# PURPOSEFUL QUESTIONING WITHIN THE STATISTICAL PROBLEM-SOLVING PROCESS—PRE K–12 GAISE II

Pip Arnold[1] and Christine Franklin[2]
[1]Karekare Education Limited, New Zealand
[2]American Statistical Association, USA
pip@karekareeducation.co.nz

*Questioning within the statistical problem-solving process supports statistical thinking and reasoning. Four statistical question purposes—investigative, data collection, analysis, and interrogative—have been identified. This paper illustrates three of these question purposes, investigative, interrogative, and analysis, and how they fit within the statistical problem-solving process through Darwin's finches' examples from the PreK–12 Guidelines for Assessment and Instruction in Statistics Education II report. Understanding and using statistical questions throughout the statistical problem-solving process allows teachers to support students as they develop into statisticians and data scientists.*

## INTRODUCTION

The *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II)* articulate that the statistical problem-solving process (Figure 1) is the foundation and core of statistical reasoning and making sense of data (Bargagliotti et al., 2020). Quoting from *GAISE II:*

> The statistical problem-solving process typically starts with a statistical investigative question, followed by a study designed to collect data that aligns with answering the question. Analysis of the data is also guided by questioning. Constant questioning and interrogation of the data throughout the statistical problem-solving process can lead to the posing of new statistical investigative questions. (p.9)

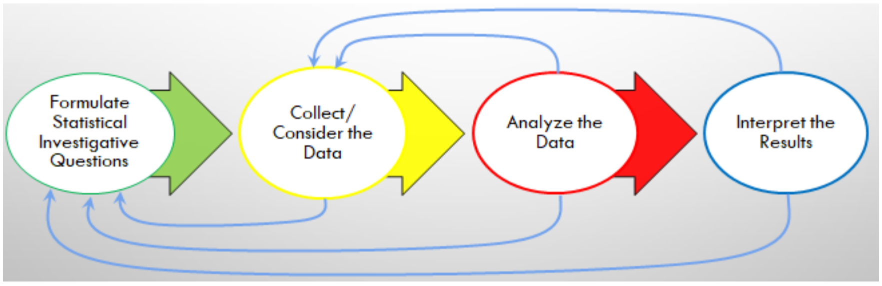


Figure 1: Statistical problem-solving process (Bargagliotti et al., 2020)

## STATISTICAL QUESTIONS

Posing investigative questions as the first phase of the process provides a starting point for explorations that necessitates the collecting of data that will vary to answer these questions. Questioning should then continue throughout all phases of the process. Understanding the types of statistical questions that occur at the different phases of the process is key to successful implementation of the statistical problem-solving process for all investigations involving data. In this paper, the purposes of statistical questions at the different phases are identified and illustrated. Arnold and Franklin (2021) identify two broad types of question: *question posing* and *question asking* and the types of questions within these categories. Question posing occurs with investigative questions at phase one or after interpreting results and with data collection questions at phase two. Question posing necessitates data collection, and to that end, questions need to be carefully constructed so it is clear what data needs to be collected. Question asking occurs at the analysis phase and throughout the process with interrogative questions (such as interrogating the data considered or if appropriate analysis tools are being utilized). Question asking is more spontaneous and relates either to the data collected to answer the posed questions (analysis questions) or the process of statistical problem-solving (interrogative questions). Investigative questions can further be classified as summary, comparison, relationship (associative), or time series questions. These different types of statistical questions will be illustrated using the dataset, *Darwin's finches* from GAISE II.

