DATA SCIENCE
TRUMAN STATE UNIVERSITY

# 7TH GRADE MATH MODULE: OUTLINE

## Rationale

The purpose of this module is to introduce students to basic concepts within data science while also providing an introductory activity for the instruction of related topics contained in the Missouri Learning Standards. This module will ask students to use sample data to make inferences about a population. It will ask students to compare distributions of data for two or more populations visually, and through statistical measures of center and variation.

This module serves as a starting point for instruction related to the following Missouri Learning Standards:

**Math:**

- 7.DSP.A.1 - Understand that statistics can be used to gain information about a population by examining a sample of the population.
    a. Understand that a sample is a subset of a population.
    b. Understand that generalizations from a sample are only valid if the sample is representative of the population.
    c. Understand that random sampling is used to produce representative samples and support valid inferences.
- 7.DSP.A.2 - Use data from multiple samples to draw inferences about a population and investigate variability in estimates of the characteristic of interest.
- 7.DSP.B.3 - Analyze different data distributions using statistical measures.
- 7.DSP.B.4 - Compare the numerical measures of center, measures of frequency and measures of variability from two random samples to draw inferences about the population.

**MoExcel Data Science Standards**

- MoExc1: **Identify** issues, problems, questions, or claims that can be addressed using large datasets.
  *The expectation is that students be able to **identify** statements, claims, or questions that can be refined into testable hypotheses.*
- MoExc2: **State** data-driven investigative questions.
  *The expectation is that students be able to **state** investigative questions based on quantitative data.*
- MoExc3: **Construct** visual representations of real-life data from publicly available datasets and **describe** patterns observed.
  *The expectation is that students are familiar with large datasets of publicly available data that allow users simple but rich manipulation of bivariate data and **describe** patterns that result from purposeful manipulation of the information.*
- MoExc4: **Suggest** and **discuss** the possible interactions among data.
  *The expectation is that students can provide and consider alternative explanations to the relationships (or lack thereof) among data.*
- MoExc5: **Identify** and **discuss** potential factors that can influence relationships between the independent and dependent variables.
  *The expectation is that students reflect on the complexity of real-life problems and consider it when attempting analyses or problem-solving. This includes identifying and accounting for different forms of control variables (intervening, confounding, or antecedent). Discussion of the differences among control variables is **not** expected.*

TRUMAN STATE UNIVERSITY

- MoExc6: **Interpret** real-life data by using patterns and relationships among data.

  *The expectation is that students are able to construct stories that provide plausible explanations for relationships that have been identified among data.*

**Standards for Mathematical Practice**

| Standard#: | Standard: |
|---|---|
| MP1 | Making sense of problems and persevere in solving them. |
| MP2 | Reason abstractly and quantitatively. |
| MP3 | Construct viable arguments and critique the reasoning of others. |
| MP4 | Model with mathematics. |
| MP5 | Use appropriate tools strategically. |
| MP6 | Attend to precision. |
| MP7 | Look for and make use of structure. |
| MP8 | Look for and express regularity in repeated reasoning. |

**Prior Knowledge & Possible Misconceptions:**

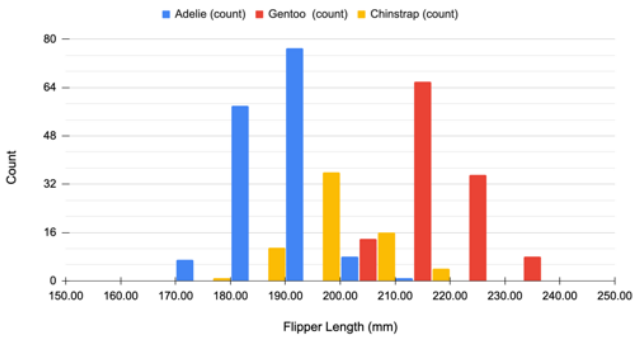
**Prior Knowledge:**

This module assumes that previous instruction has covered the 6th grade data standards, including:

- 6.DSP.A.1 - Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers.
- 6.DSP.A.2 - Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread and overall shape.
- 6.DSP.A.3 - Recognize that a measure of center for a numerical data set summarizes all of its values with a single number, while a measure of variation describes how its values vary from a single number.
- 6.DSP.A.4 - Display and interpret data.
    a. Use dot plots, histograms and box plots to display and interpret numerical data.
    b. Create and interpret circle graphs.
- 6.DSP.A.5 - Summarize numerical data sets in relation to the context.
    a. Report the number of observations.
    b. Describe the nature of the attribute under investigation, including how it was measured and its units of measurement.
    c. Give quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context of the data.
    d. Analyze the choice of measures of center and variability based on the shape of the data distribution and/or the context of the data.

