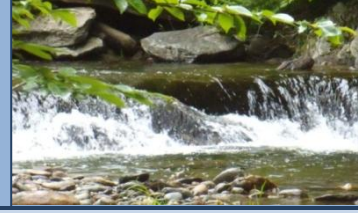
Data Analysis Tutorial



Overview

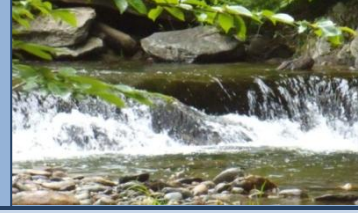
The following is a series of tutorials that have been created to help Streams Project participants understand and analyze Project data.

Module 1: **What is science?**
Module 2: **Understanding Streams Project Data**
Module 3: **Refining and Retrieving Data**
Module 4: **Data Exploration**
Module 5: **Using Statistical Analysis to Explore Relationships**
Module 6: **Summarizing Results and Drawing Conclusions**

** There are several videos in this tutorial. To watch the videos, download the QuickTime Player, if it's not already on your computer: http://www.apple.com/quicktime/download/

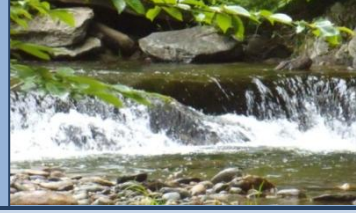
Funding for this tutorial was provided by NSF EPS Grant #0701410

Data Analysis Tutorial



Module 1: What is science?

Module 1



What is science?

"The whole of science is nothing more than a refinement of everyday thinking" – Albert Einstein, Physics and Reality, 1936

This module helps you understand and review:

1) **What science is**
2) **What makes a question scientific**
3) **How the scientific community makes advances**
4) **What type of study is the Streams Project**
5) **How to find your research hypothesis or question**

These topics might seem simple.
Taking a few minutes to contemplate them will help you develop or refine your research.

Module 1

What is science?

Science is…

…both **a body of knowledge** and **an ongoing process of discovery !** The process of discovery doesn't stop: one answer can lead to new and exciting questions. [1]

How can scientists find new topics to research?
Nature is filled with mysteries! Scientists are on an age-old quest to solve them. Scientists have been investigating nature's mysteries for centuries.

Many times scientists follow the same main topic, and chip away at the answer over many years or generations. Check out the example of evolution, where the driving questions is "what is the history of life on Earth?"

What's *natural*? [1]

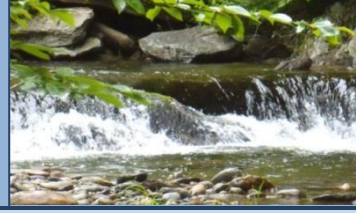
In everyday language, you may use or hear the word *natural* to refer to food, cleansers, remedies, or other products.

In science, *natural* has a much broader meaning. It refers to any element of the physical universe – whether made by humans or not, including matter, the forces that act on matter, energy, the constituents of the biological world, including humans and society.

Science can study things like the human smile, decision-making, and precipitation patterns because they are part of the physical universe, which we also call the *natural world*.

[1] Understanding Science. UCMP.
[2] Ambrose & Ambrose. Pp 7

Module 1



How science advances

"All questions start with an observation." – Ambrose & Ambrose p9

Ideas that scientists have about how to solve the mysteries of nature can be translated to ***expectations***.

Example: If Professor Champlain has an idea that temperature affects the rate of algae growth, the Professor could say "I expect algae to grow faster at higher temperatures."

Scientists make ***observations*** to test their ideas. They use those ***observations*** to determine whether their ***expectations*** and ***ideas*** hold true.

Example: To find out if algae grows faster at higher temperatures, Professor Champlain grows algae in two aquariums, one cool and one warm, and then observes or measures which one has more algae.

While scientists have access to many powerful tools, science and scientist have limits. The checklist helps us understand what kinds of mysteries science can help solve.

What makes a question scientific:

- ☑ Focuses on the natural world
- ☑ Aims to explain the natural world
- ☑ Uses testable ideas
- ☑ Relies on evidence
- ☑ Involves the scientific community
- ☑ Leads to ongoing research
- ☑ Benefits from scientific behavior

Module 1

How science advances:

Hypothesis vs Theory

Scientists use the existing body of knowledge to make hypotheses about observed phenomena.

All science advances by rejection of a hypothesis.[2]

What's the difference between a **theory** and a **hypothesis**?[1]

	In everyday language		In science
Theory	A hunch; intuition	vs	A powerful explanation that applies to a broad range of observations.
vs			
Hypothesis	A guess	vs	A proposed explanation of a fairly narrow set of phenomena. Based on reason and prior information.

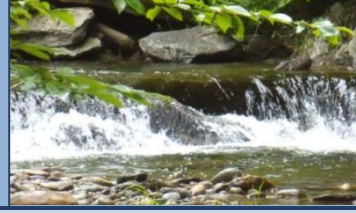
It is essential that a hypothesis is testable!

That means that your hypothesis has to be about something that :

- ❑ can be observed and
- ❑ is specific enough to address

[1] Understanding Science. UCMP.
[2] Ambrose & Ambrose. Pp 7

Module 1



How science advances: Surveys vs manipulations

To test their hypotheses, many scientists use:

experimental manipulations

Typically highly- controlled, where one or more factors is regulated by the experimenter

Observation: More algae seems to show up on the lake shores when we have a heat wave.
Hypothesis: Higher temperatures increase algae growth.

Study design: Build experimental chambers where algae can grow, such as aquariums, that are specially designed to regulate temperature. Use the chambers to increase, or manipulate, the temperature of some of the aquariums and observe if algae grew faster in the aquariums with higher temperature.

OR

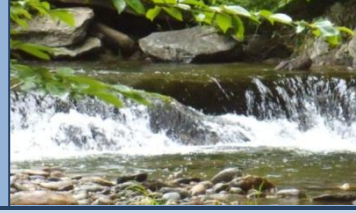
observational studies or surveys

In environmental science, often conducted in an uncontrolled habitat, such as a lake or forest reserve

Observation: Sand shiner minnows are more frequently seen in the shallow waters on sandy beaches of the lake.
Hypothesis: Sand shiners prefer sandy beaches to other habitats, such as those with lots of underwater vegetation or logs.

Study design: Place minnow traps in a range of habitats in the lake: sandy, muddy, vegetated, and in places with or without logs. Deploy the traps for a summer, identify and count the species caught in the traps over regular time intervals throughout the summer. At the end of the summer, when all the data are collected, compare the number of sand shiners caught in each type of habitat.

Module 1



What type of study is the Streams Project?

It's a survey study!

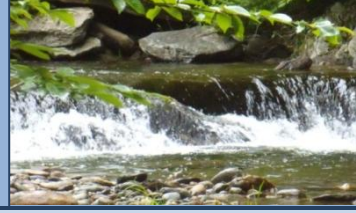
The Streams Project is a survey because throughout the past several months, you and other participants have collected data that tells us about streams. We did not manipulate the streams and they were in uncontrolled habitat.

The types of data that we collect, such as phosphorus and bacteria, were chosen so that you and other scientists could ask insightful questions about relationships among:

- water quality, including macroinvertebrates
- precipitation patterns
- land use

While the Streams Project doesn't manipulate any part of the environment that we monitor, we chose stream sites that are different. For example, one of the sites you monitored was in a watershed that was largely forested and other was probably in a watershed that had more agriculture or urban development.

Module 1



How do you find your question?

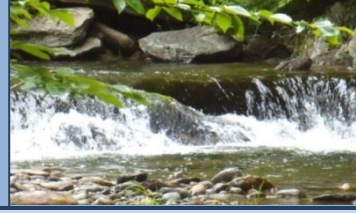
You might be wondering…

"But what kinds of questions can ask when I can't manipulate any of the parameters that I'm interested in (i.e. phosphorus concentration, stream substrate, precipitation)?"

The answer is simple: There are plenty of questions you could ask about streams because the Streams Project database was designed to allow you to research myriad questions.

Take it slow: Start by thinking broadly about the basics: water chemistry, macroinvertebrates, weather, geography, geology….

Module 1



How do you find your question: Brainstorm!

Where do you start?

Remember, science starts with an observation.

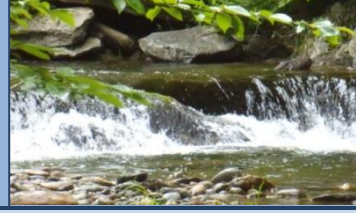
To find your question, you've got a job!

Before moving on to the next slide go through the following steps:

1. Don't worry about what data are available yet. We'll save that for later.
2. Sit down with your team and **recall your visits to your stream sites**, writing down **anything interesting you noticed** in or around the stream site. Do you have any special or strong memories? Even a funny story that you remember might lead to an observation that you can turn into a question. Talk about it with your team!
3. **Go visit your stream sites again and the macroinvertebrate Web pages.** It may jog an old memory or observation. You might notice something you never had before.
4. Write all your thoughts and topics down. Think in terms of **differences and relationships** among the parameters we measure.
5. Draw connections (literally, draw lines!) to show which topics or parameters might be related (e.g. stream discharge and phosphorus concentration).

This is the beginning or your map of ideas!
We'll turn them into scientific and answerable questions.

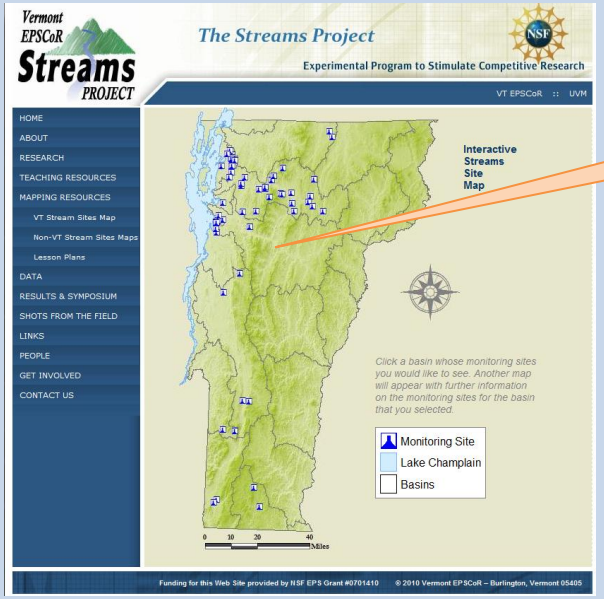
Module 1



How do you find your question: Think geographically!

Now that you've thought about how several parameters are related, let's make sure we **explore the geographic component** of the Streams Project.
Keep in mind that you can incorporate data from other group's stream sites into your own research.

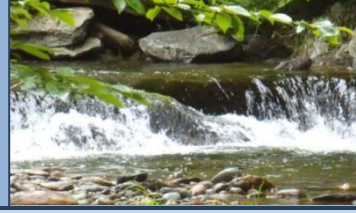
1. Visit the Streams Project mapping Web page: http://www.uvm.edu/~streams/?Content=pages/map_watersheds.inc
2. Explore the geographic range of the sites that other Streams Project participants monitor



Click on watersheds to view more detailed information about stream sites. In this example, we chose to look at the Winooski Watershed.

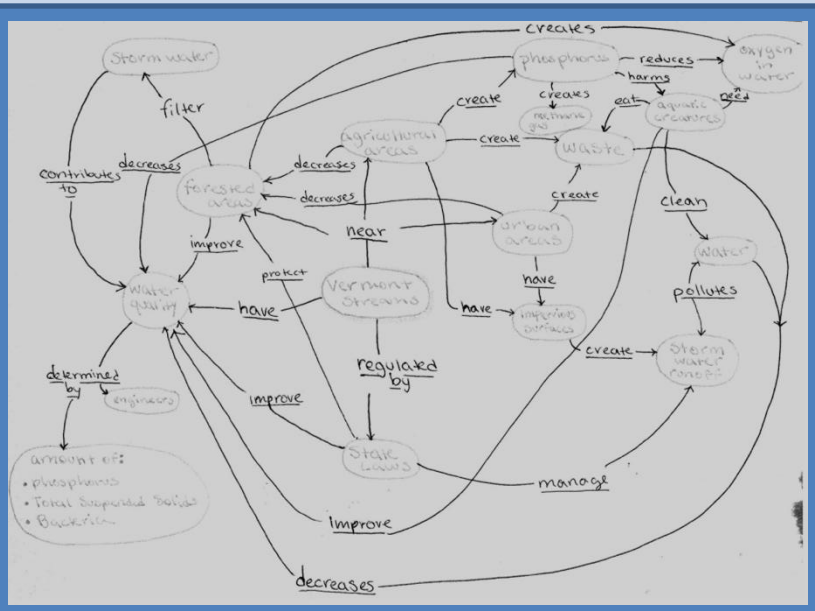
3. As a Vermonter you have some knowledge about water bodies, like streams, throughout the state. Have you ever heard a news article about water quality or water use? Use that local knowledge and experience to your advantage as you brainstorm about water quality issues throughout the state.
4. Add topics and thoughts that address geography, geology and water quality to the concept map or web that you made in the previous exercise.

Module 1



How do you find your question: Refine your observations.

Review your concept map or web with your team:



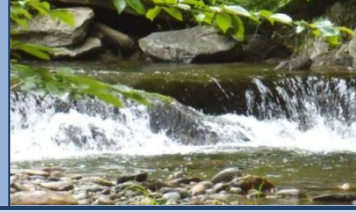
Example of a concept map or web of ideas[3]

Draw (literally!) stronger, bolder connections between parameters that you think might have interesting differences or relationships.

Talk about what kind of differences or relationships you might expect to see among the parameters. Why might you expect to see those relationships?

[3]This concept map originated from a freshman engineering course at UVM in Fall 2009 taught by Dr. Nancy Hayden. The students who created it are: Liana Schneidman, Sebastian Downs, Will Chandler, Tom Brayden.