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org 

## DARWIN'S FINCHES

*Context*

Students can take opportunities presented in subject areas such as science to engage in the statistical problem-solving process. GAISE II provides science contextual examples, including Darwin's finches. Level B students (ages 10–14) can engage in investigations that describe a single variable, compare a variable across two groups e.g., species, look at associations between paired numerical variables, and explore a variable across time. Darwin's finches in the Galapagos Islands have evolved over the course of two million years from one common ancestor, the Grass quit. Within the genus *Geospiza* there are nine different ground finches. The examples used will focus on two specific ground finches: the Medium Ground Finch (MGF), *Geospiza fortis,* and the Common Cactus Finch (CCF), *Geospiza scandens.* In these examples students are working with secondary data, data that has been collected by someone else. In GAISE II there are five examples of possible student approaches to working with the Darwin's finches' dataset. In this paper we are going to look at how statistical questions are enacted in the teaching and learning sequence as students work through the statistical problem-solving process.

*Considering Secondary Data—Interrogating the Dataset*

In this scenario the teacher has provided a dataset for students to work with. After introducing the context to the students, the teacher gets the students to interrogate the dataset using *interrogative questions* (Arnold et al., 2022) such as: *What was the purpose of collecting the data? Who collected the data? Who funded the data collection and research? When was the data collected? Where was the data collected?* And *Who was the data collected from?* This initial interrogation of the dataset as a whole will add to the contextual understanding of the dataset, where it came from, and what the data was collected for. Having established broadly that the data was collected on the Galapagos Islands by Peter and Rosemary Grant (2013) as part of their longitudinal studies of Darwin's finches, the students can then interrogate each of the variables in the dataset to understand what they measure and how they are measured. Table 1 displays part of the dataset that is used in the example. Interrogative questions such as *How was the variable measured? What are the units (if any) of the variable? What possible outcomes can the variable take? What type of variable is it?* provide important information to support students starting their investigation and posing investigative questions to explore.

Table 1. Part of the dataset—Darwin's finches' data GAISE II

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **band** | **species** | **beak.length** | **beak.depth** | **year** | **drought** |
| 2 | **15** | fortis | 9.5 | 8 | 1975 | before |
| 3 | **4378** | fortis | 10.64 | 9.66 | 1981 | after |
| 4 | **361** | scandens | 12.9 | 8.6 | 1975 | before |
| 5 | **935** | scandens | 13.4 | 9 | 1975 | before |
| 6 | **5345** | scandens | 14.95 | 9 | 1984 | after |

Although teachers can provide this information for their students, students sourcing this contextual information make a deeper connection to the context of the data they are exploring. A recent presentation by Drs. Peter and Rosemary Grant (Carnegie Science, 2022) includes information about where, what, and how they collected the data (see presentation from 14:50 mins to 20:30 mins). Even before watching the video or searching for what the variables are, students can predict what they think they are. The variables in the dataset are: the band represents the ID number of an individual finch; the species is a categorical variable with two categories for the CCF (scandens) or MGF (fortis); the beak length is a continuous quantitative variable representing the measurement of the length from the base of the beak to the tip, measured in millimeters (mm); the beak depth is a continuous quantitative variable representing the height of the widest point in the beak, also measured in millimeters (mm); the year variable is an ordinal variable that identifies the year in which the measurements were taken; lastly, the drought variable is a categorical variable with two categories indicating whether the measurements were taken before or after 1977, the year in which there was a major drought.

*Exploring the Data–Posing Investigative Questions*

As students are finding out about the dataset, they will probably start to wonder about the data and what it might show. For example, they might wonder if the drought has anything to do with beak length or beak width, or if different species have different beak lengths and beak depths. They may also wonder about whether there is an association between beak length and beak depth, and they may start by wondering about beak length and beak depth and what the distributions look like. Students can start exploring the dataset, which is available at https://tinyurl.com/Finches-CODAP. Together the students and their teacher can pose *investigative questions* that they can explore with the variables they have in the dataset. For example: *What are the beak lengths of the CCD and the MGF?* (Figure 2a). Investigative questions are the questions that we ask of the data.