**Possible Misconceptions:**

1. Students will need to be careful comparing distributions of data visually when the samples are different sizes. For example, the sample size was less for chinstrap penguins, but the distribution of flipper length for all three species has approximately the same shape.




2. Students often confuse “sample” and “population.” Consistently emphasize that the population consists of ALL individuals or items we are interested in, while the sample is the smaller group that we actually have collected data from.

# 7th Grade Data Science Math Module

**Example**: Penguins

**Question**: Can we count all the penguins?

**Data**: Penguin Counts

**Goal**: Students will see the need for sampling.

**Materials & Tech Requirements**: The teacher will need an internet-capable device to display websites. If students have their own devices, they can individually count penguins on the **Penguin Watch site** or explore **MAPPPD**.

**Discussion:**

Pull up the **map of Antarctica**. Ask students: "Can you count the penguins in this picture?" They may say it is impossible or that there aren't any penguins in the picture. Tell them that there are LOTS of penguins in Antarctica, and that it IS impossible to accurately count all of the penguins there. The climate is too harsh for humans and there are too many penguins in remote areas. If students ask why we would want to count penguins, tell them that the number of penguins and the change in their numbers over time is an indicator of how healthy the entire Antarctic ecosystem is. Instead of trying to count absolutely every penguin in Antarctica each year (which statisticians would call the "population"), scientists CAN count smaller groups of penguins (called "samples") and use this information to get an idea of how penguin numbers are changing overall.

Now ask students to count the penguins in this **image**. See if students in the class get similar numbers. While not everyone may get the exact same number, they should be very close. Ask them if this was doable. It was much easier than trying to count ALL the penguins in Antarctica!

Show the **Penguin Watch Project Info Video** (2:30 min). Then pull up the Penguin Watch **website** and work through classifying a few images together (or have students work through some on their own device). Remind students that while humans are counting penguins in some images, data scientists have trained computers to count penguins in images on their own. The human work helps make the computer algorithm more accurate. *Note: More educational and school outreach resources from the project are* **here**.

After classifying several images, explain that in the Penguin Watch project, many photos are used to try to count penguins in one area. This is the idea of sampling. Instead of counting all penguins in Antarctica, which is impossible, "samples" of penguins in an area are counted, and numbers can be compared from one breeding season to the next. Over time, this can show trends in the overall population.

**MAPPPD** is a tool that summarizes penguin count data from research papers. Pull up the tool and show students some of the data. You can click on a site or search the map. Using any species to search will bring up a table that you can sort by count. Pick a site with many counts (recommended: **Ardley Island** and **Llano Point**) and hit "explore data." Be sure to use the dropdown menu to look at multiple species. Ask students to discuss what trends they see. Notice that chinstrap and adélie penguin counts are decreasing, while gentoo penguin numbers have been increasing. If students have their own devices, you can give them a few minutes to explore the site on their own. A summary of data and more information is in this **report**. Show students the percentage changes from 2019 to 2020, and that the overall data matches the trends from the samples at Ardley and Llano.

While penguin counts are very useful, information about penguin mass, size, and measurements can also help scientists learn more about each species. We will be looking at some of this data throughout the unit.

## Suggestions for Unit Integration

### Sampling

After discussing the definition of a sample and population, show students the **adélie penguin** data set. Talk through what each variable represents in the data set. **This image** may be useful to help explain the bill measurements. Tell students that there are three samples in this data set (at three different locations). What is the size of each sample? What population is being considered? Would these samples be representative of the entire population? Why or why not?

### Using Data from Multiple Samples to Draw Inferences about a Population

Tell students that they will be working with the sample **data for adélie penguins**. Scientists want to make generalizations about the size and measurements of the species. While they have not measured EVERY adélie penguin, taking data from a sample of penguins can help them describe the species in general. The more samples that they take, or the larger the sample size is, the more accurate their generalizations about the population will be. The adélie penguin size data contains three samples from three different islands. If you did not already discuss what each variable represents, you will want to do so.