Module 1

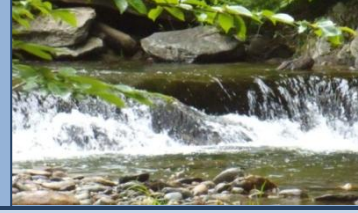


Refining and Retrieving Data

SUMMARY

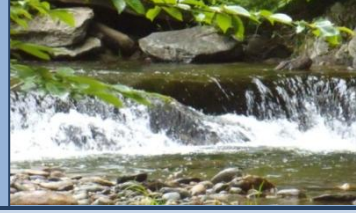
- Science is a body of knowledge and ongoing process.
- Science tests ideas and expectations about observable phenomena.
- “Hypothesis” and “theory” have a specific meaning in science that may differ from how we use those words in everyday language.
- Science advances by rejecting hypotheses.
- Scientists can test hypotheses through experimental manipulations or observational studies (a.k.a. surveys).
- The Streams Project is a survey study.
- You should have a well-developed concept map or web of ideas about differences and relationships about water quality, land use, geography, and/or macroinvertebrates.

Data Analysis Tutorial



Module 2: Understanding Streams Project Data

Module 2



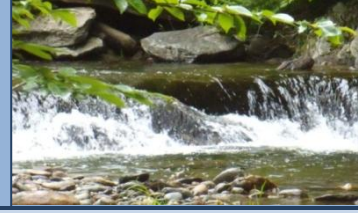
Understanding Streams Project Data

"The whole of science is nothing more than a refinement of everyday thinking" – Albert Einstein, Physics and Reality, 1936

This module helps you understand:

1) **How Streams Project data are collected**
2) **What Streams Project data are available**
3) **How to use the available data to specify your question**
4) **Why you need a hypothesis**
5) **How to finalize your question and hypothesis**

Module 2



How data are collected

You've made valuable contributions to the Streams Project data set throughout the past summer and school year! Check out some of the other people who have contributed data to the Streams Project:

Participants

High Schools
collect water and macroinvertebrate samples in streams near their schools

UVM
Undergraduates carry out independent projects

Baccalaureate College Faculty & Undergraduates
Complementary research projects

Module 2

How data are collected: A diagram of Streams Project data flow

Participants collect water and macroinvertebrate samples and make habitat observations in the field

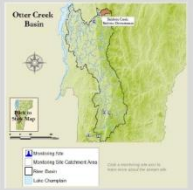
Water samples are sent to the Streams Project analytical lab at UVM

Undergraduates analyze the samples

Bacteria

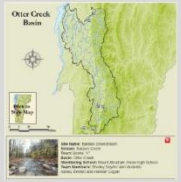
Total Phosphorus

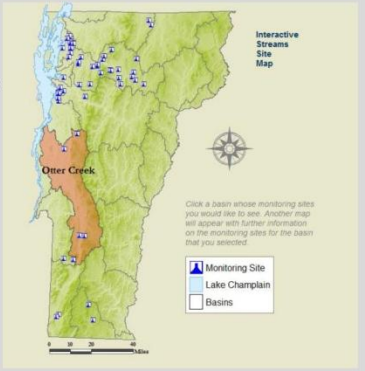
Total Suspended Solids



Undergraduates work with the Project's GIS specialist to analyze land use data and patterns. The GIS specialist also ensures that land use data are user-friendly.

High school groups monitor sites in two watersheds with contrasting land uses (e.g., agricultural, urban, forested).

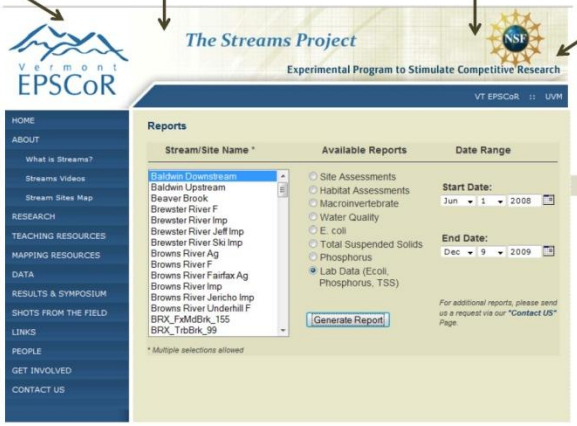




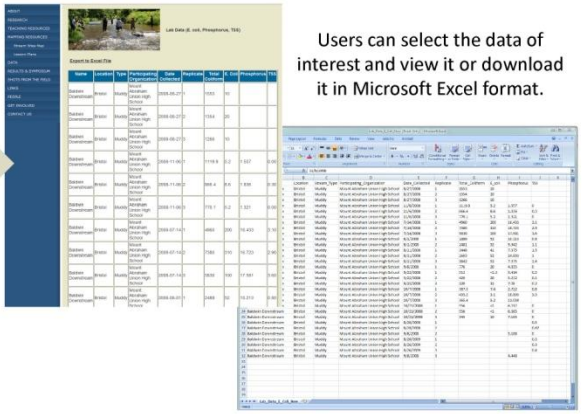
Students enter field data online

High school groups send a sub-sample of macroinvertebrates to St. Michael's College

Stream-specific, web-based macroinvertebrate identification keys are created by undergraduate interns under the direction of a science adviser. High school groups use the keys to guide them as they identify benthic macroinvertebrates.



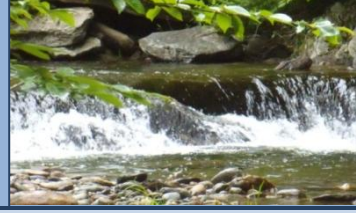
All participants contribute the data they collect to an online publicly-accessible database. Students and other researchers retrieve data from the database and use it to complete their individual research projects.



Users can select the data of interest and view it or download it in Microsoft Excel format.

Contribution to a common database provides students with the option to expand the scope of their analyses and produce more powerful results than any one team could have produced alone.

Module 2

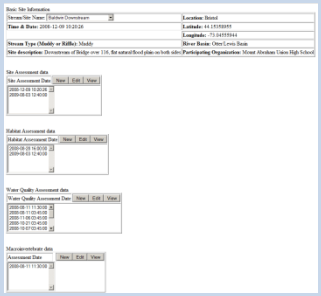


How Streams Project data are collected

The following data are available in the database:

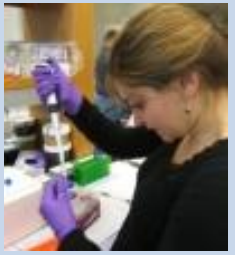
- **Site Assessment Data**
- **Habitat Assessment Data**
- **Water Quality Assessment Data**
- **Macroinvertebrate Data**

Gathered by our high school and undergraduate researchers in the field and entered into the database through this online interface:



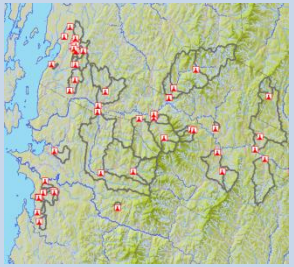
- **E.coli Lab Data**
- **TSS Lab Data**
- **Phosphorus Lab Data**

Analyzed by our lab technician and undergraduate interns at our UVM water quality lab and entered directly into the database by our database specialist.

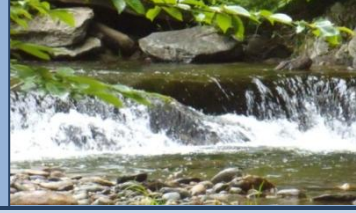


- **GIS Assessment Data**

Analyzed by our GIS specialist and undergraduate interns at UVM and entered directly into the database by our database specialist.



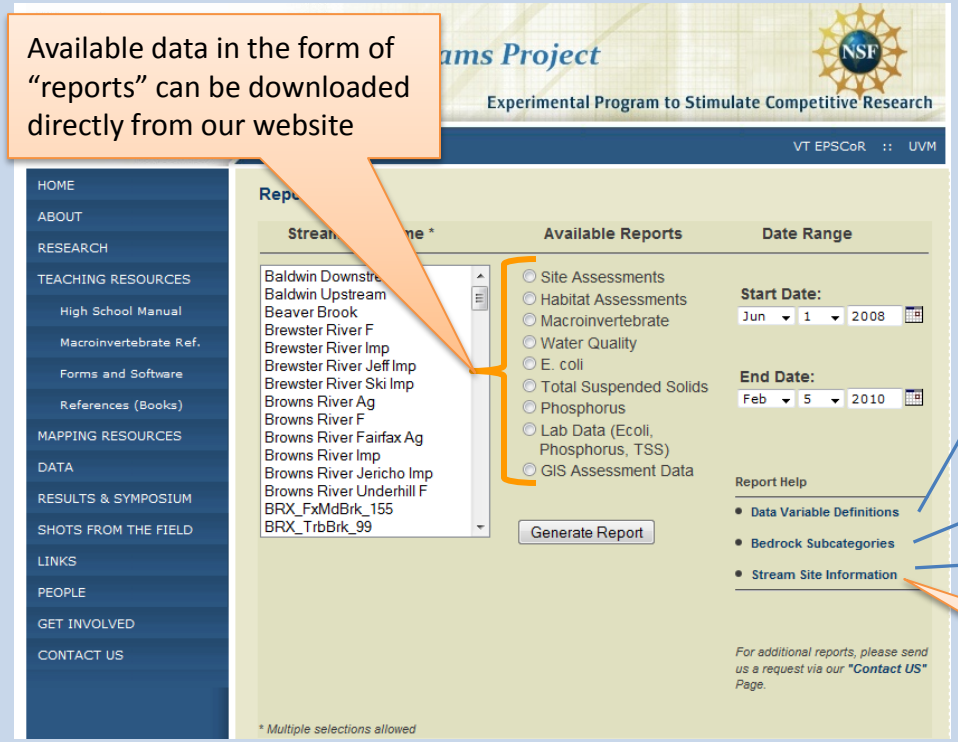
Module 2

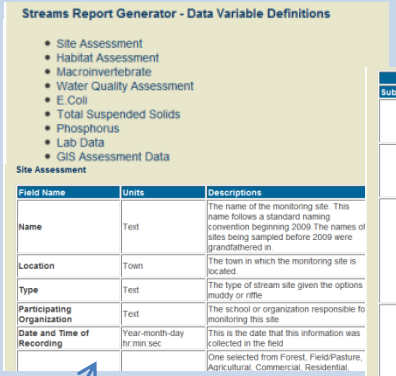


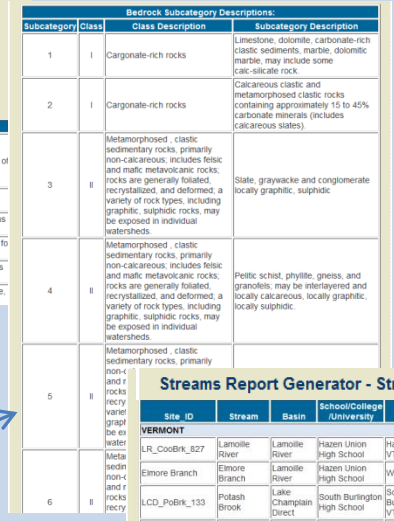
The Streams Project Database

Below is the web interface through which you will download your data. All the data sources mentioned on the previous page are listed as “Available Reports”:

Available data in the form of “reports” can be downloaded directly from our website



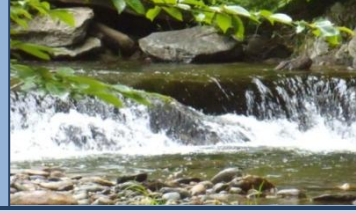




These three “Help” pages give detailed definitions of each variable. They can help you understand what each variable means and are a valuable initial resource.

We will show you how to retrieve data in the next module.

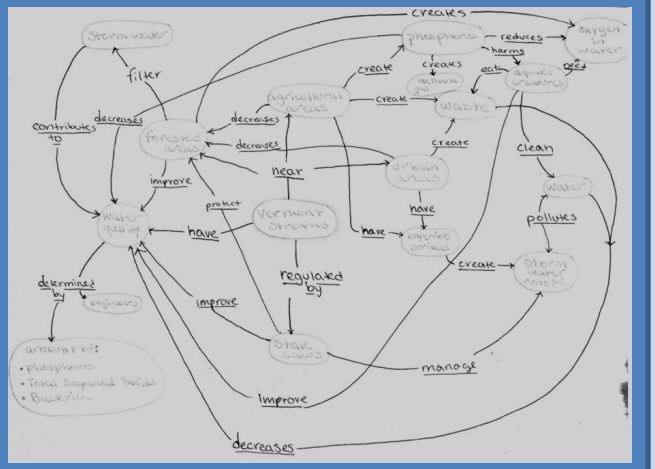
Module 2



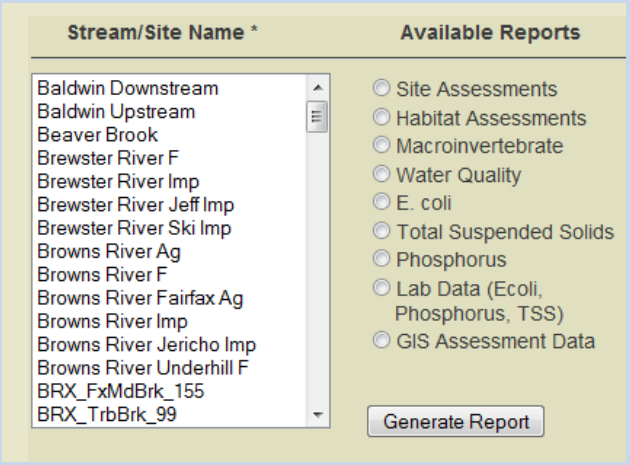
How to use the available data to specify your question

Let's learn how to turn those observations into ***answerable*** scientific questions!

