Carnegie Foundation for the Advancement of Teaching

1.1.1

# The Statistical Analysis Process

## INTRODUCTION

### What is Statistics? Why do we Study Statistics?

Statistics is about using **data** to answer questions. Data is information that we collect from our world. Data involves facts and observations that we make. Before scientists created statistics and before people used data, they would use opinions and hunches to explain how the world worked. A lot of times these explanations were wrong.

For example, people once believed that the sun revolved around the earth. When people started making observations and using measures, they discovered that this was incorrect. The earth revolves around the sun. Data helps us draw better conclusions. In this example, data helped us see that the earth goes around the sun, not the other way around.

Data can help us answer many types of questions.

- Students can use data to help pick a college that is best for them.
- Teachers use data to find the best ways to educate their students.
- Medical professionals use data to learn if new treatments actually work.
- Voters need data about their society and planet to create a better democracy.
- Politicians use data to better represent the people who elect them.

In statistics, we gather, summarize, and analyze data to search for answers to our questions.

Let's begin with an exercise that will help us think about how statistics can help us answer a question that we may have. Remember that data involves facts, observations, and measures about a particular topic or idea.

How significant is your birthday? Many people believe that habits, hobbies, likes, dislikes and many other things are affected by the time of year you were born. One such theory involves *chronotypes*. The term "chronotype" refers to a person's natural tendency to be most active and alert at certain times of the day. People who are most active early in the day are often labeled as morning people, early birds or larks. Those who are at their best in the evening are called night people or owls. Now imagine that you want to answer the following question: Can someone's birth month influence his or her chronotype?

Statway® Version 3.0, © 2016 The Carnegie Foundation for the Advancement of Teaching

1 If you wanted to gather information about birthdates and chronotypes, what kind of information would you look for? How could you use data to answer this question?

**Statistical analysis** is the process of looking at data to learn about something bigger. Looking at data allows us to make **generalizations** about **populations** that are large and difficult to understand. We can think of the statistical analysis process in four steps.

## Steps in a Statistical Analysis

**Step 1: Ask a question that can be answered by collecting data.**

**Step 2: Decide what to measure and then collect data.**

**Step 3: Summarize and analyze the data.**

**Step 4: Draw a conclusion and communicate the results.**

We will now do an activity that will help us learn about the statistical analysis process. In this activity we use the statistical analysis process to investigate a question about whether we can use birth months to predict chronotypes.

## Chronotypes and Birthdates

*Can someone's birth month influence his or her chronotype?* In this investigation, we will focus on exploring whether or not there is a relationship between birthdates and body clocks. That is, we are trying to determine if being a morning person, night person or neither/both is related to the month when you were born.

**Notes**: Chronotype theory states that there are four birth months that correspond to each chronotype (you will find out which ones later in the lesson). For the following activities, let's assume the chance of being born in each month is about the same.

You and your classmates will first answer a set of questions that will help each of you to determine whether you are a morning person, an evening person or neither/both.

What was your chronotype? (Morning Person, Night Person, Neither/Both)

Result of Activity_____.

Let's take a closer look at the **statistical analysis** process. We want to investigate if the month when you were born can predict your chronotype. Before going into great detail, let's think about those steps.

## TRY THESE

2 Answer each of the following questions about the four steps of the statistical analysis process.

### Step 1: Ask a question that can be answered by collecting data.

A What question are we trying to answer in our investigation?

### Step 2: Decide what to measure and then collect data.

B What information can we get from each student to answer this question? How are the data related to the question we are trying to answer?

### Step 3: Summarize and analyze the data.

C After each student learns which chronotype is predicted by his or her birth month, some students will find that the predicted chronotype matches the one they selected and others will not. How could we *summarize* the results for all individual students?

### Step 4: Draw a conclusion and communicate the results.

D Once the data are summarized, how can we use the summary to answer the research question? What would we expect to learn from your data? Can we generalize, or **infer**, the results to something bigger than our class?