Then assign each student (or have students work with a partner) a sample - Dream Island, Biscoe Island, or Torgersen Island. Ask students what the size of their sample is. Then have students use Google sheets to calculate statistical measures for their assigned sample. You may choose which variables you wish to have them calculate measures for. It may be easiest to start with a single measure, like body mass, and calculate and compare the three sample results as a class before discussing any other variables.

*Directions for calculations in Google sheets:*

- To calculate the mean, type =AVERAGE(C:C) where C is the column of data you wish to use. You can click this column instead of typing it.
- To calculate the median, type =MEDIAN(C:C) where C is the column of data you wish to use. You can click this column instead of typing it.
- To calculate the mode, type =MODE(C:C) where C is the column of data you wish to use. You can click this column instead of typing it.
- To calculate the range, type =MAX(C:C)-MAX(C:C)where C is the column of data you wish to use. You can click this column instead of typing it.
- To calculate the interquartile range, type =QUARTILE(C:C, 3)-QUARTILE(C:C, 1) where C is the column of data you wish to use. You can click this column instead of typing it.
- To calculate the mean absolute deviation, type =AVEDEV(C:C)where C is the column of data you wish to use. You can click this column instead of typing it.
- Reminder: It is easiest to do calculations for one column, then highlight the cell and click the blue dot in the corner to drag the formula across and calculate the rest of the columns.

Here are the results that students should get:

| Measure | Island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g |
|---|---|---|---|---|---|
| Mean | Dream | 38.50 | 18.25 | 189.73 | 3,688.39 |
| Median | Dream | 38.55 | 18.40 | 190.00 | 3,575.00 |
| Mode | Dream | 36.00 | 18.50 | 190.00 | 3,900.00 |
| Range | Dream | 12.00 | 5.70 | 30.00 | 1,750.00 |
| IQR | Dream | 3.63 | 1.40 | 7.50 | 706.25 |
| MAD | Dream | 2.04 | 0.91 | 5.06 | 380.11 |

| Measure | Island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g |
|---|---|---|---|---|---|
| Mean | Biscoe | 38.98 | 18.37 | 188.80 | 3,709.66 |
| Median | Biscoe | 38.70 | 18.45 | 189.50 | 3,750.00 |
| Mode | Biscoe | 37.80 | 18.90 | 187.00 | 3,800.00 |
| Range | Biscoe | 11.10 | 5.10 | 31.00 | 1,925.00 |
| IQR | Biscoe | 3.03 | 1.38 | 8.25 | 587.50 |
| MAD | Biscoe | 2.03 | 0.95 | 5.40 | 384.92 |

| Measure | Island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g |
|---|---|---|---|---|---|
| Mean | Torgersen | 38.95 | 18.43 | 191.20 | 3,706.37 |
| Median | Torgersen | 38.90 | 18.40 | 191.00 | 3,700.00 |
| Mode | Torgersen | 36.70 | 17.60 | 190.00 | 3,700.00 |
| Range | Torgersen | 12.50 | 5.60 | 34.00 | 1,800.00 |
| IQR | Torgersen | 4.45 | 1.90 | 8.00 | 662.50 |
| MAD | Torgersen | 2.42 | 1.09 | 4.85 | 358.23 |

Compare the measures from the three samples. Point out that the values are very similar across the three samples, but they are **not** exactly the same. All samples will have some differences, but as long as the penguins in each sample were selected at random at each location, they should give similar results that accurately represent the population of all adélie penguins. Selecting a representative sample is key to get accurate results.

**Comparing Distributions for Multiple Populations**

*Comparing Through Visuals*

The **penguin size dataset** contains data for three different species of penguins: adélie, gentoo, and chinstrap. Take a few minutes to show students the **species profiles** for the gentoo, adélie, and chinstrap penguins. Then show students the **histograms** displaying the distribution of mass, bill measures, and flipper length by species. Ask students to compare species measurements based on the histograms. For example: "Which species appears to be the heaviest? Lightest?" There are some guiding questions on the slides. The slides also give histograms for the mass of male and female penguins for each species. *Optionally: If you have the time, you could ask students to use the data to create their own histogram or another type of chart for the data.*

