setosa sepal width* and

  μ_V is the true population mean for *Iris Virginica sepal width*.
  - We do not observe these population means of course but we can still test with the data…

- The Null hypothesis is

  $\boldsymbol{H_0: \mu_S = \mu_V}$ *There is no (statistically significant) difference in the mean sepal widths of the two species.*

The corresponding alternative hypothesis is: $\boldsymbol{H_1: \mu_S \neq \mu_V}$
There is difference in the mean sepal widths of the two species.

- We need to check if the conditions (for any test) are met. Here, we check:
  - **Independence:** mainly from the design of the study, investigate with the researcher. Here we just assume that the species of Iris in our data set are independent.
  - **Normality**: The populations' sepal widths are normally distributed.
  - **Assumption of Homogeneity of Variance for Student's two sample t-test:** The variances of the two populations should be approximately equal.
    - Note that formal statistical tests for the assumptions 2 and 3 exist but is not unusual to just do informal checks.

  - Let's study the density plots for the sepal widths for all species.

```
ggplot(iris, aes(Sepal.Width, color = Species)) +
geom_density() + theme(legend.position = "top")
```

- Let's explicitly check the variances:
  - One can extract the sepal widths for both species (we did examples like this before).
  - Or we can use the `by()` function; it automatically applies a function to each level of a factor(s).
    - `by()` is similar to BY processing in SAS® statistical software.

```
by(iris$Sepal.Width, iris$Species, function(x)
var(x))
```

- Without performing a formal test, the variances do seem unequal.
  - We should use Welch's two sample t-test that allows for unequal variances.

- Now let's consider the actual code for the two-sample t-test for difference of means (independent samples, variances are not assumed equal):

```
# drop the versicolors; or you can extract x,y
iris2<- iris[iris$Species!="versicolor",]

tst<-t.test(Sepal.Width ~ Species, data=iris2)
```


Let us study the test results and the model fit.

- The p-value is basically the probability to observe the data under the null hypothesis.
- We compare the p-value to pre-determined significance level α.
  - intuitively, α is the maximum probability to wrongfully reject the null hypothesis that we can tolerate.

- For our test result, the p-value is $5 \times 10^{-9} \ll \alpha = 0.05$.
  - We reject the null, conclude the alternative hypothesis… There is (very) strong evidence to (reject the null hypothesis and) conclude that… (the alternative).

Here we conclude: *"There is very strong evidence that the mean sepal widths are different for the Setosa and Virginica species of iris."*

- By default, the method R uses is Welch's two sample t-test that allows for unequal variances.
  - As a side note, although inappropriate (assumption for equal variance is seemingly not met), the original Student's t-test for two means will give almost the same result.
  - The Student's t-test was developed by Gosset in 1908. Gosset was working at the Guinness Brewery, Ireland and published under the pseudonym "Student".

```
t.test(Sepal.Width~Species, var.equal=T, data=iris2)
```


- Exercise: *versicolor* and *virginica* look with much more similar sepal widths.

Let's test this with the Student's two sample t-test…

***

Note: if you are unsure about whether the variances are equal, it is always safer to use Welch's t-test. This will look better when you report your result.