**Language Tip**

**To *infer* means to use statistical evidence to make a conclusion that applies to a larger group than the sample.**

## NEXT STEPS

We have now outlined how we might approach each of the steps in a statistical analysis in order to answer the question of whether someone's birth month can influence their chronotype. Now let us see whether we can carry out those steps. We have actually completed steps 1 and 2. We have asked a question and we have collected data (by having everyone record the chronotype that matches them best).

We now prepare to *summarize and analyze* our data. To do this, we need to think about how the data can be used to answer our question. We will use *probability* to help in this. Talk about the following questions with your group to begin this process.

3 In the first part of this lesson you classified yourself as a morning person, a night person, or neither. Suppose that those chronotypes *are not* related to birth months.

A Despite the fact that chronotypes and birth months are not related, will the choices students make still match the prediction from time to time? Explain why you think this.

B Out of all students in your class, about what **fraction** of choices should match predictions? Why do you think this?

4 Suppose that chronotypes *are* related to birth months.

A If chronotypes are related to birth months, do you expect the **fraction** of student choices that match the prediction to be *greater* or *less than* the fraction that you chose in Question 3B?

B What fraction of the class would need to choose chronotypes matching the prediction for you *to be convinced* that someone's birth month can predict their chronotype? Why do you think this?

5 Imagine that half the students in your class select chronotypes matching the prediction. Does this *guarantee* that the theory could be true? If not, give another explanation for why so many students picked the matching chronotypes.

## NEXT STEPS

### Using Probability to Discover What May Happen by Chance

**Language Tip**
***Chance variation*** **describes the type of differences we would naturally expect to see between different samples.**

If birth month has nothing to do with personality traits, then we expect the fraction of students in the class who pick the chronotype that corresponds to their birth month to be around 1/3. But how far above 1/3 would would the fraction need to be in order to convince us that the opposite is true, that is, that a person's birth month can predict their chronotype? In the next activity, we will answer this question by learning about the values which are most probable. Probable values are those that are most likely to occur by **chance variation**.

6 Your instructor has given you three cards, and asked you to write "Match" on one and "No Match" on the other two. Mix the cards and choose one of them. Complete the following information.

- My card says:
- Number of students in the class:
- Number of students who randomly selected the card that said "Match":
- Fraction of the students who randomly selected "Match":
- **Proportion** of the students who randomly selected "Match" (written as a decimal):

*Note:* In statistics, a **proportion** is a number between 0 and 1. It represents a portion out of the total. We usually give proportions as decimals or percents. We can calculate a decimal proportion by dividing the **numerator** of a fraction by the **denominator**. For example, if the fraction of students who picked a "Match" card is (5/7), then you divide 5 by 7. The proportion would be 0.714. To change to a percent we multiply by 100 or move the decimal 2 places to the right. 0.714 = 71.4%.

7 Is the proportion of students in the class who picked a "Match" card equal to 1/3? If not, is it greater than or less than 1/3?

8 If the class repeats this process a second time, would we get exactly the same proportion of "Match" cards for the class? Why or why not?

We will repeat the process of observing *chance variation* a large number of times. This will help us understand what kinds of match proportions are consistent with picking cards at random.

9 First, write the *class proportion* of "Match" cards in the following table under **Observed Proportion** next to Trial 1. Second, work with your classmates to repeat the process of picking cards and entering the observed proportion in the table until you have proportions for 10 trials. (*Observed Proportion* means the proportion we see from the data we collect in each trial.)

| Trial | Observed Proportion |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |

10 Answer these questions after your instructor has constructed a **dotplot** which shows the observed "Match" proportions for the trials where students picked one of three choices at random.

A What was the smallest "Match" proportion observed?

B What was the largest "Match" proportion observed?

C Did the "Match" proportion differ much from trial to trial?

11 Where is the dotplot centered? Why do you think this is?

12 Summarize what the dotplot tells you. Why did your instructor create this graph?

A **distribution** of data (like "Match" proportion) shows what the possible values are and how often each of the values occurs. We can compare the distribution of "Match" proportions in the dotplot with the results of the first activity where you selected your chronotype. In both cases you picked one item out of three. First you chose one chronotype out of three. Next you randomly selected one card out of three. If birth month does not predict chronotype well, each activity should result in similar proportions—the chronotype you selected should match the chronotype for your birth month about one time in three, and you would have picked a "Match" card about one time in three.