Recall your concept map or web of ideas that you made in Module 1. It should look something like this:

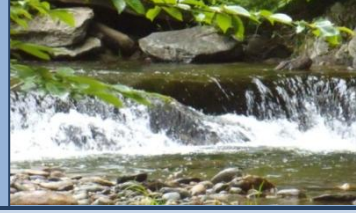


1. Look at each of the related ideas on your concept map/web.
2. Evaluate which ideas on the map have data in the database.
3. Cross out the ideas for which no data are available.



For example, if you're interested in how fish abundance is related to temperature, cross that idea out since the Streams Project doesn't collect data on fishes. However, before you write that idea off completely, ask yourself if there is some other type of data that we DO have that could substitute for fishes, like macroinvertebrates. Re-write the relationship using macroinvertebrates in place of fish!

Module 2

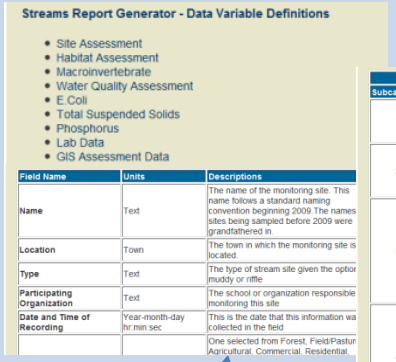


How to use available data to specify your question

At this point, you probably have at least 2 or 3 ideas about what differences or relationships you want to investigate.

Specify the following for each idea from your concept map:

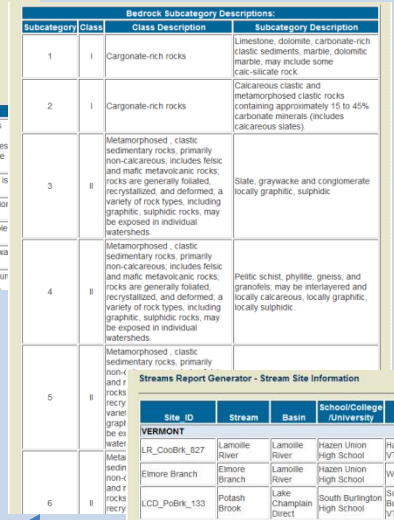
1. **The time range.** Do you want to investigate:
   - across multiple years?
   - different months in the same year?
   - the same month in different years?
2. **The spatial extent.** Do you want to investigate:
   - different watersheds?
   - different catchments?
   - streams with different surrounding land uses?
3. **The parameters of stream health.** Do you want to investigate:
   - nutrients, like phosphorus?
   - total suspended solids?
   - macroinvertebrates?
   - discharge?
   - temperature?
   - other parameters we measure?



Streams Report Generator - Data Variable Definitions

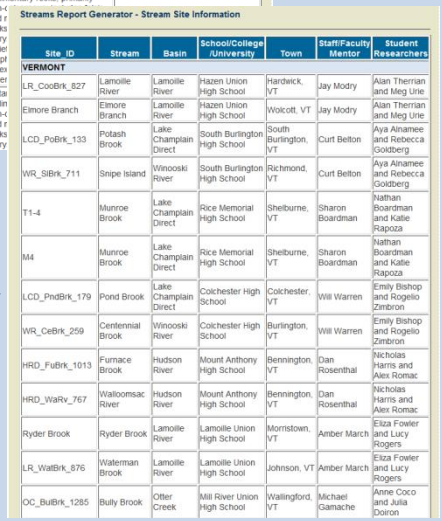
- Site Assessment
- Habitat Assessment
- Macroinvertebrate
- Water Quality Assessment
- E.Coli
- Total Suspended Solids
- Phosphorus
- Lab Data
- GIS Assessment Data

Field Name	Units	Descriptions
Name	Text	The name of the monitoring site. This name follows a standard naming convention beginning 2009.The names sites being sampled before 2009 were grandfathered in
Location	Town	The town in which the monitoring site is located.
Type	Text	The type of stream site given the optior muddy or riffle
Participating Organization	Text	The school or organization responsible monitoring this site
Date and Time of Recording	Year-month-day hr:min:sec	This is the date that this information wa collected in the field
		One selected from Forest, Field/Pastur Agricultural, Commercial, Residential,



Bedrock Subcategory Descriptions:

Subcategory	Class	Class Description	Subcategory Description
1	I	Cargonate-rich rocks	Limestone, dolomite, carbonate-rich clastic sediments, marble, dolomitic marble, may include some calc-silicate rock.
2	I	Cargonate-rich rocks	Calcareous clastic and metamorphosed clastic rocks containing approximately 15 to 45% carbonate minerals (includes calcareous slates).
3	II	Metamorphosed , clastic sedimentary rocks, primarily non-calcareous; includes felsic and mafic metavolcanic rocks; rocks are generally foliated, recrystallized, and deformed; a variety of rock types, including graphitic, sulphidic rocks, may be exposed in individual watersheds.	Slate, graywacke and conglomerate locally graphitic, sulphidic
4	II	Metamorphosed , clastic sedimentary rocks, primarily non-calcareous; includes felsic and mafic metavolcanic rocks; rocks are generally foliated, recrystallized, and deformed; a variety of rock types, including graphitic, sulphidic rocks, may be exposed in individual watersheds.	Pelitic schist, phyllite, gneiss, and granofels; may be interlayered and locally calcareous, locally graphitic, locally sulphidic.
5	II	Metamorphosed , clastic sedimentary rocks, primarily non-c and r rocks recry varie graph be ex wate	
6	II	Meta sedin non-c and r rocks recry	

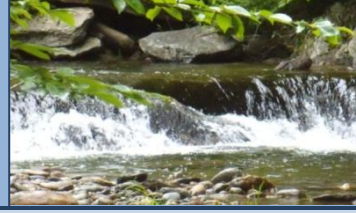


Streams Report Generator - Stream Site Information

Site_ID	Stream	Basin	School/College /University	Town	Staff/Faculty Mentor	Student Researchers
VERMONT						
LR_CooBrk_827	Lamoille River	Lamoille River	Hazen Union High School	Hardwick, VT	Jay Modry	Alan Therrian and Meg Urie
Elmore Branch	Elmore Branch	Lamoille River	Hazen Union High School	Wolcott, VT	Jay Modry	Alan Therrian and Meg Urie
LCD_PoBrk_133	Potash Brook	Lake Champlain Direct	South Burlington High School	South Burlington, VT	Curt Belton	Aya Ainamee and Rebecca Goldberg
WR_SlBrk_711	Snipe Island	Winooski River	South Burlington High School	Richmond, VT	Curt Belton	Aya Ainamee and Rebecca Goldberg
T1-4	Munroe Brook	Lake Champlain Direct	Rice Memorial High School	Shelburne, VT	Sharon Boardman	Nathan Boardman and Katie Rapoza
M4	Munroe Brook	Lake Champlain Direct	Rice Memorial High School	Shelburne, VT	Sharon Boardman	Nathan Boardman and Katie Rapoza
LCD_PndBrk_179	Pond Brook	Lake Champlain Direct	Colchester High School	Colchester, VT	Will Warren	Emily Bishop and Rogelio Zimbron
WR_CeBrk_259	Centennial Brook	Winooski River	Colchester High School	Burlington, VT	Will Warren	Emily Bishop and Rogelio Zimbron
HRD_FuBrk_1013	Furnace Brook	Hudson River	Mount Anthony High School	Bennington, VT	Dan Rosenthal	Nicholas Harris and Alex Romac
HRD_WaRv_767	Walloomsac River	Hudson River	Mount Anthony High School	Bennington, VT	Dan Rosenthal	Nicholas Harris and Alex Romac
Ryder Brook	Ryder Brook	Lamoille River	Lamoille Union High School	Morristown, VT	Amber March	Eliza Fowler and Lucy Rogers
LR_WatBrk_876	Waterman Brook	Lamoille River	Lamoille Union High School	Johnson, VT	Amber March	Eliza Fowler and Lucy Rogers
OC_BuiBrk_1285	Bully Brook	Otter Creek	Mill River Union High School	Wallingford, VT	Michael Gamache	Anne Coco and Julia Doiron

Tip: You may want to refer to the "Help" pages that define each variable. They can help you determine the scale and intervals on which data are collected.

Module 2



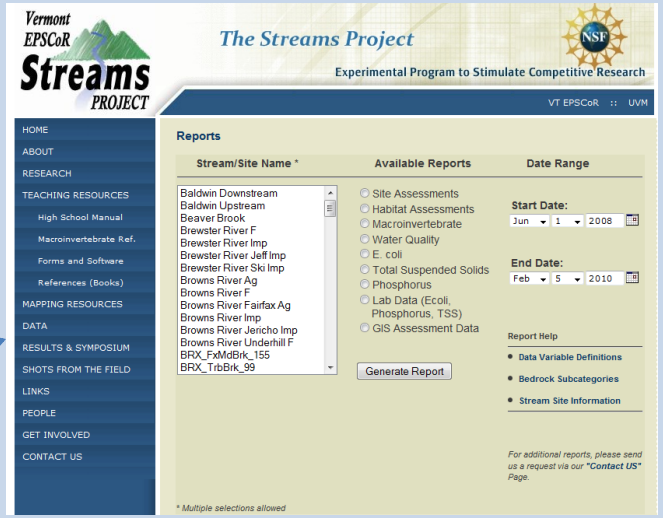
How to use available data to specify your question

At its essence, research addresses **a question** about the phenomena of interest. In the case of the Streams Project, the phenomena are several aspects of stream health. The scientific community refers to this question as a **hypothesis** which you, the scientist, **test** by collecting data and analyzing it with statistics.

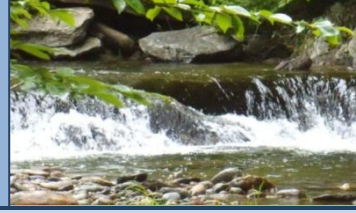
Use these specifications you made about time, space, and stream health on the previous slide to turn your idea into a question.

Take the previous example about our interest in the relationship between macroinvertebrate abundance and temperature. You can turn it into a question by saying, "Are macroinvertebrates more abundant in streams with higher temperatures?"

Check back with the database to make sure that the data you need to answer your question are available. Again, you may want to use the "Help" pages that define each variable.



Module 2



Why you need a hypothesis

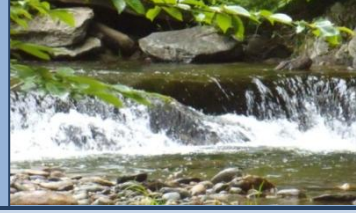
Throughout this module we will use the term **"hypothesis"** to refer to your question. In statistics, a hypothesis is really composed of two hypotheses: a "**null hypothesis (H_o)**" and an "**alternative hypothesis (Ha).**" The hypothesis is your question phrased in the format:

H_o = null hypothesis = *The hypothesis of **no** difference or **no** real relationship between or among parameters of interest.*

Ha = alternative hypothesis = *The hypothesis that there **is** a difference or real relationship between or among parameters of interest.*

Statistical tests allow you to determine if there are or are not differences or relationships in the parameters of interest. Such statistical tests are structured in a way that requires both a null and alternative hypothesis. Module 5 demonstrates this.

Module 2



Finalizing your hypothesis

From here, you should be able to take your questions and turn them into hypotheses!

Take the following examples:

Question 1: Are macroinvertebrates more abundant in streams with higher temperatures?

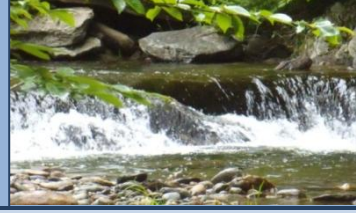
For this question we would write our hypothesis as the following:

H_o = *There is no difference between the number of macroinvertebrates in warm sites and cooler sites.*

Ha = *There is a difference between the number of macroinvertebrates in warm sites and cooler sites.*

Continued…..

Module 2



Finalizing your hypothesis

Take the following example:

Question 2: Are the levels of phosphorus and TSS related?
For this question we would write our hypothesis as the following:

H_o = *There is no relationship between the levels of phosphorus and TSS.*
Ha = *There is a relationship between the levels of phosphorus and TSS.*

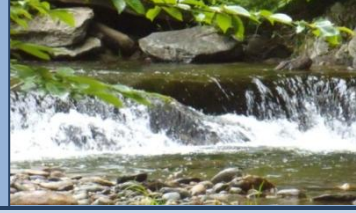
The above examples are very simple. You may choose to ask a similarly simple question, but don't be afraid to add layers of complexity:

Question 2: Are the levels of phosphorus and TSS related at my urban site, but not at my forested site?

For this question we would write our hypothesis as the following for each site:

H_{o_1} = *There is no relationship between the levels of phosphorus and TSS at my urban site.*
Ha_1 = *There is a relationship between the levels of phosphorus and TSS at my urban site.*
H_{o_2} = *There is no relationship between the levels of phosphorus and TSS at my forested site.*
Ha_2 = *There is a relationship between the levels of phosphorus and TSS at my forested site.*

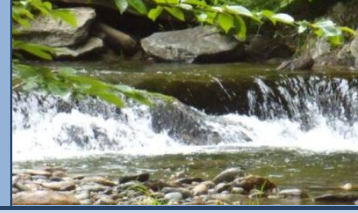
Module 2



Finalizing your hypothesis

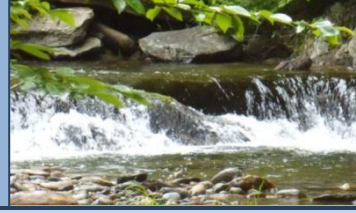
SUMMARY

- High school teams, undergraduate students, and faculty researchers all contribute data to the Streams Project database.
- All quality controlled data are available online.
- The “Help” pages define what each variable means and can offer valuable insight into how you can ask your research question.
- Reconcile your ideas with the data available in the streams database, using the “Help” pages.
- Define the time, spatial extent, and parameters of the ideas you have on your concept map.
- Turn your idea into a question and a null and alternative hypothesis.
- Hypotheses are necessary for conducting statistical tests.