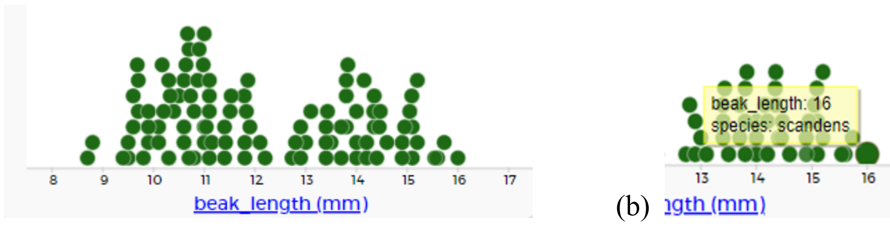


Figure 2. (a) Dot plot showing beak length for the Common Cactus finch and Medium Ground finch (b) Zoomed in view showing information about the beak length when the dot at 16 is selected

*Describing Distributions—Asking Analysis Questions*

Describing distributions involves considering overarching statistical concepts: contextual knowledge, distribution, graph comprehension, variability, and signal and noise (Arnold, 2013). Using *I notice …* (Shaughnessy, 1997) as a starter encourages students to notice features of the data visualization. For example, they might notice that the smallest beak length is 8.7 mm, and the longest beak length is 16 mm. By clicking on the data points in the graph, they also notice that the smallest beak length is from the fortis species, and the longest beak length is from the scandens species (Figure 2b). The teacher can prompt students for a statistical description of the shape of the distribution, for example, asking the *analysis questions*: *What is the shape of the distribution? How many peaks does the distribution have?* Analysis questions are questions we ask about data distributions to help us describe the distribution and answer the investigative question. Students notice, either from the teacher prompting (analysis questions) or without prompting, that there appear to be two groups of beak lengths. The distribution is bimodal; one group of beak lengths is between 8.5 mm and 12.5 mm and another group of beak lengths is between 12.5 mm and 16 mm.

Students might wonder if the two groups are due to species and decide to drop the species variable onto the dot plot (Figure 3a). They notice that the scandens (CCF) beak lengths are almost all greater than the fortis (MGF) beak lengths. A decision is made to pose a new investigative question: *Does the beak length of the CCF tend to be greater than the beak length of the MGF?* (Figures 3b and 3c).

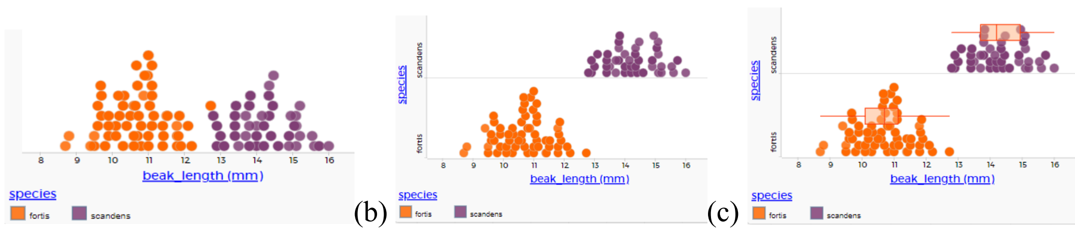


Figure 3. (a) Color coded by species (b) Split by species (c) With box plots

The teacher could ask the students about what environmental features might account for why the CCF has a longer beak length than the MGF (see Carnegie Science, 2020 video from 20:30 mins to 22:30 mins).

*Modeling Thinking and Building Experiences Faster—Analysis Questions*

Additional investigative questions are posed and explored. For example, knowing that the beak length for CCF was longer than the beak length for the MGF, students might wonder: *Does the beak depth of the CCF tend to be longer than the beak depth for the MGF?* (Figure 4a). They also wonder: *What association, if any, is there between beak length and beak depth for MGF and CCF?* (Figure 4b).