13 Look at the dotplot and think back to the chronotypes you and your classmates selected. How could this graphical display help you decide whether your class's proportion of correct chronotypes would support the birth month theory?

14 Use the dotplot to answer the following question. If students in the class were picking a chronotype at random, it would be unusual to see a "Match" proportion as large as _____.

15 Suppose that the proportion of students in your class who picked the chronotype that matched their birth month was 0.40. Would this convince you that chronotype and birth month are related ? Explain your answer using the dotplot.

16 How large does the proportion of students who picked a matching chronotype have to be to convince you there is a relationship between chronotype and birth month? Look back at your answer to questions 14 and 15 to help you answer this question.

## Draw a Conclusion and Communicate the Results

Now it is time to take a look at the *actual* class data gathered on chronotypes. You will decide if the results provide evidence that there is a relationship between chronotype and birth month. Fill in your answers:

Chronotype that best matched you:

Chronotype that matches your birth month:

Did the chronotype you selected match the prediction from your birth month?

Number of students in the class:

Number of students who have matching chronotypes for their birth month:

To provide a *summary* of your class's data, compute the proportion of correct matches:

Decision time! It is now time to make an inference. An inference is a decision based on the evidence we have gathered.

17 Does the class proportion of matches provide convincing evidence that chronotypes are related to birth month? Why or why not? (*Hint*: Look at your class's data and compare it to the dotplot from Question 10.)

## NEXT STEPS

The task you have just completed illustrates the **statistical analysis process**, which we have described in four steps. These are given again below.

### Steps in a Statistical Analysis

1. Ask a question that can be answered by collecting data.
2. Decide what to measure and then collect data.
3. Summarize and analyze the data.
4. Draw a conclusion and communicate the results.

A statistical investigation is an ongoing process. Often, researchers analyze results of one study and this leads them to think of other research questions. Then they conduct more research. Also, researchers start to think more carefully about how they collect data and this may also lead them to think of ways to improve the data collection process.

18 In the table below, identify each step of the Statistical Analysis Process for the Chronotype investigation.

| Steps in Statistical Analysis | For the Chronotype Investigation |
|---|---|
| 1. Ask a question that can be answered by collecting data. | |
| 2. Decide what to measure and then collect data. | |
| 3. Summarize and analyze the data. | |
| 4. Draw a conclusion and communicate the results. | |

## TRY THESE

Consider how you would identify the steps in statistical analysis in the the following two studies:

**Language Tip**
***Convincing evidence*** **is information that provides very strong support for a conclusion. The evidence is beyond what would be expected due to chance variation.**

### Study 1—A Study about a Population

A group of researchers studied women who had visited a fertility clinic. The researchers wondered if fewer than half of the women who visit the clinic would want to choose the sex of their future child.[1]

The researchers mailed a survey to women who had visited the clinic. The survey asked women if they would choose the sex of their future child, if they were able to do so. A total of 561 women responded to the survey. Of these 561 women, 229 said that they wanted to choose the sex of their future child.

The researchers did statistical analysis on the data. Based on their statistical analysis of these data, the researchers concluded that there is **convincing evidence** that *fewer than half* of the women who visit the clinic would choose the sex of a future child. This conclusion is based on the following observation:

> If, in reality, at least half of women who visit a fertility clinic would like to choose the sex of a future child, it would be very unusual to observe a percentage as low as 41% in a sample of 561 women who visited the clinic (229/561 ≈ 0.41).