Module 3: Refining and Retrieving Data

Module 3



Refining and Retrieving Data

The Streams Project database holds all of the data from your field forms, the lab analysis of phosphorus, E.coli, Coliform and total suspended solids (TSS), and GIS-analysis of spatial data for each monitoring site and associated catchment area.

For your independent research you most likely don't want or need all of this data. This module helps you to do the following:

1) **Review the data in the database**

2) **Decide what data you need**

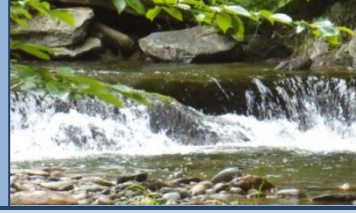
3) **Determine the time frame for which you'd like data**

4) **Determine the spatial extent for which you'd like data**

5) **Download the data you need from the Streams Project database**

6) **Refining your data**

Module 3



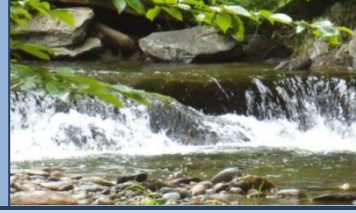
Refining and Retrieving Data

1) Review the data in the database

The database contains the following data discussed in module 2, available on the website in the form of “reports”:

- **Site Assessment Data**
- **Habitat Assessment Data**
- **Water Quality Assessment Data**
- **Macroinvertebrate Data**
- **E.coli and Coliform Lab Data**
- **TSS Lab Data**
- **Phosphorus Lab Data**
- **GIS Assessment Data**

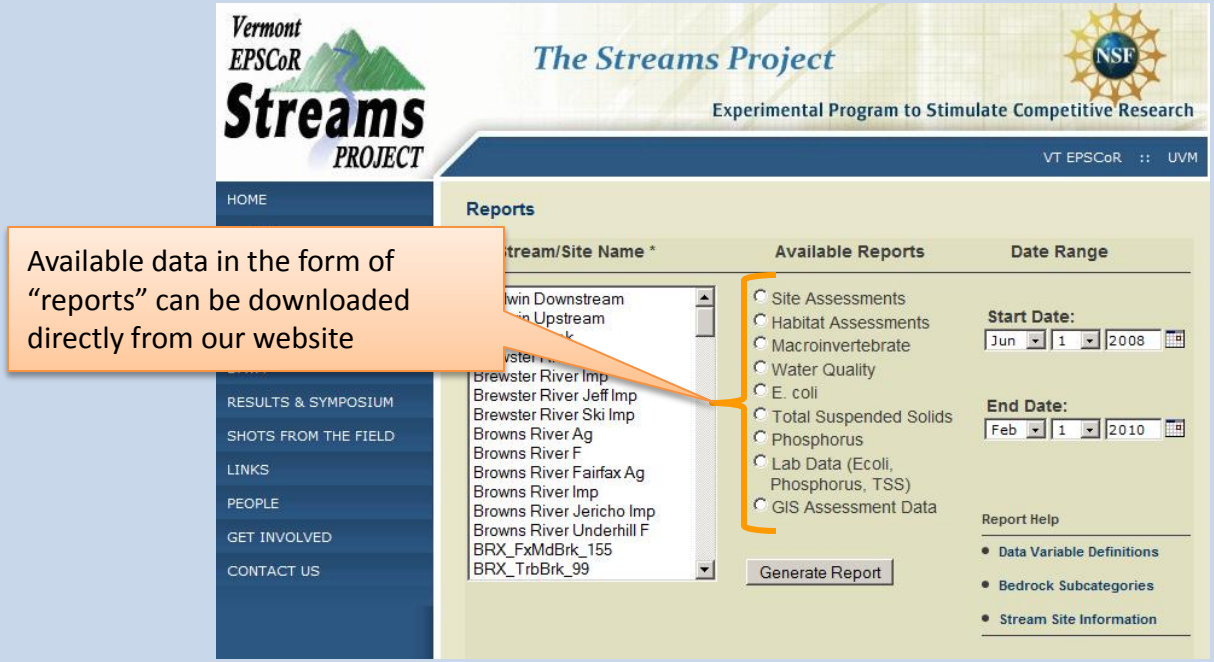
Module 3



Refining and Retrieving Data

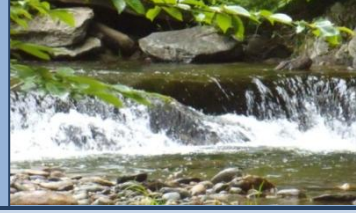
1) Review the data in the database

The following is the web interface through which you will download your data. All the data sources mentioned on the previous page are listed as “Available Reports”:



For more information about what is in each of these reports, download the “Data Variable Definitions” file in the bottom right hand corner of the data download webpage under “Report Help.” This file describes each field for each report including units of measurement.

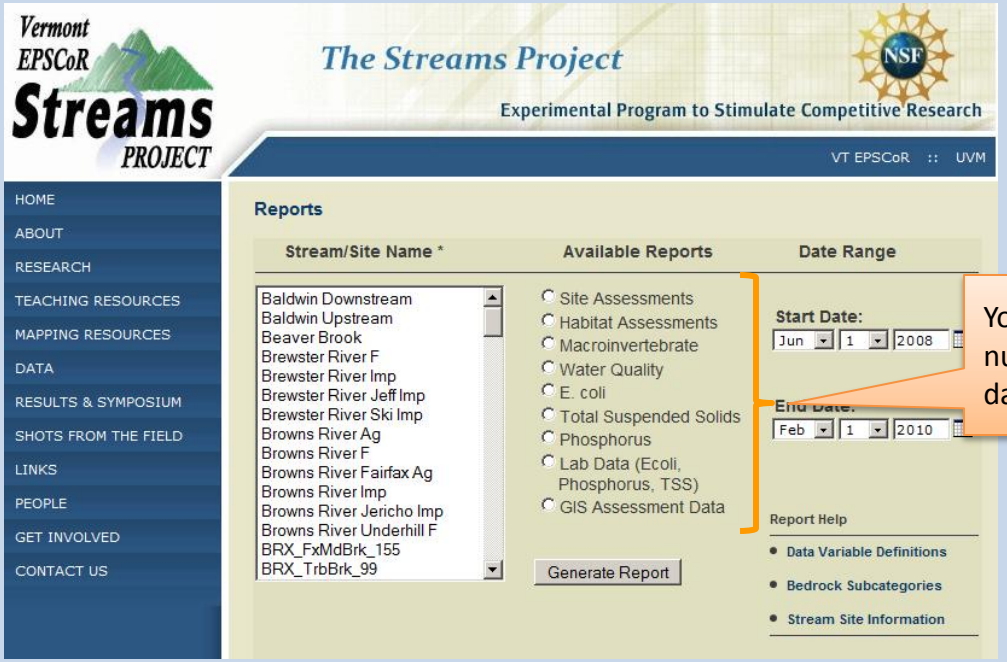
Module 3



Refining and Retrieving Data

2) Decide what data you need

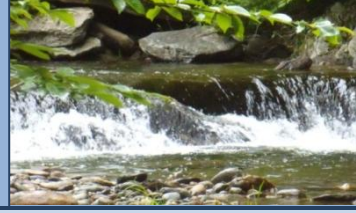
Most likely you will not need to download all of the available reports to do your data analysis. Think about your central question and in which report the data to answer this question is located.



You can download any number of reports to get the data you need

Again, For more information about what is in each of these reports, download the “Data Variable Definitions” file in the bottom right hand corner of the data download webpage under “Report Help.”

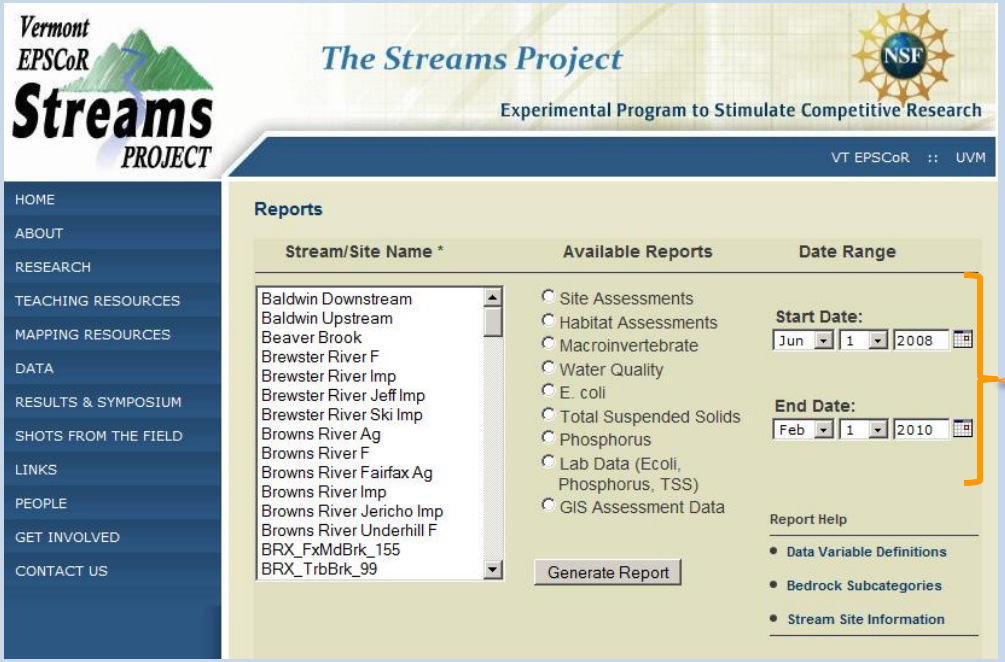
Module 3



Refining and Retrieving Data

3) Determine the time frame

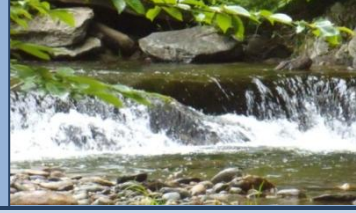
The database contains all data collected from June 2008 through the current date. As it is important to do throughout, review the question you are trying to answer – is a specific time interval best suited for answering this question?



You can specify the date for which you would like data here.

Most of this data was collected during June – December of each year. Think about if you want data from all years and all months or just a specific interval of time.

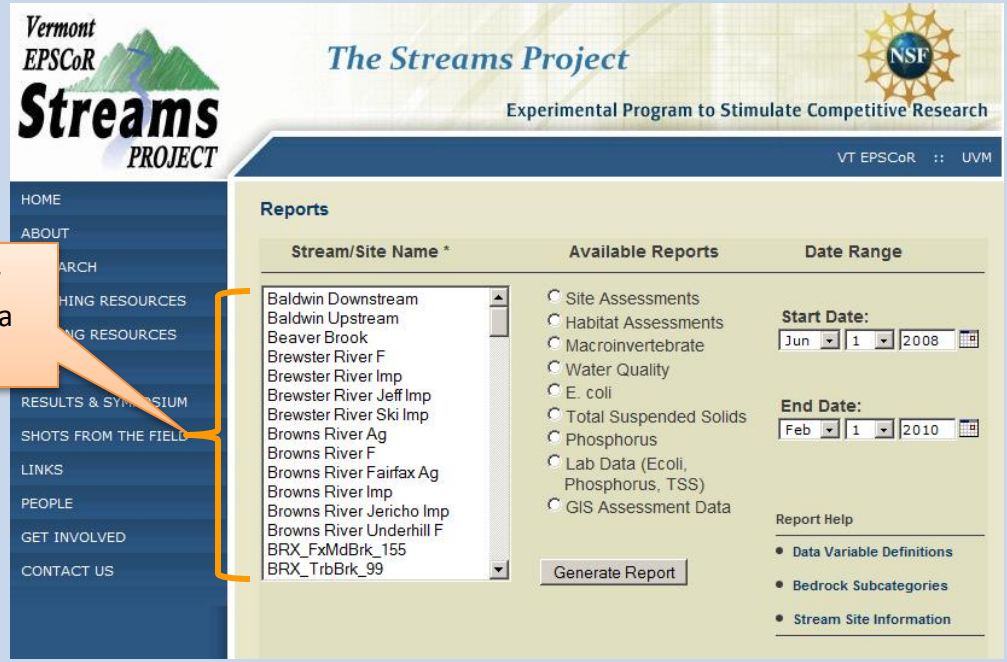
Module 3



Refining and Retrieving Data

4) Determine the spatial extent

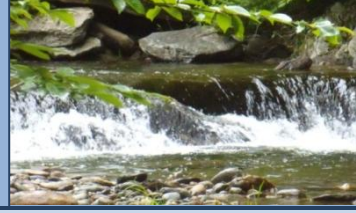
The database contains data from all streams sites being monitored in Vermont, New York, Connecticut and Puerto Rico. Use the Stream Sites Map to look up stream site names and help you narrow down the sites for which you would like data. Though you may decide you want to use all the data which is fine!!



Select which the sites for which you would like data here.

To select a stream site, highlight the name. To select all sites, highlight all the names. To select just a handful of sites, highlight the first and then hold down “Ctrl” while you highlight the remaining names.

Module 3



Refining and Retrieving Data

5) Download the data

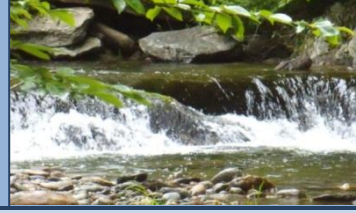
Now that you've reviewed the database and thought about the data you need, the time frame, and spatial extent for which you need this data, you are ready to download the data.