(a) (b)

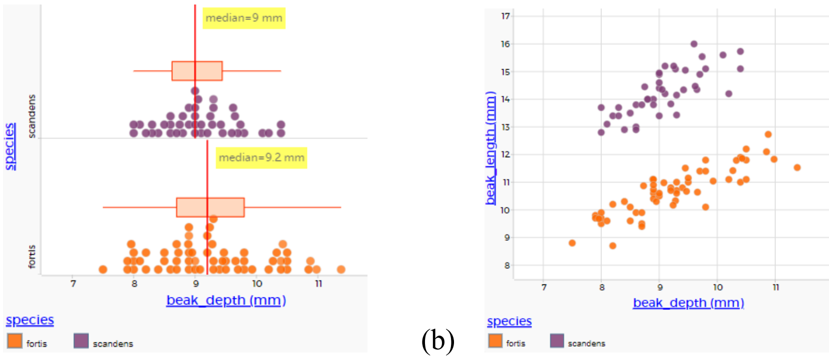


Figure 4. (a) Dot plot showing beak depth for CCF and MGF (b) Scatter plot showing beak length and beak depth for CCF and MGF

Statistical thinking "is the touchstone at the core of the statistician's art" (Wild & Pfannkuch, 1999, p. 223), and to quote Wild and Pfannkuch further, "for most of us, it [statistical thinking] has been much more a product of experience, war stories and intuition than it is of any formal instruction that we have been through" (p. 223). Using analysis questions allows the teacher to both model statistical thinking and support students to build experiences faster. Teachers with their students can develop a set of analysis questions to support describing distributions for different situations—summary, comparison, association, and time series. Taking the overarching statistical concepts into consideration, features of distributions can be identified and then analysis questions created. These analysis questions are questions we ask ourselves as we start to describe the distribution. Table 2 gives examples of analysis questions for summary, comparison, association, and time series situations.

How can we use analysis questions to model thinking? Teachers can use the analysis questions to *think out loud*, and by answering the analysis questions they start to develop a description of the distribution. For example, *What association, if any, is there between beak length and beak depth for MGF and CCF?* (Figure 4b). The teacher observes the data visualization with their students and together they notice that there are two distinct clusters of points for the association between beak depth and beak length. One cluster, the top one in purple is the scandens species, the CCF, and the bottom cluster (orange) is the fortis species, the MGF. The teacher notes that for their students, they have already attended to contextual knowledge (variables and who the data is from), and they have started to decode the visual shape by identifying the two clusters of points. The teacher then indicates that they will describe the MGF cluster first. The teacher asks *is the trend for the MGF linear? Is the trend positive or negative? Is the association strong or weak?* The quadrant count ratio (QCR) is found for MGF to be 0.63. Answering this collection of questions, the teacher notes with the students that there is a positive linear association between the beak length and beak depth for the MGF. The QCR of 0.63 also shows that there is a fairly strong positive association between MGF beak length and MGF beak depth. Similarly, the class describes the CCF cluster. *Is the trend positive or negative? Is the association strong or weak?* The QCR is found for CCF to be 0.77. Answering this collection of questions, the teacher notes with the students that there is a positive linear association between the beak length and beak depth for the CCF also. The QCR of 0.77 also shows that there is a fairly strong positive association between CCF beak length and CCF beak depth. Analysis questions help to identify key features to describe for different analysis situations and support teachers to model statistical thinking and support students to engage in statistical thinking and to build experiences faster.

Table 2. Examples of analysis questions for summary, comparison, association, and time series situa