19 Now that you know the details from Study 1, complete the four steps of the statistical analysis process in the following table.

| Steps in Statistical Analysis | Study 1 |
|---|---|
| 1. Ask a question that can be answered by collecting data. | |
| 2. Decide what to measure and then collect data. | |
| 3. Summarize and analyze the data. | |

[1]Tarun Jain et al., "Preimplantation Sex Selection Demand and Preferences in an Infertility Population," *Fertility and Sterility* 83, no. 3 (2005): 649-58.

| | |
|---|---|
| 4. Draw a conclusion and communicate the results. | |

## Study 2—A Study about an Experimental Treatment

Researchers wanted to know if people think a task will be hard to accomplish when the instructions are difficult to read.[2] To answer this question, researchers randomly divided twenty student volunteers into two groups of 10 students each. Researchers gave instructions to each group of students using different fonts (see below). Instructions for one group were written in a large upright font. The other group was given the *same* instructions but in a font that used *hard-to-read italics*. Researchers asked students to read the directions and say how many minutes they thought the task would take. Researchers did this in order to figure out if the fonts used for the instructions made a difference.

This is the easy-to-read upright font that was used in the study.

*This is the hard-to-read italic font that was used in the study.*

The first group of students, those that read the instructions printed in the easy font, had an average time estimate of 8.23 minutes. The other group, the group that read the instructions in the *hard-to-read italic* font, had an average time estimate of 15.1 minutes.

Researchers concluded that such a large difference between the averages was not likely to have occurred by chance. There was evidence that people think a task will be harder when instructions are difficult to read.

20 Complete the four steps of the statistical analysis process for Study 2 in the following table.

| **Steps in Statistical Analysis** | **Study 2** |
|---|---|
| 1. Ask a question that can be answered by collecting data. | |
| 2. Decide what to measure and then collect data. | |
| 3. Summarize and analyze the data. | |

[2]Hyunjin Song, "The Effects of Processing Fluency on Judgment and Processing Style: Three Essays on Effort Prediction, Risk Perception, and Distortion Detection" (PhD diss., The University of Michigan, 2009).

| | |
|---|---|
| 4. Draw a conclusion and communicate the results. | |

21 These two studies both follow the same general process but they are different in some ways. What are two ways that these studies are different?

STUDENT NAME _____DATE _____

## TAKE IT HOME

In the lesson, your class looked at two different sets of data:

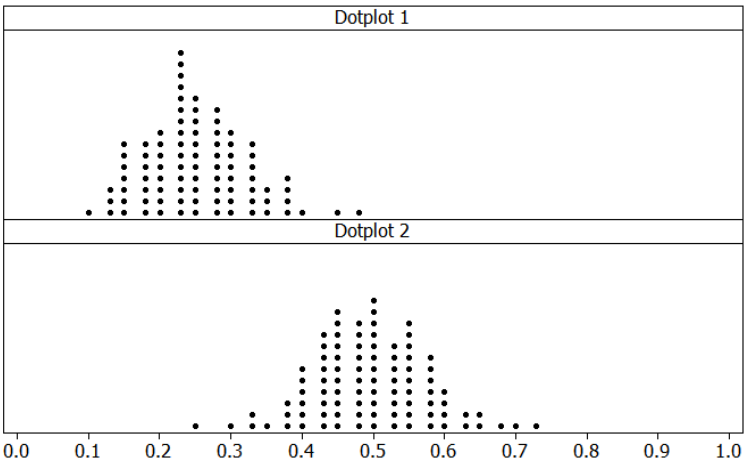
- Hypothetical Data: These were the data collected from the match card activity and the dotplot, where you were given three cards (two "No Match" cards and one "Match" card). These were *hypothetical* data because we used a computer to create much of the data. The dotplot the class looked at showed the data on a graph. The hypothetical data allowed us to see what the proportion and dotplot would look like when there was *no relationship* between chronotype and birth month. We saw that if there is no relationship between birth month and chronotype then about 1/3 (or 0.33) of the class will pick the matching chronotypes.
- Actual Data: These were the data collected when you answered questions to determine your chronotype and found the chronotype that matches your birth month.

You compared the *hypothetical* results to the *actual* result for the class. You did this to see whether the proportion in the class was consistent with **chance variation**. If the class result was unlikely to occur by chance alone this provided evidence to support the chronotype and birth month theory.

1 Imagine an investigation in which 40 students were asked to determine their chronotype by choosing from among *four* chronotypes instead of three: morning person, night person, neither, or both.