The following video will walk you through a quick example of how to download data using the data download website below:

WATCH DOWNLOAD VIDEO

Data download website: http://www.uvm.edu/~streams/?Content=pages/download_data8.inc

Module 3



Refining and Retrieving Data

6) Refining your data

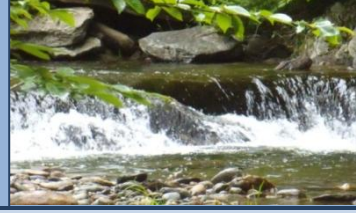
The data that you download from the website in many cases will not be ready for analysis as is. Please consider the following when preparing your data for analysis:

- You will note for E.coli, Coliform, TSS, and Phosphorus lab datasets, that all replicates numbers of data are included in the downloadable reports. You will have to refine these numbers, before using them in your analysis.

- You should be aware that there may be gaps because not everyone sampled on the same dates, so keep this in mind when you are compiling your dataset.

- Think about how your question might require that you structure or condense your data in a certain way. For example, if you want to regress land use and TSS, you might need to average your values of TSS for each site before running a regression analysis, as there will only by one land use value per site which is not dependent on the same time scale as TSS throughout your sample period.

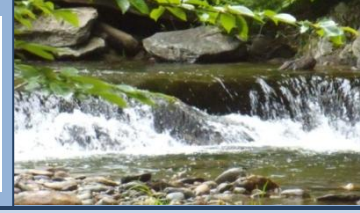
Module 3



Refining and Retrieving Data

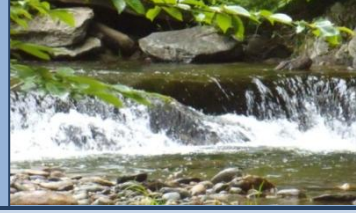
SUMMARY

- Data comes from field forms entered online, lab analysis on water quality samples gathered in the field, and GIS analysis done for the catchment area of each site.
- The following data is available as a "report" for download off the Streams Project website: Site Assessment Data, Habitat Assessment Data, Water Quality Assessment Data, Macroinvertebrate Data, E.coli Lab Data, TSS Lab Data, Phosphorus Lab Data, GIS Assessment Data
- Review your question to determine what data you need for your analysis
- Use the excel file Data_Field_Descriptions.xls to help you interpret the reports available for download.
- Consider the time interval for which you need data: all available data or a specific time period
- Consider the sites for which you'd like data: all available sites, or just a selection. Use the Stream Sites Map to help you determine which sites you'd like to focus on if not all.
- Download your data at http://www.uvm.edu/~streams/?Content=pages/download_data8.inc
- The data that you download is in a raw format, review guidelines in this module when refining your data for analysis



Module 4: Data Exploration

Module 4



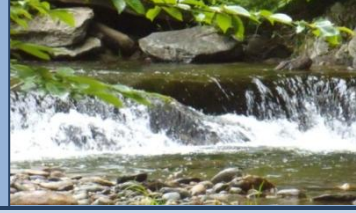
Data Exploration

Now that you have your data downloaded from the Streams Project database, the detective work can begin! Before computing any advanced statistics, we will first use descriptive statistics to examine the distribution of your data.

The following topics are covered in this module:

1) **Overview of descriptive statistics**
2) **Central tendency**
3) **Dispersion**
4) **Visualizing your data**

Module 4

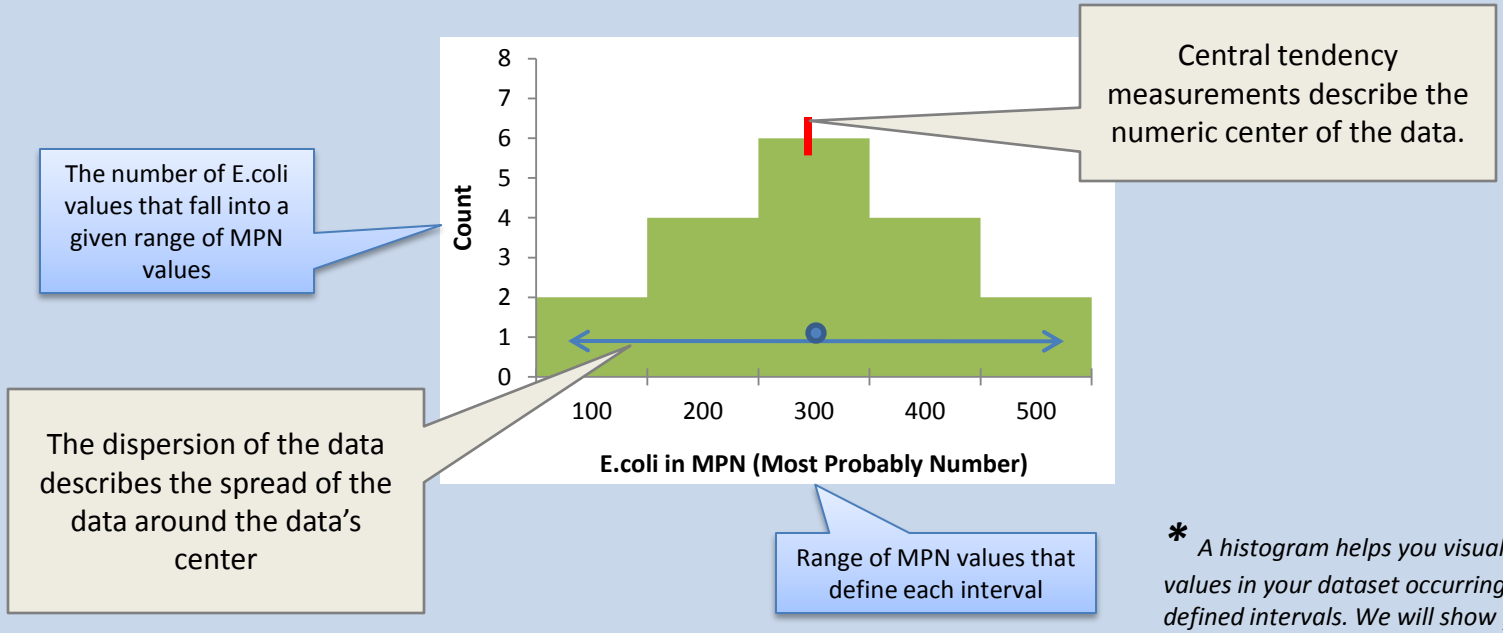


Data Exploration

1) Overview of Data Exploration

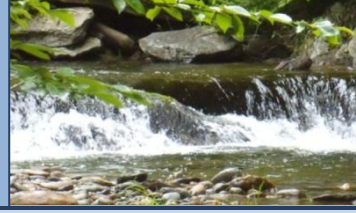
Calculating the descriptive statistics outlined in this module may be the extent of your analysis or the first step towards a more in-depth analysis as outlined in Module 5.

Descriptive statistics help you describe your data in terms of its distribution. To examine your data's distribution you will need a measure of **central tendency** and **dispersion**. These measures are illustrated in the frequency histogram below which shows the count of data for each interval of 100 MPN (Most Probably Number) of E.coli helping you visualize the frequency of data occurring over the range of values in the dataset:



* *A histogram helps you visualize the frequency of values in your dataset occurring over a series of defined intervals. We will show you how to create a histogram later in the module.*

Module 4

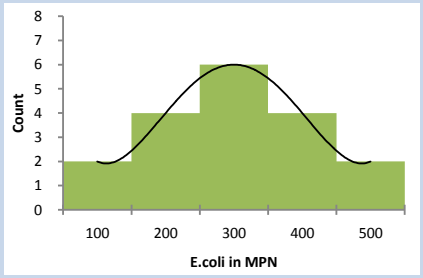


Data Exploration

1) Overview of Data Exploration

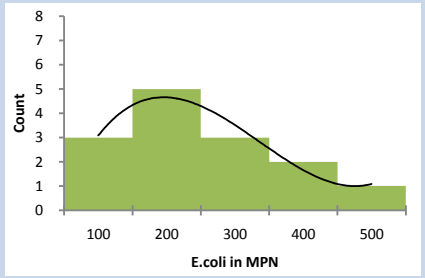
Before we calculate measure of central tendency and dispersion, let's look at what we mean by distribution. The ideal distribution of data is called the "normal distribution." For a normal distribution, all measures of central tendency (you will see there are several!) are the same, and there is an equal number of observed data points on either side of these measures of central tendency.

A histogram is a great way of initially visualizing your data's distribution because you can get a sense of the central tendency and dispersion of data around that center before calculating any statistics. The following histograms illustrate normal distributions as well as non-normal, or "skewed," distributions for comparison:



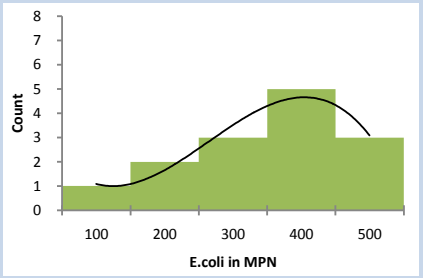
Data is equally dispersed on either side of the data's center (peak of the histogram)

Normal Distribution



There is more data on the lower end of the data range with values tapering to the higher end

Left-Skewed

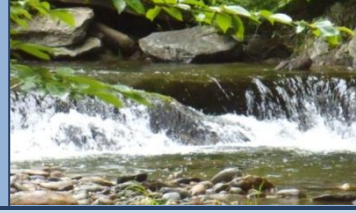


There is more data on the higher end of the data range with values tapering to the lower end

Right-Skewed

For advances statistical tests, it is important to determine if your distribution is normal or otherwise, as this will affect the type of statistical test you chose to use. The statistical tests in this tutorial assume your data is normally distributed.

Module 4



Data Exploration

2) Central tendency

Ambrose et al. (2002:22) describe "central tendency" as "what usually happens…"

If we measure E.coli at all our stream sites, what amount of E.coli do we usually measure?

Having a measurement that reflects what usually happens allows us to compare individual sample data points to the "usual" value of data observed represented by a measure of central tendency. The following are three measurements of central tendency:

Mean = ***the sum of observed data points divided by the number of data records***

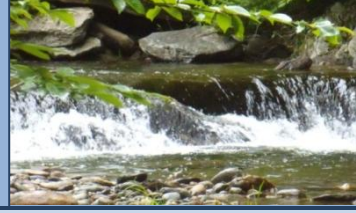
Median = ***the middle data point (or average of the two middle data points if there is an even number of observations) when all data points are lined up in either ascending or descending order***

Mode = ***The most frequently occurring data point value in a dataset***

On the next page you will see examples of how these three measurements would be calculated for a sample E.coli dataset.

Continued…

Module 4



Data Exploration

2) Central tendency

Data Point #	E.Coli (MPN)
1	410
2	263
3	310
4	476
5	388
6	417
7	345
8	402
9	379
10	379

Sample dataset of E.coli values with units MPN

Given our sample dataset, we would calculate the mean, median, and mode as follows:

Mean = (410+263+310+476+388+417+345+402+379+381)/10 =

377

Watch an additional example in Excel

Median = 263, 310, 345, 379, **381, 388,** 402, 410, 417, 476 > (381+388)/2 =

385

Watch and additional example in Excel

Mode = 379 which occurs twice while other values occur only once

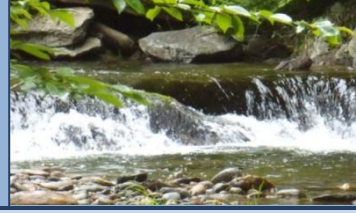
379

Watch an additional example in Excel

These values all suggest that the center of our data, or the most frequent values of E.coli measured in our streams, falls around 370 – 390 MPN.

Continued…

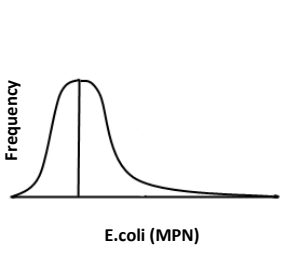
Module 4



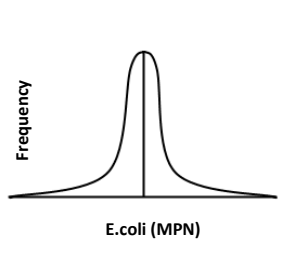
Data Exploration

3) Dispersion

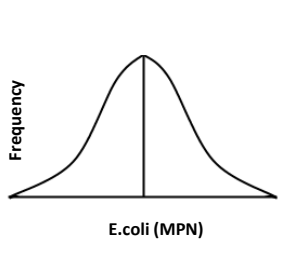
We have a measure of central tendency for our sample dataset (let's use the mean), but how do we describe how the rest of the data falls around our mean? If the following bell curves (think of them as smoothed out histograms) represent data with the centerline as your measure of central tendency, how do we describe the different ways that the data falls around their centerline?



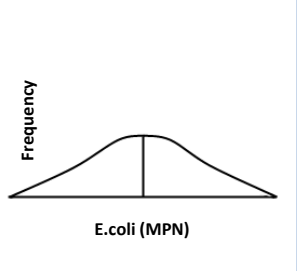
Is the data distributed unevenly about the mean, with a few values far from the mean in one direction?



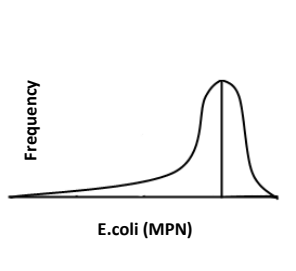
Is the data distributed evenly on both side of the mean but concentrated around on mean?



Is the data distributed evenly on both side of the mean with most data within a standard distance?



Is the data distributed evenly on both sides of the mean dispersed over a greater range without much concentration by the mean?