*Comparing Through Statistical Measures*

The **penguin size dataset** contains data for three different species of penguins: adélie, gentoo, and chinstrap. If you did not already, take a few minutes to show students the **species profiles** for the gentoo, adélie, and chinstrap penguins. Statistical measures of center and variation for each species are given in the tables below.

| Adelie | bill_length_mm_a | bill_depth_mm_a | flipper_length_mm_a | body_mass_g_a | body_mass_g |
|---|---|---|---|---|---|
| Mean | 38.79 | 18.35 | 189.95 | 3,700.66 | 3,688.39 |
| Median | 38.80 | 18.40 | 190.00 | 3,700.00 | 3,575.00 |
| Mode | 32.10 | 15.50 | 172.00 | 2,850.00 | 3,900.00 |
| Range | 13.90 | 6.00 | 38.00 | 1,925.00 | 1,750.00 |
| IQR | 4.00 | 1.50 | 9.00 | 650.00 | 706.25 |
| MAD | 2.23 | 0.94 | 4.83 | 371.50 | 380.11 |

| Gentoo | bill_length_mm_a | bill_depth_mm_a | flipper_length_mm_a | body_mass_g_a | body_mass_g |
|---|---|---|---|---|---|
| Mean | 47.50 | 14.98 | 217.19 | 5,076.02 | 3,688.39 |
| Median | 47.30 | 15.00 | 216.00 | 5,000.00 | 3,575.00 |
| Mode | 46.50 | 15.00 | 215.00 | 5,000.00 | 3,900.00 |
| Range | 18.70 | 4.20 | 28.00 | 2,350.00 | 1,750.00 |
| IQR | 4.25 | 1.50 | 9.00 | 800.00 | 706.25 |
| MAD | 2.44 | 0.81 | 5.30 | 423.43 | 380.11 |

| Chinstrap | bill_length_mm_a | bill_depth_mm_a | flipper_length_mm_a | body_mass_g_a | body_mass_g |
|---|---|---|---|---|---|
| Mean | 48.83 | 18.42 | 195.82 | 3,733.09 | 3,688.39 |
| Median | 49.55 | 18.45 | 196.00 | 3,700.00 | 3,575.00 |
| Mode | 51.30 | 17.30 | 195.00 | 3,950.00 | 3,900.00 |
| Range | 17.10 | 4.40 | 34.00 | 2,100.00 | 1,750.00 |
| IQR | 4.73 | 1.90 | 10.00 | 462.50 | 706.25 |
| MAD | 2.76 | 0.95 | 5.53 | 294.10 | 380.11 |

Ask students to compare different measures by species. Possible questions to ask:

- Which species has the shortest bill? How do you know?
  (the adelie penguin has the smallest mean, median, and mode for bill length)

- Which species appears to have the most variability in body mass? How do you know?
  (the gentoo penguin has the largest range, IQR, and MAD)

- Which two species appear to have the most similar bill depth? How do you know?
  (the chinstrap and adelie have similar measures of center; the mode of the adelie penguins is smaller, but mode is not a particularly useful measure for numerical data, since it only gives the most frequent measure)

- Are there any measures of central tendency that do not seem particularly representative when compared to the rest of the data?
  (the mode for body mass is a poor representation for center when compared to the median and mean)

*Optionally: If you have the time, you could ask students to do all of their own calculations for each species instead of giving them to students. Data could also be compared for male and female penguins of the same species.*

**Sources and Links**

GISGeography. (2022, June 11). *Antarctica map and satellite imagery*. GIS Geography. Retrieved August 16, 2022, from **https://gisgeography.com/antarctica-map-satellite-image/**

Horst, A. (n.d.). *Palmer Penguins: A great intro dataset.* GitHub. Retrieved August 16, 2022, from **https://github.com/allisonhorst/palmerpenguins**

*MAPPD*. Mapping Application for Penguin Populations and Projected Dynamics. (n.d.). Retrieved August 16, 2022, from **https://www.penguinmap.com/mapppd/**

*Penguin Watch*. zooniverse.org. (n.d.). Retrieved August 16, 2022, from **https://www.zooniverse.org/projects/penguintom79/penguin-watch**

Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus *Pygoscelis*). PLoS ONE 9(3): e90081. doi:**10.1371/journal.pone.0090081**