A If there is *no relationship* between chronotype and birth month, about what fraction of the students would you expect to pick the chronotype that matches their birth month? Why do you think this?

B Researchers computed the proportion of students who selected the chronotype that matched their birth month by chance. This process was repeated a large number of times to generate data that was used to construct one of the dotplots below. Which dotplot do you think is the one that was constructed this way?




C Why did you pick this dotplot?

D What proportion of the 40 students needs to match correctly to provide convincing evidence that there is a connection between birth month and chronotype? Explain your reasoning. (*Hint*: Use your answer from Question 1B.)

2 Read the following study description:

The United States Government recommends that to stay physically fit, middle-aged adults (ages 40 to 60) need to burn 150 to 400 calories per day doing exercise. Researchers at Minnesota State University, Mankato, wanted to learn whether middle-aged adults who used the Wii Fit video game exercised enough to meet the government's fitness recommendations.[3] The Wii Fit is a video game that includes exercises.

The researchers taught 20 middle-aged adult volunteers how to use the Wii Fit video game. On the day after they were trained, the adults exercised for 20 minutes with the Wii Fit. Researchers measured the

[3]B. Guderian, "The Cardiovascular and Metabolic Responses to Wii Fit Video Game Playing in Middle-Aged and Older Adults," *The Journal of Sports Medicine and Physical Fitness* 50, no. 4 (2010): 436-42. http://www.ncbi.nlm.nih.gov/pubmed/21178930.

total amount of energy each of the adults in the study used in calories. They found that the average energy used was 116 calories for the 20 minute session.

Based on the results of the study, the researchers concluded the Wii Fit video game could be a helpful form of exercise for middle aged adults. But, for exercise with Wii Fit to meet the government's recommendation, the researchers stated that the length of the exercise session should be increased from 20 minutes to 30 minutes.

A Complete the four steps of statistical analysis for the study in the table.

| **Steps in a Statistical Investigation** | **Study** |
|---|---|
| 1. Ask a question that can be answered by collecting data. | |
| 2. Decide what to measure and then collect data. | |
| 3. Summarize and analyze the data. | |
| 4. Draw a conclusion and communicate the results. | |

B Based on the results of this study, what next steps do you recommend to the researchers investigating this issue?

Carnegie Foundation for the Advancement of Teaching

# 1.1.1-S

# Chronotype Chart (Supplement)

**Are you a morning person, a night person, or neither one or the other?** Circle the answers that best apply to you. Then calculate your results on the following page.

| Characteristic[1] | 1 | 2 | 3 |
|---|---|---|---|
| Most alert | Around noon | Around 6 PM | Equally alert at noon and 6 PM |
| Most productive | Morning | Evening | Equally productive in morning and evening |
| Alarm clock | Don't need it | Need multiple alarms | One alarm is enough |
| Favorite exercise time | Morning | Evening | Equally likely to exercise in morning and evening |
| Mood | Declines steadily all day | Rises steadily all day | Same throughout the day |
| Favorite meal | Breakfast | Dinner | I like breakfast and dinner equally |
| Coffee use | None or a small amount | Quite a bit | Somewhere in the middle |
| Class Time | I would prefer a class at 8 AM | I would prefer a class at 7 PM | No preference between a class at 8 AM or 7 PM |
| Morning behavior | Full of energy | Out of steam | Somewhere in the middle |
| Evening behavior | Out of steam | Full of energy | Somewhere in the middle |
| Travel | I get bad jet lag | Adapt quickly to time zone changes | I get a bit of jet lag |

[1]Source: Adapted from http://www.nasw.org/users/llamberg/larkowl.htm.

**Compute your results**:

Count how many circled answers you have in each column and enter the information below.

Column 1 Total: _____

Column 2 Total: _____

Column 3 Total: _____

**Interpretation**:

If your Column 1 Total is the highest, this indicates that your chronotype is a morning person.

If your Column 2 Total is the highest, this indicates that your chronotype is a night person.

If your Column 3 Total is the highest, this indicates that your chronotype is neither a night nor morning person.

*Tiebreaker. If you have a tie between your two top columns, simply select the result with which you most identify.