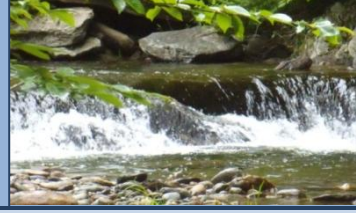


Is the data distributed unevenly about the mean, with a few values far from the mean in one direction?

Now that you have a visual on what we mean by "dispersion," the following page helps you calculate statistics used to quantify the nature of the dispersion around the mean.

Continued…

Module 4



Data Exploration

3) Dispersion

The following are measurements of dispersion used to quantify the spread of data about the data's center:

Range = ***The highest measurement – the lowest measurement in the dataset***

Variance = ***A cumulative measure of individual data points' distance from the mean. The following equation is used to calculate variance:***

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

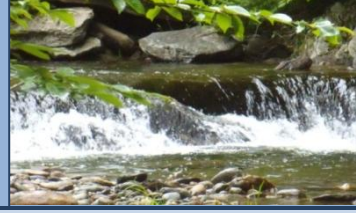
$\sum$ *means you add everything that follows together. Remember to pay attention to parentheses – they're important!*

Where:

s^2 *= variance*
x *= individual values in of the dataset*
n *= the number of data points in the dataset*

Continued…

Module 4



Data Exploration

3) Dispersion

Standard Deviation = ***Somewhat of an average deviation of the data from the mean. Standard deviation is calculated as the square root of the variance:***

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

$\sum$ *means you add everything that follows together. Remember to pay attention to parentheses – they're important!*

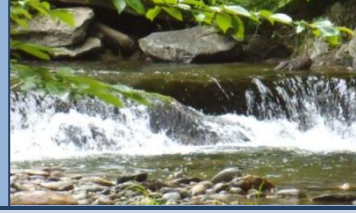
Where:

s = *standard deviation*
x = *individual values in of the dataset*
n = *the number of data points in the dataset*

On the next page you will see examples of how these three measurements would be calculated for our sample E.coli dataset.

Continued…

Module 4



Data Exploration

3) Dispersion

Data Point #	E.Coli (MPN)
1	410
2	263
3	310
4	476
5	388
6	417
7	345
8	402
9	379
10	379

Sample dataset of E.coli values with units MPN

Given our sample dataset, we would calculate the following values for range, variance, and standard deviation:

Range = 476 (highest value) – 263 (lowest value) **= 213**

Watch an additional example in Excel

Variance =

$$s^2 = \frac{\sum 410^2 + 263^2 + 310^2 \ldots - \frac{(\sum 410 + 263 + 310 \ldots)^2}{10}}{10 - 1}$$ **= 3528.1**

Watch an additional example in Excel

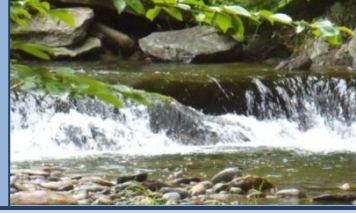
Standard deviation =

$$s = \sqrt{\frac{\sum 410^2 + 263^2 + 310^2 \ldots - \frac{(\sum 410 + 263 + 310 \ldots)^2}{10}}{10 - 1}}$$ **= 59.4**

Watch and additional example in Excel

The standard deviation, variance and mean are all metrics commonly used in more advanced statistical tests, some of which will be described in Module 5.

Module 4



Data Exploration

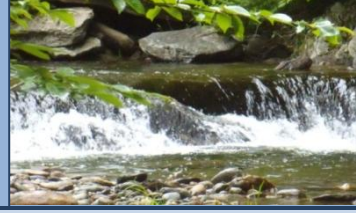
4) Visualizing your data

In the first part of this module you learned how to calculate two types of parameters to help you describe the distribution of your data:

- **Measure of central tendency**
- **Dispersion**

When you are presenting these parameters, it is helpful to provide a visual of your data's distribution in addition the numbers. The remainder of this module will help you create a histogram and/or a boxplot depicting your data's distribution. The module will conclude by helping you describe your combined visual and statistical results.

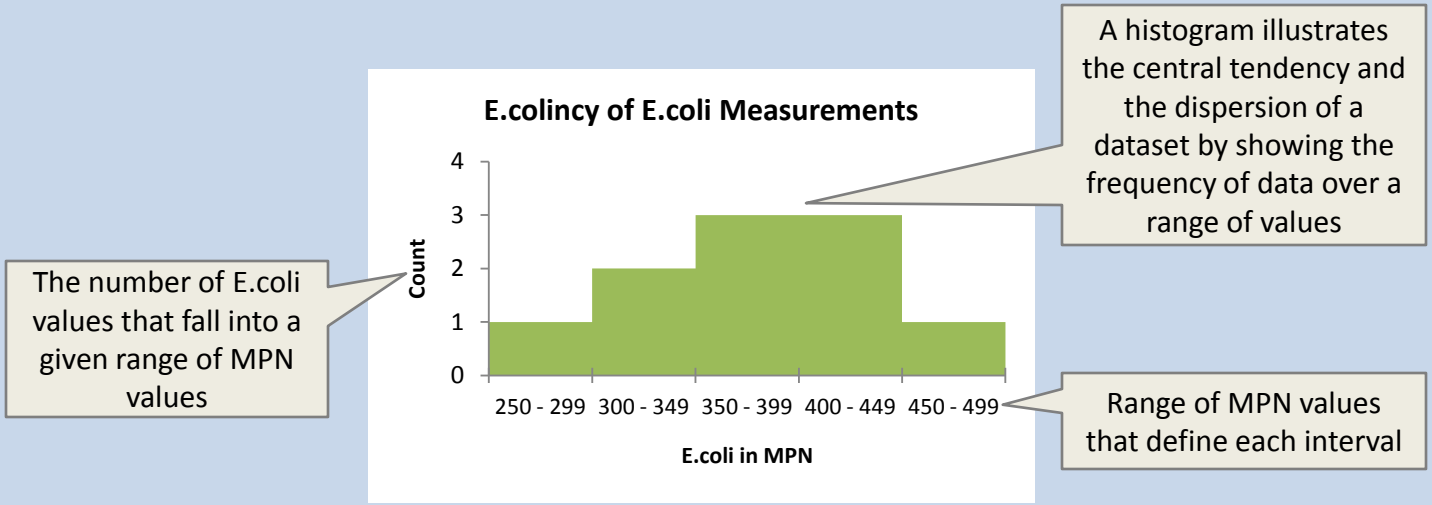
Module 4



Data Exploration

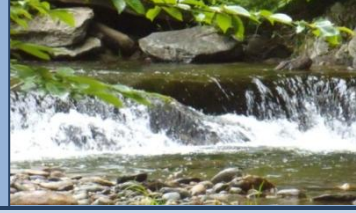
4) Visualizing your data

First, we will revisit our **frequency histogram**, which shows the frequency of values in your dataset over a series of intervals that covers the range of your dataset. See the link below to see how to create a histogram for your data in excel.



Click on the video icon to watch a video on how to create a histogram using Microsoft Excel

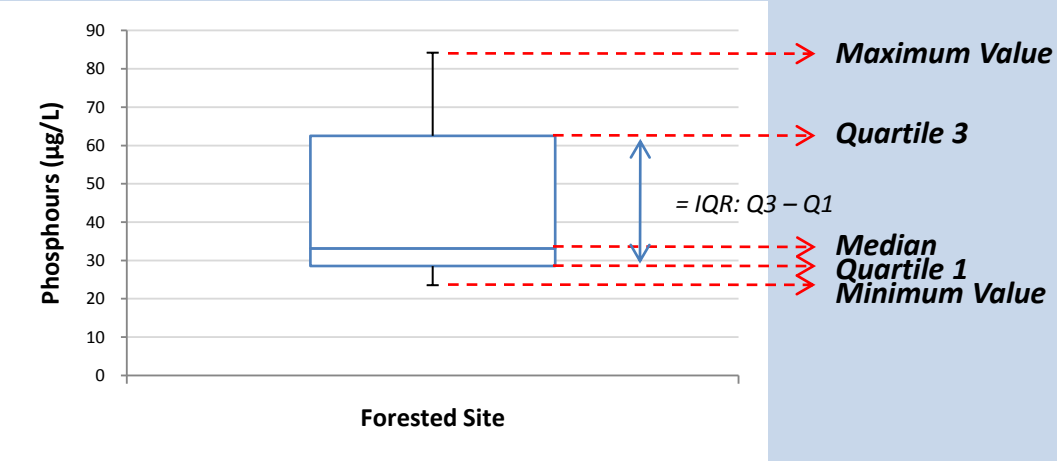
Module 4



Data Exploration

4) Visualizing your data

A box plot also illustrates the distribution of your data. A box plots is made up of the following values derived from your dataset: median, minimum value, maximum value, quartile 1 value, and quartile 2 values. The following graph illustrates these components with descriptions following:

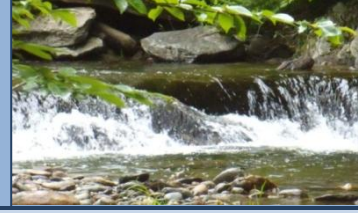


Definitions:

Maximum Value = the largest value in the dataset
Minimum Value = the smallest value in the dataset
Median = the middle value of the dataset when all values are lined up either ascending or descending
Quartile 1 = the median value of all values less than and excluding the median of the entire dataset
Quartile 3 = the median value of all values greater than and excluding the median of the entire dataset
Inter-Quartile Range (IQR) = Quartile 3 – Quartile 1

Click on the video icon to watch a video on how to create a box-plot using Microsoft Excel

Module 4

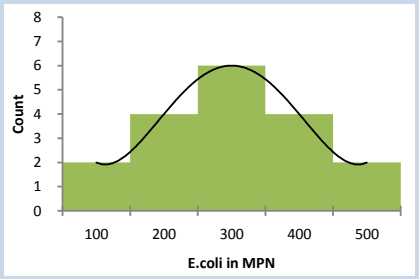


Data Exploration

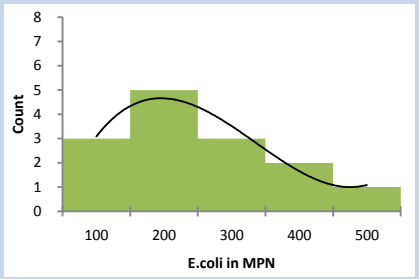
4) Visualizing your Data

Remember to discuss whether or not your data is normally distributed, or perhaps is skewed to the left or right. Your graphs, in addition to your descriptive statistics, help you communicate your findings, so be sure to include both!

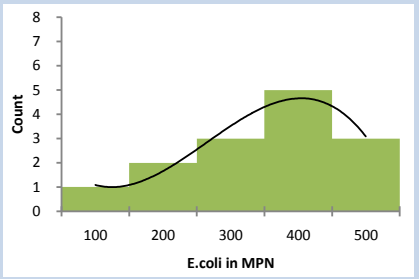
Histogram



Data is equally dispersed on either side of the data's center (peak of the histogram)

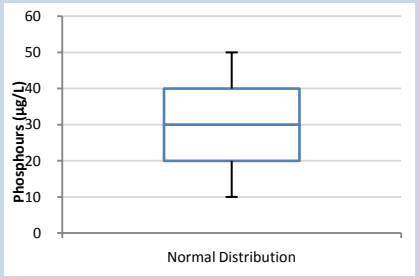


There is more data on the lower end of the data range with values tapering to the higher end

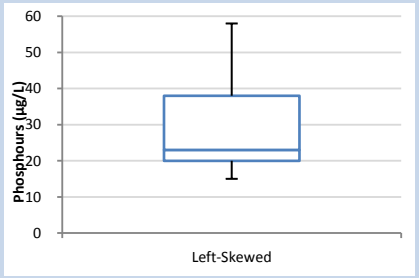


There is more data on the higher end of the data range with values tapering to the lower end

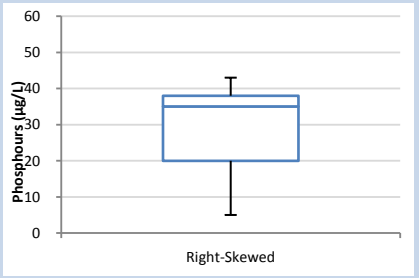
Box Plot



Data is equally dispersed on either side of the median (center line of the box)



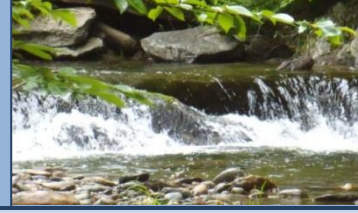
The median is visibly closer two one the lower end of values



The median is visibly closer two one the higher end of values

Normal Distribution | **Left-Skewed** | **Right-Skewed**

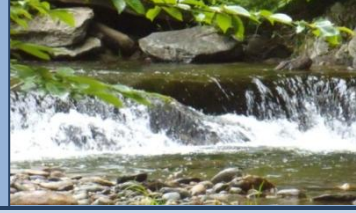
Module 4



Data Exploration

It is often the case that after exploring your data with descriptive statistics you want to modify or refine your central question – that's fine!

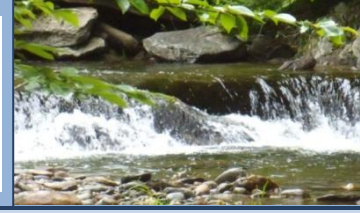
Module 4



Data Exploration

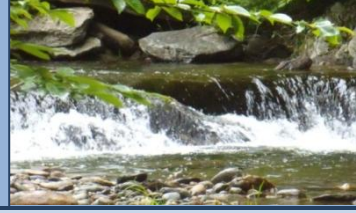
SUMMARY

- Descriptive statistics help describe your data's distribution
- A measure of central tendency and dispersion are needed to describe your data's distribution statistically
- Ideally your data fits the descriptions of a normal distribution with data distributed evenly on either side of the measure of central tendency.
- The following are measures of central tendency: mean, median and mode
- The following are measure of dispersion: range, variance, and standard deviation
- Histograms and box plots can help you illustrate your data's distribution
- Your descriptive statistics, histograms and/or box plots together help you describe the nature of your data
- After exploring your data using descriptive statistics it's good to reflect on your question and modify or refine it as needed.