| *Overarching statistical concept* (characteristic) | Summary and Comparison Situations | Association Situations | Time |
|---|---|---|---|
| *Contextual knowledge*<br>Variable, group, or population of interest | What is the variable we are describing? Are there any units for the variable? Who are these data from? | What are the variables we are describing? Are there any units for the variables? Who are these data from? | What<br>descr<br>the va<br>perio |
| *Distributional*<br>Aggregate view, symmetry, modality, individual cases | Is the shape of the distribution(s) symmetrical? If not, is it left or right skewed? What is the maximum value? Minimum value? | Is the trend linear? What is the shape of the scatter? | Is the<br>decre<br>What<br>the va<br>highe |
| *Graph comprehension*<br>Decoding visual shape, unusual features | What is the shape of the distribution? How many peaks does the distribution have? Are there any clusters or groupings? Are there any extreme (outlying) values? | Is the trend positive or negative? Is the association weak or strong? Are there any clusters or groupings? Are there any extreme values? | What<br>increa<br>Are a<br>than c<br>value |
| *Variability*<br>Spread, density | Describe the interval of values from highest to lowest. What is the interquartile range? Which group is more spread out? How does the position of the majority of the values for group A compare to the position of the majority of the values for group B? | Is there a lot or a little vertical spread in the scatter plot? How close are the points to the $y = x$ line (as appropriate)? Is there a lot of scatter? Are there clusters? | Is the<br>patter<br>patter<br>Wher |
| *Signal and noise*<br>Center | What is the median/mean value? What is the difference between the median values for the two groups?<br>Describe the middle 50% of data.<br>Where does the data peak? | Is the trend linear? How well does the "line of best fit" fit the data? What is the quadrant count ratio and what does it tell us? | Overa<br>increa |

## CONCLUSION

The foundational pillar of statistics and data science is the statistical problem-solving process and its four components. Questioning continually throughout the statistical problem-solving process is essential. From school level and beyond, students and practitioners will work with this same problem-solving process, picking up new tools for their toolboxes as they develop the conceptual understanding and statistical habits of mind to appreciate the beauty of statistics but also to be healthy skeptics of statistical information. Quoting from *GAISE II:*

> It is critical that statisticians—or anyone who uses data—be more than just data crunchers. They should be data problem solvers who interrogate the data and utilize questioning throughout the statistical problem-solving process to make decisions with confidence, understanding that the art of communication with data is essential. (Bargagliotti et al., 2020, p.8)

Teachers play a key role in supporting students to learn about and pose and ask statistical questions. Best practice in modeling and precise use of language will support students to pose more effective statistical questions in the future than they have in the past. Teachers can provide guidance on how students can engage with the statistical content and the different purposes of questioning throughout the statistical problem-solving process. Until teachers provide opportunities to distinguish the different types of statistical questions, students typically call on the only statistical questions they have been exposed to, analysis-type questions asked of statistical displays in textbooks. We encourage all statistics educators to be more careful in their language use and to model using the different types of statistical questions.

## REFERENCES

Arnold, P. (2013). *Statistical investigative questions. An enquiry into posing and answering investigative questions from existing data* [Doctoral dissertation, The University of Auckland]. ResearchSpace@ Auckland.

Arnold, P. (in press). *Statistical investigations: Te tūhuratanga tauanga.* NZCER Press.

Arnold, P., Bargagliotti, A., Franklin, C., & Gould, R. (2022). Bringing complex data into the classroom. *Harvard Data Science Review, 4*(3). https://doi.org/10.1162/99608f92.4ec90534

Arnold, P., & Franklin, C. (2021). What makes a good statistical question? *Journal of Statistics and Data Science Education*, *29*(1), 122–130. https://doi.org/10.1080/26939169.2021.1877582

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education II (GAISE II). A framework for statistics and data science education.* American Statistical Association; National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Carnegie Science. (2022, February 9). *40 years of evolution of Darwin's finches—Drs. Peter and Rosemary Grant* [Video]. YouTube. https://www.youtube.com/watch?v=bLJP9kpymYs&t=142s

Grant, P. R., & Grant, B. R. (2013). *40 years of evolution. Darwin's finches on Daphne Major Island* [Data set]. Dryad. https://doi.org/10.5061/dryad.g6g3h

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph & K. Carr (Eds.), *People in mathematics education. Proceedings of the 20th annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 6–22). University of Waikato. https://merga.net.au/Public/Public/Publications/Annual_Conference_Proceedings/1997_MERGA_CP.aspx