Module 1: Probability & Distributions

# Reading

*Read*

* Reading for this week:
* (OpenIntro) Chapter 2: Summarizing Data
* (OpenIntro) Chapter 3: Probability

*Watch*

* Kahn Academy: Probability -- <https://youtu.be/uzkc-qNVoOk>
* The last banana -- <https://youtu.be/Kgudt4PXs28>

# Discussion for this week:

*Watch:*

Why you should love statistics -- <https://youtu.be/ogeGJS0GEF4>

*Prompt:*

Find a case of probability (percent/proportion/chance) in your work or the news and analyze the claim. What is the basis for the number (are there data)? Who is making the claim? Where did they get their sample from? Are the values surprising or intuitive?

POST RESPONSE TO [#dan602\_discuss](https://stonehillmpsda21.slack.com/archives/C01A1MZUGDB) in Slack

# Probability and Distributions:

*Goal:* Use SAS Studio to visualize and describe sample distributions.

*Methods:*

Book Data

1. Load book dataset STATS::SALARY (refer to the first workshop for guidance on loading book data into the STATS library).

\*Refer to Chapter 6 (Cody) for help finding these options in the Summary Statistics menu in SAS Studio\*

1. Create summary statistics of the variable Weekly\_Salary that include mean, median, standard deviation, and 99% confidence interval. Report the stats table here.
2. Using the “Distribution Analysis” task menu: Set up an analysis of the distribution to perform a test of normality to create a histogram and normal Q-Q plot for the Weekly\_Salary data. Add summary statistics for mean, skewness, and kurtosis.   
     
   How would you describe the distribution (histogram)? What do the values for skewness and kurtosis tell you about the distribution? (See pg 51 [Cody] for definitions and interpretation)  
     
   What does the test for normality (use Kolmogorov-Smirnov) tell you about the distribution? A p-value of >0.05 suggests that this distribution IS normally distributed, does that seem right when you look at the histogram (See author’s aside on page 66-67 [Cody] to aid in your interpretation)?
3. Keeping your code from (B) separate, create a new analysis that does the same, but adds “Age\_Group” as a classifying variable. What do you notice is different?
4. Keeping your code from (B) separate, create a new analysis that does the same, but adds “Education” as a classifying variable. What do you notice is different? Which do you think is more explanatory for the Salary parameter, Age\_Group or Education?
5. Keeping your code from (B) separate, create a new analysis that does the same, but adds “Age\_Group” AND “Education” as a classifying variable. What do you notice is different from the output of B,C, and D?
6. Many of the inferential statistics that we will cover (t-tests, ANOVA, etc.) rely on assumptions of normality. Most books will say that when your sample is not drawn from a normal distribution you cannot use these tests. Given the results of the above analyses, is this always the case if you only consider the result of the tests for normality (Kolmogorov-Smirnov)? How can we know if it’s OK to use these tests?

**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

COVID Data Analysis

1. Download data:   
   US State data: <https://covidtracking.com/data/download/all-states-history.csv>   
   World data (complete): <https://covid.ourworldindata.org/data/owid-covid-data.csv>   
   See description of data: <https://covidtracking.com/about-data/data-definitions>
2. Import data into SAS Studio (Suggestion: Import into a personal library; e.g., STATS\_IMP; ELSE import into WORK). Name these as:  
     
   US State data -> COVID\_STATES  
   World data -> COVID\_WORLD
3. Working with the COVID\_STATES data:

Create a new dataset that is:

* Filtered to look only at data in August, 2020 (> 2020-08-01) (Use “Tasks -> Data -> Filter”)
* LOG transformed for the positiveIncrease column (Use “Tasks -> Data -> Transform Data”)
  1. *Set 1:* Distribution of daily increase (positiveIncrease) in COVID cases from August
  2. *Set 2:* Distribution of LOG transformed daily increase (tr1\_positiveIncrease) in COVID cases from August
  3. *Set 3:* Same as Set 1 and Set 2 BUT first filter only for the state == “MA”
  4. *Set 5:* Calculate the percent positive tests for Massachusetts in August (Define a custom transformation using: positiveIncrease / totalTestsPeopleViralIncrease \* 100)
  5. *Set 4:* Distribution of daily tests conducted per state (Go back and filter to 3-5 states of interest; Use state as a classifying variable)

*Questions:*

1. (Set1 & Set2) Which is a better representation of the postiveIncrease data, counts (original) or LOG transformed? Why? Which would be suitable for inferential statistics with a T-Test (requires normality)?
2. (Set 3) Same questions as for #1. Also, which is easier to interpret for the state of Massachusetts? By your best guess, what value for postiveIncrease would be considered a “bad day” for the state?
3. (Set 4) Describe the distribution of percent positive values for the state. What values are “extreme”? What is the (approximate) expected value?
4. (Set 5) Describe the distribution of tests conducted for each state. Discuss whether and why there might be differences you can discern from these data.

**Challenge (optional):** Investigate case or death distributions in the COVID\_WORLD data and report your findings.

**Turn in answers to questions and graphs ONLY. Keep program files for your reference.**