Module 5: Statistical Analysis

Module 5



Statistical Analysis

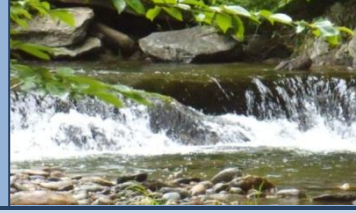
To answer more complex questions using your data, or in statistical terms, to test your hypothesis, you need to use more advanced statistical tests.

This module reviews the formulation of a central question, or hypothesis, and then describes three major categories of statistical tests:

1) **Questions and Hypotheses**
2) **Differences**
3) **Correlations**
4) **Regressions**

For each category, examples of the types of questions/hypothesis the test might help answer are given, along with directions on how to compute these statistical tests and create graphs and figures to illustrate your results.

Module 5



Statistical Analysis

1. Questions and Hypothesis

Central to any scientific research is a question that the research is trying to address. Scientific literature transforms this question into the form of a statement called a hypothesis which will be tested by your research.

Throughout this module we will use the term **"hypothesis"** to refer to your question that has been rephrased to make a statement. In statistics, a hypothesis is really composed of two hypotheses: a "**null hypothesis (H_o)**" and an "**alternative hypothesis (Ha).**" Take the following question as an example:

Are the levels of phosphorus recorded for my forested and urban sites different?

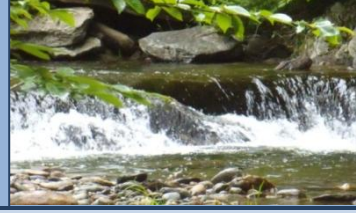
For this question we would write our hypothesis as the following:

H_o = *There is no difference between the levels of phosphorus at my forested site compared to my urban site.*

Ha = *There is a difference in the level of phosphorus at my forested site compared to my urban site.*

Continued…

Module 5



Statistical Analysis

1. Questions and Hypothesis

H_o = *There is no difference between the levels of phosphorus at my forested site compared to my urban site.*

Ha = *There is a difference in the level of phosphorus at my forested site compared to my urban site.*

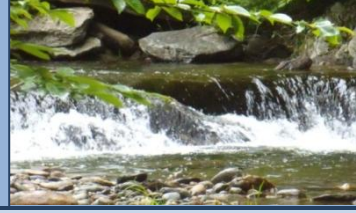
To test your hypothesis you will chose an appropriate statistical test which this module will walk you through. The results of this test will either be ***significant*** enough so that you will "reject your null hypothesis in support of your alternative hypothesis" or ***insignificant*** such that you "cannot reject your null hypothesis in favor of your alternative hypothesis."

Translated in terms of our example question that means:

Insignificant test result = *Your data does not provide enough evidence to show that there might be a difference between the two sites.*

Significant test result = *The results support the idea that there is a difference in the level of phosphorus between your two sites.*

Module 5



Statistical Analysis

2. Differences

Testing for differences allows us to statistically determine if the distributions, means or variances of multiple datasets are different.

Our example question about phosphorus is a question of differences:

Are the levels of phosphorus recorded for my forested and urban sites different?

And our hypotheses were as follows:

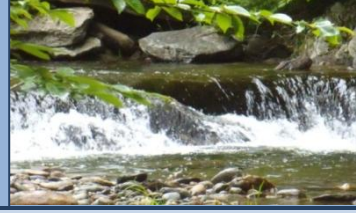
H_o = *There is no difference between the levels of phosphorus at my forested site compared to my urban site.*

Ha = *There is a difference in the level of phosphorus at my forested site compared to my urban site.*

The following statistical test can be used to test your hypothesis:

- **Two-sample t-test**

Module 5



Statistical Analysis

2. Differences

Two-sample t-test

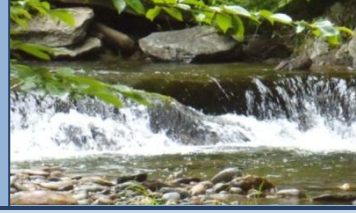
- *What it tests:*
  - The two-sample t-test is a statistical test that allows you to determine if the mean of two datasets are statistically different. It does this by using the mean and variance in a complex equation to produce a test statistic, known as "t." The value for this test statistic is compared to critical value of "t" which shows how likely the relationship between your two datasets is to occur under normal circumstances.

Click on the video icon to watch a video on how to use the t-test to calculate a P-value using Microsoft Excel

- *Interpreting the Output:*
  - The output that you will get from running a t-test in excel is the probability ("p-value") of getting the t-statistic calculated for your datasets. As a general rule, if your p-value is less than the critical value of .05 it means your results are significant and therefore support your alternative hypothesis which states that there is a difference in the distributions of your two datasets. The significance of the critical value of .05 is not explained in this tutorial, but we encourage you to explore further outside of what is offered here.

Continued…

Module 5



Statistical Analysis

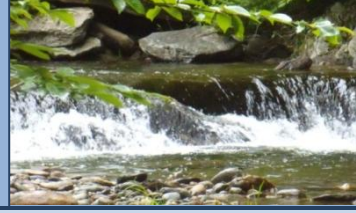
2. Differences

Two-sample t-test

- *Talking About Results*:
  - If you get a significant test statistic ($p<.05$), let's say for our question about the difference in phosphorus levels at your forested and urban sites, the results of your analysis support your alternative hypothesis that there is a difference in the phosphorus levels measured at these two sites.
  - If you get a significant test statistic that is >.05, you cannot reject your null hypothesis that there is no difference in the phosphorus levels measured at these two sites.
  - Your analysis can only say that there is or is not a statistically significant difference; this statistic *does not* explain what is causing the difference between the two datasets.
  - If you establish that there is a difference, you might look at other variables in your datasets such as land use or geology to help you speculate about what might potentially be causing these differences. Be sure to mention these ideas when you are describing your results!

Continued…

Module 5

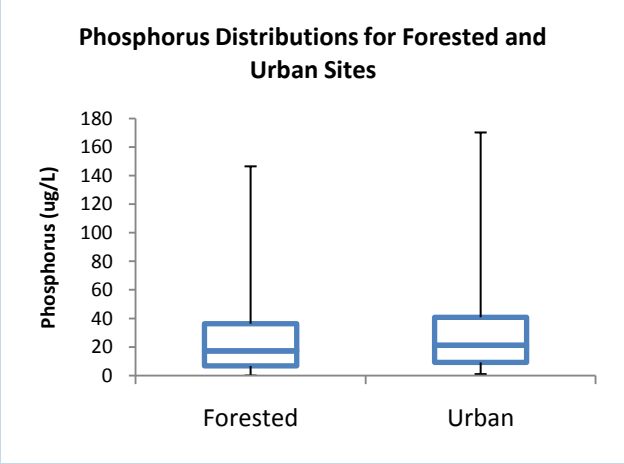


Statistical Analysis

2. Differences

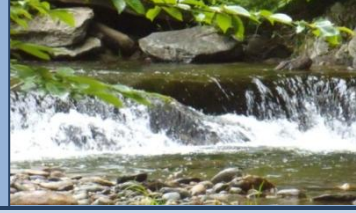
Two-sample t-test

- *Visualizing Results*:
  - A side-by-side box plot can be used to illustrate the results of your two sample t-test. First review how to create a single box plot in Module 4.
  - While a side-by-side box plot is used to compare the distribution of two datasets, it can also help you visually compare the central tendencies of multiple datasets as the middle line of the box represents the median value of the dataset which should be about equal to your mean.



Click on the video icon to watch a video on how to create a box-plot

Module 5



Statistical Analysis

3. Correlations

Testing for correlations allows us to statistically determine if there is a relationship between two variables in a dataset, and if so, the nature of the relationship (positive = they increase together or negative = one decreases while the other increases).

The following is an example of a question of correlation:

Is there a relationship between the level of E.coli in the water and water temperature?

And our hypotheses would be as follows:

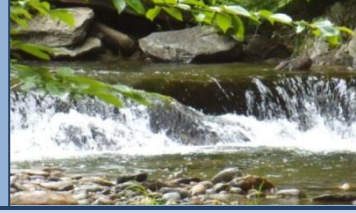
H_o = *There **IS NO** relationship between E.coli and water temperature measured at a stream site.*
Ha = *There **IS** a relationship between E.coli and water temperature at a stream site.*

To test for correlation, the following statistical test would be used:

- **Spearman's Rank Correlation**

Continued…

Module 5



Statistical Analysis

3. Correlations

Spearman's Rank Correlation

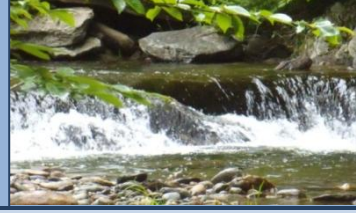
- *What it tests:*
  - Spearman's Rank Correlation is a statistical test that allows you to determine if there is a relationship between two variables in a dataset. It does this by using the mean in a complex equation to produce a correlation coefficient referred to as "R."

Click on the video icon to watch a video on how to calculate a correlation coefficient using Microsoft Excel

- *Interpreting the Output:*
  - The output that you will get from doing a correlation in excel is the correlation coefficient "R." The closer your correlation coefficient is to 1 or -1 the stronger the relationship between your two variables. If your correlation coefficient is negative than your two variables are inversely related (one increases as the other decreases). If your correlation coefficient is positive, then your two variables are positively correlated (they both increase together).

Continued…

Module 5



Statistical Analysis

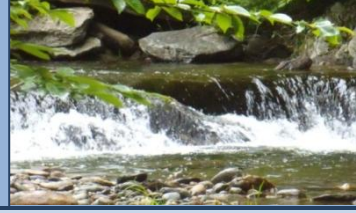
3. Correlations

Spearman's Rank Correlation

- *Talking About Results*:
  - If your correlation coefficient (R) is close to 1 or -1, let's say for our question about E.coli being related to water temperature, the results of your analysis support your hypothesis that there is a relationship between your two variables (E.coli and water temperature).
  - There is no critical threshold that says your correlation coefficient either is or isn't significant; we talk about the results as showing the strength of the relationship on this scale from 0 to 1 and 0 to -1.
  - Be careful: the correlation coefficient *does not* prove with 100% certainty that these two variables are related, and *does not* show cause in effect, though if you suspect that the value of one variables might be dependent on the value changes in the other you should read on about regression analysis!

Continued…

Module 5

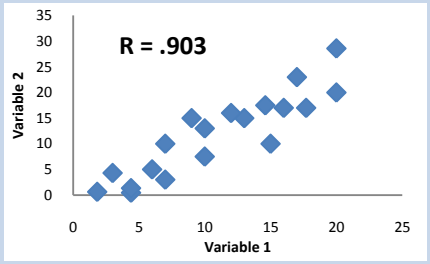


Statistical Analysis

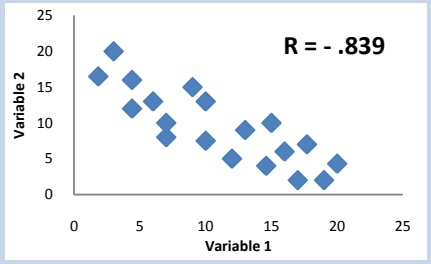
3. Correlations

Spearman's Rank Correlation

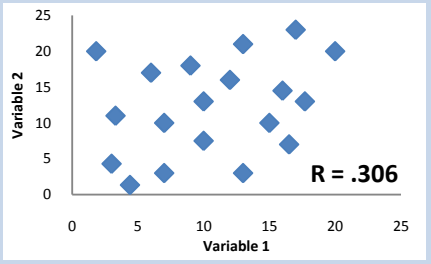
- Visualizing Results:
  - Correlations are best illustrated using a scatter plot
  - You might also use a scatter plot earlier on in your analysis when you are beginning to ask questions of correlations which you then might test using Spearman's Rank Correlation.
  - Scatter plots are made by plotting one variable against the other variable – the following three scatter plots illustrate the types of relationships you might see between two variables who may or may not be correlated:



Significant, positive correlation



Significant, negative correlation

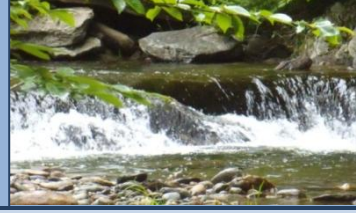


No significant correlation

  - Include your correlation coefficient (R) on the graph.

Click on the video icon to watch a video on how to create a scatter plot using Microsoft Excel

Module 5



Statistical Analysis

4. Regressions

A regression analysis is very similar to a test of correlation. The difference is that with a regression analysis we are looking to see if the values of one variable in our dataset, identified as the dependent variable (Y), increase or decrease as the values of another variable, identified as the independent variable (X), increase or decrease. If a change in ***X*** *does* cause a change ***Y***, the variables would be said to have a linear ***dependent*** relationship.

The following is an example of a question that can be answered through a regression analysis:

Does an increase in agricultural land use cause an increase in the amount of TSS in the water?

And our hypotheses would be as follows:

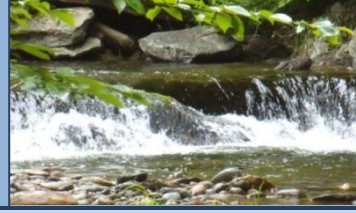
H_o = *The amount of TSS measured **DOES NOT DEPEND** on the amount of upstream agricultural land use.*
Ha = *The amount of TSS measured **DEPENDS** on the amount of upstream agricultural land use*

To test for a linear, dependent relationship the following statistical test would be used:

- **Regression Analysis: simple linear regression**

Continued…

Module 5



Statistical Analysis

4. Regressions

Regression Analysis: Simple Linear Regression

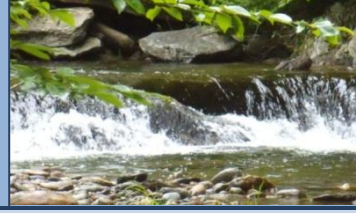
- *What it tests:*
  - A regression analysis is a statistical test that allows you to determine if there is a dependent relationship between two variables in a dataset. First you have to designate one variable as the dependent variable (Y), and the other as the independent (X). To do this, use common sense – would the amount of agricultural land use depend on the amount of TSS in the water? Or is it more likely that the amount of TSS depends on the amount of agricultural land use? Your variables are then organized into X-Y pairs. For example, at site B there is X-amount of agricultural land upstream, and the TSS reading at this site was Y, (etc. for all sites). The relationship between these two variables is represented by the linear equation $Y = aX + b$, and the strength of the relationship measured by the coefficient of determination "R^2."

Click on the video icon to watch a video on how to to calculate R^2 using Microsoft Excel

- *Interpreting the Output:*
  - The output that you will get from doing a regression analysis in excel is the coefficient of determination "R^2." The closer your R^2 value is to 1 the greater the dependent relationship between your two variables. If you read about correlations, R^2 is your R-value squared!

Continued…

Module 5



Statistical Analysis

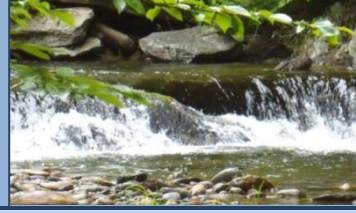
4. Regressions

Regression Analysis: Simple Linear Regression

- Interpreting Results:
  - The closer your R^2 value is to 1, the stronger the linear, dependent relationship between your two variables. This your Y variable being dependent on your X variable.
  - Looking at our question about agricultural land use and TSS, the closer to 1 your R^2 is, the more support there is for our alternative hypothesis that the amount of TSS measured depends on the amount of agricultural land use upstream.
  - Just as with a correlation, this relationship can be either positive or negative depending on the slope (b) of your linear equation: a negative sign means your Y variable decreases in response to an increase in your X variable, and a positive sign means your Y variable increases in response to an increase in your X variable.
  - Be careful: this analysis *does not* show cause in effect, but it does show dependence of one variable on another, and the nature of that dependence (positive of negative).

Continued…

Module 5

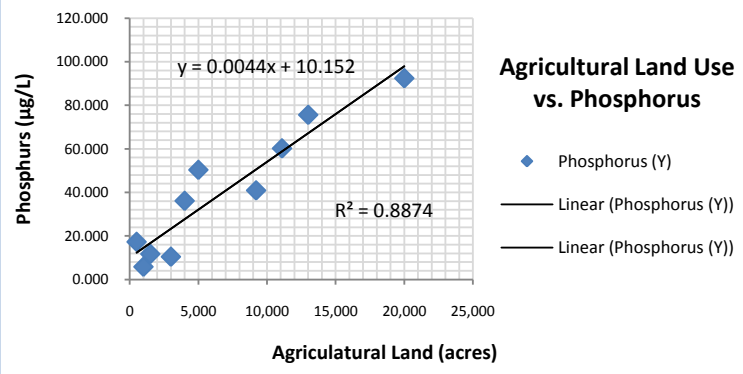


Statistical Analysis

4. Regressions

Regression Analysis: Simple Linear Regression

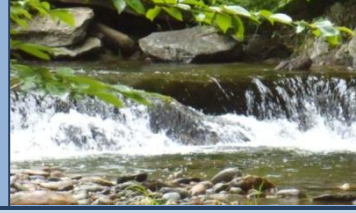
- Visualizing Results:
  - A scatter plot with a "best fit" line is used to illustrate the results of your regression analysis
  - These graphs are made by plotting the independent variable on the X-axis and the dependent variable on the Y-axis. The "best fit" line represents your linear equation y = aX + b.



  - Your equation gives you a line that represents a type of average describing the relationship between your two variables.
  - You should also add your R^2 value to the graph as well.

Click on the video icon to watch a video on how to create a graph of your regression analysis results using Microsoft Excel

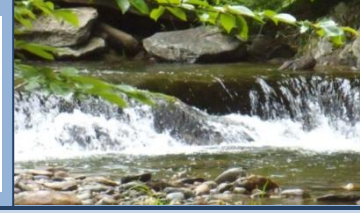
Module 5



Statistical Analysis

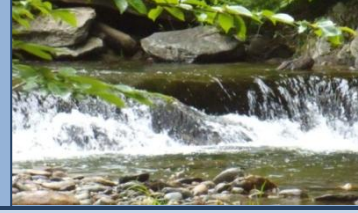
SUMMARY

- The questions you are trying to answer should be phrased as a hypothesis
- If your hypothesis asks if two datasets are different, then you should use a Two-sample t-test to determine if your two datasets are statistically different
- If your hypothesis asks if two variables in a dataset are correlated, then you should use Spearman's Rank Correlation to determine the strength of the relationship between these two variables.
- If your hypothesis asks is one variable is dependent on another variable, then you should run a linear regression analysis to determine if your dependent variable (X) is dependent on your independent variable (Y).



Module 6: Summarizing Results and Drawing Conclusions

Module 6



Summarizing Results and Drawing Conclusions

You have your results! Now what?

The following is the 6th Module in a series of tutorials that have been created to help your streams project participants understand and analyze their data.

1) **What is the difference between results and discussion?**

2) **What has been learned (Discussion)?**

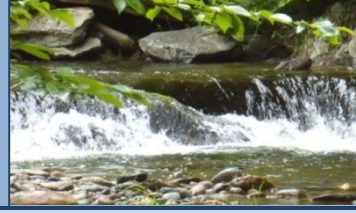
3) **What do your results mean (Discussion, continued)?**

4) **Can your results be applied to the world? If so, how?**

5) **What new questions do you have?**

6) **How can you effectively communicate your research findings to others?**

Module 6



Summarizing Results and Drawing Conclusions

1. Questions and Hypothesis

What is the difference between results and discussion?

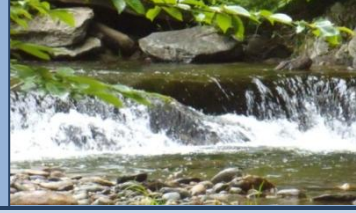
Results

Results are simply your findings. A results section of a scientific paper or talk is strictly for narrating your findings, without trying to interpret for evaluate them. This is often done using graphs, figures, and tables. If you found a notable correlation between two variables (phosphorus and land use, for example), this should be included in your results. Speculating why this correlation exists, however, belongs in the discussion section.

Discussion

You may have seen a “discussion” section in a scientific research paper. Discussion means interpreting your results and trying to explain what they mean.

Module 6



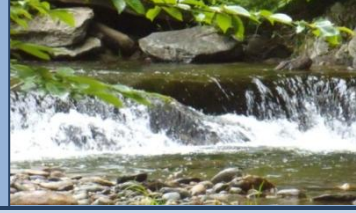
Summarizing Results and Drawing Conclusions

2. Discussion: What has been learned?

You have your results! Now what? Follow the steps below and try to answer the questions asked as they apply to your results.

- In the results section of your paper or talk, summarize your results, both in written form and visually, using graphs and charts.
- Ask yourself, what has been learned from this experiment?
- Do your results support or disprove your hypothesis? If your results do not support your hypothesis, why do you think this is the case?
- Every study has limitations. These limitations are very important to acknowledge. What are the limitations of your study?
    - Were there any weaknesses or errors in your study design or data that may have influenced your results?
    - Are you able to prove causation (that one thing is causing another), or association (that one thing is related to another), or differences (that one dataset is different from another)? This is determined by the types of data analysis you used (refer to Module 5 for details).

Module 6



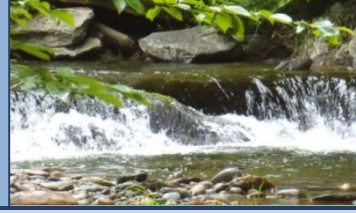
Summarizing Results and Drawing Conclusions

3. Discussion, continued: What do your results mean?

Follow the steps below and try to answer the questions asked as they apply to your results.

- Are your results consistent with past studies? If not, why do you think this might be the case?
- Has any new valuable information been learned?
- Even if the results were not as you predicted/hypothesized, this can be a valuable finding. Disproving your hypothesis can be just as significant as supporting it! This may lead to you revising your hypothesis and future research studies.
- Your conclusions must be justified by your results.

Module 6



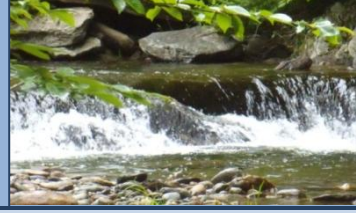
Summarizing Results and Drawing Conclusions

4. Can your results be applied to the world? If so, how?

Follow the steps below and try to answer the questions asked as they apply to your results.

- What field do your findings pertain to (biology, hydrology, geology, land use planning, etc.)?
- How do your findings contribute to this field? All research studies add to the overall understanding and body of knowledge of a given topic or field.
- Who might be interested in your findings? Other students? Researchers? Professionals in the field?
- Who might be able to apply your findings? Examples include a Watershed Planner or River Restoration Scientist who work for the Vermont Agency of Natural Resources, a Land Use Planner who works for the Central Vermont Regional Planning Commission, or a community volunteer who works with a local watershed group.
- Do not overstate the importance of your findings. Remember that all studies have limitations. Your results and conclusions may have been different if you used a different study site or larger dataset, for example.

Module 6



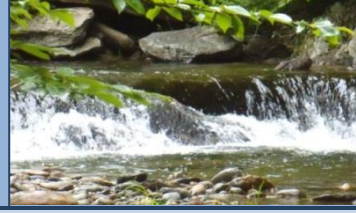
Summarizing Results and Drawing Conclusions

5. What new questions do you have?

Your findings might also help to drive future research studies by generating new questions. Follow the guidance below and try to answer the questions asked as they apply to your results.

- Most research uncovers more questions than answers. This is one of the most important benefits of science!
- What questions did your study generate?
- What might you do differently if you were to repeat your study again?
    - Did you have all the data you needed to test your hypothesis?
    - Did you ask an answerable question?
- What research questions would you suggest other students studying your stream site(s) ask in the future?

Module 6

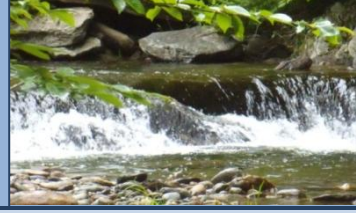


Summarizing Results and Drawing Conclusions

6. How can you effectively communicate your research findings to others?

Effectively communicating your findings to others can be the most important and most challenging step of scientific research. Follow the steps below and try to answer the questions asked as they apply to your results.

- Presenting your work (in written or verbal form) allows others to learn from your work and contributes to the overall body of knowledge in your field. It will likely be a learning process for you too!
- Who will your audience be (peers, scientists, professionals, general public)? Are they already familiar with your topic, in general? Present the material in a way that is appropriate for your audience.
- Be as clear as possible. Label and describe all figures. Focus on your most important findings. Use your data and results to justify your conclusions.
- Be careful how you describe your results. Did you really prove your hypothesis or did you just find evidence supporting it?
- Ask the audience for questions or comments. They may have a different and equally valid interpretation of your results.



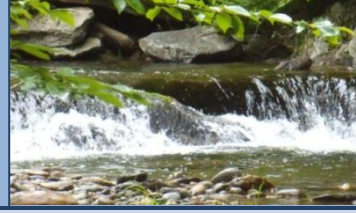
Summarizing Results and Drawing Conclusions

Test Your Understanding of Module 6

See if you can answer the questions below, with the help of your peers and your teacher. If you have questions or would like more information on any of the topics covered in this Module, contact Streams Project staff for assistance.

- Let's pretend that your group found a relationship between riparian buffers and TSS and thought that it may be because buffers help to stabilize stream banks and hence, reduce the amount of sediment entering streams. When presenting your research findings, does this information belong in the results or discussion section of your paper/talk?
- What determines if you're able to prove causation (that one things causes another)?
- Why shouldn't you be disappointed if you disprove your hypothesis?
- If a scientist from California called you to ask about applying your findings to an urban stream in San Francisco, what would you tell him/her?
- Name at least two new questions generated by the findings of your study.
- What are the benefits of sharing your research findings with others?

Module 6

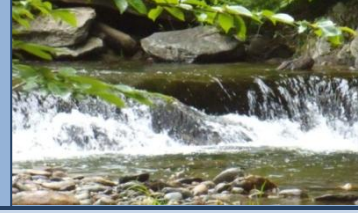


Summarizing Results and Drawing Conclusions

SUMMARY

- **What is the difference between results and discussion?**
- **What has been learned?** Do your results support or disprove your hypothesis?
- **What do your results mean?** Has any new valuable information been learned?
- **Can your results be applied locally or regionally? If so, how?** What do your findings contribute to your field of research?
- **What new questions do you have?** How could your findings contribute to future research studies focused on this topic?
- **How can you effectively communicate your results to others?** How can you best summarize and interpret your findings for different audiences?

Resources & References



These modules were created using materials, information, and ideas from the sources cited here. Please also consider consulting these additional resources yourself as you develop your project.

- Ambrose, H., & Ambrose, K. (2002). *A Handbook of Biological Investigation*. Winston-Salem, NC: Hunter Textbooks, Inc.

- Shuttleworth, M. (2008). *The Scientific Method: A website about research and experiments.* Retrieved at: http://www.experiment-resources.com/index.html

- (n.d.) *Understanding Science: how science really works.* Retrieved at: http://undsci.berkeley.edu/